National Cancer Institute

Statistical Detection of Association Hotspots in Highly Dimensional Genomic Data

Jianxin Shi

Jianxin.Shi@nih.gov

Biostatistics Branch

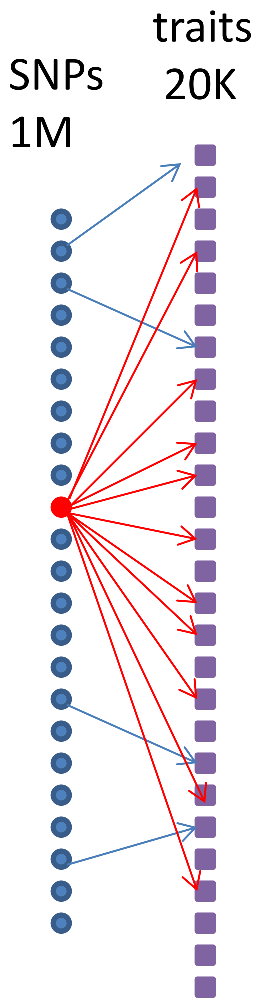
Division of Cancer Epidemiology and Genetics

National Cancer Institute

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institues
of Health

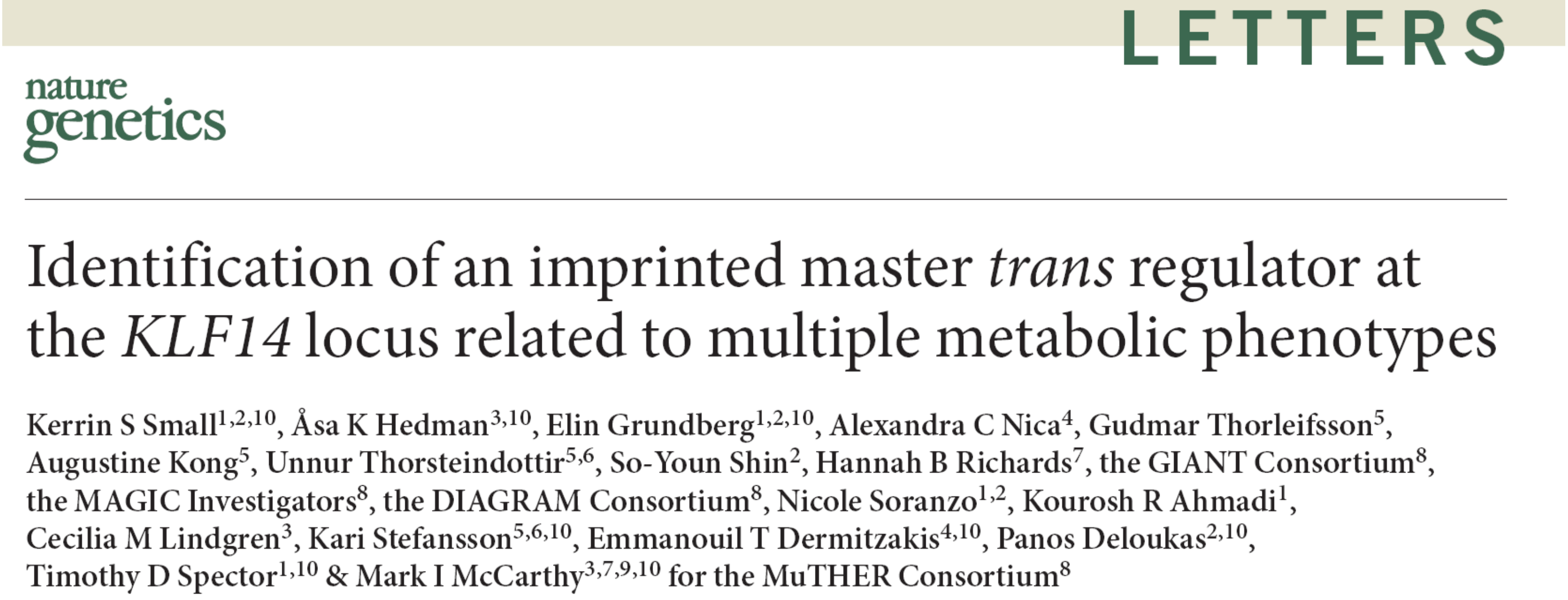
cis QTLs v.s. *trans* QTLs



Master regulators (MR)
trans-regulate many targets

	cis	*trans*
locations	SNP and target gene are physically close	SNP and target gene are far away or on different chromosomes
Effect sizes	big	small
Multiple testing	Search locally, small multiple testing burden	Search genome-wide, huge multiple testing burden
Power	good	Poor

- **SNP -> gene expression**
- **SNP -> DNA methylation**
- **CpG -> gene expression**
- **SCNA -> gene expression in tumors**

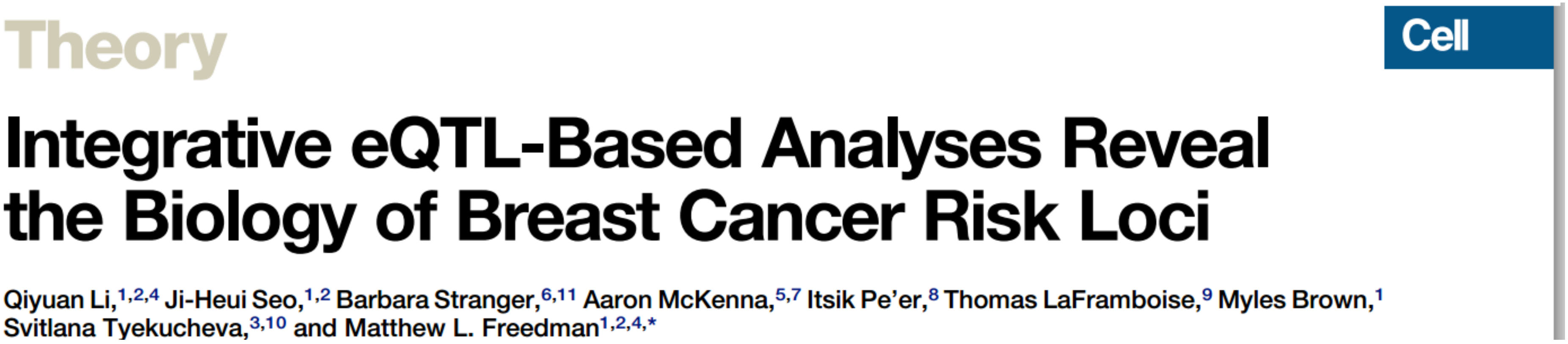


LETTERS

nature genetics

Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes

Kerrin S Small[1,2,10], Åsa K Hedman[3,10], Elin Grundberg[1,2,10], Alexandra C Nica[4], Gudmar Thorleifsson[5], Augustine Kong[5], Unnur Thorsteindottir[5,6], So-Youn Shin[2], Hannah B Richards[7], the GIANT Consortium[8], the MAGIC Investigators[8], the DIAGRAM Consortium[8], Nicole Soranzo[1,2], Kourosh R Ahmadi[1], Cecilia M Lindgren[3], Kari Stefansson[5,6,10], Emmanouil T Dermitzakis[4,10], Panos Deloukas[2,10], Timothy D Spector[1,10] & Mark I McCarthy[3,7,9,10] for the MuTHER Consortium[8]



Theory

Cell

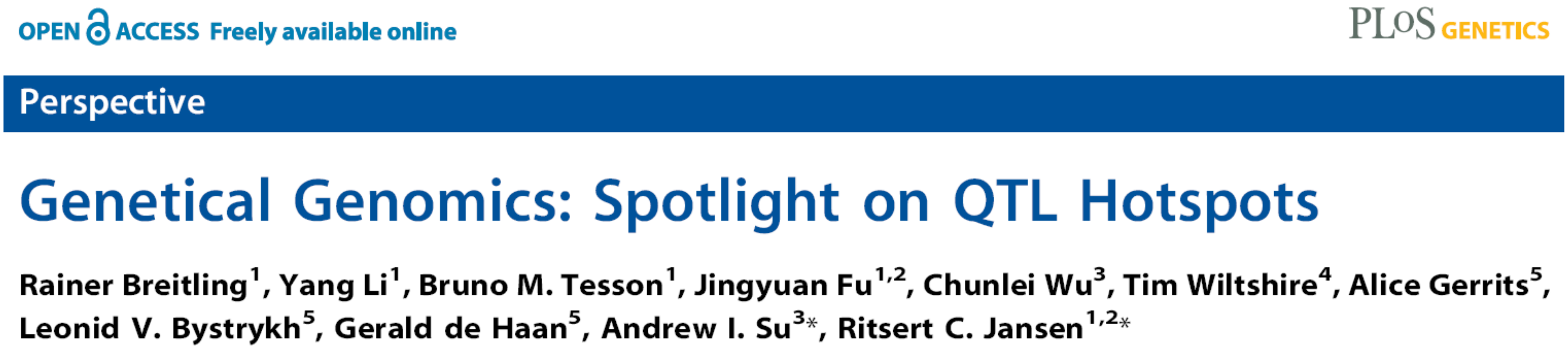
Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci

Qiyuan Li,[1,2,4] Ji-Heui Seo,[1,2] Barbara Stranger,[6,11] Aaron McKenna,[5,7] Itsik Pe'er,[8] Thomas LaFramboise,[9] Myles Brown,[1] Svitlana Tyekucheva,[3,10] and Matthew L. Freedman[1,2,4,*]

Provide evidences that three breast cancer SNPs cis-act MYC, ESR and KLF4, which further affect downstream genes.

Methods for Detecting Master Regulators

- **First method**
  - Choosing a *P*-value threshold (stringent, E-10) to detect significant *trans*-QTLs. For each SNP, count the number of traits significantly associated with the SNP.
- **Second Method**
  - For each SNP, let P_k be the *P*-value between the SNP and trait *k*. Choose a liberal threshold p_0 (e.g. 0.001) and count $M=\#\{P_k<p_0\}$. If *M* is large, the SNP is a master regulator.
  - Assuming traits are independent, calculate significance using Poisson approximation.
  - Many eQTL studies reported master regulators.



OPEN ACCESS Freely available online

PLoS GENETICS

Perspective

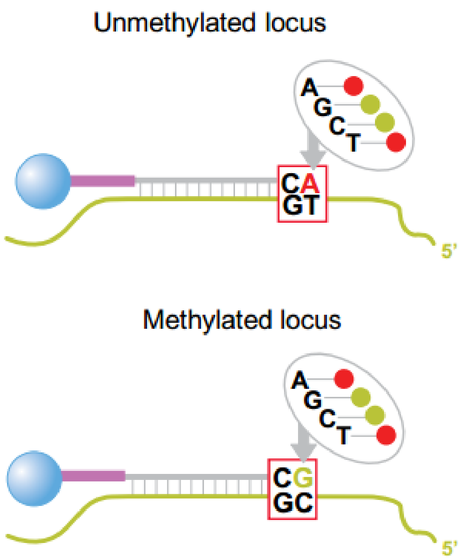
Genetical Genomics: Spotlight on QTL Hotspots

Rainer Breitling[1], Yang Li[1], Bruno M. Tesson[1], Jingyuan Fu[1,2], Chunlei Wu[3], Tim Wiltshire[4], Alice Gerrits[5], Leonid V. Bystrykh[5], Gerald de Haan[5], Andrew I. Su[3]*, Ritsert C. Jansen[1,2]*

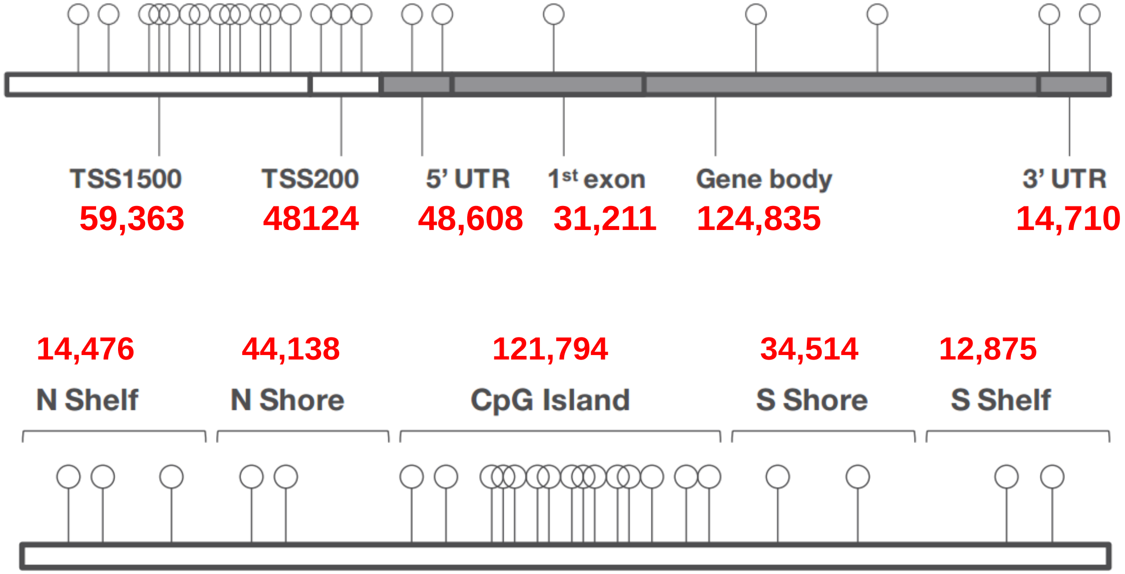
- **One example study**
  - Wu et al. (Plos Genetics, 2008) detected ~1600 master regulators in mouse adipose tissue eQTL study.
- **A formal permutation test**
  - Permute genotype; keep correlation in traits unchanged.
  - Test statistic is $M=\#\{P_k<p_0\}$.
  - Based on permutations, the best SNP has a *P*-value 0.23!
- **Correlation in traits and statistical inference**

meQTLs Analysis in EAGLE Normal Lung Tissue

EAGLE: Environment And Genetics in Lung cancer Etiology, DCEG/NCI
Dr. Landi and Dr. Caporaso at DCEG/NCI
210 tumor adjacent normal fresh frozen tissues in EAGLE
500K common SNPs and 340K CpG probes after QC.



Illumina HumanMethylation450 BeadChip



Distribution of CpG probes in platform

Shi et al., Nature Communications, 2014

RNA-seq eQTLs Analysis in GenRED Blood Sample

GenRED: Genetics of Recurrent Early-Onset Depression

RNA-Seq in 922 blood samples and GWAS SNPs

Dr. Douglas Levinson, Stanford University

~700K common SNPs and 15,000 genes after QC.

Identified ~11K *cis* eQTLs and ~110 *trans* eQTLs.

One master regulator was detected.

Battle et al., Genome Research, 2013

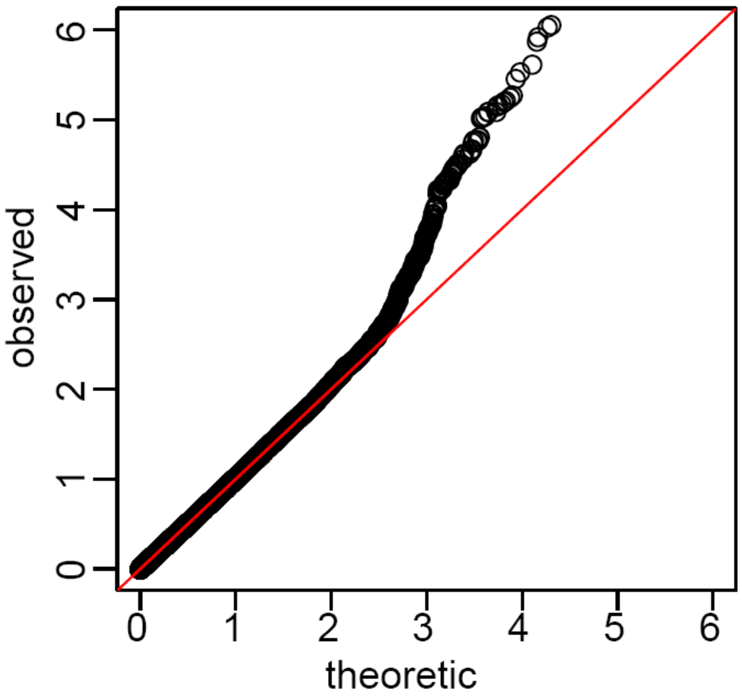
A Statistical Framework for Detecting Master Regulators

- **Notations**
  - Consider one SNP and *K* genes not in its *cis*-region.
  - Let Z_k be the score statistic for testing the association between the SNP and gene *k*. Let $u_k=E(Z_k)$. Under null, $Z_k \sim$ N(0,1).

- **Statistical question**
  - Given $\{Z_1, \ldots, Z_K\}$, we test null hypothesis H_0: $u_k=0$ for $k=1,\ldots$,K.
  - If H_0 is rejected, a master regulator is detected. We use FDR to identify associated genes for the master regulator.

Testing H_0 for Independent Traits

QQ plot of –log(p)



Siegmund and Zhang's test

$$T = \sum_{k=1}^{K} \log(1 - w + w\exp(z_k^2/2))$$

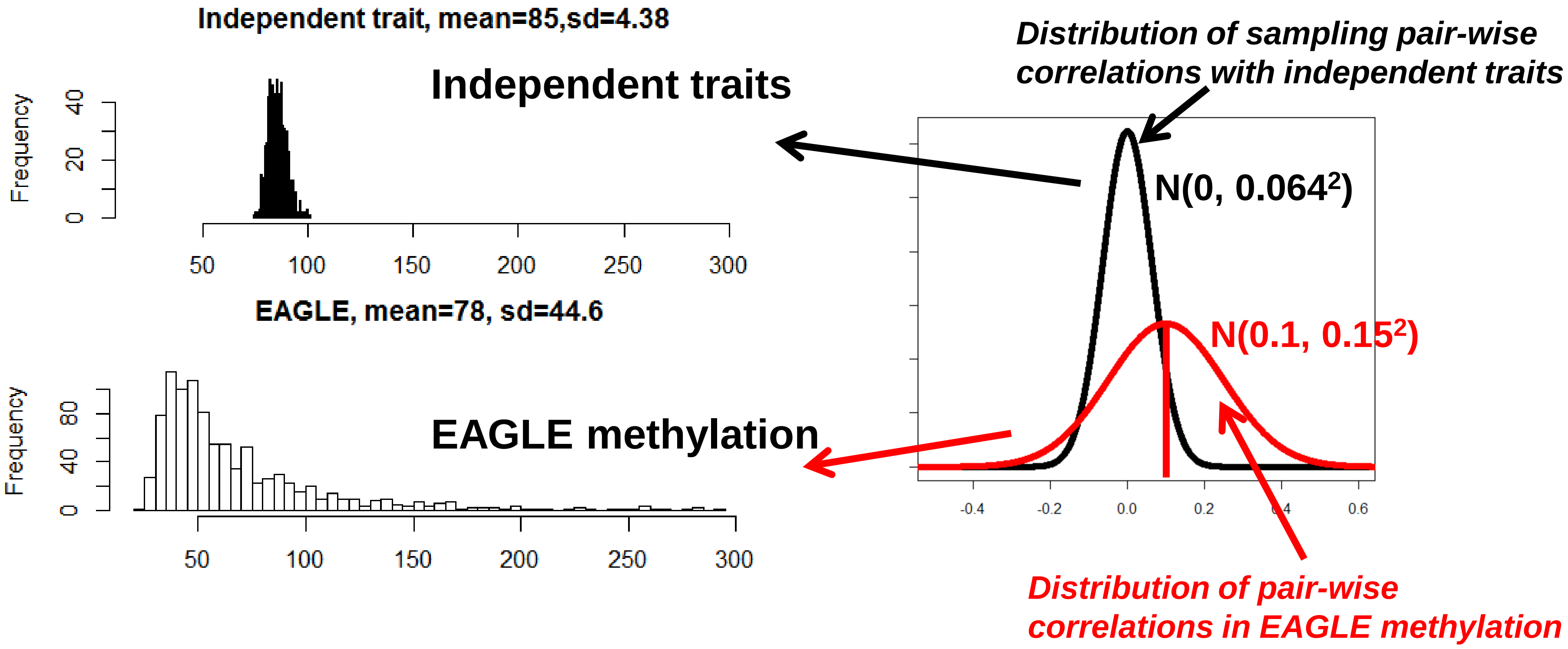
w **defines the proportion of genes associated with the SNP.**

If a small proportion of genes are associated, we choose a small ***w*****.**

Or choose a series of ***w*** **and correct for multiple testing.**

Long Range Correlations in Traits Damages MR Detection

Null distribution for T based on permutations



Independent traits

EAGLE methylation

Distribution of sampling pair-wise correlations with independent traits

$N(0, 0.064^2)$

$N(0.1, 0.15^2)$

Distribution of pair-wise correlations in EAGLE methylation

Correlations in traits make the variance of test T extremely large.
The power is typically close to zero.

Permutations can protect against false positives.
But we need power!

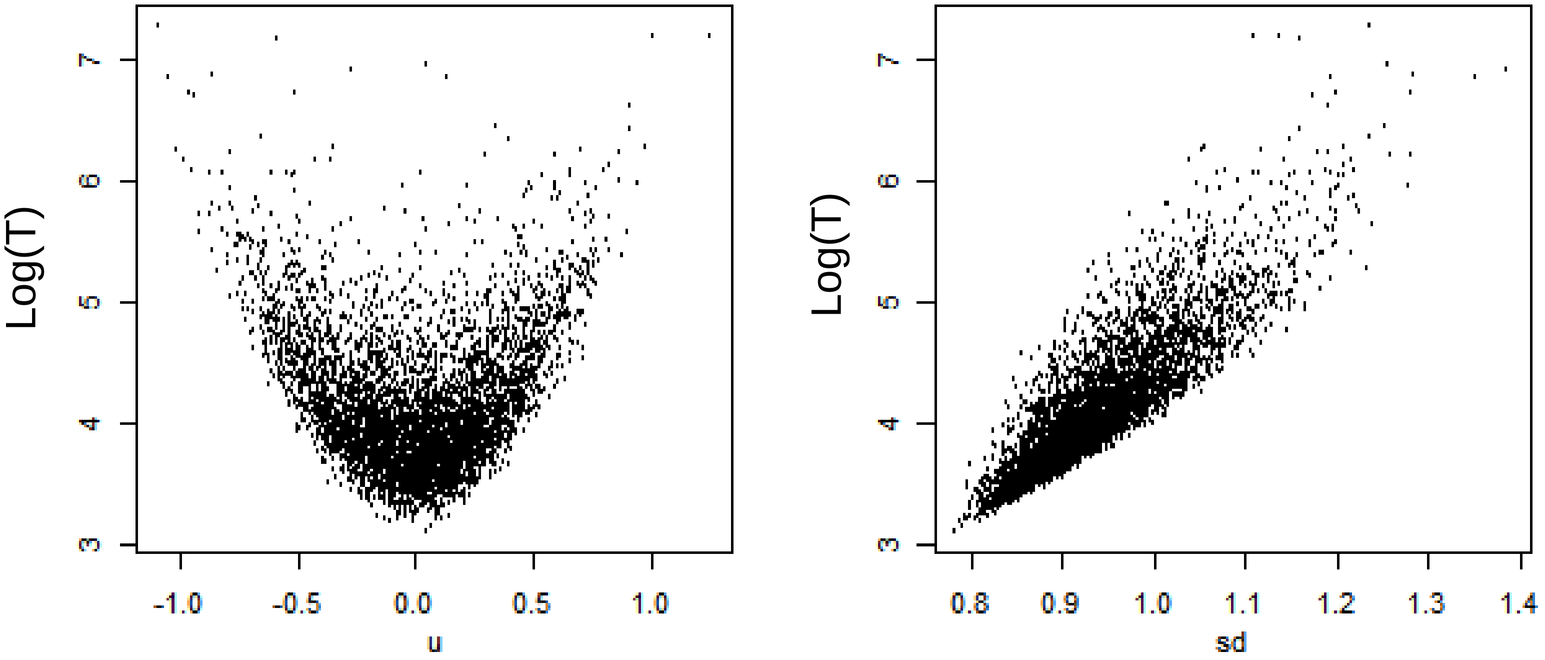
Long Range Correlations in EAGLE and GenRED

	Empirical distribution of correlations	Expected distribution of correlations if independent
EAGLE methylation	$N(0.10, 0.149^2)$	$N(0.0, 0.064^2)$
GenRED RNA-seq	$N(0.02, 0.065^2)$	$N(0.0, 0.032^2)$

We have corrected for sex, age, plates and PCA vectors capturing some hidden confounding factors.

Improve Power by Reducing Var(T)

- **Permutations**
  - 1M permutations, keeping trait correlation structure.
  - For permutation *t*, calculate $(z_1^t, \cdots, z_K^t)$, u_t, σ_t and T_t statistic.



- This observation suggests that we can reduce Var(T) by adjusting for the features of the empirical distribution of $(Z_1, \cdots, Z_K)$.

Detecting Master Regulatory SNPs by Correcting for Empirical Null Distribution

- Consider one SNP and *K* traits.
- Using original data to calculate $(Z_1,\cdots,Z_K)$, μ, σ and T.
- Run 1M permutations keeping correlation structure. For permutation *t*, calculate $(z_1^t,\cdots,z_K^t)$, u_t, σ_{t} and T_t statistic.
- Run linear regression (or smoothing)

$$\log(T_t)=\alpha+\beta_1\mu_t+\beta_2\sigma_t+\varepsilon_t$$

- Define a new test as

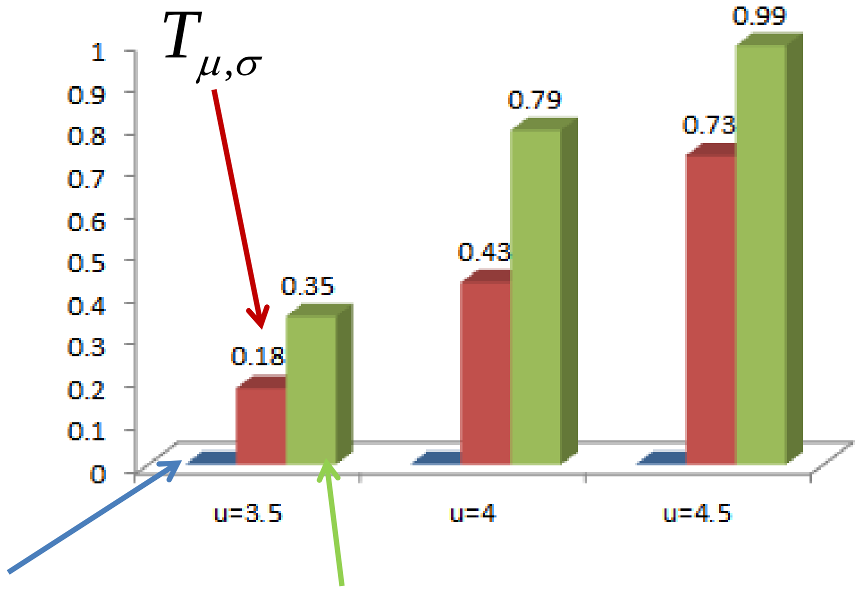
$$\log(T_{\mu,\sigma})=\log(T)-\hat{\alpha}-\hat{\beta}_1\mu-\hat{\beta}_2\sigma$$

- Significance of $T_{\mu,\sigma}$ is evaluated using permutations.

Power Simulation in Realistic Correlated Traits

- Using 340K CpG traits of 210 samples from EAGLE data.
- 10 or 20 traits associated with a SNP. Other traits are from data.
- Null distribution is based 1M permutations.

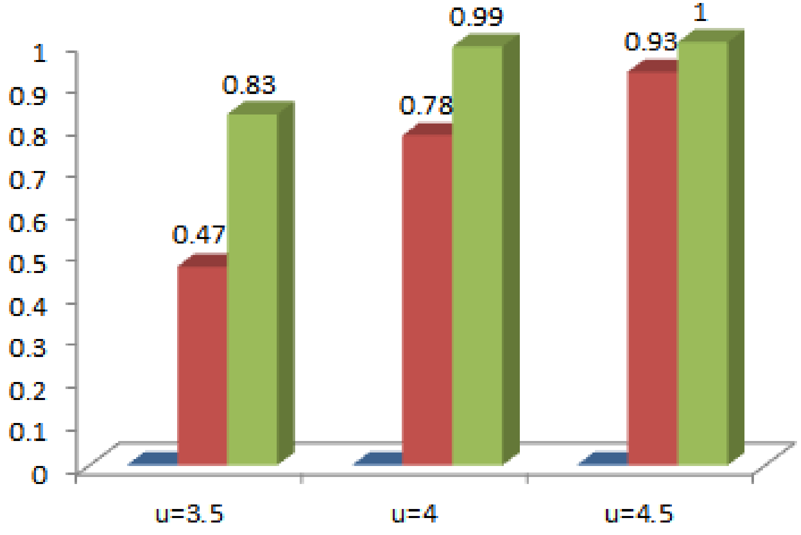
10 associated CpGs



T

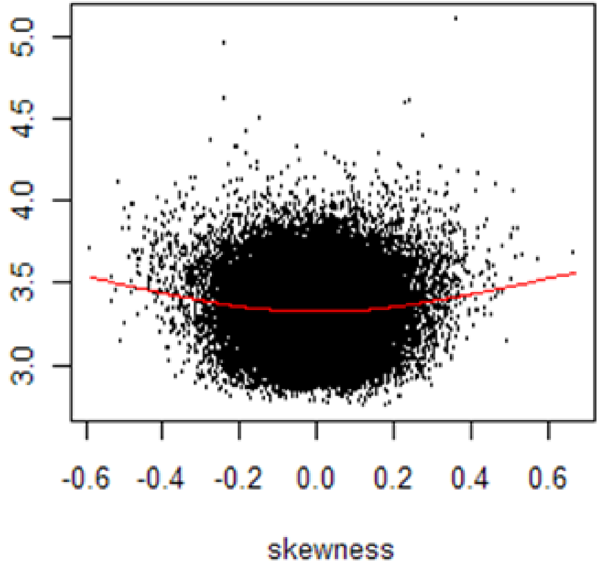
Traits are independent

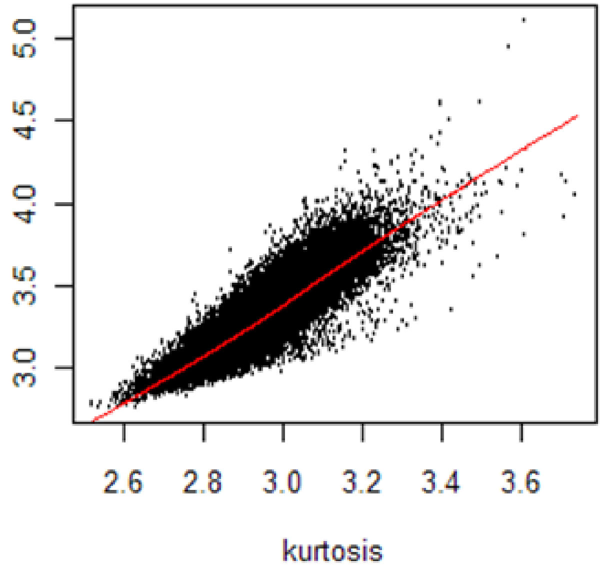
20 associated CpGs



$\mu=E(Z_k)$ is effect size; $\alpha=0.001$.

Correcting for Skewness and Kurtosis Further Improves Statistical Power



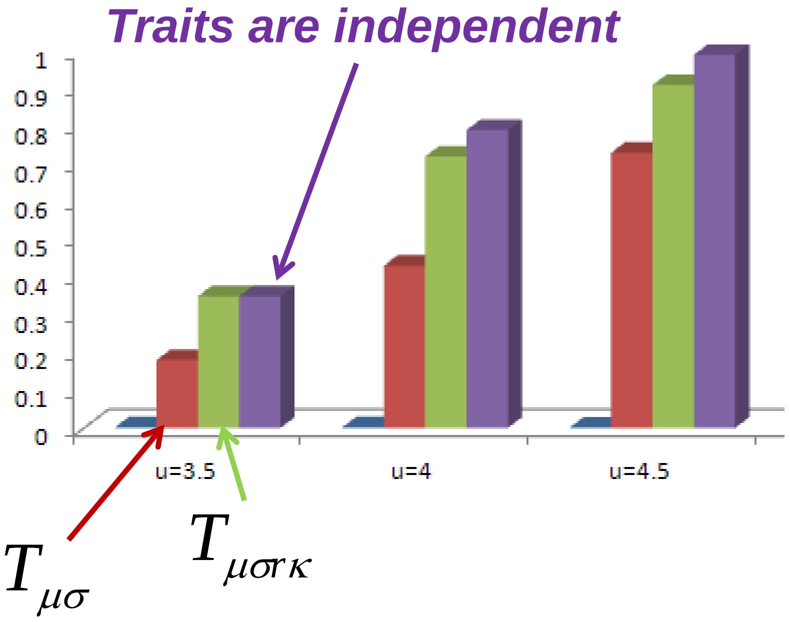


Linear regression:

$$\log(T_t) = \alpha + \beta_1\mu_t + \beta_2\sigma_t + \beta_3 r_t + \beta_4\kappa_t + \varepsilon_t$$

First four moments corrected test:

$$\log(T_{\mu\sigma r\kappa}) = \log(T) - \hat{\alpha} - \hat{\beta}_1\mu - \hat{\beta}_2\sigma - \hat{\beta}_3 r - \hat{\beta}_4\kappa$$


Traits are independent

$T_{\mu\sigma}$

$T_{\mu\sigma r\kappa}$

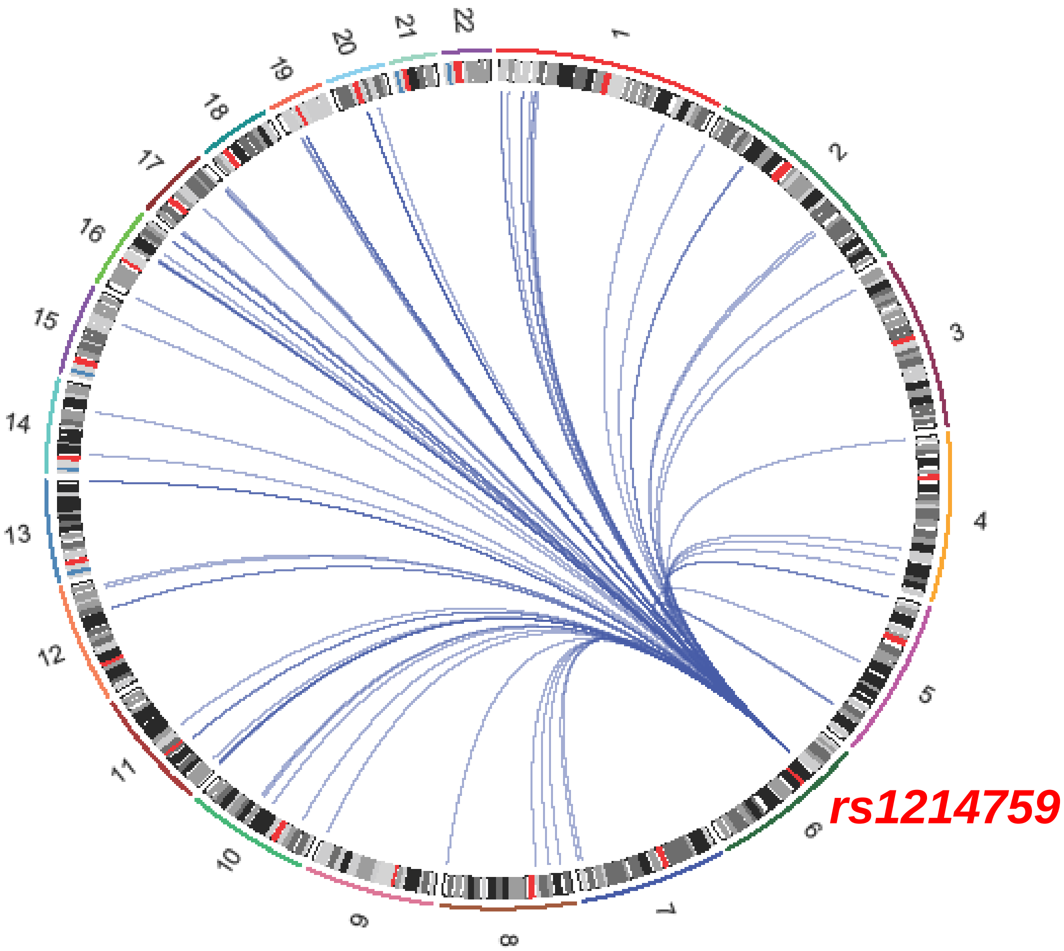
EAGLE Methylation QTL Study

- **210 normal lung samples, 340K CpG probes.**
- **rs1214759 (6p21.1) was detected as a candidate**
  - P=3.8×10^{-6} , did no reach genome wide significance 5×10^{-8}.
  - 80 CpG probes associated with rs1214759 at P<10^{-5}
  - No (SNP,CpG) pair achieved significance 2×10^{-10} to be detected.
- **Can we replicate?**

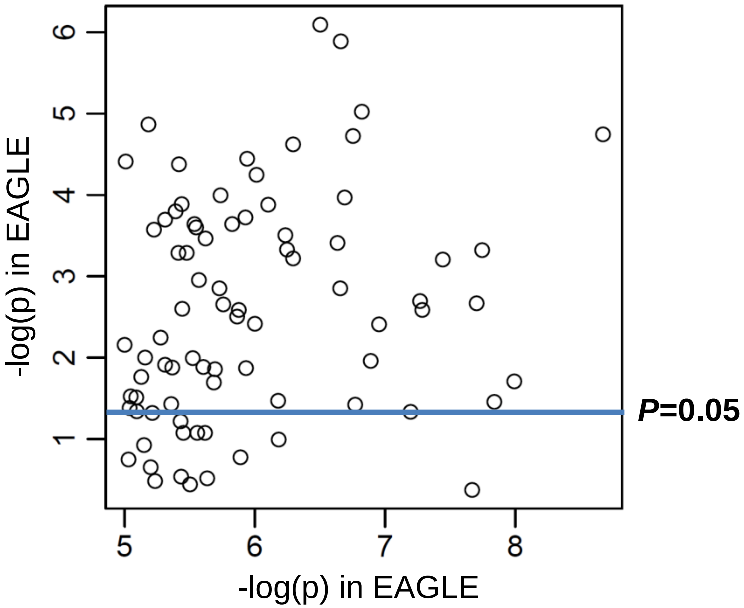
How many CpG probes are associated with rs1214759?

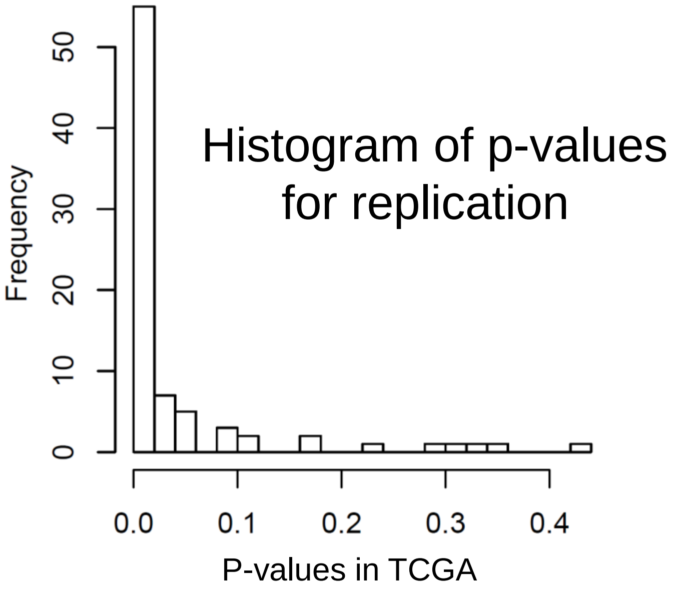
	CpGs expected under Null	CpGs in EAGLE	
P<10^{-5}	34K*10^{-5}=3.4	80	

EAGLE discovery sample



TCGA lung validation 65 samples





EAGLE Methylation QTL Study

- **210 normal lung samples, 340K CpG probes.**
- **rs1214759 (6p21.1) was detected as a candidate**
  - P=3.8×10^{-6} , did no reach genome wide significance 5×10^{-8}.
  - 80 CpG probes associated with rs1214759 at $P<10^{-5}$
  - No SNP/CpG pair achieved significance 2×10^{-10} to be detected.
- **Can we replicate?**

How many CpG probes are associated with rs1214759?

	CpGs expected under Null	CpGs in EAGLE	CpGs replicated TCGA lung (n=65)
$P<10^{-5}$	34K*10^{-5}=3.4	80	53

Replication criterion: same direction and P<0.05.

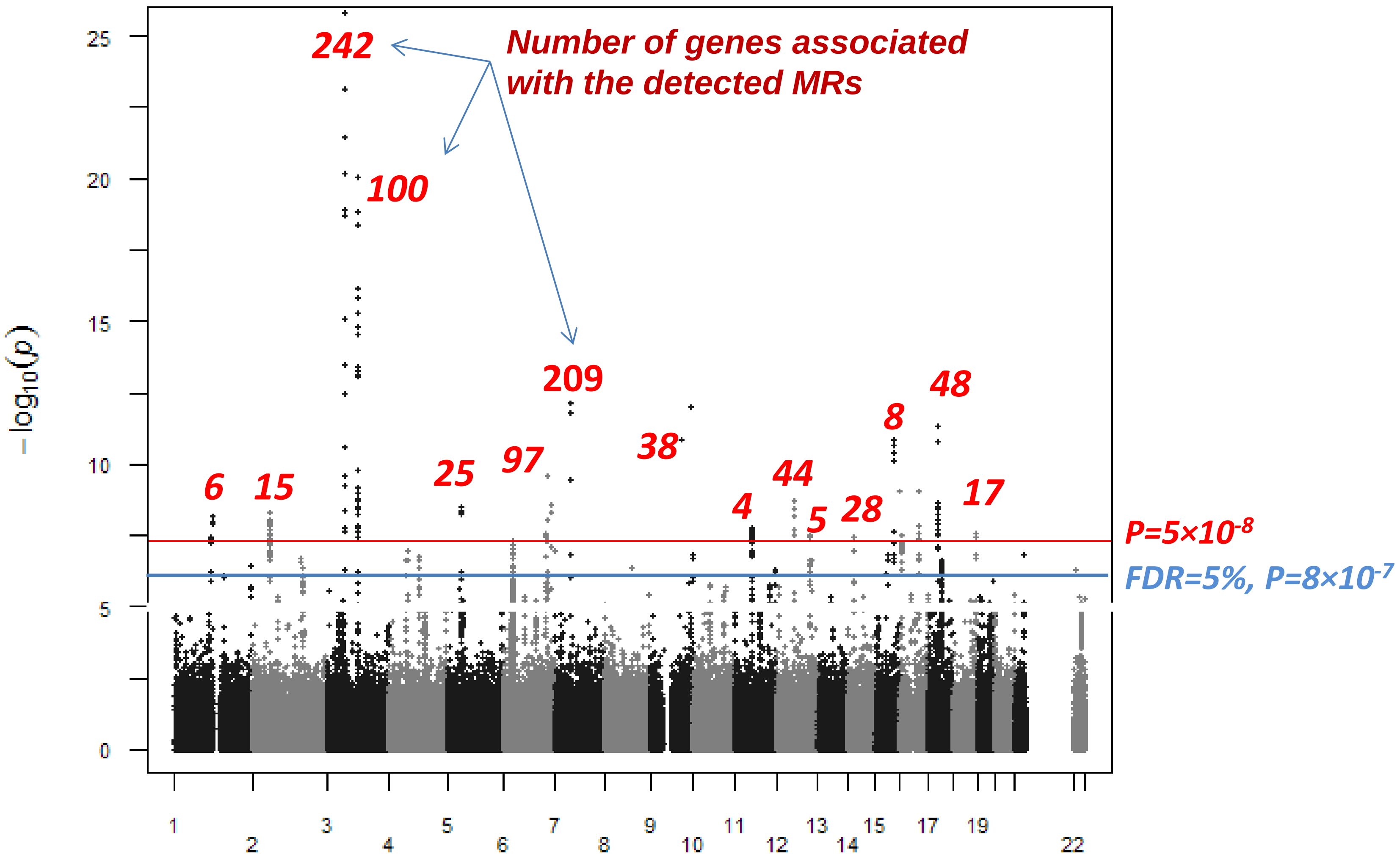
GenRED RNA-Seq of 922 Blood Samples Internal Validation Results

615 samples as discovery; 13 master regulators at 5E-8; 22 at FDR=5%.
307 samples as validation

SNP	*P*-discovery	*P*-validation	# detected genes	Same direction in validation	*P*<0.05 in validation
rs1354034	1.80E-24	**8.30E-22**	178	167	139
rs497953	8.20E-19	**2.80E-08**	69	67	53
rs10251980	1.60E-11	**4.60E-07**	123	116	73
rs13289095	2.10E-11	**8.20E-03**	26	26	19
rs13218225	7.80E-11	**7.80E-02**	56	43	12
rs6580981	1.00E-09	**1.30E-01**	31	27	17
rs8056400	3.10E-09	**8.40E-08**	8	8	7
rs8073060	6.00E-09	**3.60E-05**	33	32	27
rs1138358	6.90E-09	**2.40E-03**	20	19	15
rs12145080	8.40E-09	**2.00E-07**	5	5	5
rs821470	1.50E-08	**2.00E-04**	10	10	7
rs9399137	2.60E-08	**7.50E-07**	7	7	6
rs13019832	2.80E-08	**9.80E-07**	5	5	5
rs12938031	1.80E-07	**4.00E-06**	2	2	2
rs16911097	1.80E-07	**2.70E-03**	26	24	20
rs12419022	2.30E-07	**6.50E-04**	4	4	4
rs10074873	2.50E-07	**4.20E-06**	12	11	7
rs8090565	3.50E-07	**1.00E-02**	5	4	4
rs4964607	3.70E-07	**2.20E-04**	3	3	3
rs11229606	3.90E-07	**1.30E-04**	4	4	4
rs12418771	4.60E-07	**5.20E-03**	4	4	4
rs3130612	6.60E-07	**4.30E-03**	3	3	3

GenRED RNA-Seq of 922 Blood Samples

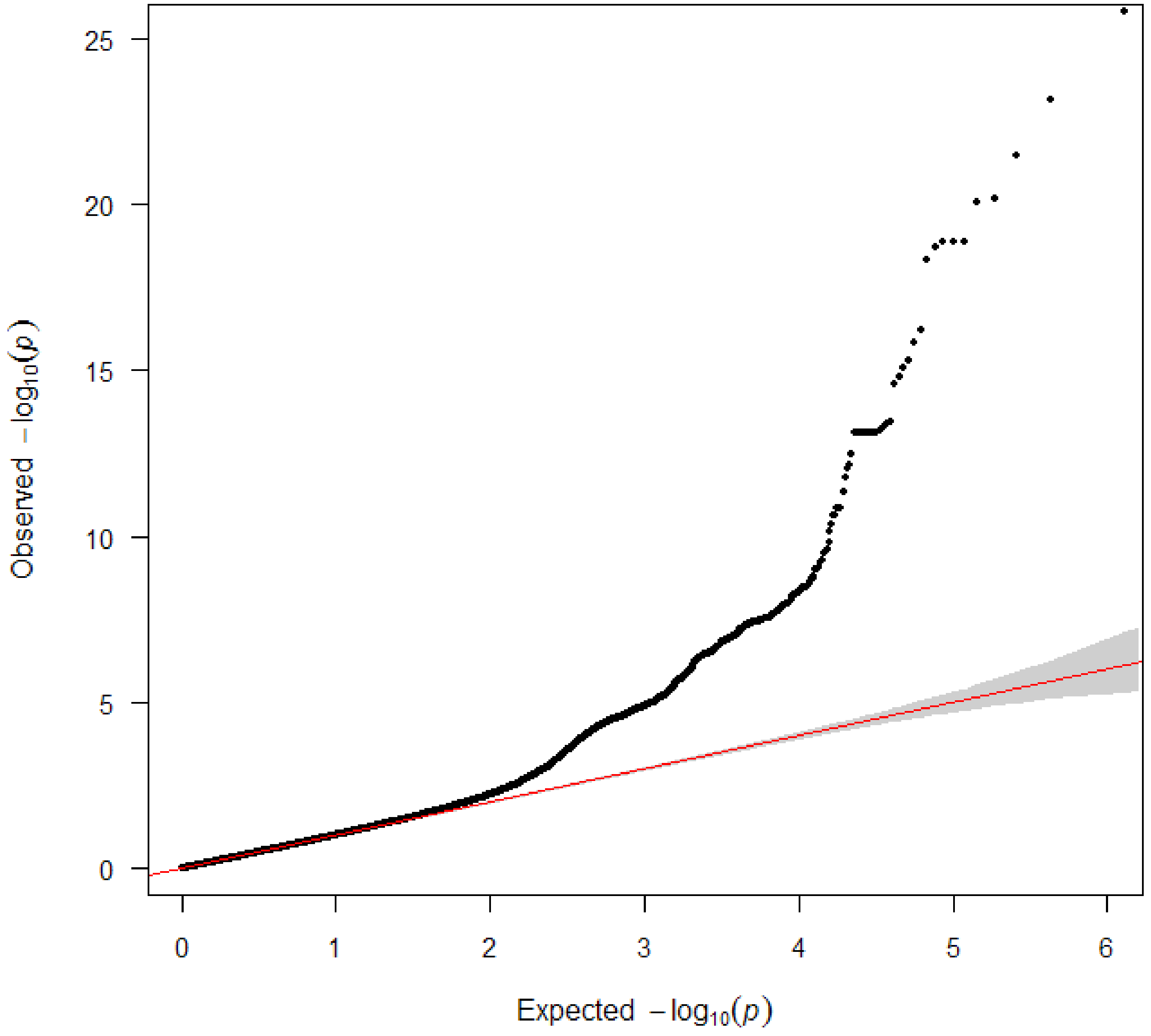
Manhattan plot of *P*-values for detecting MRs



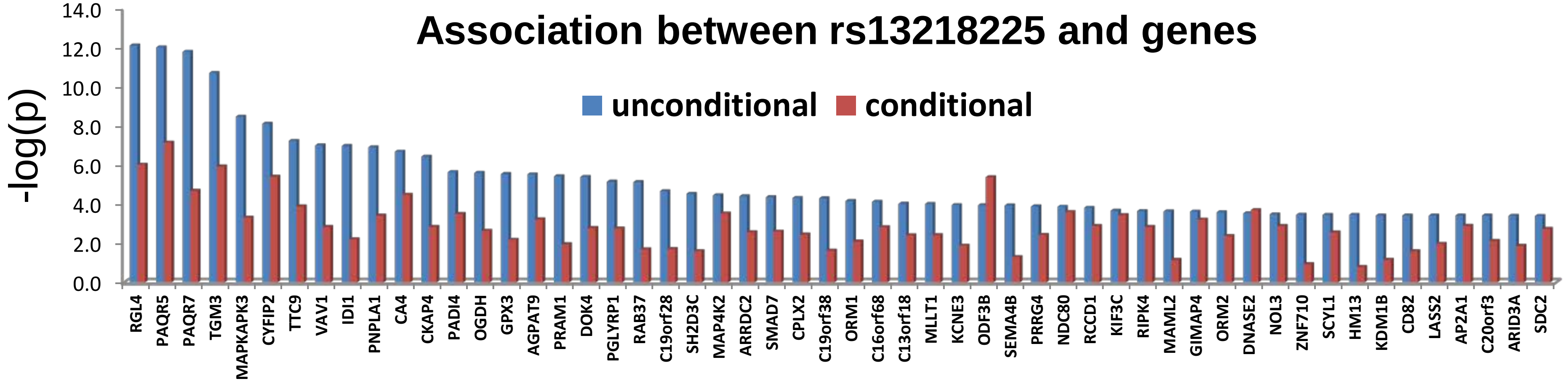
22 master regulators detected at *P*<5E-8;

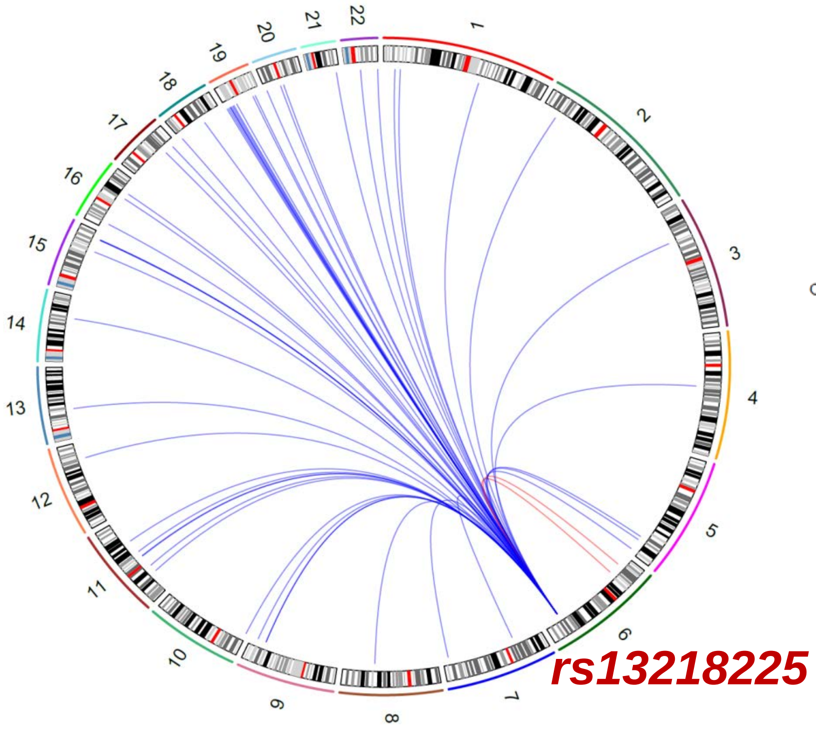
33 master regulators detected at *FDR*<5% or *P*<8.0E-7;

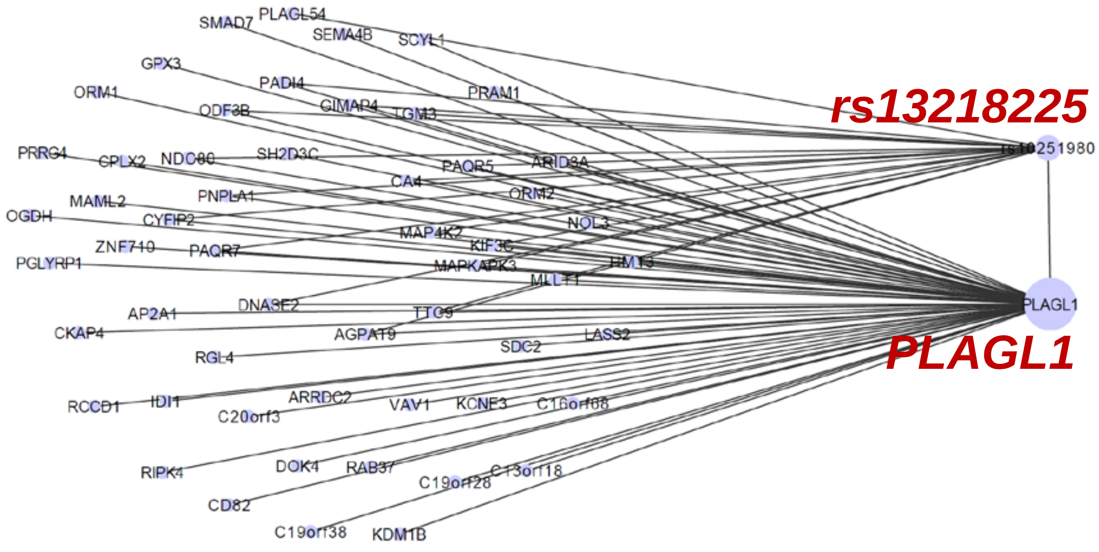
QQ plot for detecting master regulators in the RNA-Seq eQTL study based on GenRED, lambda=0.95



trans-Associations Mediated by *PLAGL1*

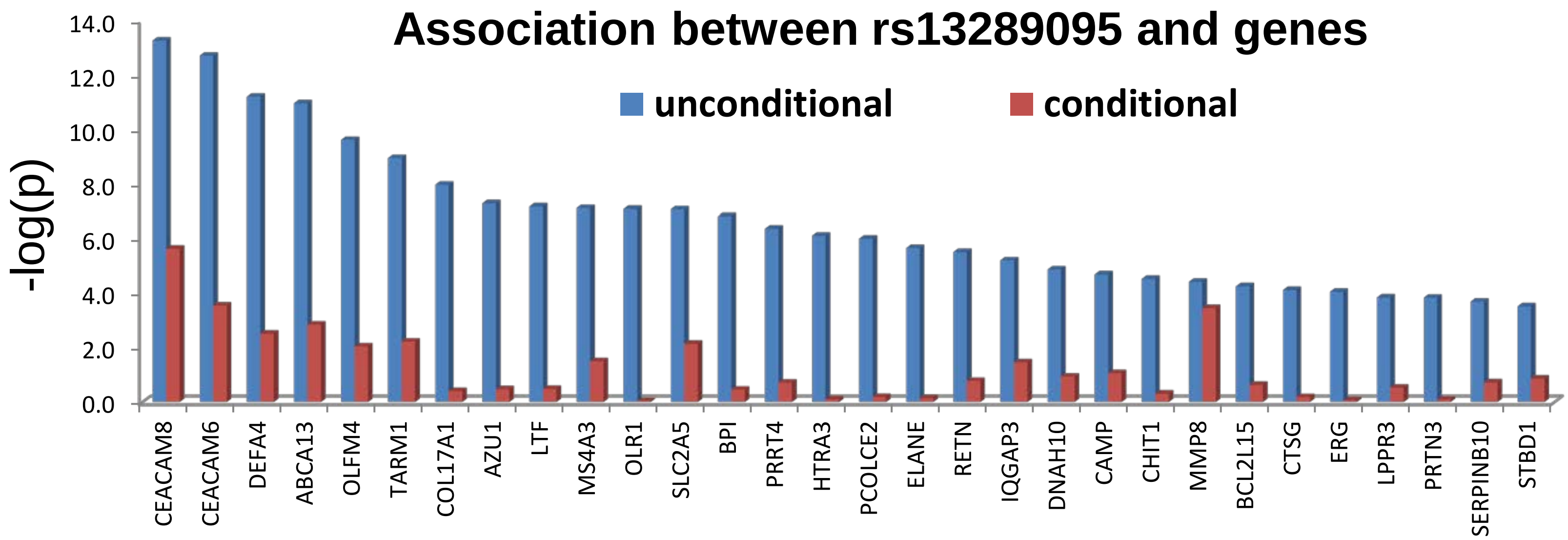


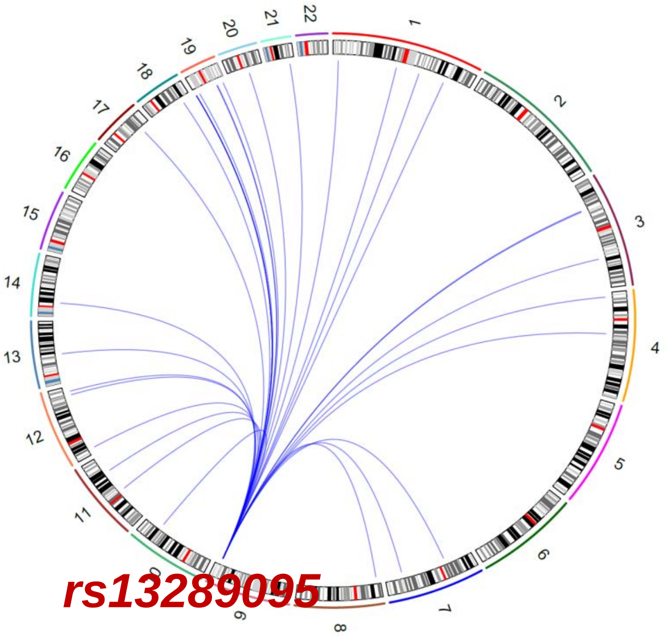


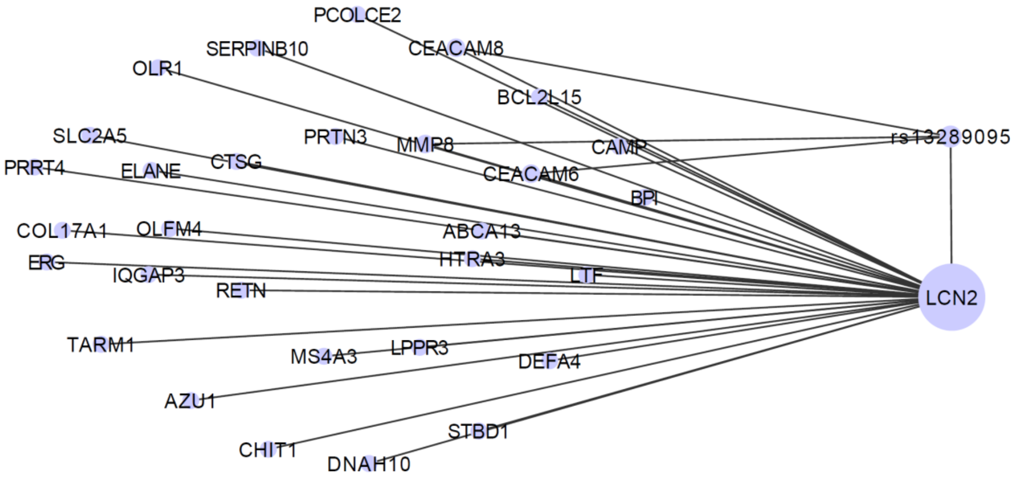


PLAGL1 **encodes a C2H2 zinc finger protein with transactivation and DNA binding.**

trans-Associations Mediated by *LCN2*

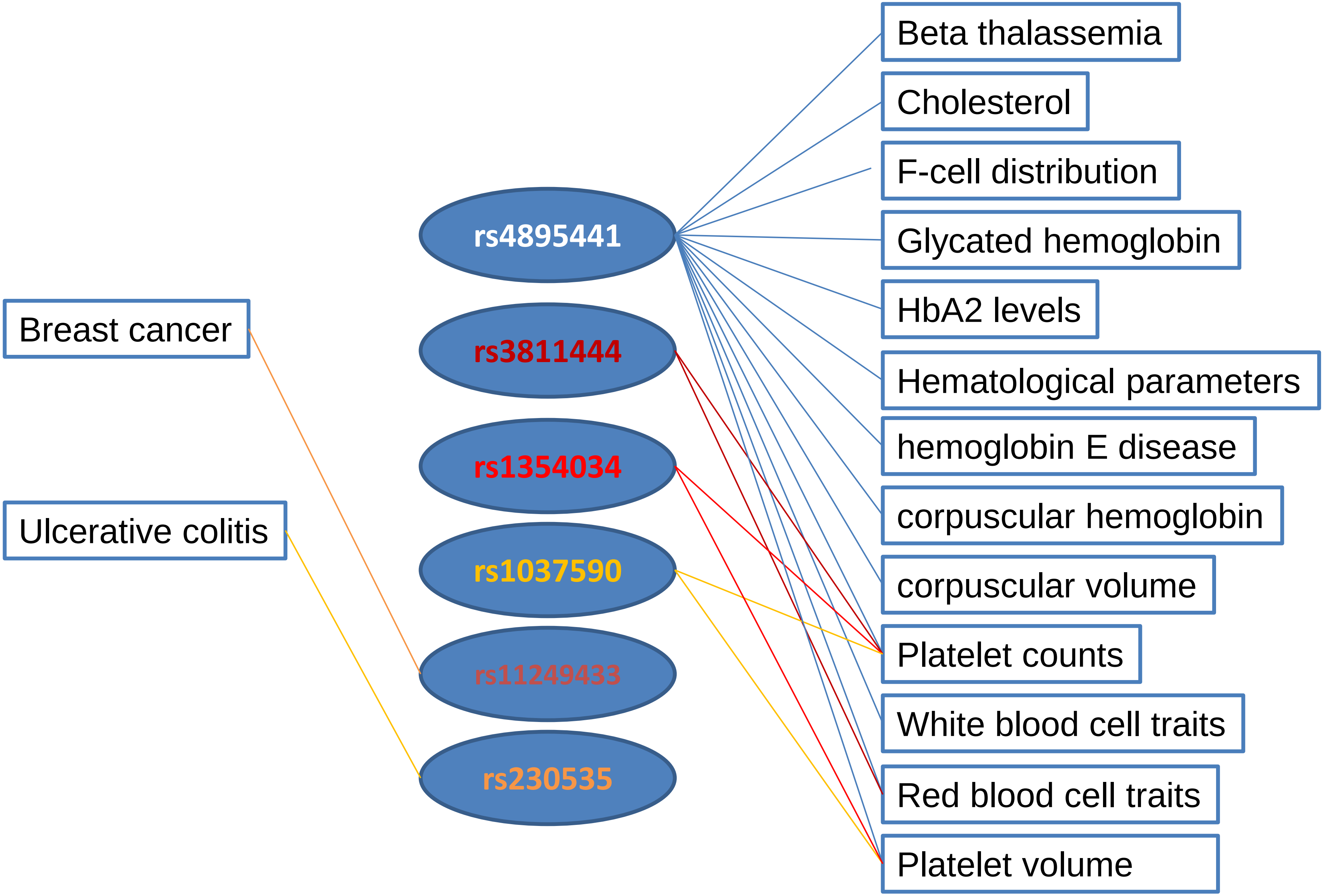

Association between rs13289095 and genes

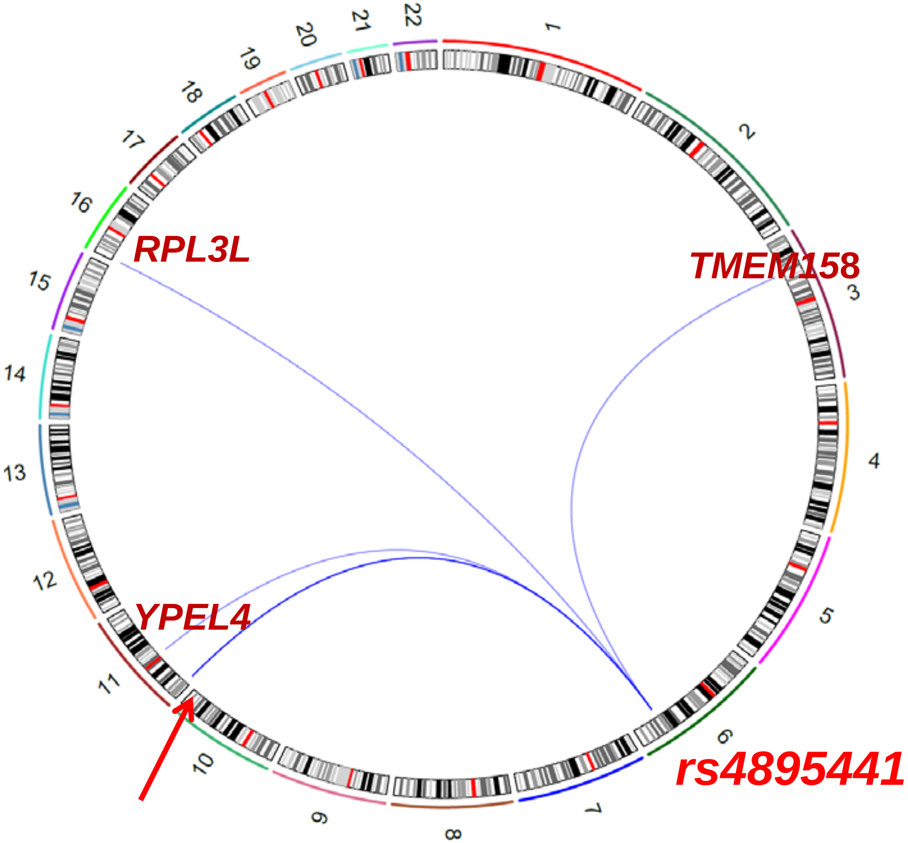




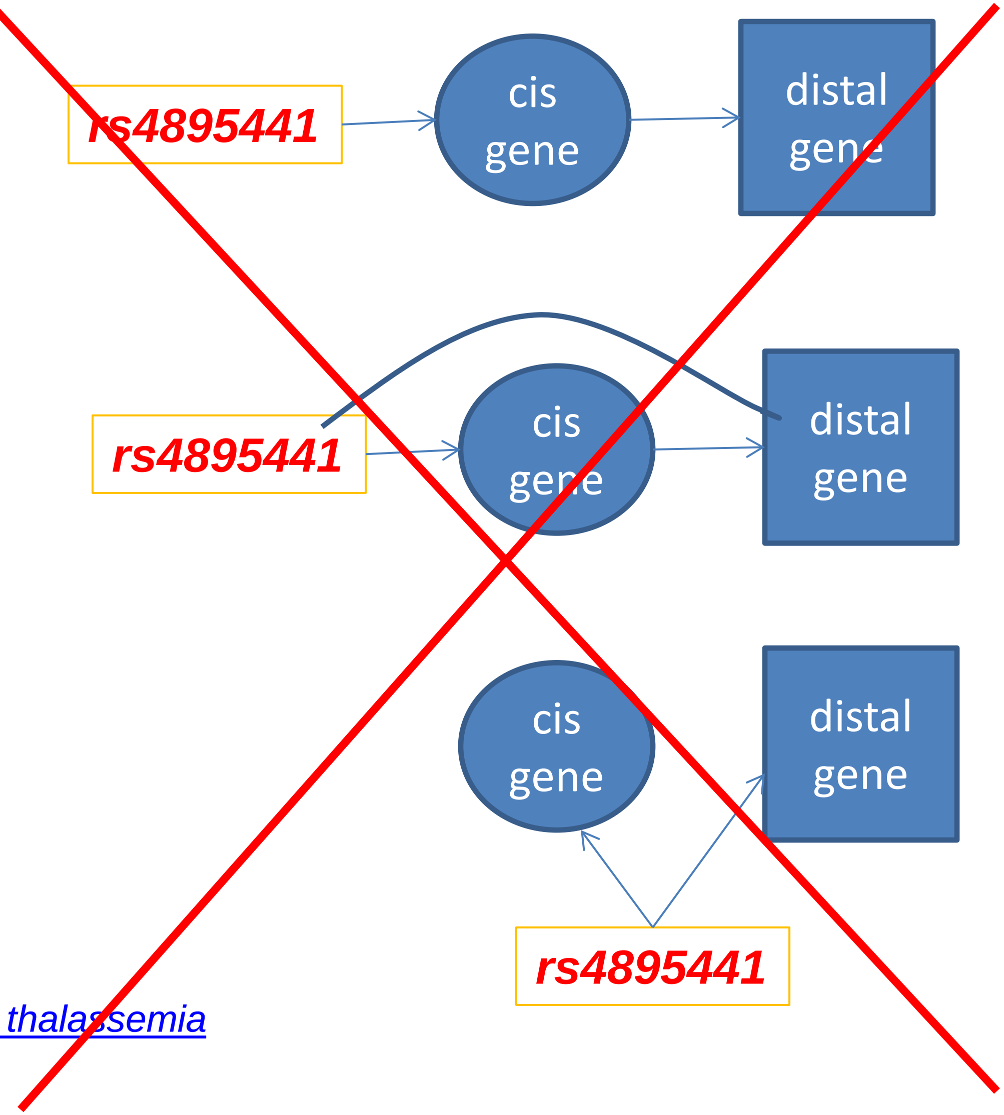
***LCN2* not transcription factor!**

Master eSNPs and GWAS Catalog





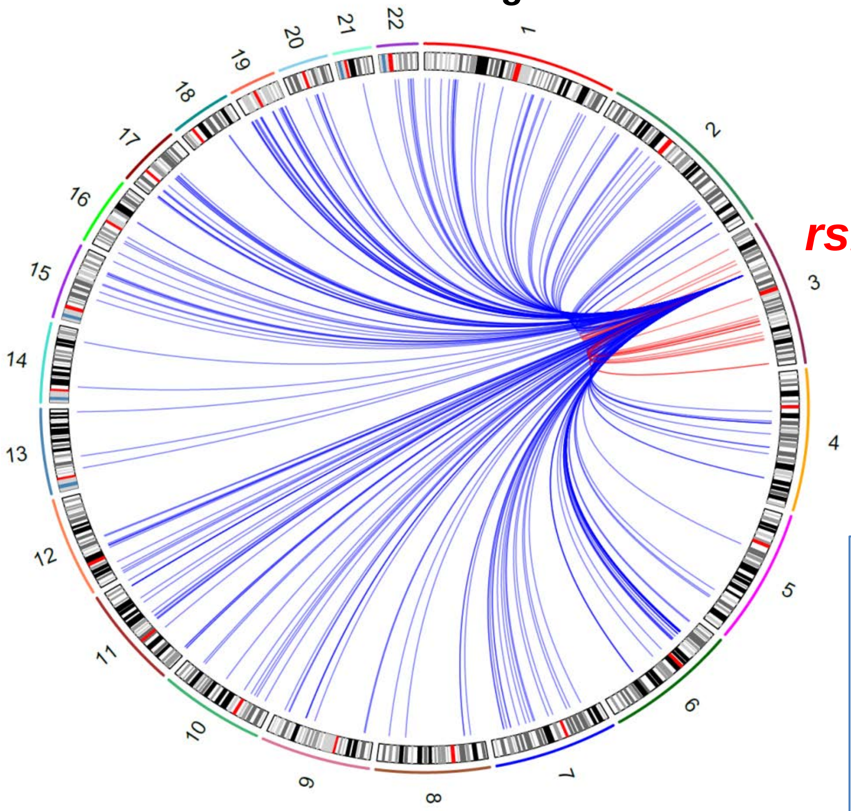
***HBE1* (Hemoglobin, Epsilon 1)**
***HBBP1* (Hemoglobin, Beta Pseudogene 1),** *beta thalassemia*
***HBG2* (hemoglobin, gamma G)**
***HBG1* (hemoglobin, gamma G)**



rs4895411 not associated with any gene in 5M.
Other biological mechanism.

***rs1354034 trans*-regulates 242 genes.**
***rs1354034* associated with platelet counts and mean platelet volume in GWAS.**

***trans*-associations not mediated by genes in local 5M region.**



rs1354034

Platelet counts (PLT)

Mean platelet volume (MPV)

Gieger et al., Nature, 2011

rs1354034 trans-regulates genes that affect platelet functions (GO enrichment):

- blood coagulation (*P*=2.9E-20)
- platelet activation (*P*=4.5E-20)
- platelet degranulation (*P*=3.2E-13)
- wound healing (*P*=3.5E-19).

Ongoing Work

- **Identify master regulators in TCGA data**
- **Predictors**
  - SNP, DNA methylation, somatic copy number aberrations, somatic gene mutation status
- **Quantitative traits**
  - Gene expression, DNA methylation in tumor samples
- **~30 cancers in TCGA**
  - Sample sizes range from 100 to ~1000.

Summary

- **Long range correlation in traits may damage the power of detecting master regulators.**

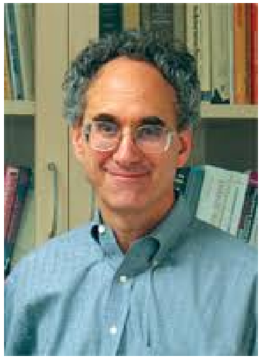
- **We developed a computationally expensive but statistically powerful approach for detecting master regulators.**

- **We identified replicable master regulators for gene expression and DNA methylation.**

- **Computation is intensive.**

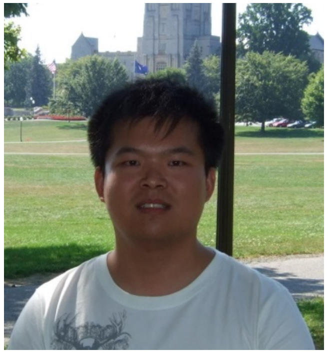
Acknowledgement

Dr. William Wheeler, NCI/DCEG contractor



Dr. Douglas Levinson, Stanford

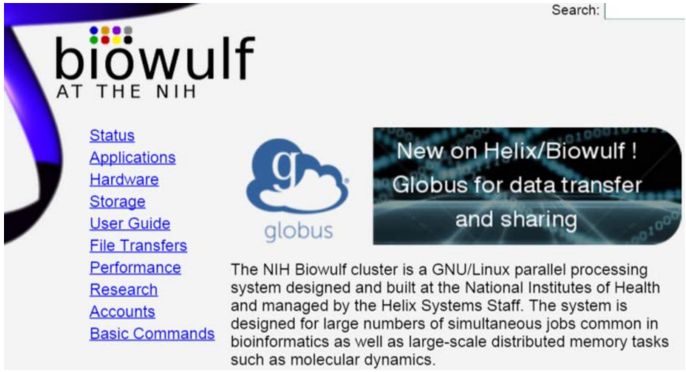
Dr. Alexis Battle, John Hopkins



Mr. Lei Song, NCI/DCEG contractor

Dr. Teresa Landi, NCI/DCEG

Dr. Nilanjan Chatterjee, NCI/DCEG



Search:

biowulf
AT THE NIH

- Status
- Applications
- Hardware
- Storage
- User Guide
- File Transfers
- Performance
- Research
- Accounts
- Basic Commands

New on Helix/Biowulf ! Globus for data transfer and sharing

The NIH Biowulf cluster is a GNU/Linux parallel processing system designed and built at the National Institutes of Health and managed by the Helix Systems Staff. The system is designed for large numbers of simultaneous jobs common in bioinformatics as well as large-scale distributed memory tasks such as molecular dynamics.