Experimental and computational methods for identifying genetic variants impact on gene regulation

Roger Piqué-Regí

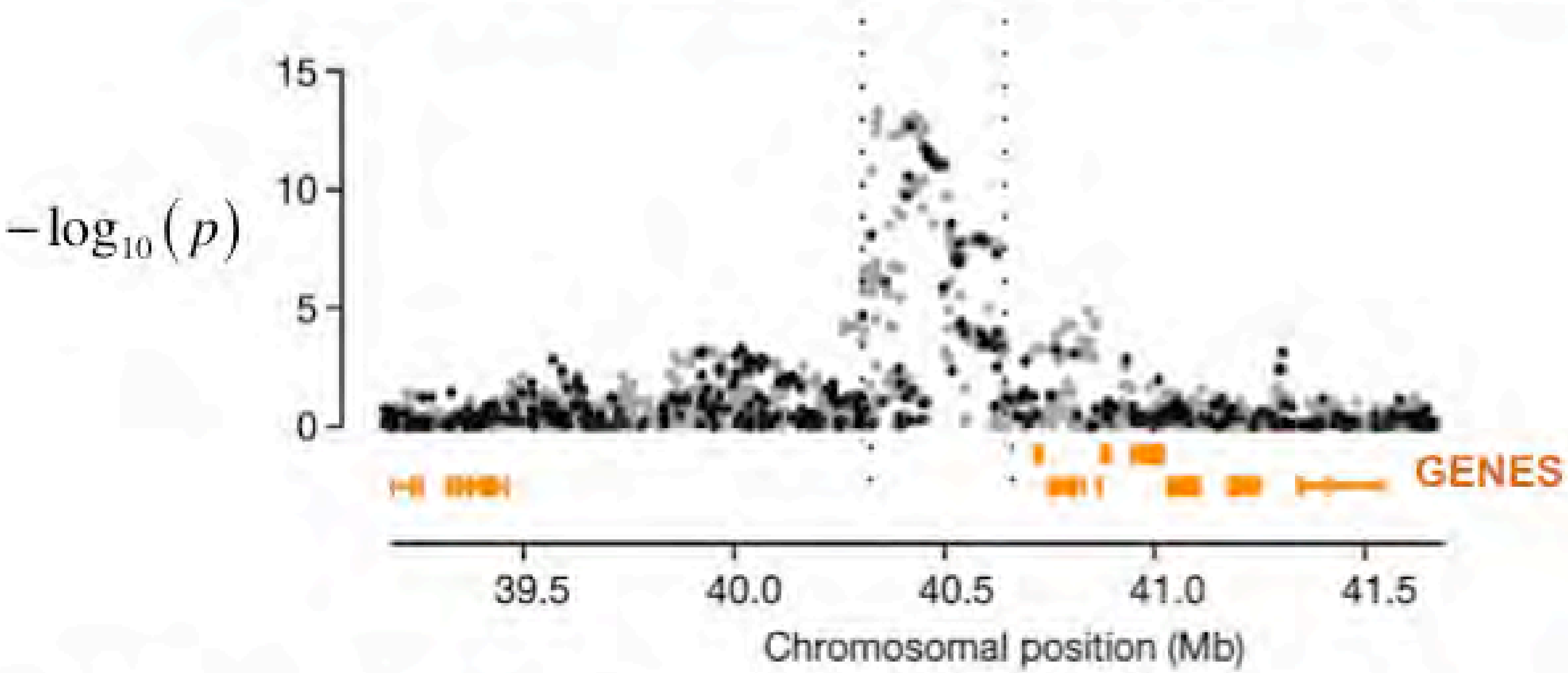
CENTER FOR MOLECULAR MEDICINE AND GENETICS

WAYNE STATE

School of Medicine

Why are we interested in gene regulatory variants?

genome-wide association hit for Crohn’s disease (from WTCCC)



- Much of the key functional variation is due to changes in gene regulation
- **Predicting the impact of genetic variation on gene regulatory sequences remains a challange**

We now know that much of human phenotypic variation (including diseases) is due to the combined effects of many loci, an example is human height:

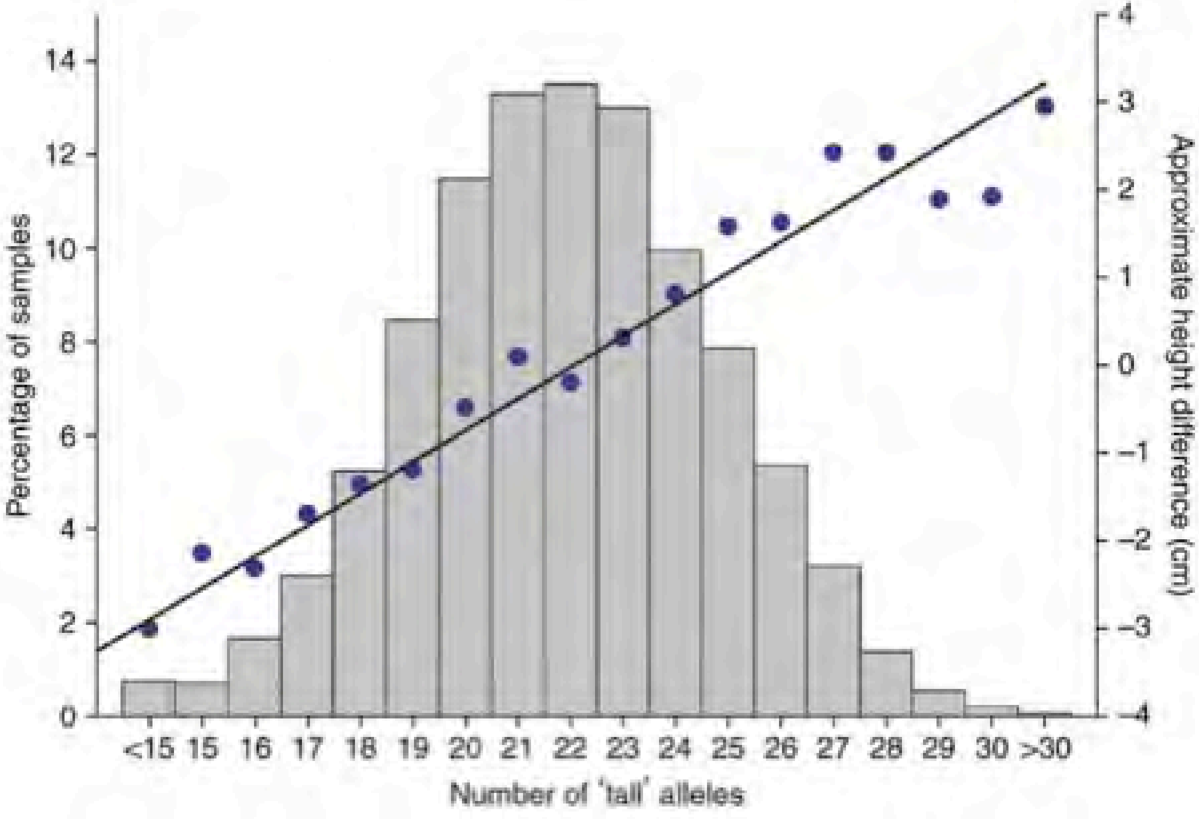
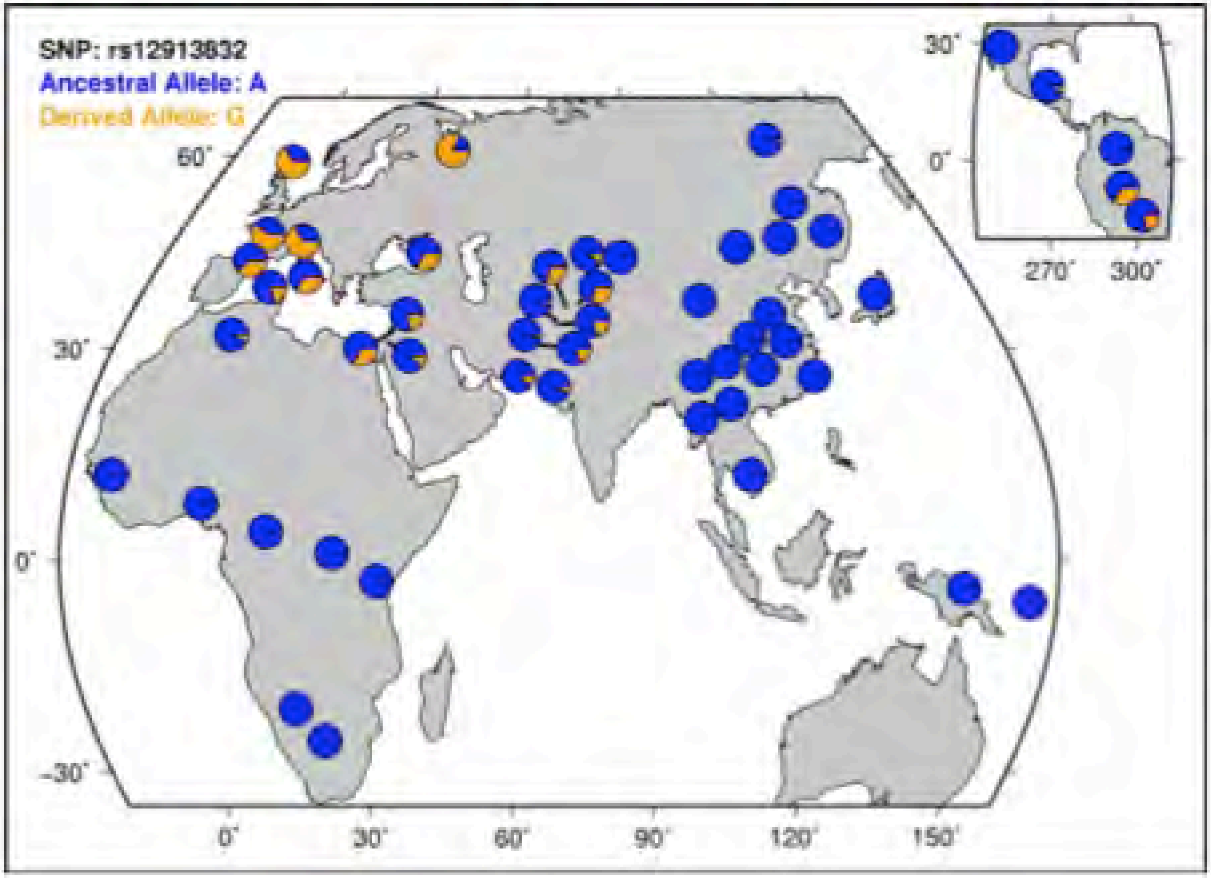


Figure from Weedon et al 2008

E.g., variation in human height is due to the combined influence of many hundreds of loci.

Each variant adds or subtracts just a few mm or less to expected height.

Some alleles perhaps conferred a **selective advantage** when moving to new environments



http://hgdp.uchicago.edu/

rs12913832 is associated with variation in eye color (GG increases the likelihood of blue eyes)

See related work by Tishkoff, Coop, Sabeti, DiRienzo and many others

- Can we identify the sequences that actively regulate gene expression in any given cell type/condition?

- Can we identify genetic variation affecting gene regulation?

- How non-coding gene regulatory variants contribute to disease and complex traits?

CSH PRESS GENOME RESEARCH

Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data

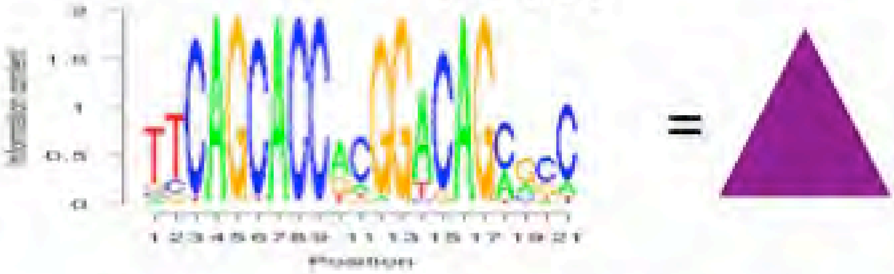
MAPPING TISSUE SPECIFIC REGULATORY SITES

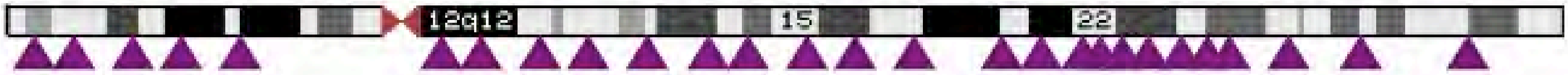
The CENTIPEDE approach

For every sequence motif known (JASPAR, TRANSFAC, PBM) or candidate word:

Step 1: Scan genome for all matches to the motif

NRSF



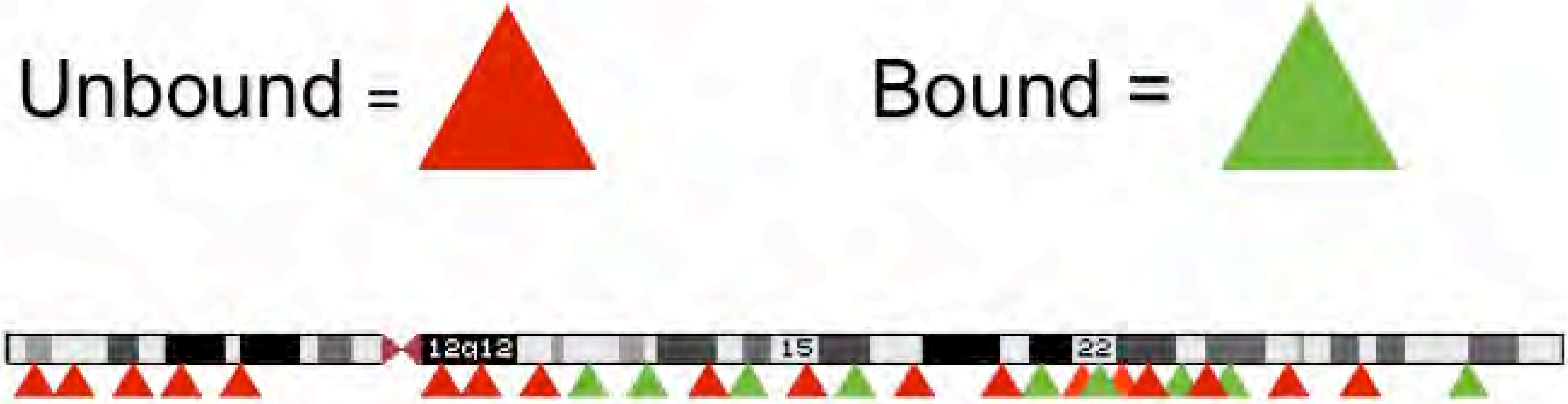


Pique-Regi, et al. GR 2011

The CENTIPEDE approach

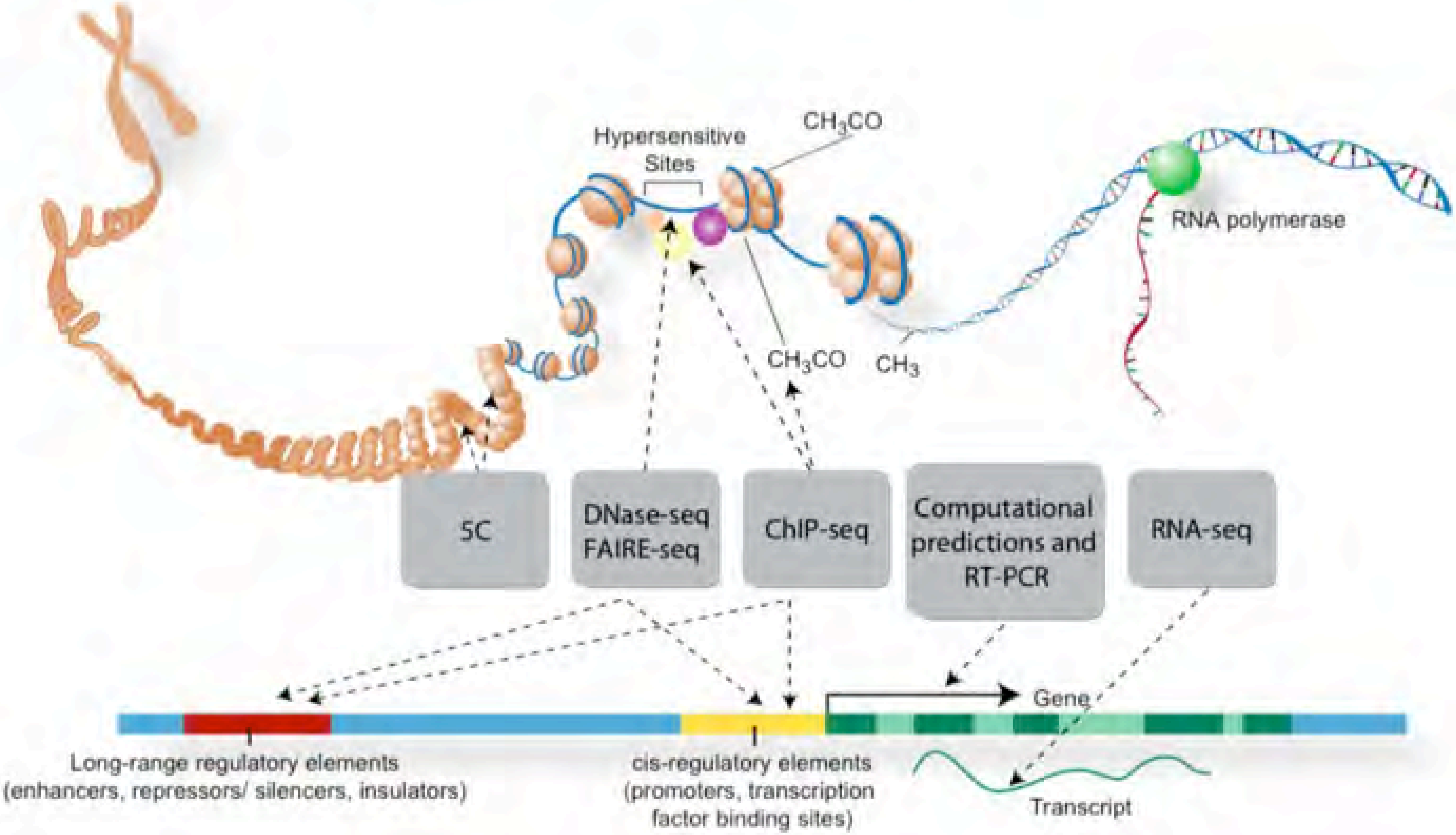
Using experimental data and existing genomic information

Step 2: Separate between bound and not bound instances for each TF using a mixture model



Each TF has its own specific model

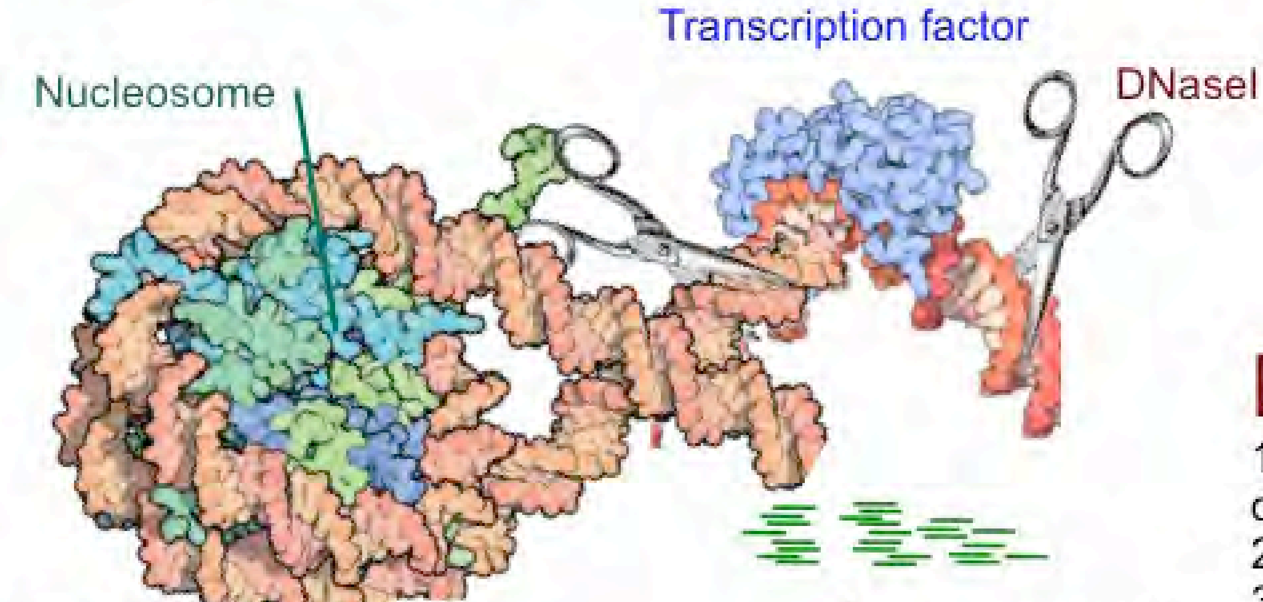
Functional genomics assays

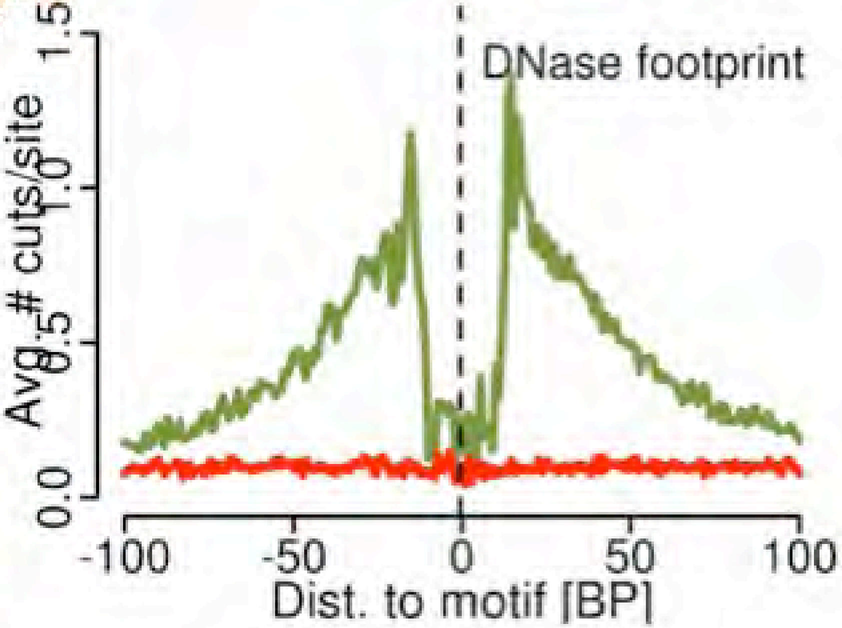

Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq
FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

ENCODE, Roadmap Epigenome and others. Image credit: ENCODE

DNaseI footprinting

Galas and Schmitz. (1979)

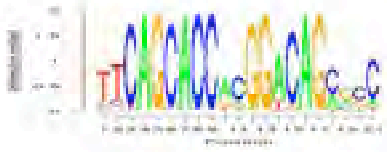




DNase-seq

1. DNaseI cuts preferentially open DNA
2. Sequence DNA fragments
3. Map to the genome
4. Fit CENTIPEDE models

NRSF motif:

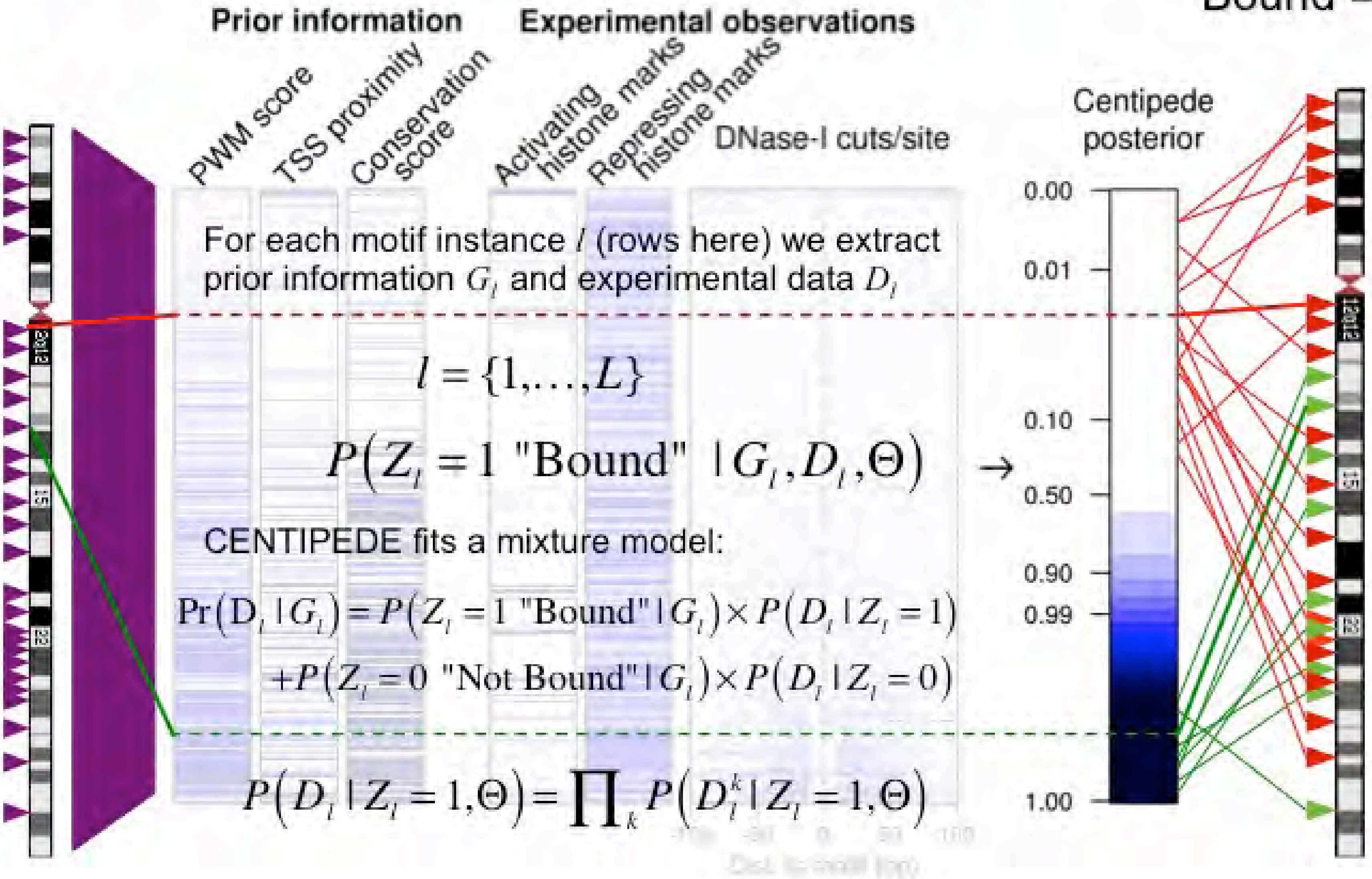


See also:
Boyle et al. (2008)
Hesselberth et al. (2009)
Chen et al. (2010)
Boyle et al. (2011)
Pique-Regi et al. (2011)

CENTIPEDE approach

Unbound = ▲

Bound = ▲



For each motif instance l (rows here) we extract prior information G_l and experimental data D_l

$$l = \{1,\ldots,L\}$$

$$P\left(Z_l = 1 \text{ "Bound" } \mid G_l, D_l, \Theta\right) \rightarrow$$

CENTIPEDE fits a mixture model:

$$\Pr\left(D_l \mid G_l\right) = P\left(Z_l = 1 \text{ "Bound"} \mid G_l\right) \times P\left(D_l \mid Z_l = 1\right) + P\left(Z_l = 0 \text{ "Not Bound"} \mid G_l\right) \times P\left(D_l \mid Z_l = 0\right)$$

$$P\left(D_l \mid Z_l = 1, \Theta\right) = \prod_k P\left(D_l^k \mid Z_l = 1, \Theta\right)$$

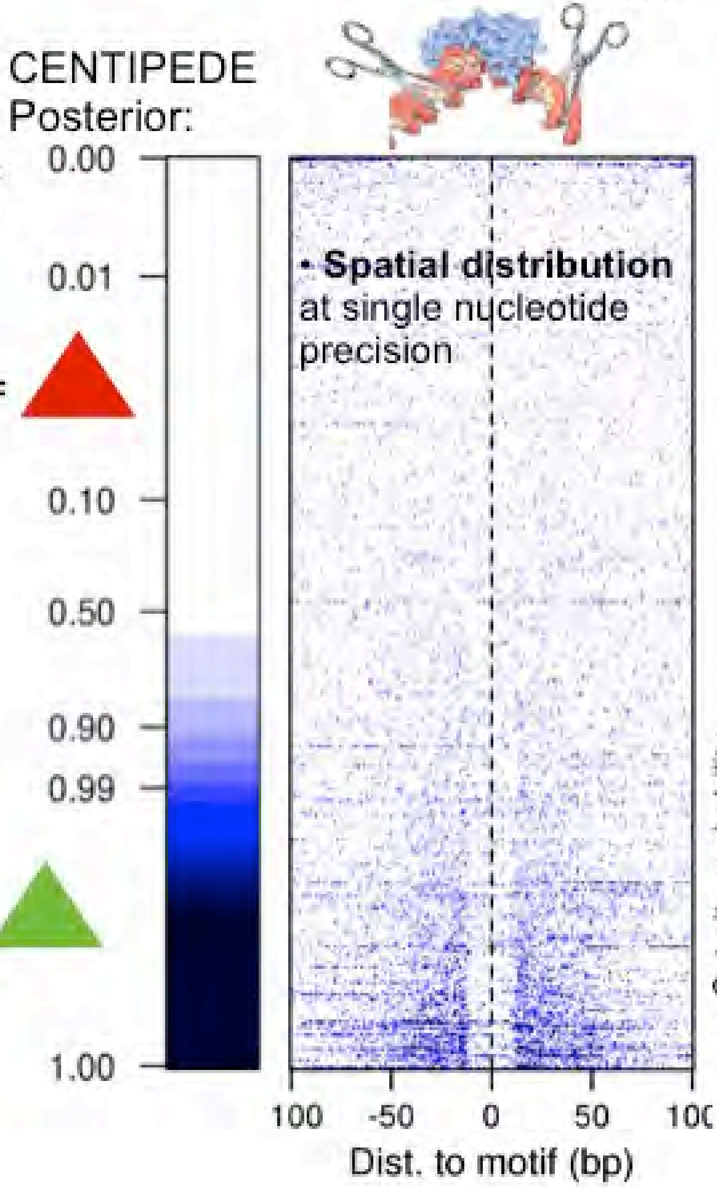
The CENTIPEDE model (e.g.,NRSF)

Dnase-seq

- **Each row is a NRSF** motif instance

Unbound = ▲

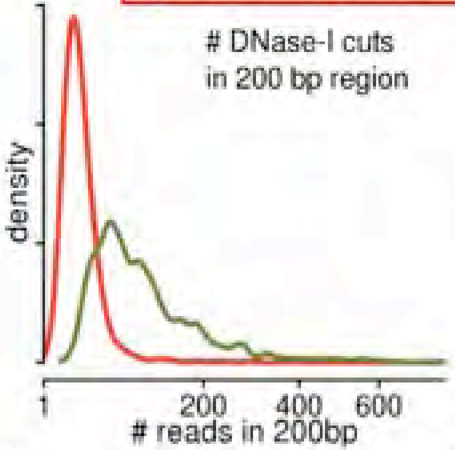
Bound = ▲



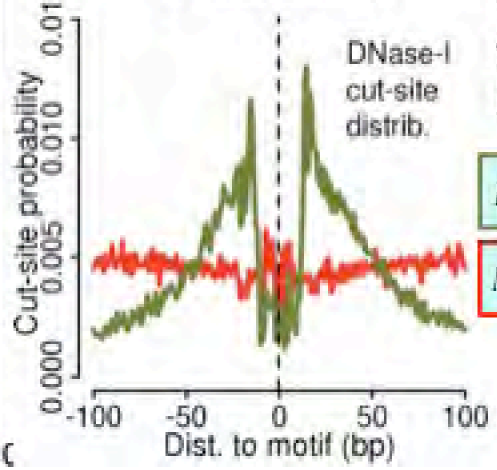
- **Spatial distribution** at single nucleotide precision

$$P(R_l \mid Z_l = 1) = NegativeBinomial(R_l \mid \alpha_1, \tau_1)$$

$$P(R_l \mid Z_l = 0) = NegativeBinomial(R_l \mid \alpha_0, \tau_0)$$



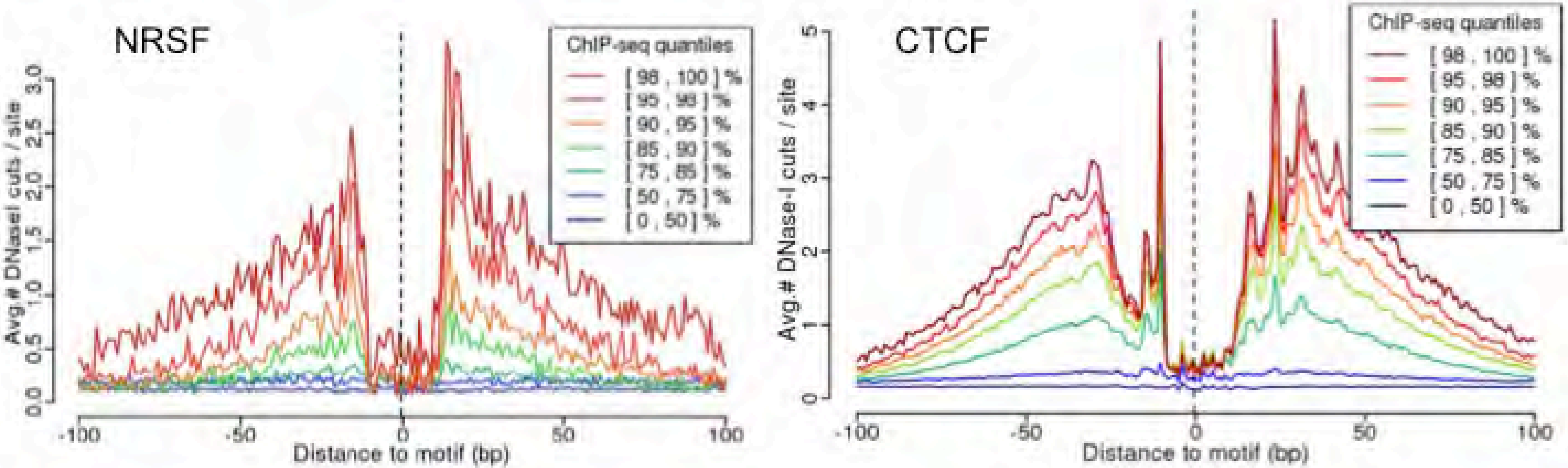
- **Specific footprint** for each factor: NRSF = 21 bp dip



$$Multinomial(X_{l,\beta} \mid R_l, \{\lambda_1, \ldots, \lambda_S\})$$

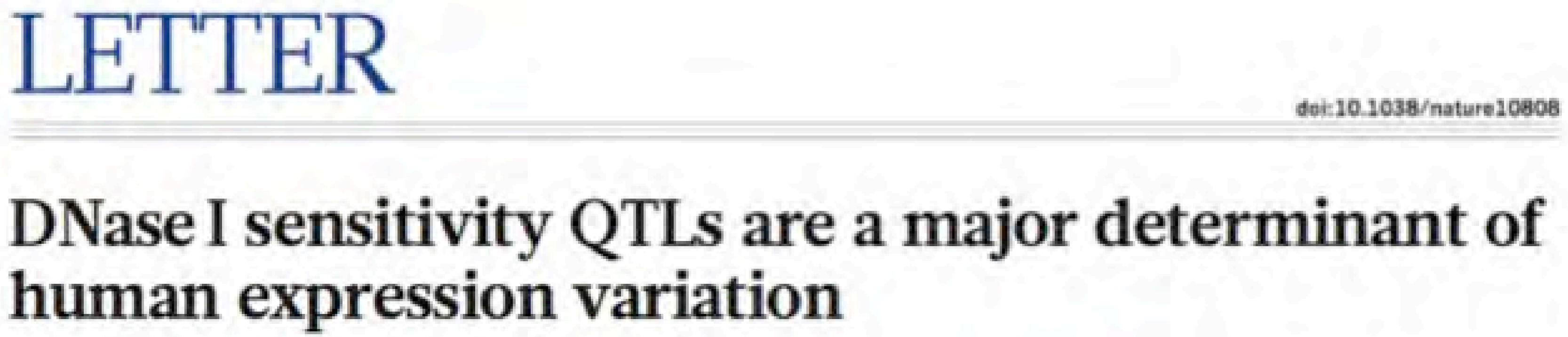
$$Multinomial(X_{l,\beta} \mid R_l, \{1/S, \ldots, 1/S\})$$

DNase-seq read depth also provides quantitative measurement of TF binding



230 PWMs + 49 novel motifs
830,000 Binding sites in ***1 assay***

Pique-Regi, et al. GR 2011, Data thanks to ENCODE



LETTER

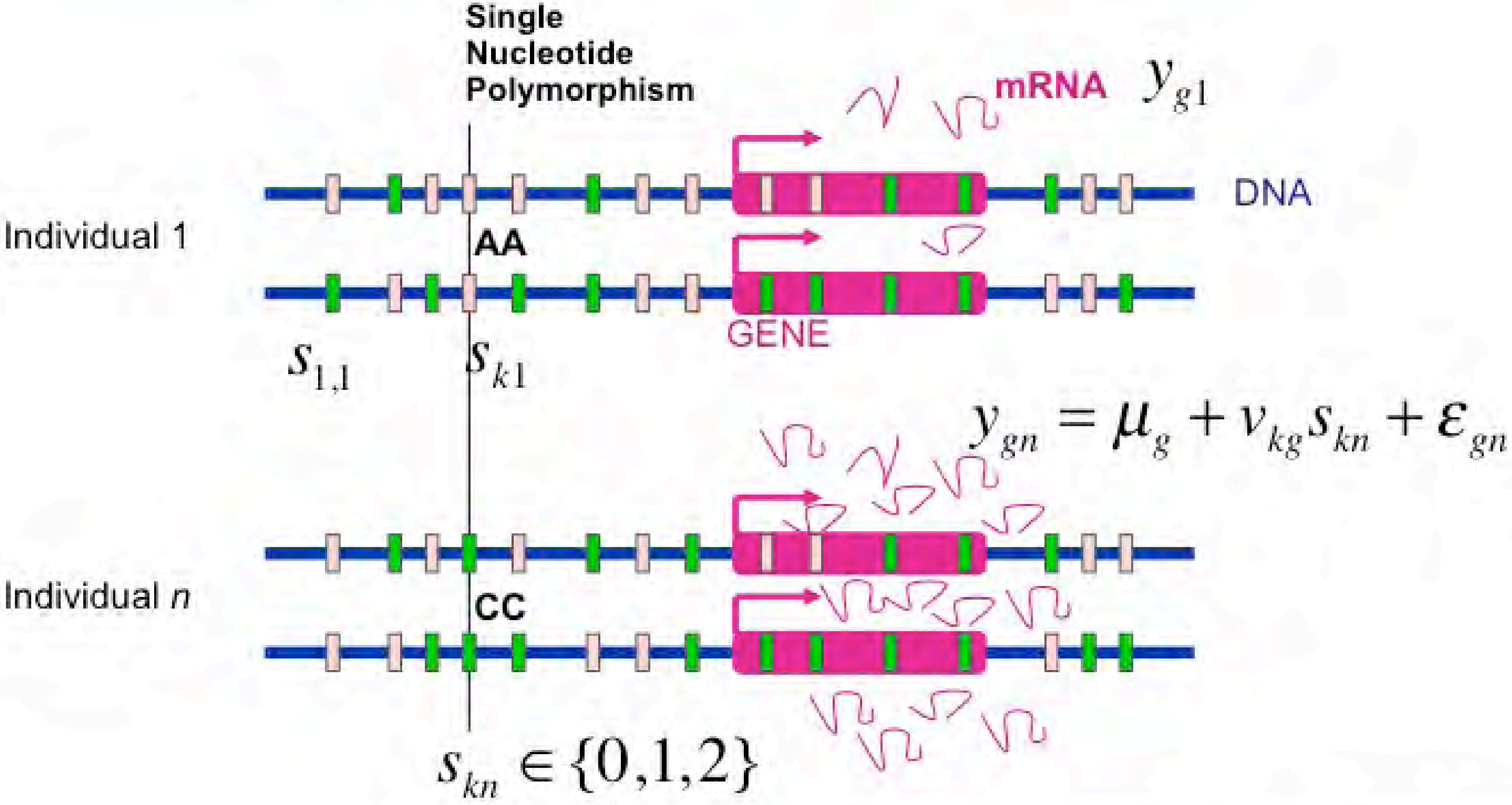
doi:10.1038/nature10808

DNase I sensitivity QTLs are a major determinant of human expression variation

FUNCTIONAL IMPACT OF REGULATORY VARIANTS

eQTLs: expression Quantitative Trait Loci

link genetic variation to changes in gene regulation

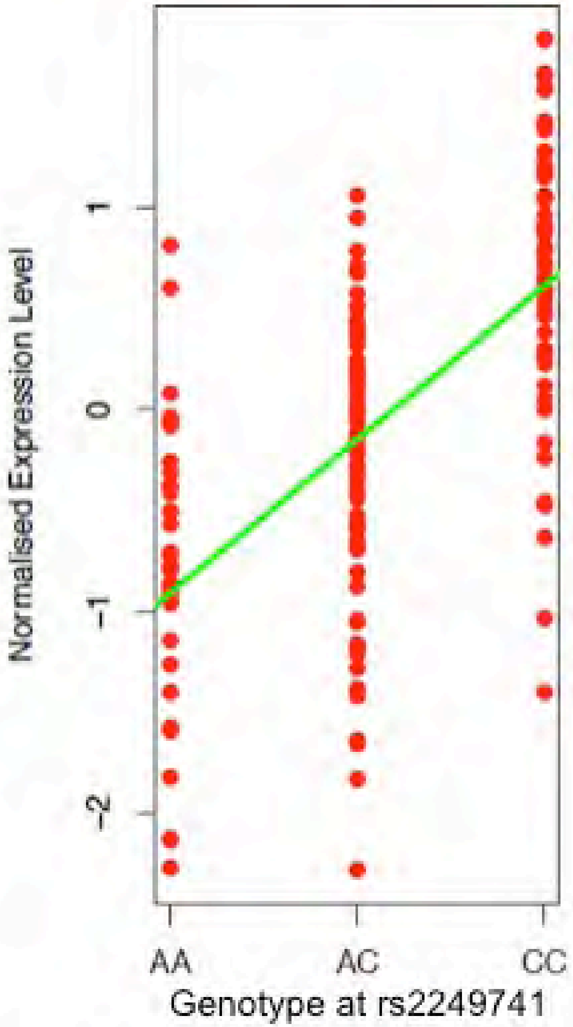


$$y_{gn} = \mu_g + v_{kg} s_{kn} + \varepsilon_{gn}$$

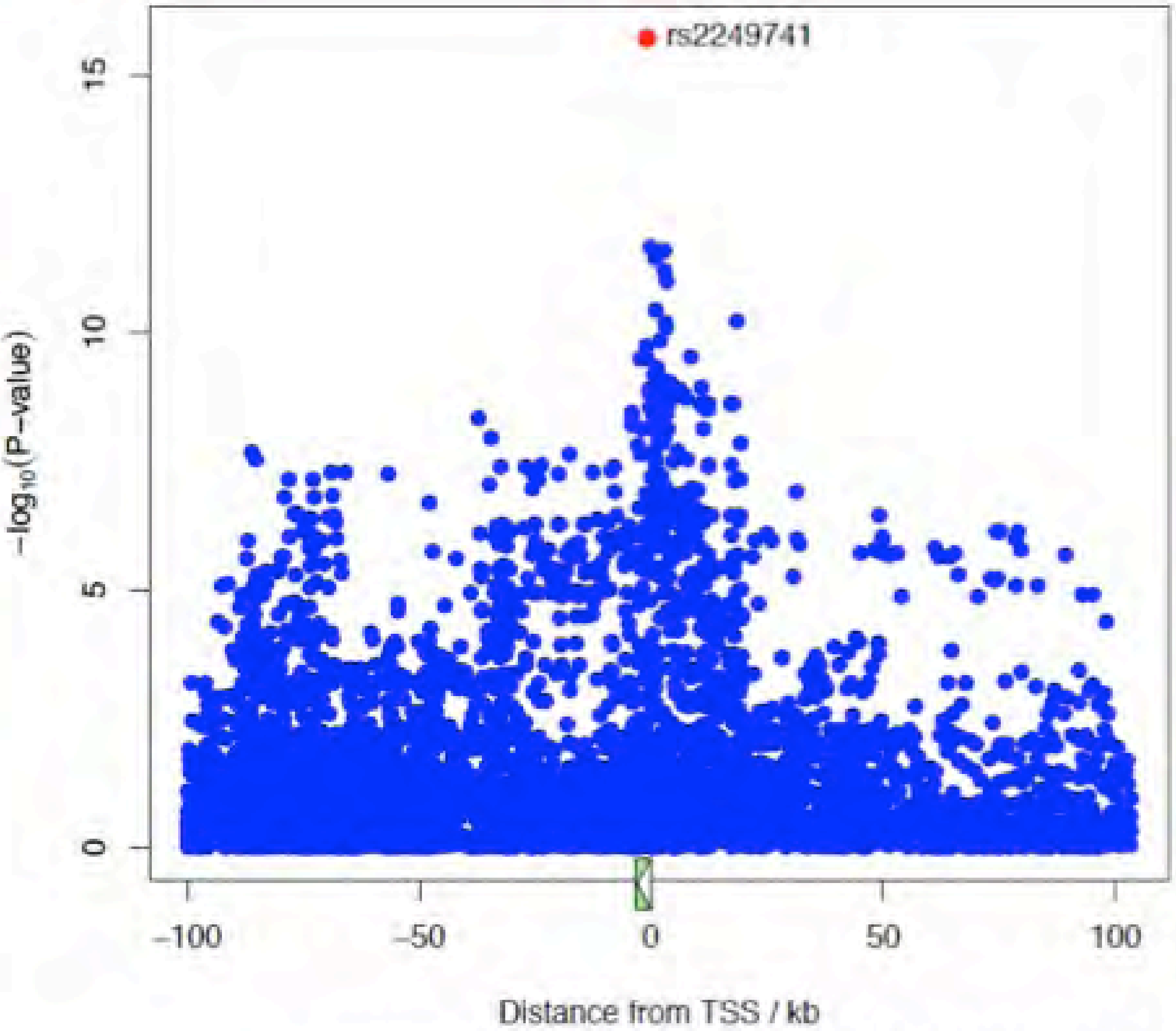
$$s_{kn} \in \{0,1,2\}$$

Related work by Leonid Kruglyak, Manolis Dermitzakis, Vivian Cheung, Eric Schadt, and others.

Example: SNPs associated with HLA-C expression

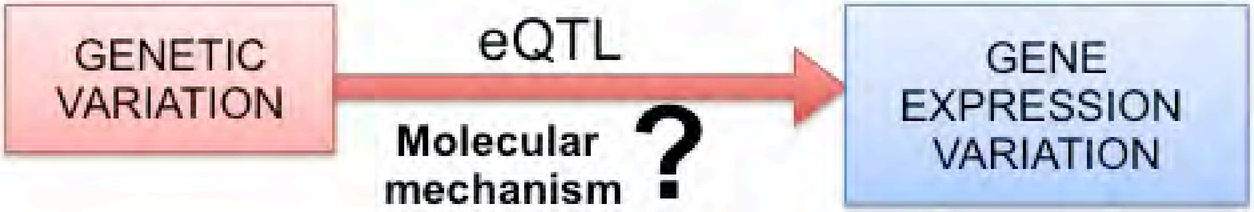


[HapMap CEU data; expression data from Stranger 2007]



same SNPs associated with HIV progression
[Goldstein group: Fellay et al (2007)]

Understanding the molecular mechanisms linking sequence changes to gene expression differences in humans

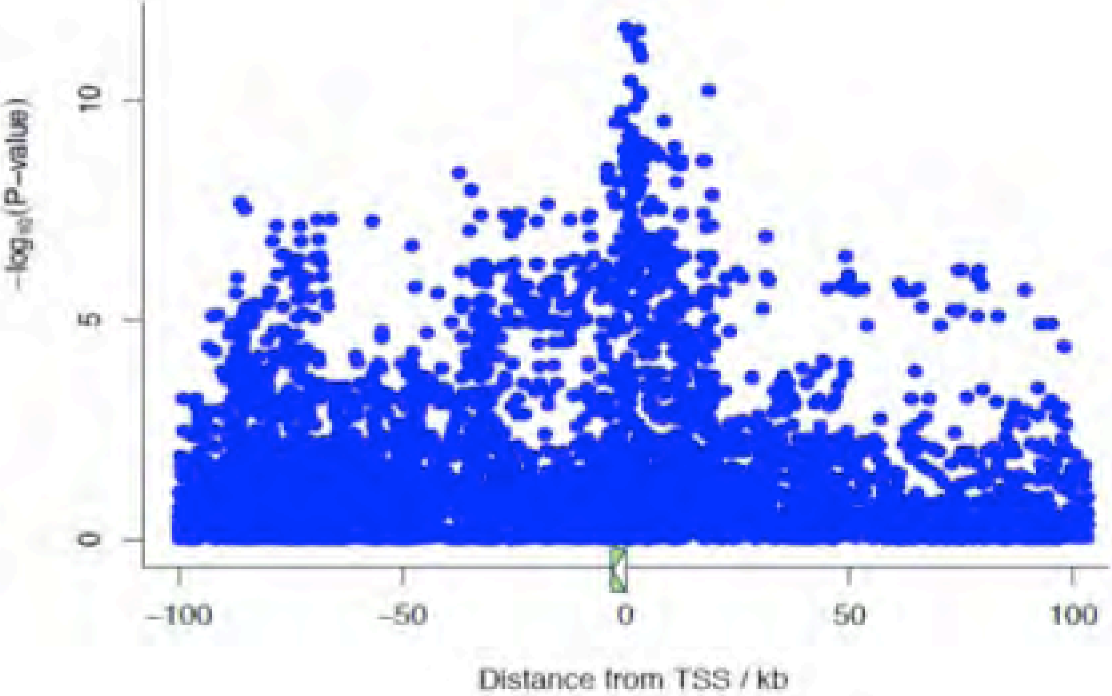


GENETIC VARIATION

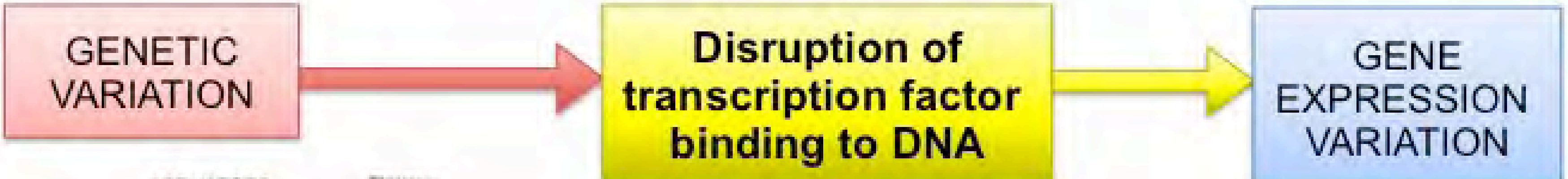
eQTL

Molecular mechanism ?

GENE EXPRESSION VARIATION



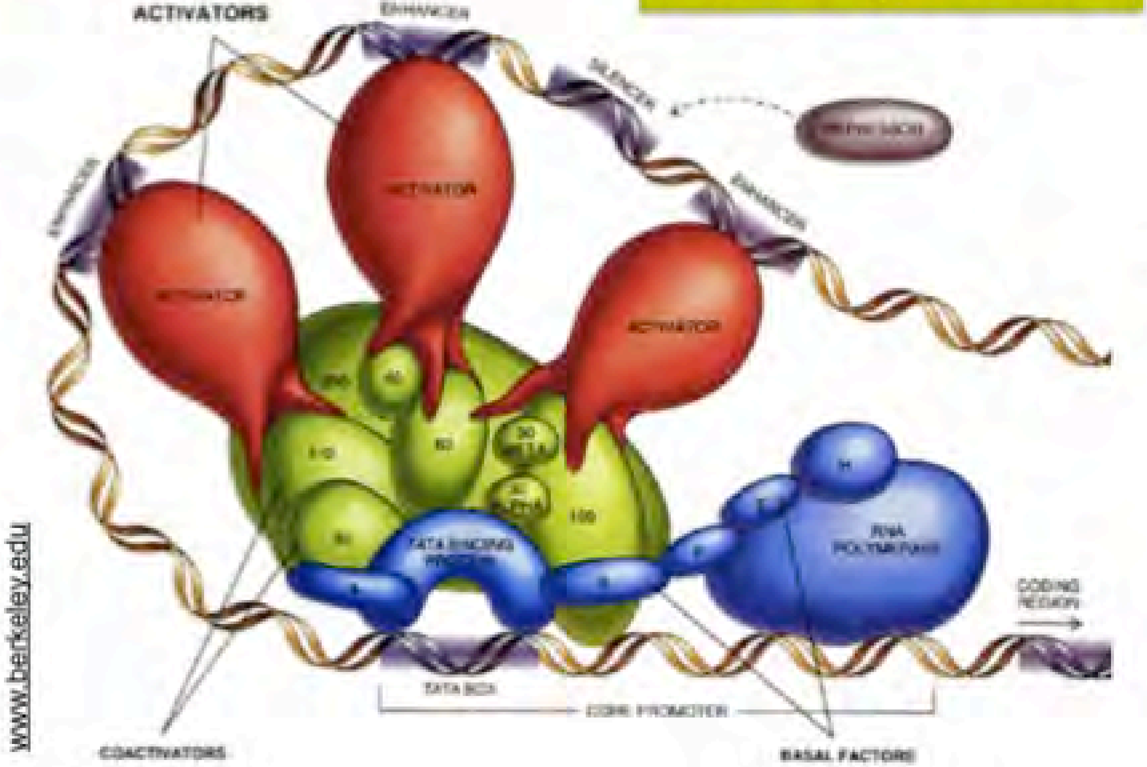
Understanding the molecular mechanisms linking sequence changes to gene expression differences in humans



GENETIC VARIATION

Disruption of transcription factor binding to DNA

GENE EXPRESSION VARIATION

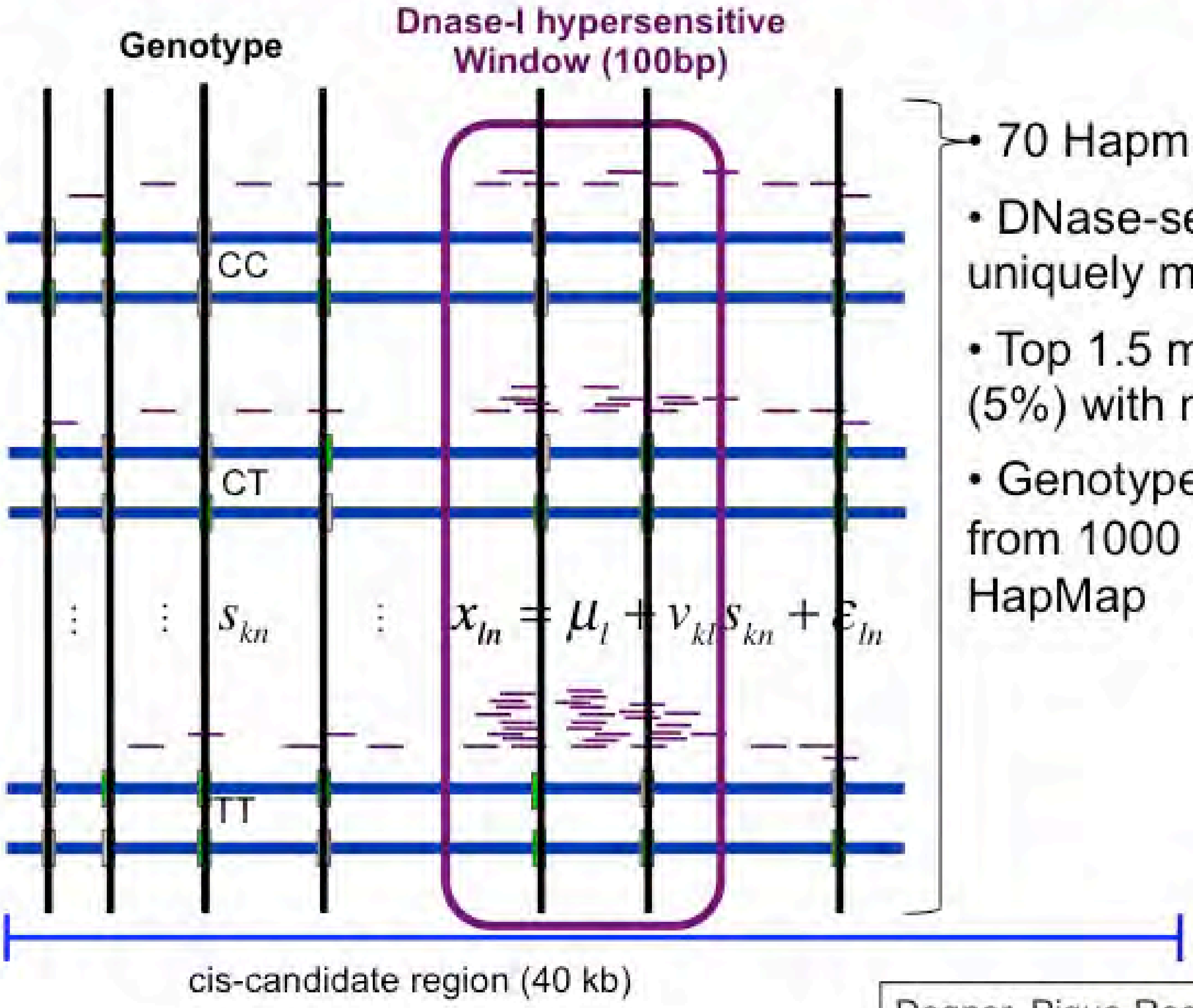


www.berkeley.edu

Alternative or co-occurring mechanisms:

- DNA methylation
- Histone positioning
- Pol II (CTD domain)
- Chromatin modifications
- Lamina Associated Domains
- mRNA splicing factors
- mRNA degradation / stability
- …

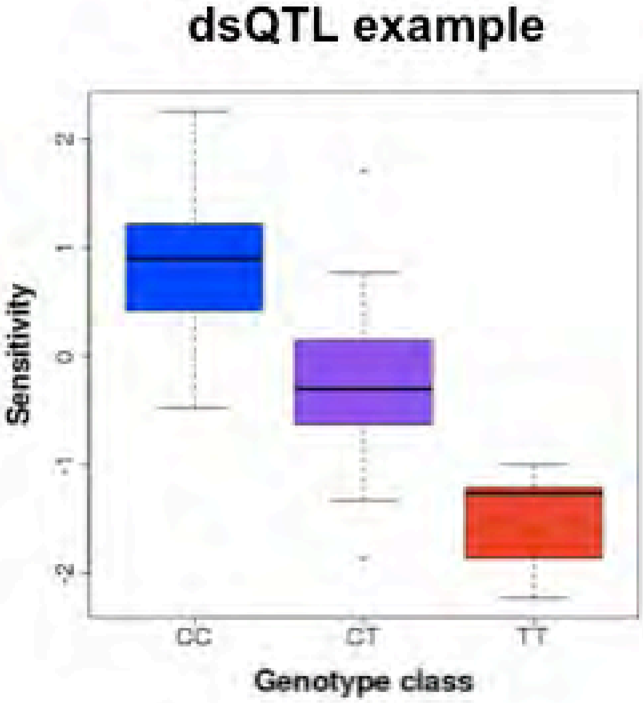
DNase I sensitivity QTLs → dsQTLs


Genotype
Dnase-I hypersensitive Window (100bp)
CC
CT
TT
s_{kn}
$x_{ln} = \mu_l + v_{kl} s_{kn} + \varepsilon_{ln}$
cis-candidate region (40 kb)

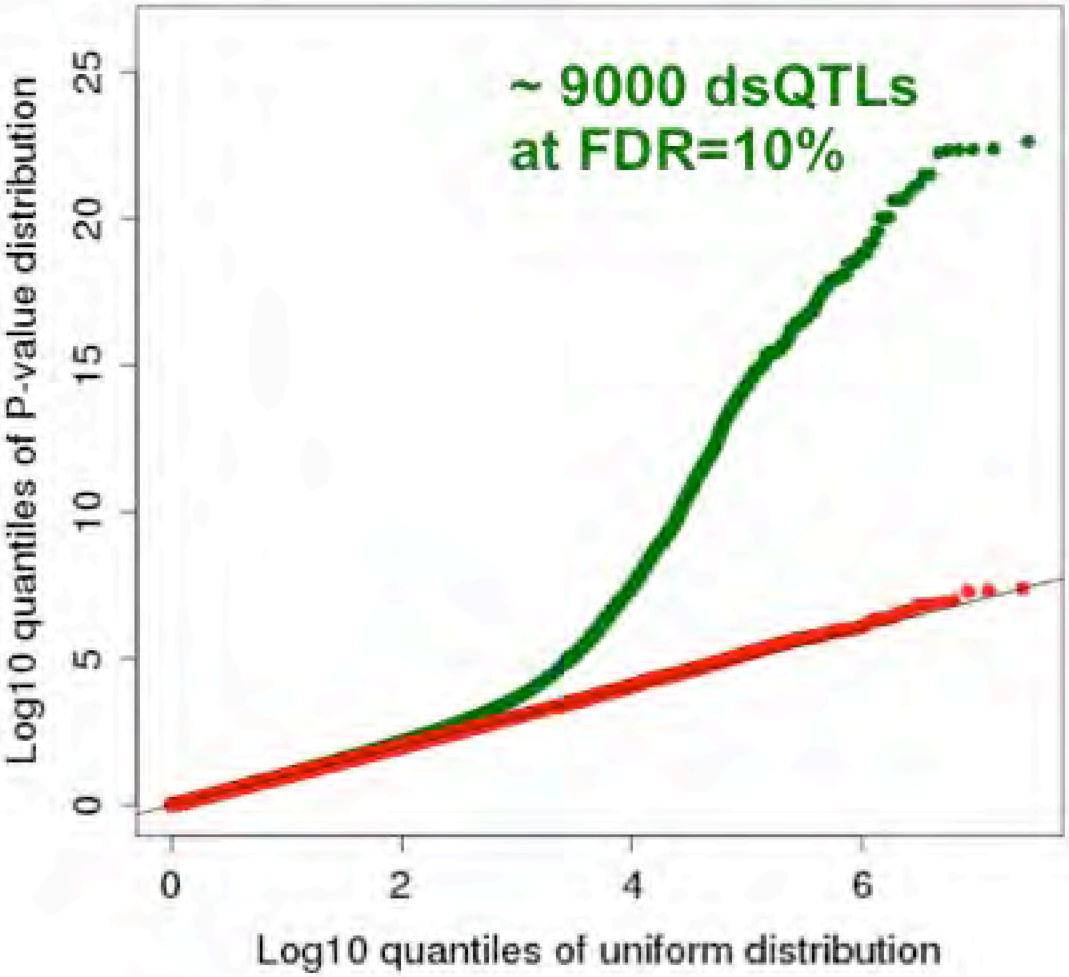
- 70 Hapmap individuals
- DNase-seq (~3.1 billion uniquely mapped reads)
- Top 1.5 milion windows (5%) with most reads
- Genotype information from 1000 Genomes and HapMap

Degner, Pique-Regi, et al. Nature 2012

Large numbers of dsQTLs

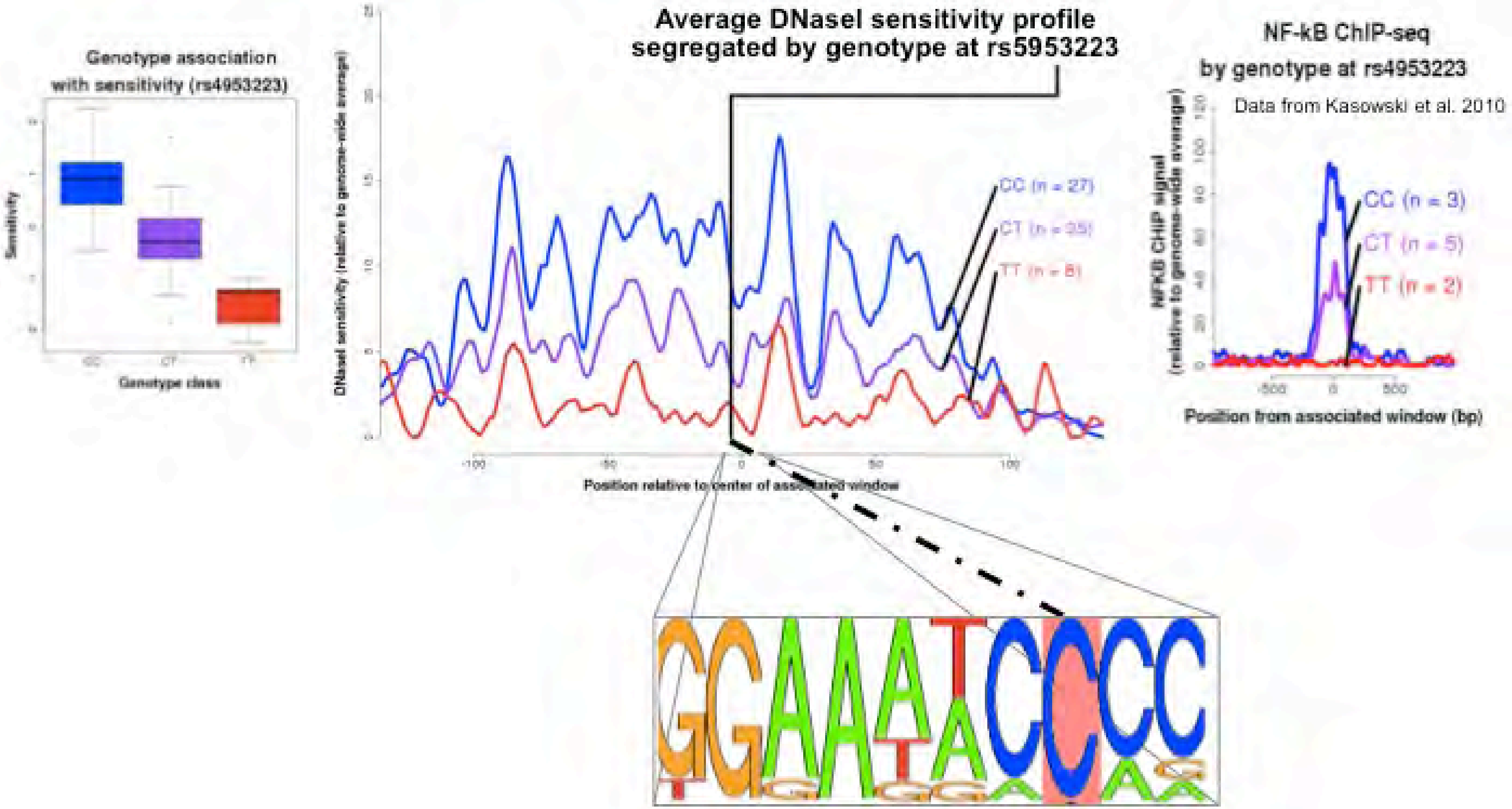


dsQTL example



~ 9000 dsQTLs
at FDR=10%

dsQTL example for NFKB binding site


Genotype association with sensitivity (rs4953223)

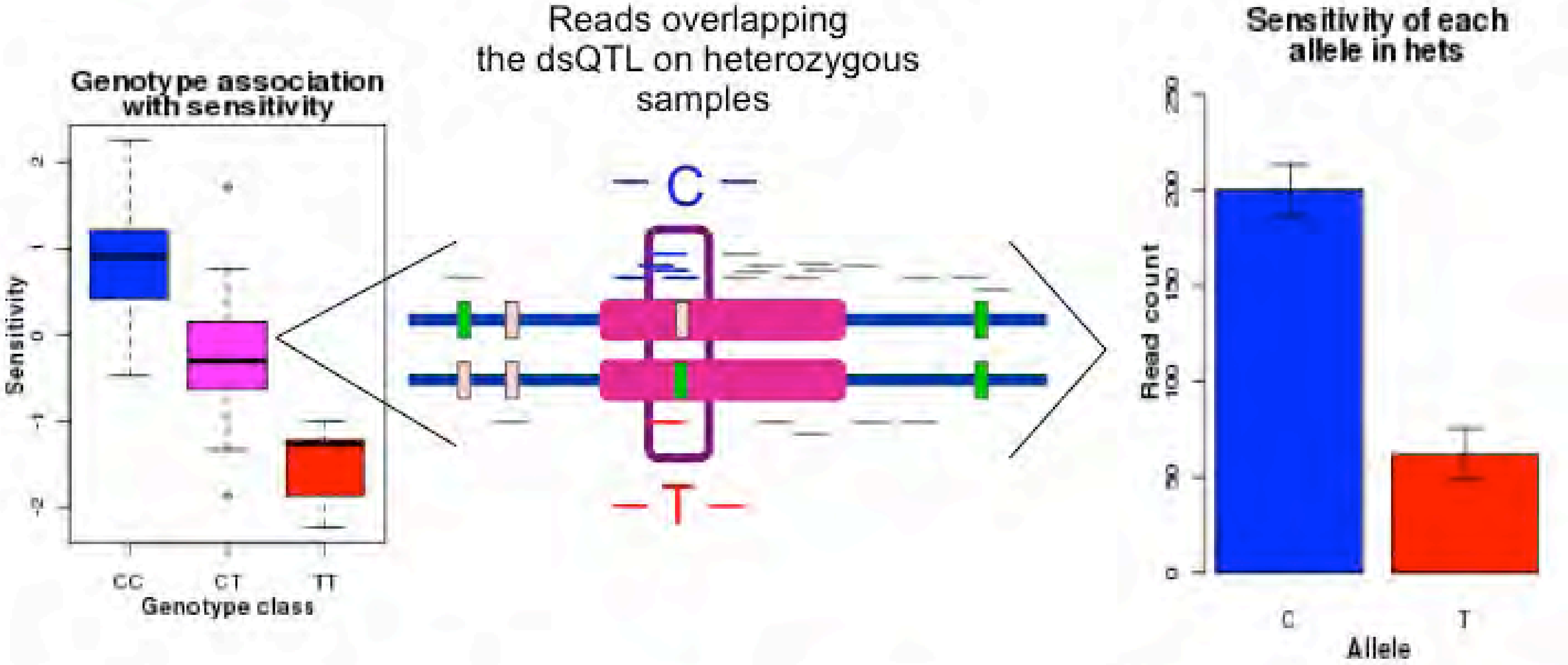
Average DNaseI sensitivity profile segregated by genotype at rs5953223

NF-kB ChIP-seq by genotype at rs4953223

Data from Kasowski et al. 2010

Degner, Pique-Regi, et al. Nature 2012

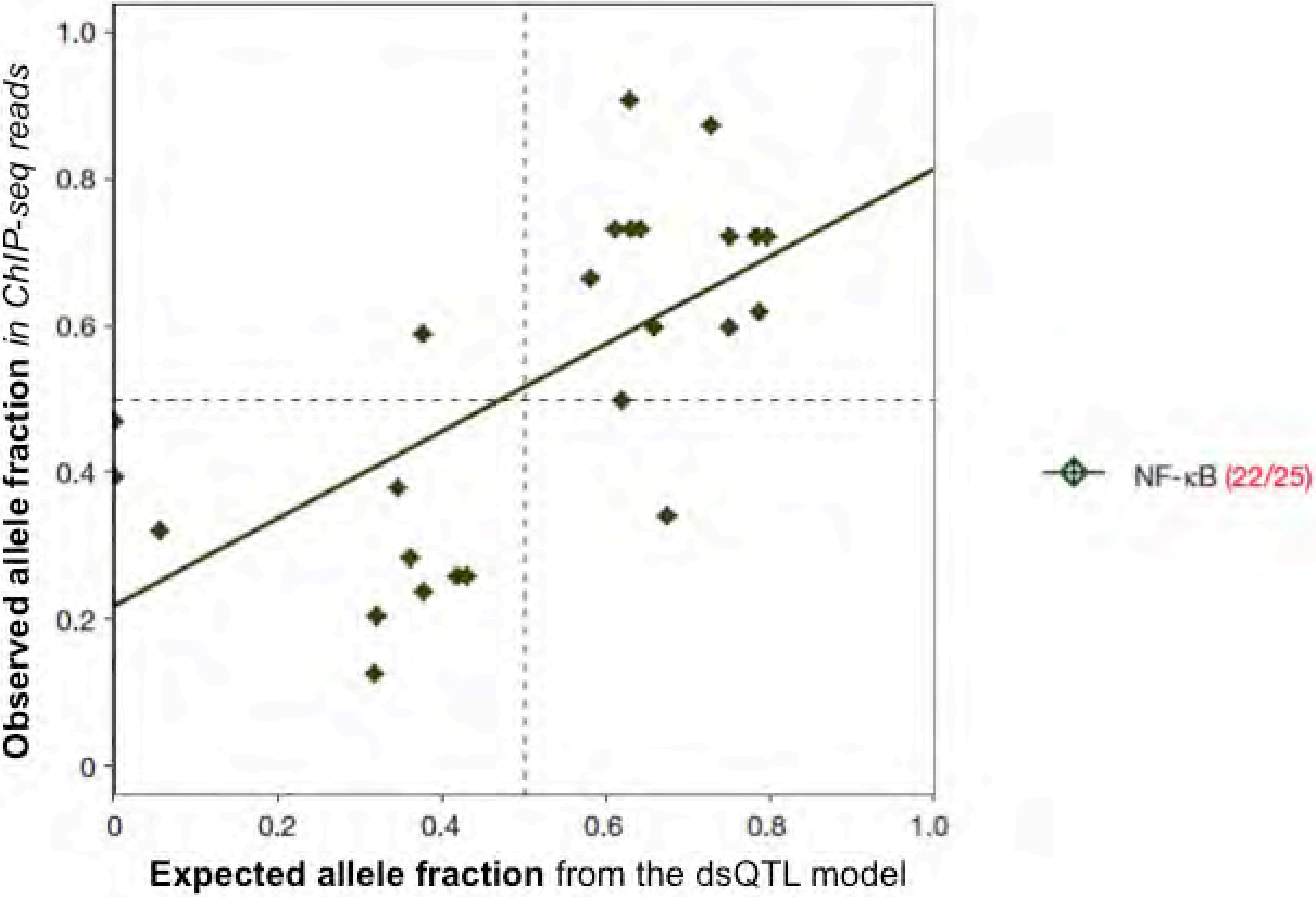
Allelic imbalances in sequencing data: e.g. DNase I sensitivity



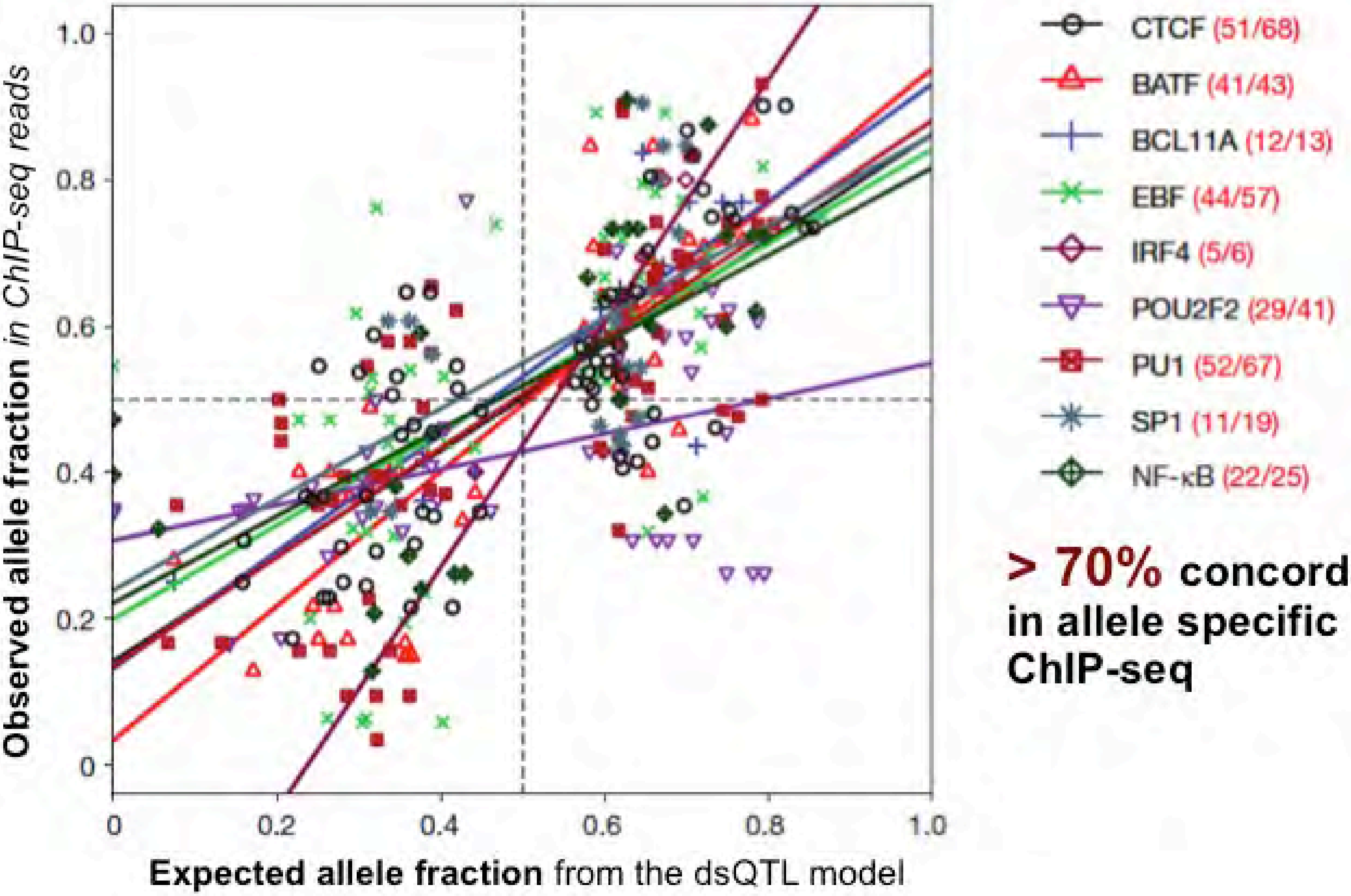
QTL-analysis from Degner, Pique-Regi, et al. *Nature* 2012
Check also poster RG50 (by Heejung Shim) for a new multi-scale analysis method

For ASB w/ DNase see also:
McDaniell et al. *Science* 2011
Reddy et al. *Nature* 2012
McVicker et al. *Sience* 2013

TF allele specific binding using ChIP-seq in one single individual

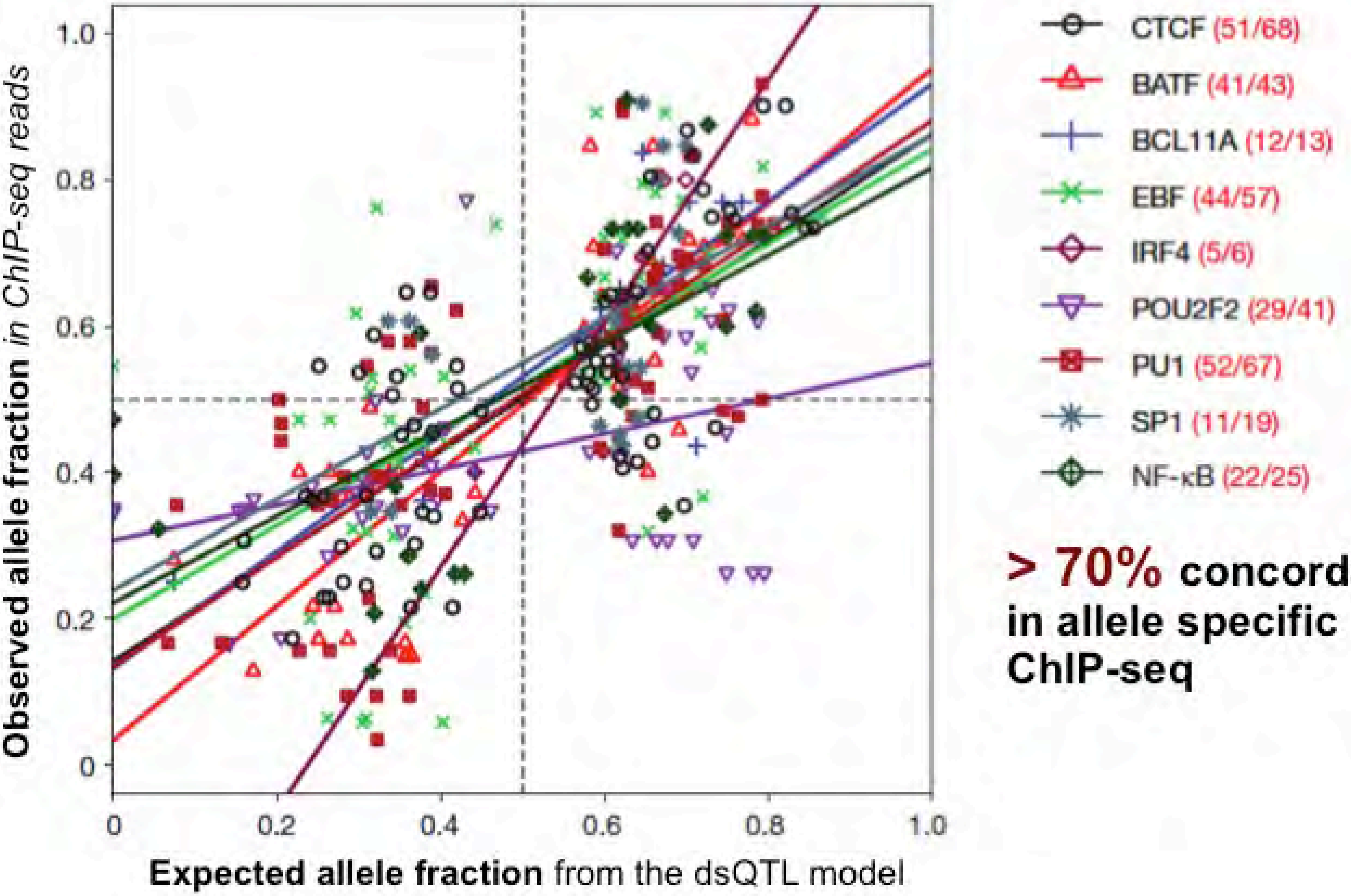


dsQTLs impact TF binding is also validated by allele specific ChIP-seq



> 70% concordance in allele specific ChIP-seq

dsQTLs impact TF binding is also validated by allele specific ChIP-seq

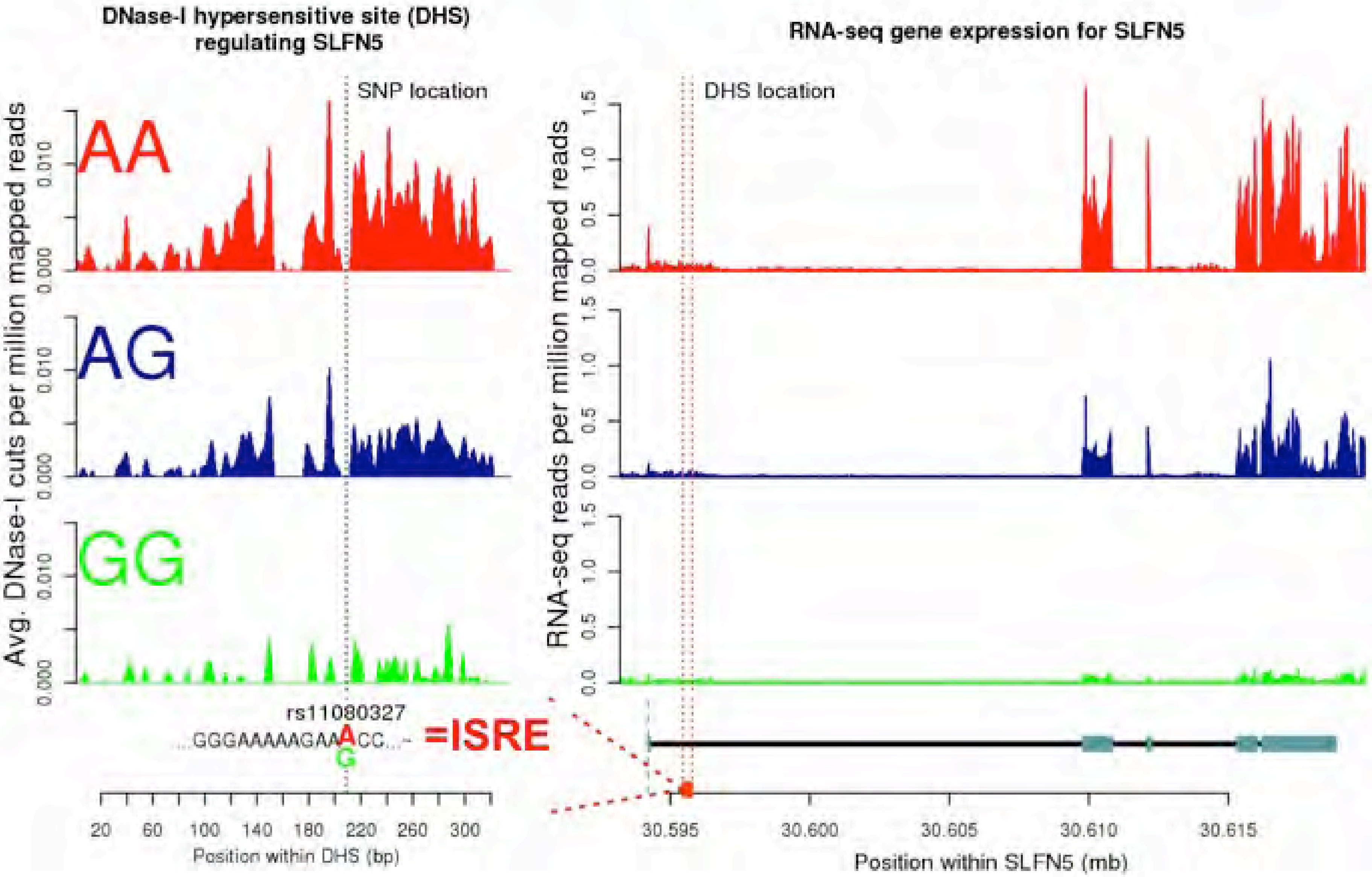


> 70% concordance in allele specific ChIP-seq

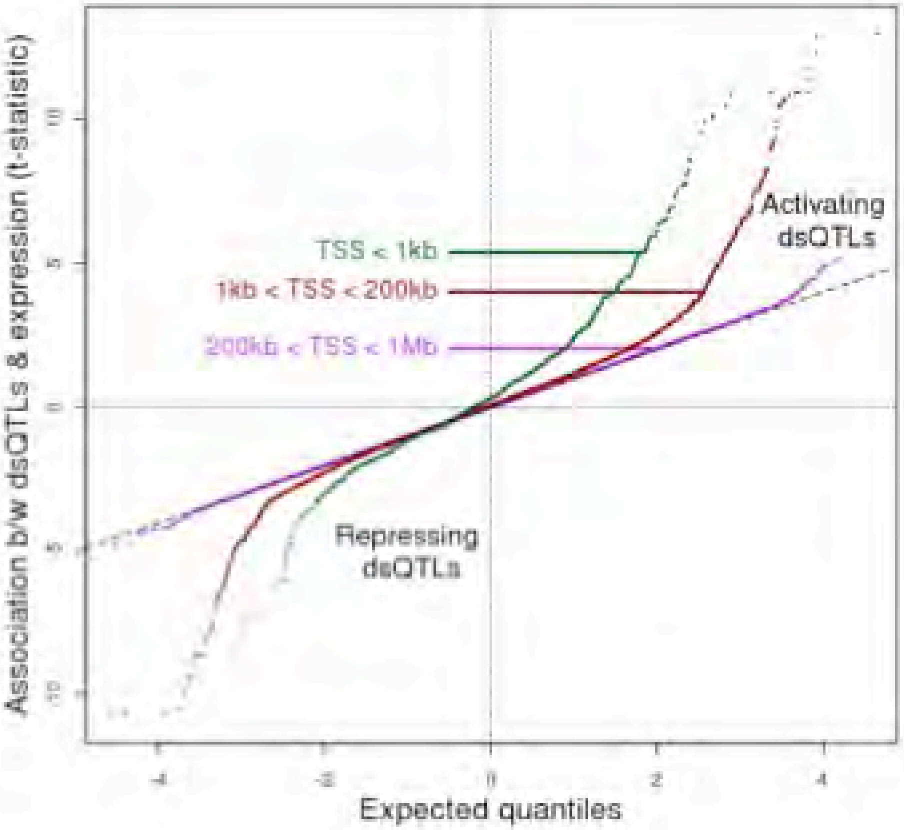
Are dsQTL SNPs associated with gene expression changes?

Yes!

dsQTLs are also eQTLs

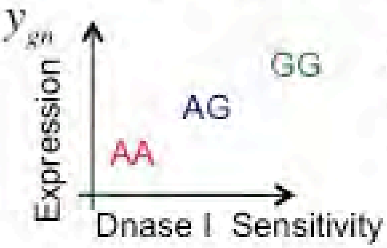


Large fraction of dsQTLs are eQTLs



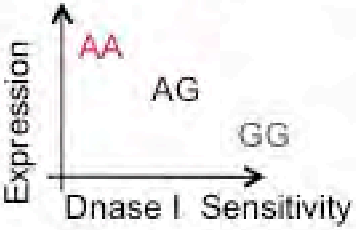
At FDR = 10%:

- 824 dsQTLs are eQTLs
- Most **(70%)** are **activating**



$$y_{gn} = \mu_g + v_{kg}s_{kn} + \varepsilon_{ln}$$
$$x_{ln} = \mu_l + v_{kl}s_{kn} + \varepsilon_{ln}$$

- Some are **repressing**

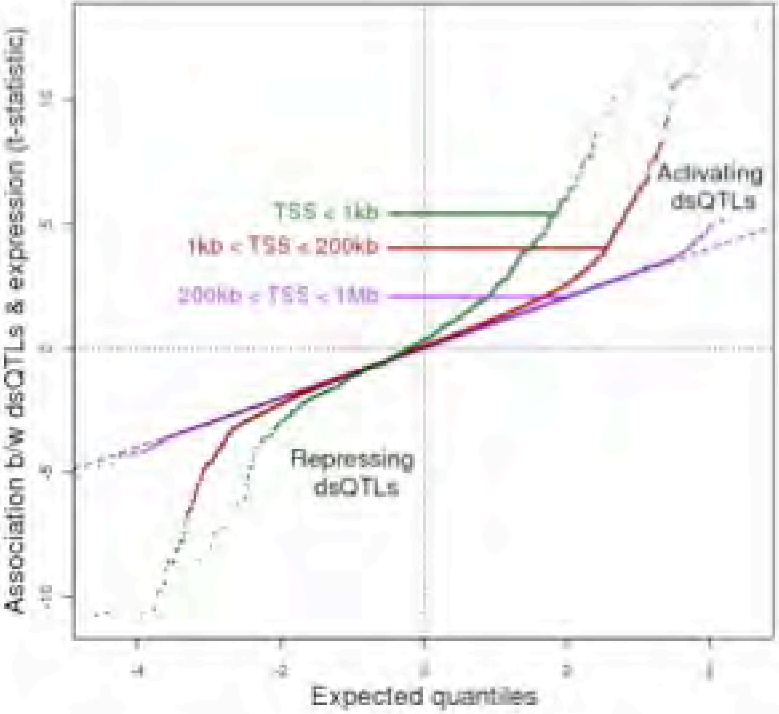


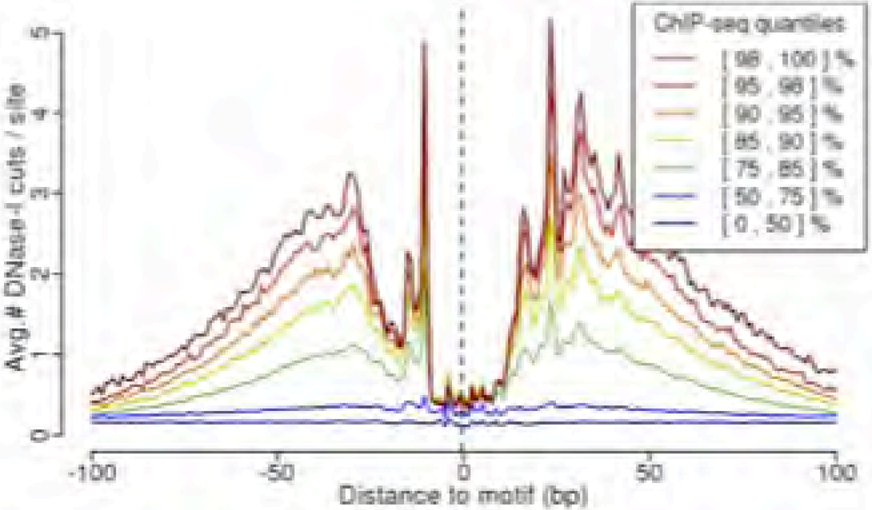
- After correcting for incomplete power, we estimate that ***55% of all eQTLs are also dsQTL***.

Summary I

→ dsQTLs give a molecular mechanism for cis-regulatory control of gene transcription

~50% of the eQTLs are estimated to be dsQTLs





→ dsQTLs tend to occur in DNase-seq footprints

>70% dsQTLs in ChIP-seq peaks are validated by ASE

→ DNase-seq footprints can localize key regulatory sequences for large set of transcription factors

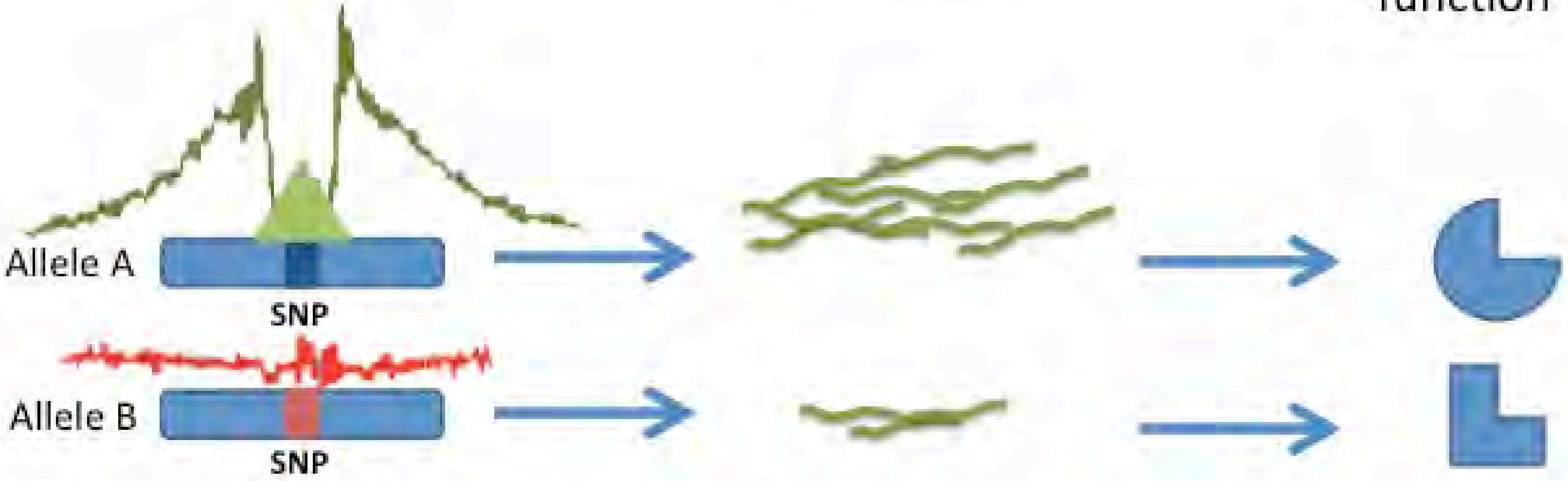
EXTENDING TO OTHER TISSUES

Identifying non-coding variants that have a function

Prediction

Validation

Downstream function

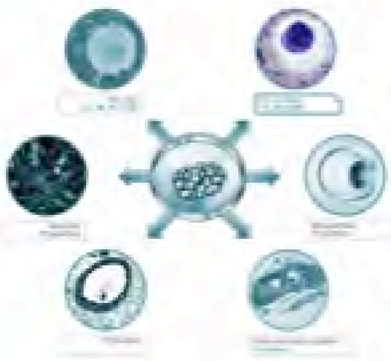


Regulatory sequence variants
Sequence + footprint model

Allele specific expression (ASE)
Allele specific binding (ASB)

Phenotype
GWAS studies

Running CENTIPEDE on > 600 tissues / cell-types (data from ENCODE and Roadmap Epigenome)



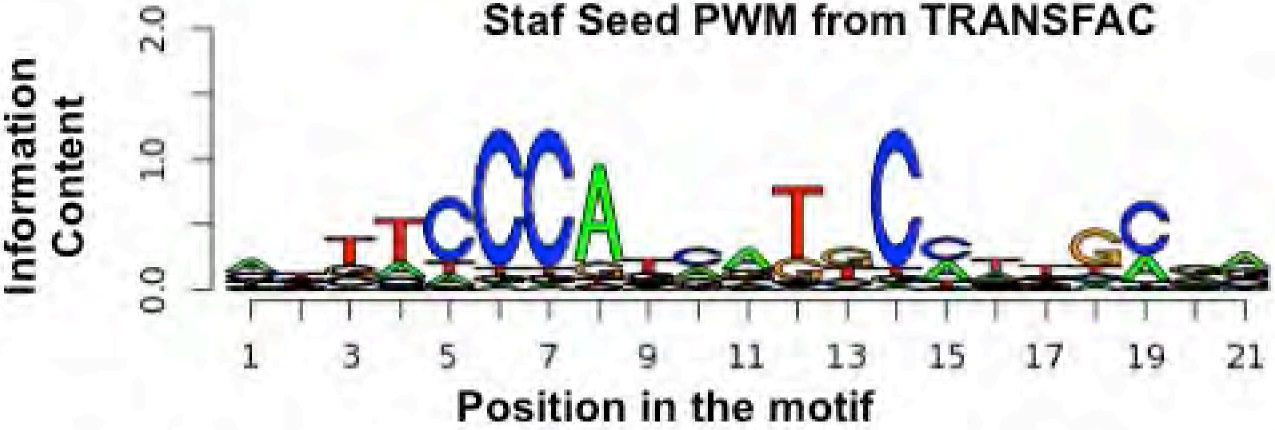
w/ Gregory Moyerbrailean

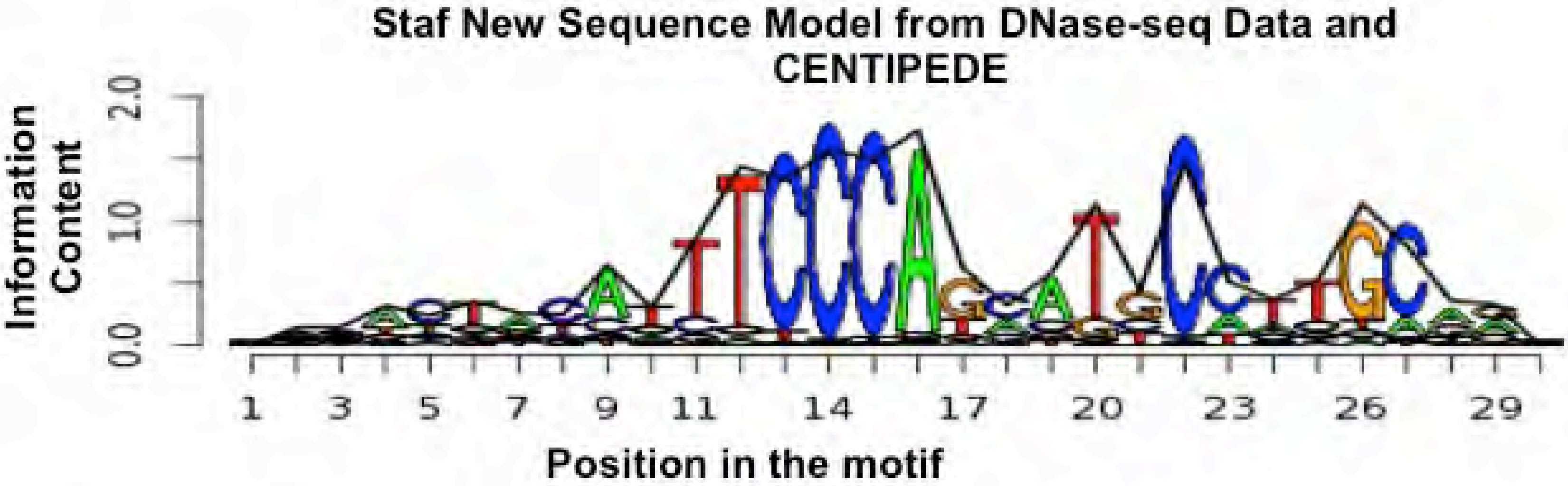
Learning a new motif with CENTIPEDE

- After an initial CENTIPEDE scan we select which motifs are actively used and on which cells.
- Then, **we relearn the sequence motif from the "active" sites** on homologous sequences not covered by the original motif (**excluding SNPs**)
- Scan the entire genome and also genetic variants from 1000 Genomes project and run CENTIPEDE again

w/ Gregory Moyerbrailean

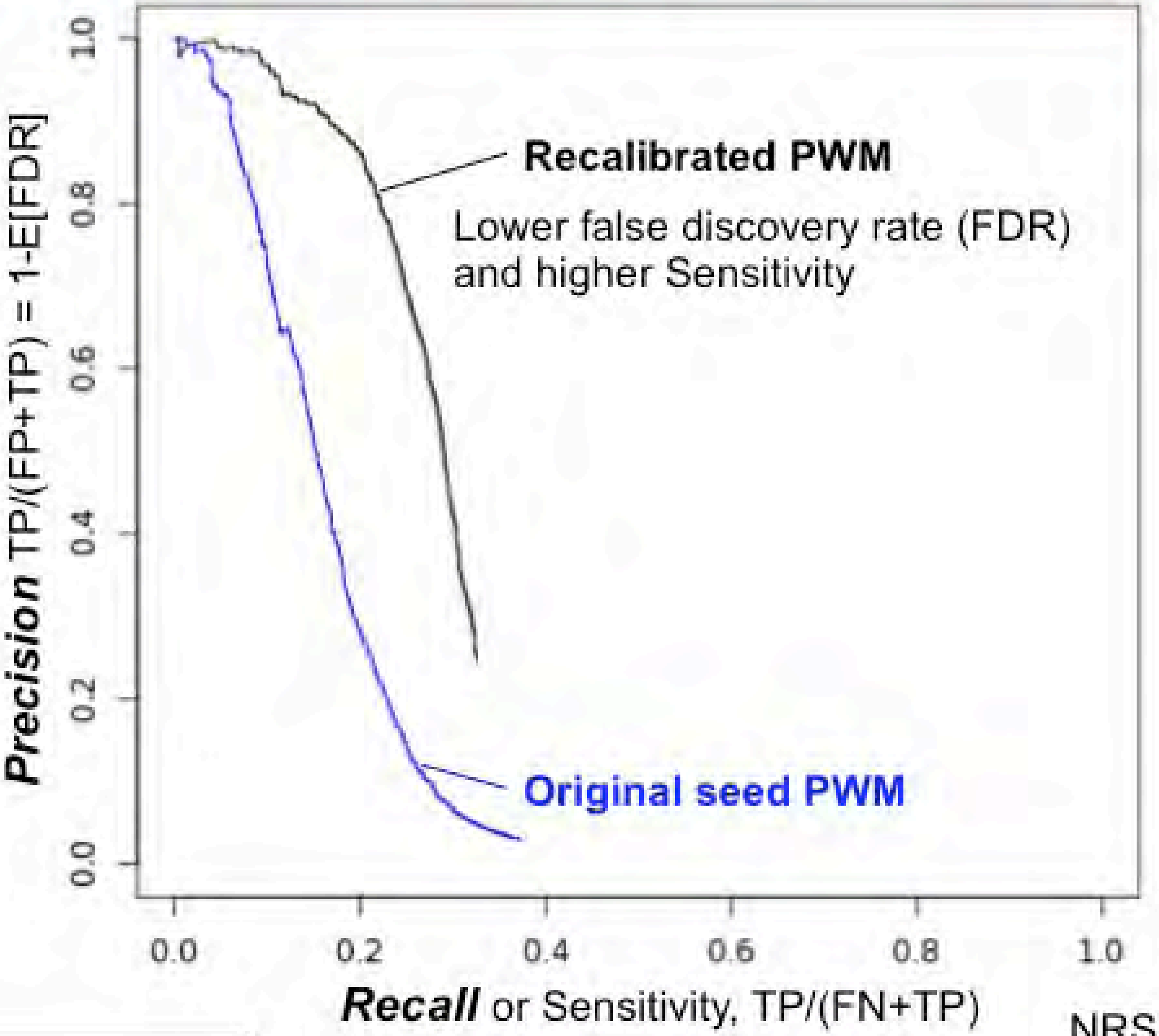
Recalibrated sequence models using CENTIPEDE footprint model





w/ Gregory Moyerbrailean

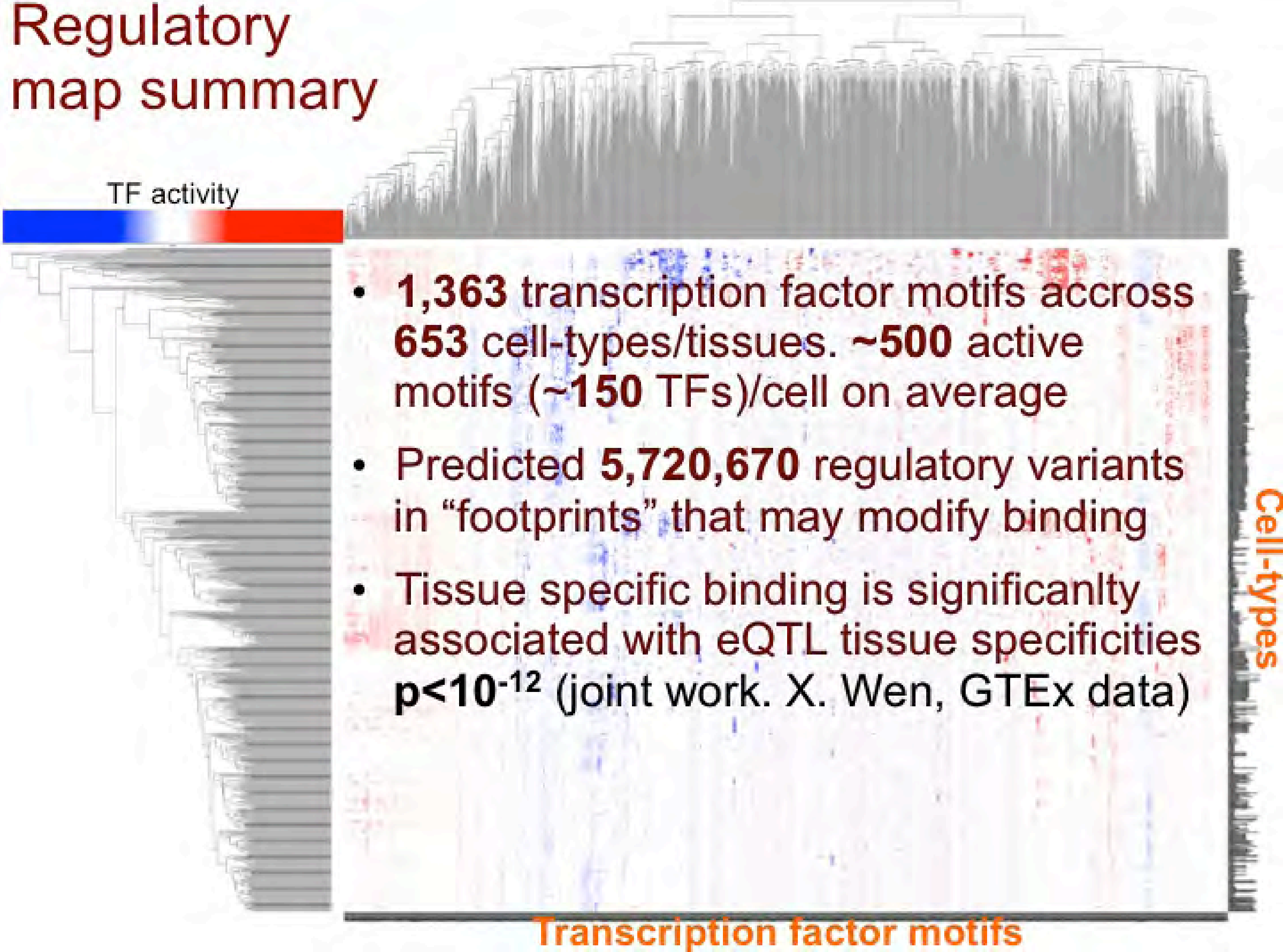
Recalibrated sequence models using CENTIPEDE footprint model (e.g. NRSF)



NRSF ChIP-seq data
from ENCODE GM12878

w/ Gregory Moyerbrailean

Regulatory map summary



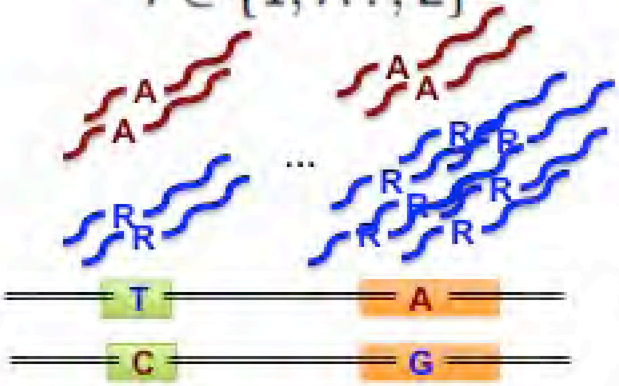
- **1,363** transcription factor motifs accross **653** cell-types/tissues. **~500** active motifs (~**150** TFs)/cell on average
- Predicted **5,720,670** regulatory variants in "footprints" that may modify binding
- Tissue specific binding is significanlty associated with eQTL tissue specificities **$p<10^{-12}$** (joint work. X. Wen, GTEx data)

VALIDATION WITH ASB

Joint genotyping and allele specific analysis **(because genotypes are not available)**

Reads overlapping SNP '*l*' $\{R_{lk}\}_{k=1}^{N_l}$

$l \in \{1, \ldots, L\}$



R_{lk} = 1 ➔ Read carries the **R**eference allele
0 ➔ Read carries the **A**lternate allele

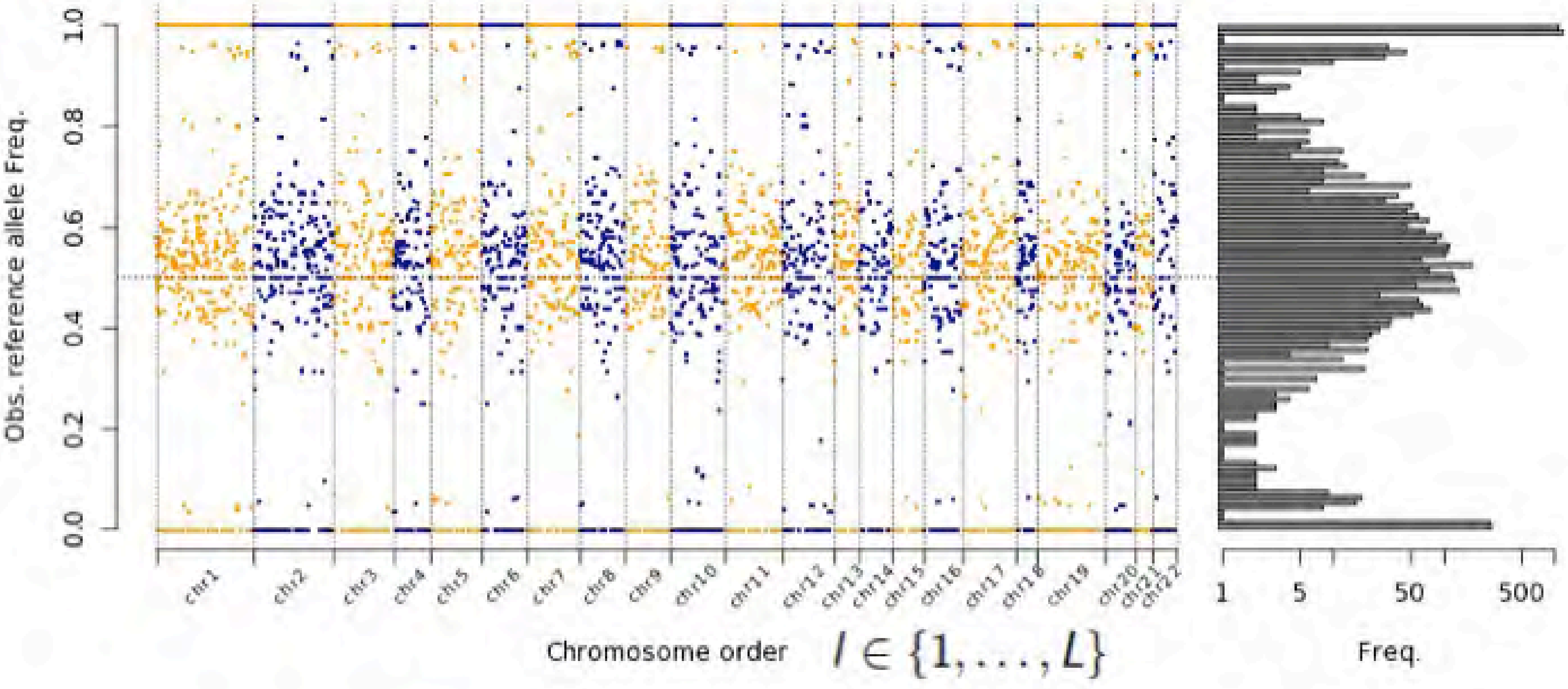
$$\Pr(\{R_{lk}\}) = \prod_l \Pr(\{R_{lk}\}_k)$$

$$= \prod_l \sum_{g \in \{0,1,2\}} \Pr(\{R_{lk}\}_k | G_l = g) \Pr(G_l = g)$$

w/ Chris Harvey

Joint genotyping and allele specific analysis. **The data:**

$\rho_l = R_l / N_l$

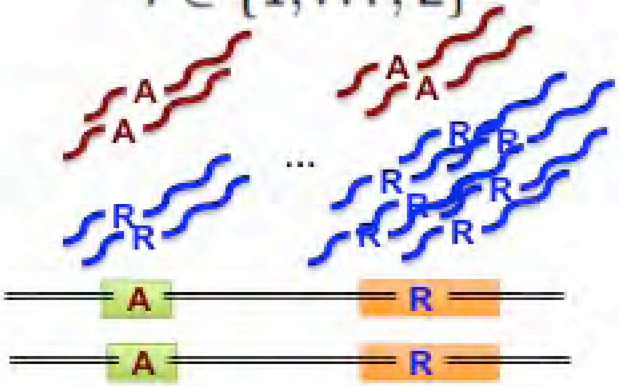


$l \in \{1, \ldots, L\}$

Joint genotyping and allele specific analysis

Reads overlapping SNP '*l*' $\{R_{lk}\}_{k=1}^{N_l}$

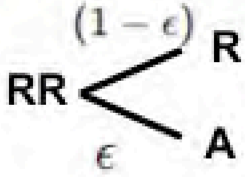
$l \in \{1, \ldots, L\}$



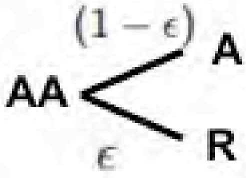
R_{lk} = 1 ➔ Read carries the **R**eference allele
0 ➔ Read carries the **A**lternate allele

$$\Pr(\{R_{lk}\}_k | G_l) = \prod_k \Pr(R_{lk}|G_l)$$

We model the read emission probabilities for the homozygous genotypes $G_l = 0$ (RR) and $G_l = 2$ (AA) as:



$$\Pr(R_{lk}|G_l = 0) \;=\; (1-\epsilon)^{R_{lk}}\epsilon^{(1-R_{lk})}$$



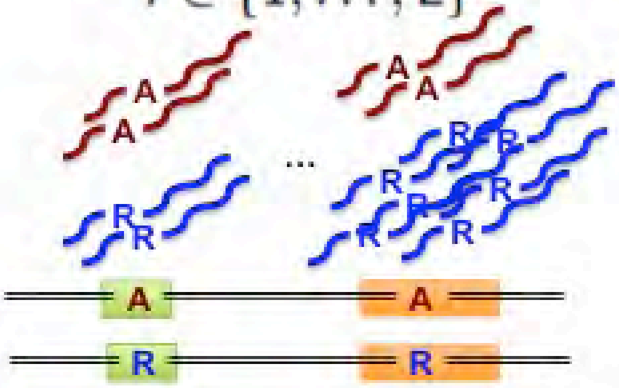
$$\Pr(R_{lk}|G_l = 2) \;=\; \epsilon^{R_{lk}}(1-\epsilon)^{(1-R_{lk})}$$

w/ Chris Harvey

Joint genotyping and allele specific analysis

Reads overlapping SNP 'l' $\{R_{lk}\}_{k=1}^{N_l}$

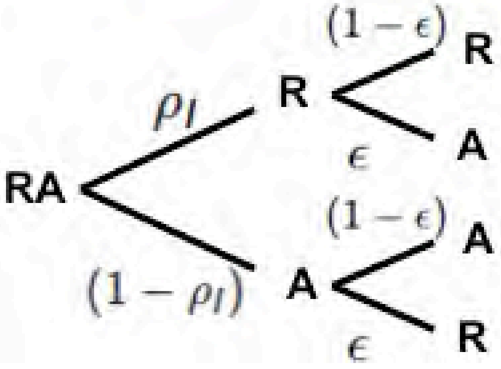
$l \in \{1, \ldots, L\}$



R_{lk} = 1 ➔ Read carries the **R**eference allele
0 ➔ Read carries the **A**lternate allele

$$\Pr(\{R_{lk}\}_k | G_l) = \prod_k \Pr(R_{lk}|G_l)$$

Under the heterozygous state: $G_l = 1 \text{ (RA)}$



$$\Pr(R_{lk}|G_l = 1) = (\rho_l(1-\epsilon) + (1-\rho_l)\epsilon)^{R_{lk}}$$
$$((1-\rho_l)(1-\epsilon) + \rho_l\epsilon)^{(1-R_{lk})}$$

Under the null hypothesis: $\rho_l = 0.5$

$$\Pr(R_{lk}|G_l = 1) = (0.5)^{R_{lk}}(0.5)^{(1-R_{lk})}$$

w/ Chris Harvey

Joint genotyping and allele specific analysis. **The algorithm**:

- Use an expectation-maximization (EM) approach to jointly estimate the model parameters and the genotypes (at 1000 Genomes SNPs)

	DNase-seq GM12878	**Joint genotyping & ASB**	
		Homozygotes	Heterozygotes
High coverage 1000 genomes	Homozygotes	11,271	**0**
	Heterozygotes	7	1,372

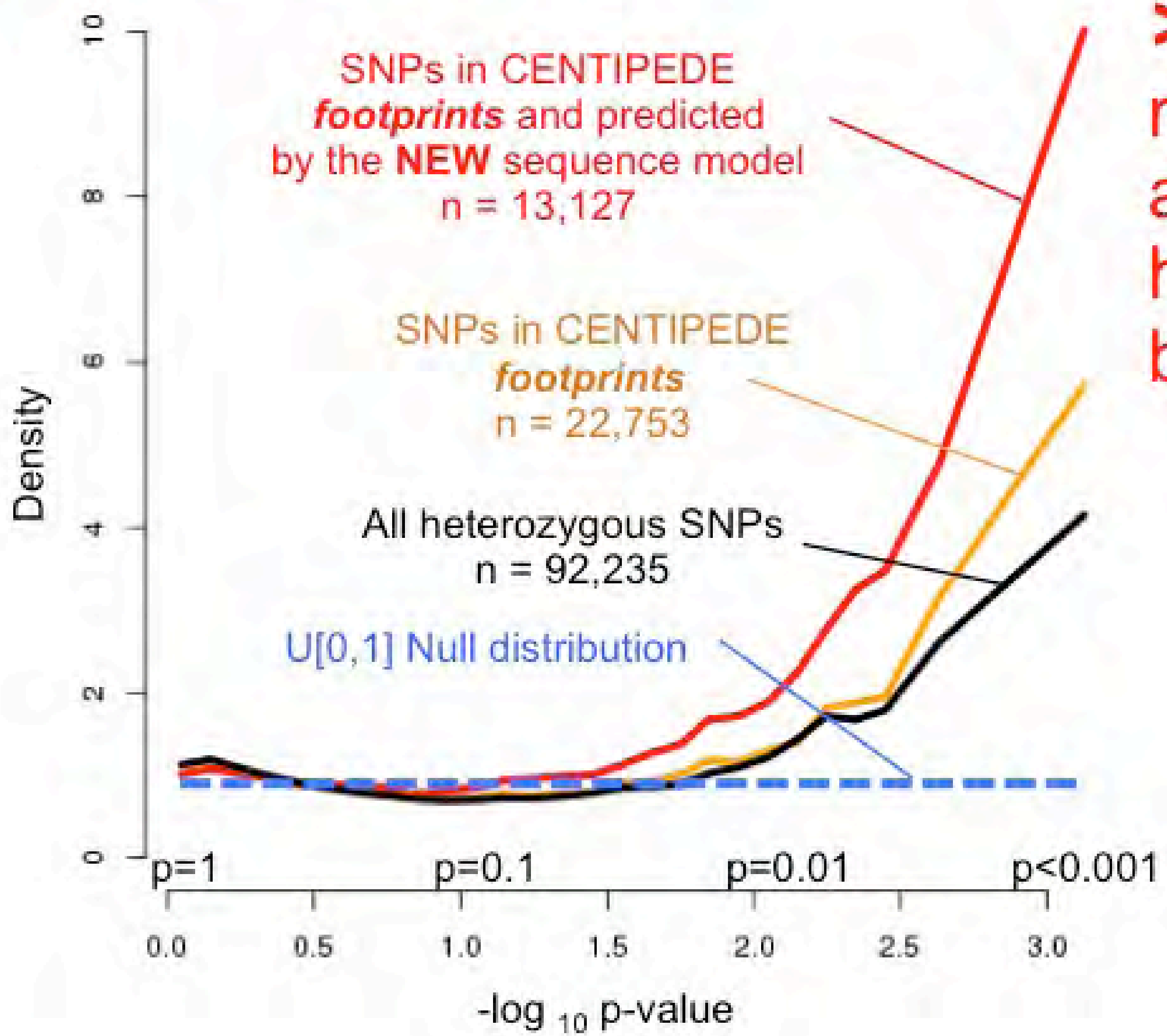
- Calculate a likelihood ratio to test for allelic imbalance:

$$\Lambda_l = -2log\left\{\frac{max\left\{P\left(R_{l,k}|G_l,\rho_l\right) : G_l \in \{0,2\}\ or\ G_l = 1\ \&\ \rho_l = 0.5\right\}}{max\left\{P\left(R_{l,k}|G_l = 1,\rho_l\right) : \rho_l \in [0,1]\right\}}\right\}$$

$$\Pr\left(R_l|N_l,\rho_l,D\right) \quad = \binom{N_l}{R_l}\frac{\Gamma(D)\Gamma(R_l+\rho_l D)\Gamma(A_l+(1-\rho_l)D)}{\Gamma(N_l+D)\Gamma(\rho_l D)\Gamma((1-\rho_l)D)}$$

w/ Chris Harvey

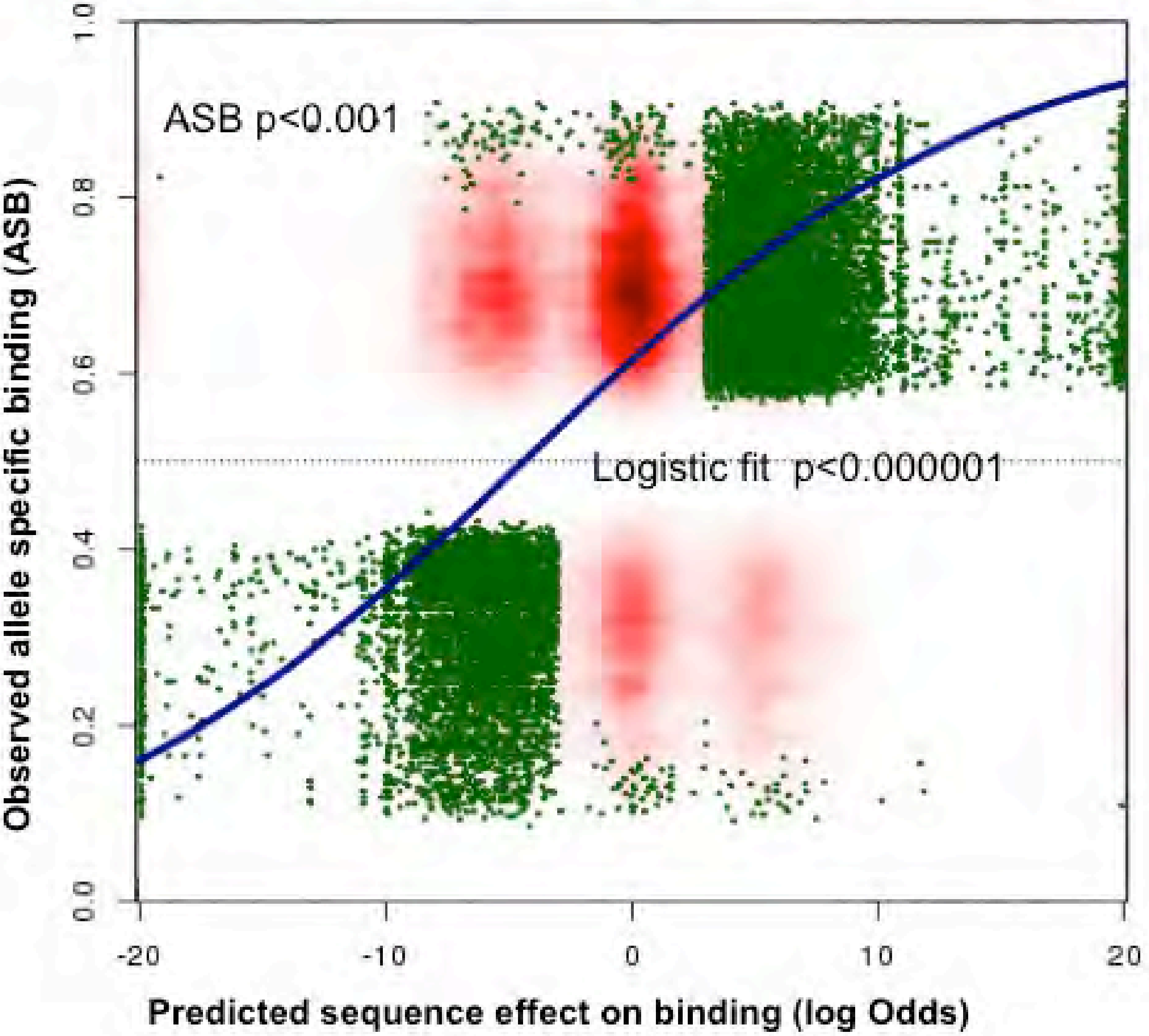
How many sequence variants show Allele Specific Binding?



>55% of predicted regulatory SNPs are estimated to have an impact on binding (ASB)

w/ Gregory Moyerbrailean

Combining sequence model and ASB empirical evidence

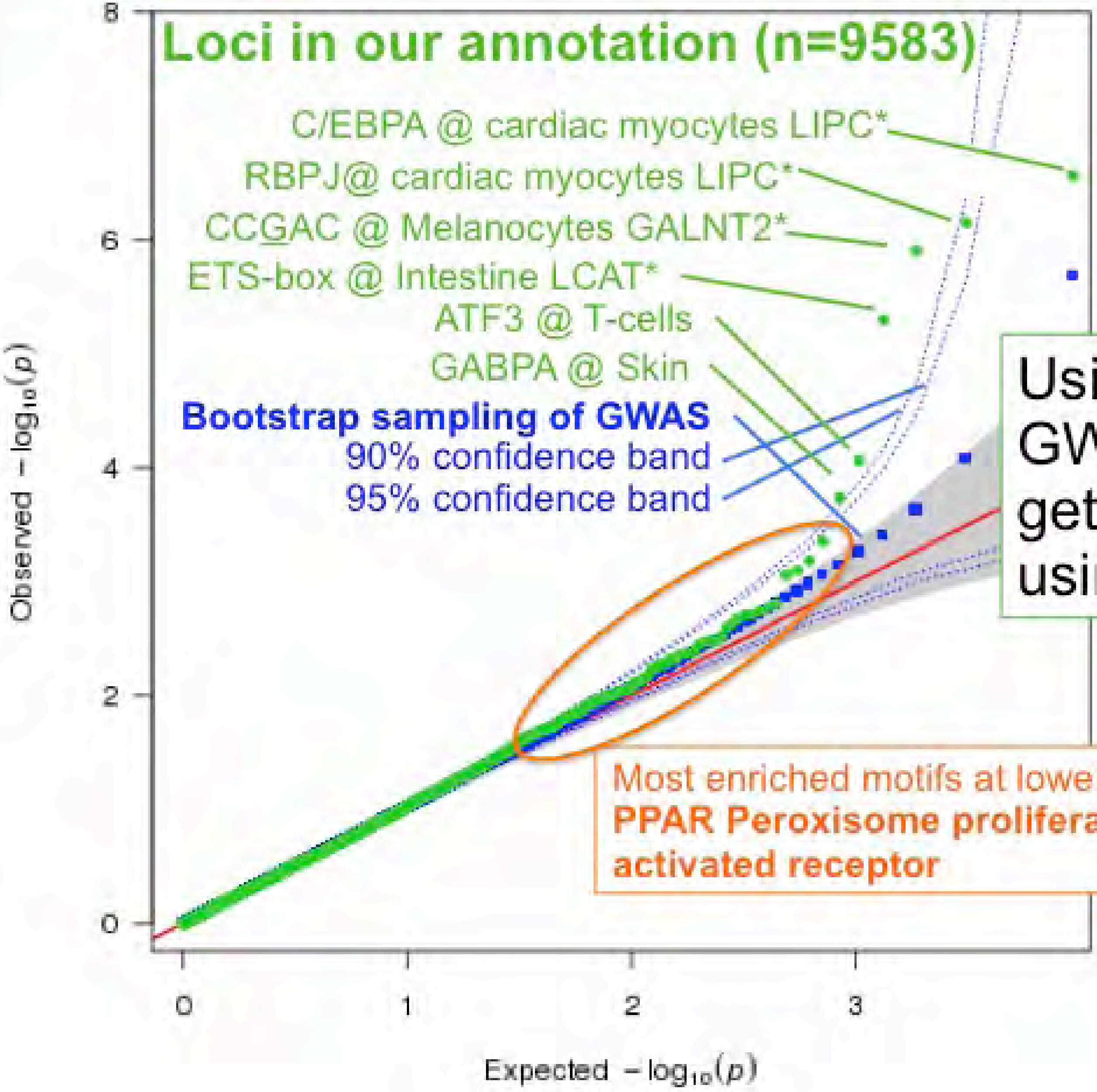


Can we integrate both?

FDR 10%:	200
FDR 15%:	687
FDR 20%:	1540
FDR 25%:	2442
FDR 30%:	3681
FDR 40%:	9538

UNDERSTANDING GWAS HITS

SNPs associated with HDL (Lipids meta-GWAS)

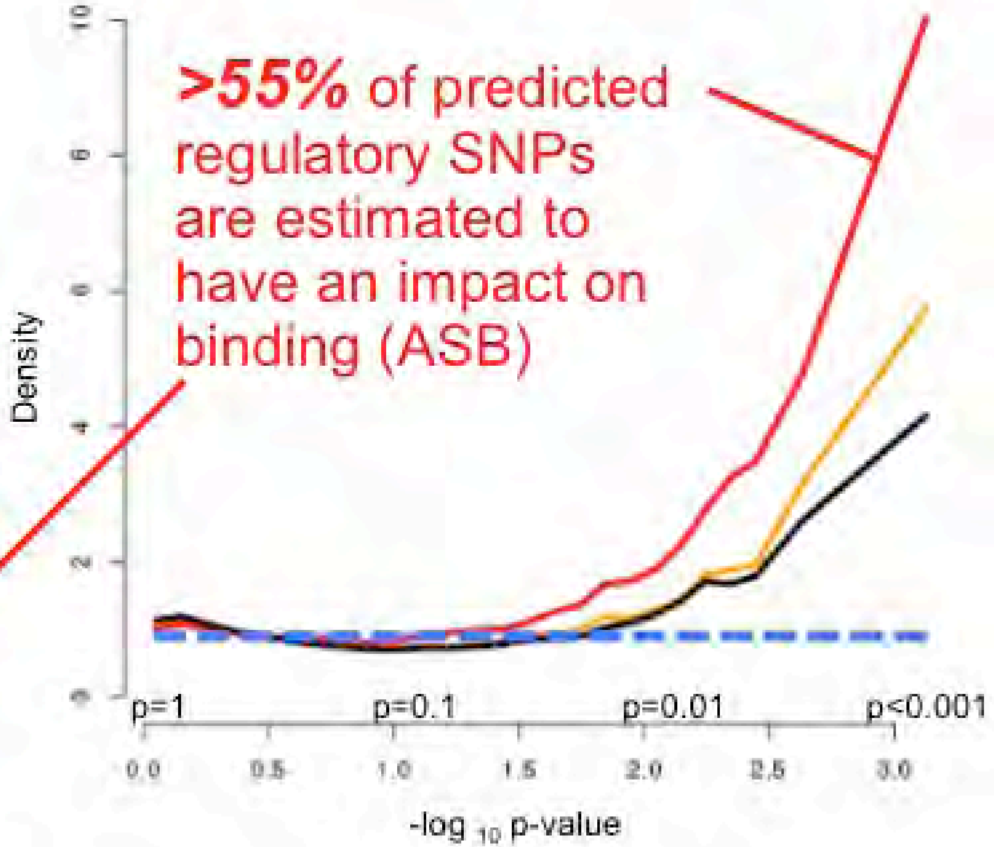


Using the complete GWAS catalog, we get **~2x** more hits than using footprints alone.

Most enriched motifs at lower p-value:
PPAR Peroxisome proliferator-activated receptor

Summary

- Tissue/condition specific regulatory maps for >600 experiments **(high res.)**
- New PWM models predict > 5,000,000 binding variants in footprints
- Joint ASB analysis & genotyping



>55% of predicted regulatory SNPs are estimated to have an impact on binding (ASB)

- Predicted regulatory non-coding SNPs that are validated with ASB are **~2x** enriched for GWAS hits than footprints alone
- Annotation provides also a “validated” motif dimension in addition to tissue specificity

Acknowledgements:

CENTER FOR MOLECULAR MEDICINE AND GENETICS

Francesca Luca

Gregory Moyerbrailean (poster - RG06)

Chris Harvey

Omar Davis

Donovan Watza

Holly Santalucia

WAYNE STATE

School of Medicine

WSU-GRID HPC

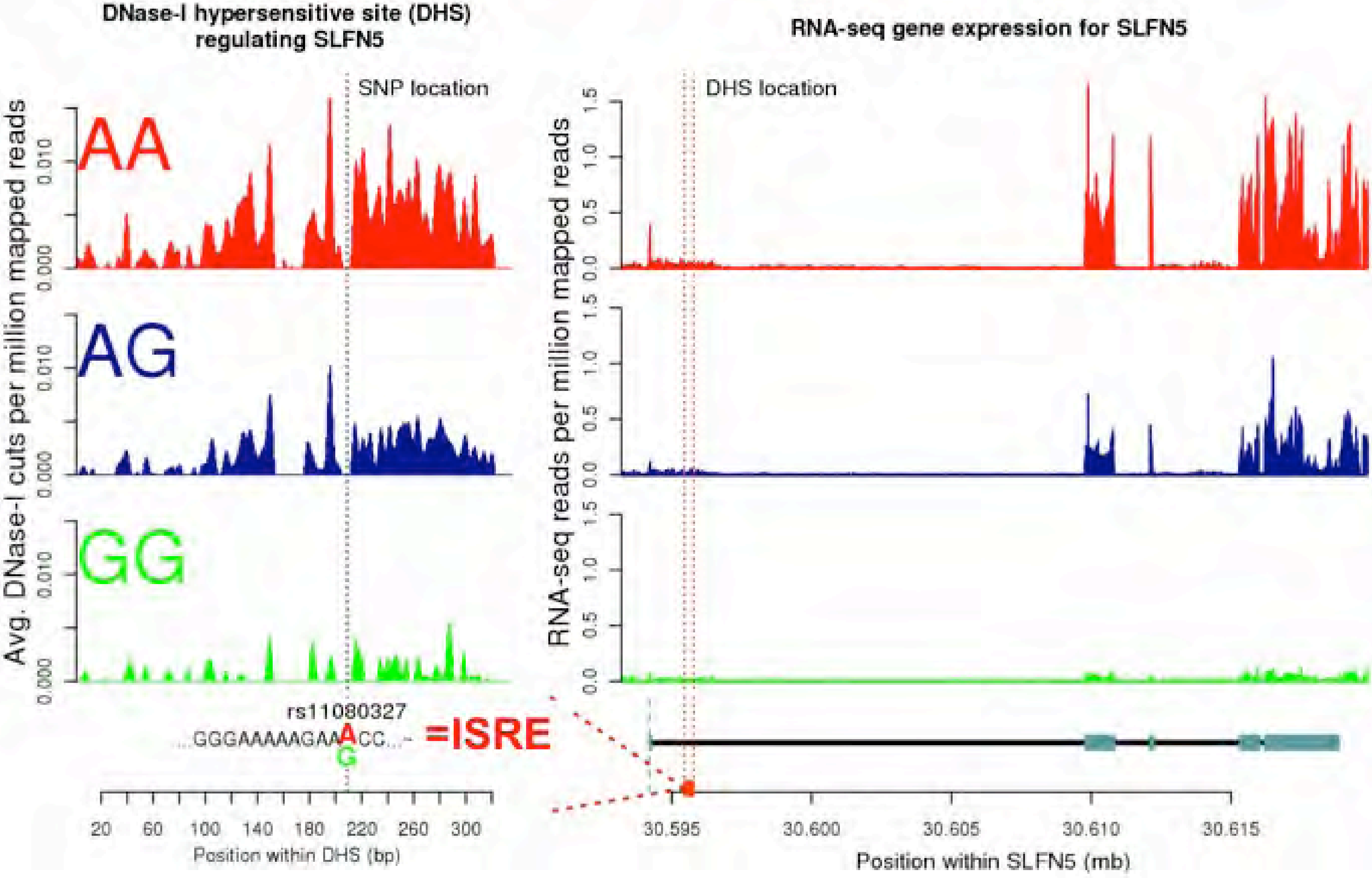
M SPH

Xiaoquan (William) Wen

Thanks for making the data available:
ENCODE, Roadmap Epigenome, GTEX, GWAS catalog, and 1000 Genomes project

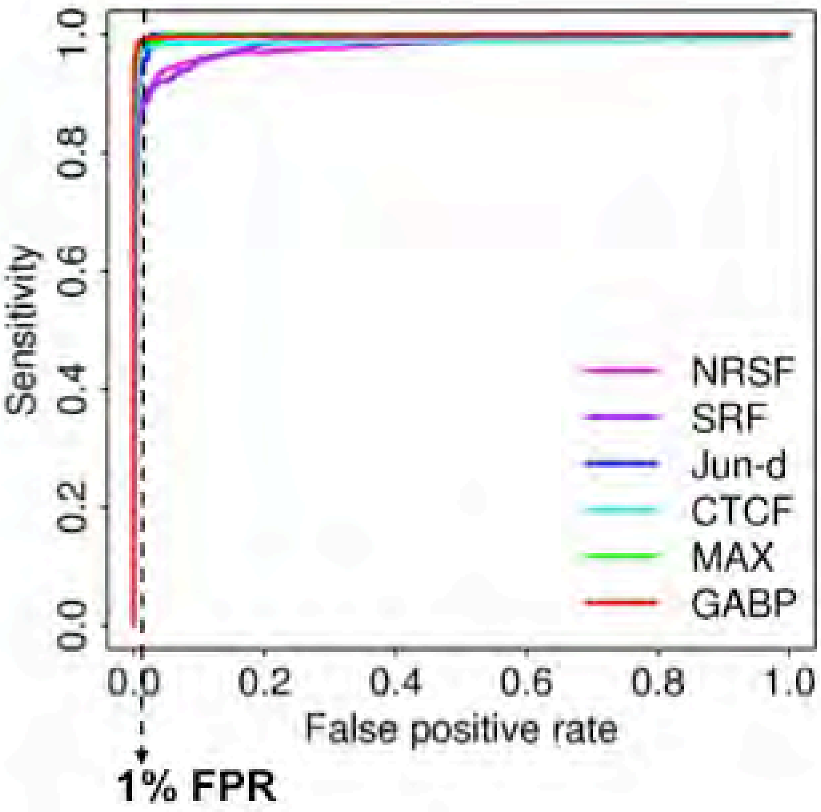
ADDITIONAL SLIDES

dsQTLs are also eQTLs

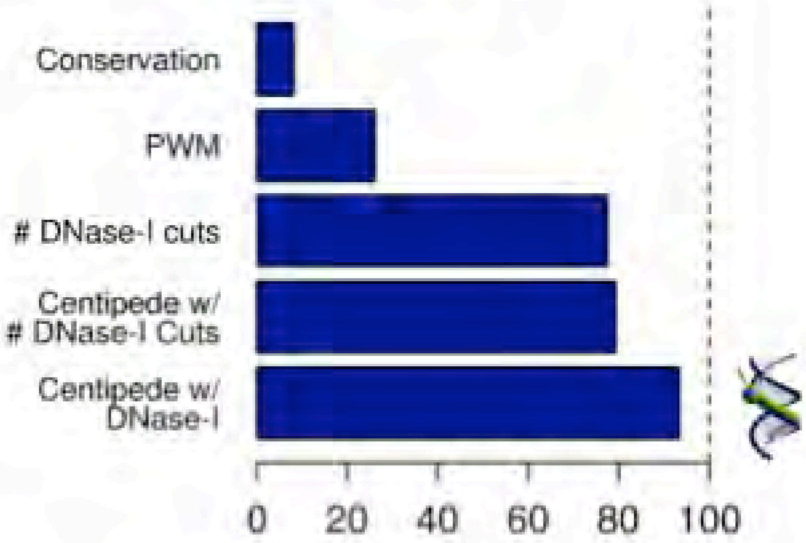


Validation with ChIP-seq (LCLs)

ROC curve



Average Sensitivity % at 1% FPR



ChIP-seq data from Myers, Bernstein and Snyder ENCODE groups

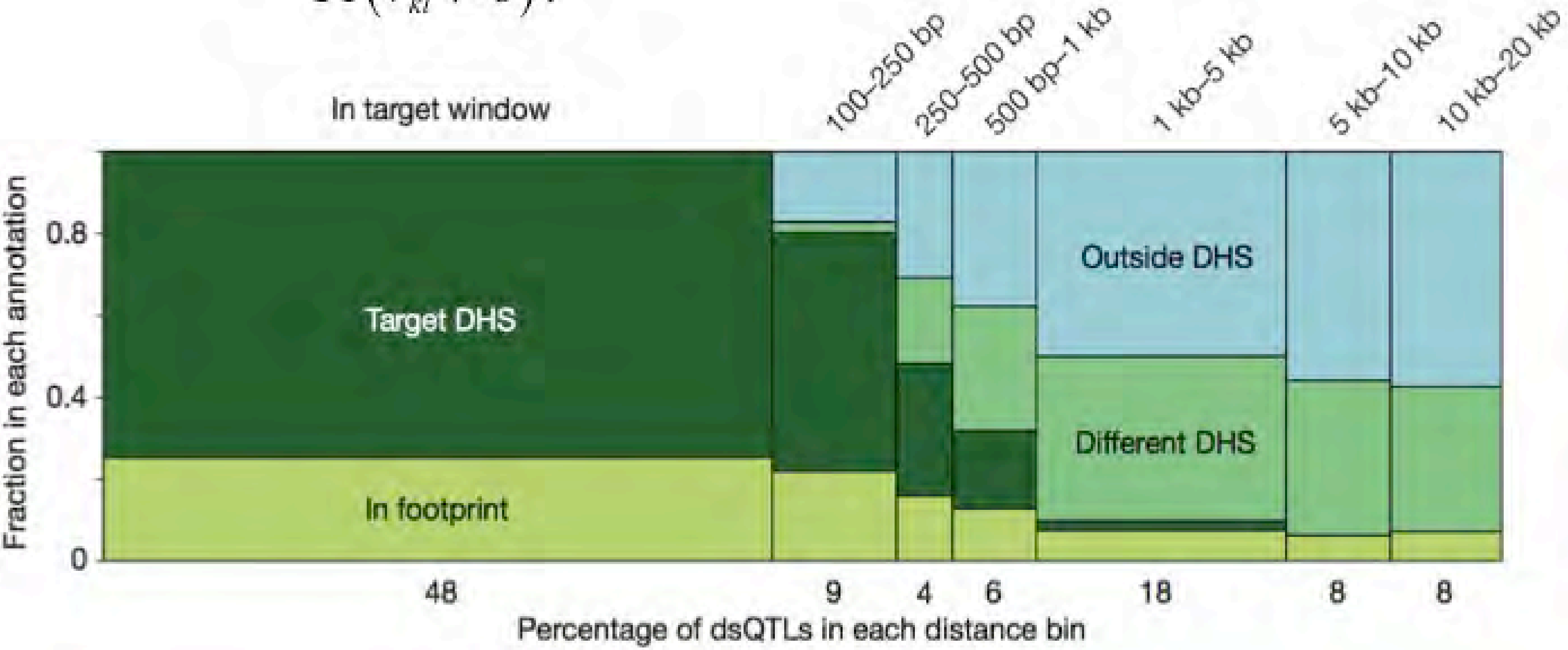
Where are these dsQTLs located?

$$x_{ln} = \mu_l + v_{kl} s_{kn} + \varepsilon_{ln}$$

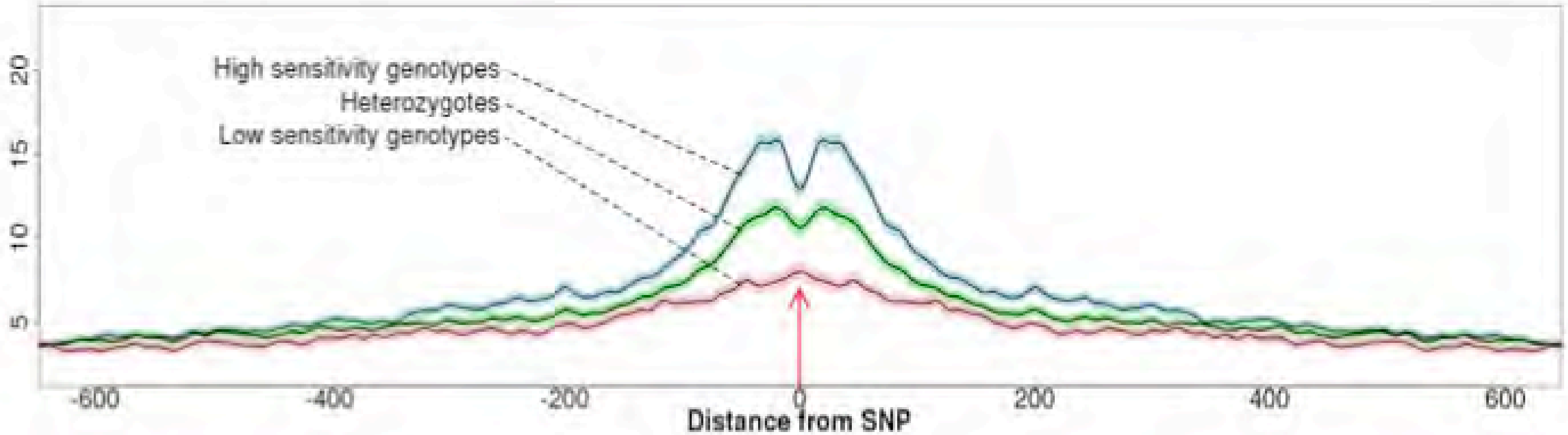
$v_{kl} \neq 0$ for only one k

Method by J-B Veyrieras et al 08

$\Pr(v_{kl} \neq 0)$?

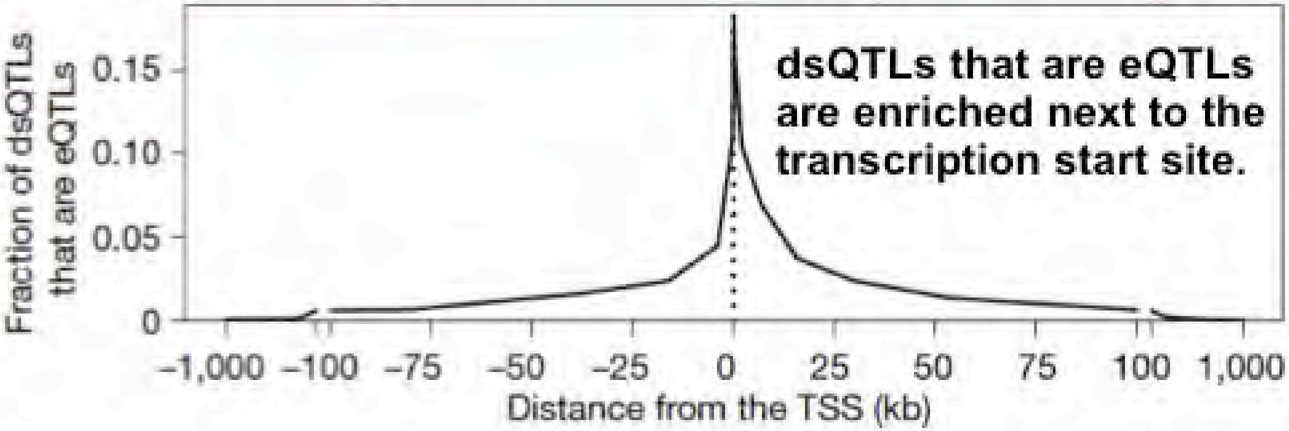


dsQTL frequently occur in DNase-seq footprints



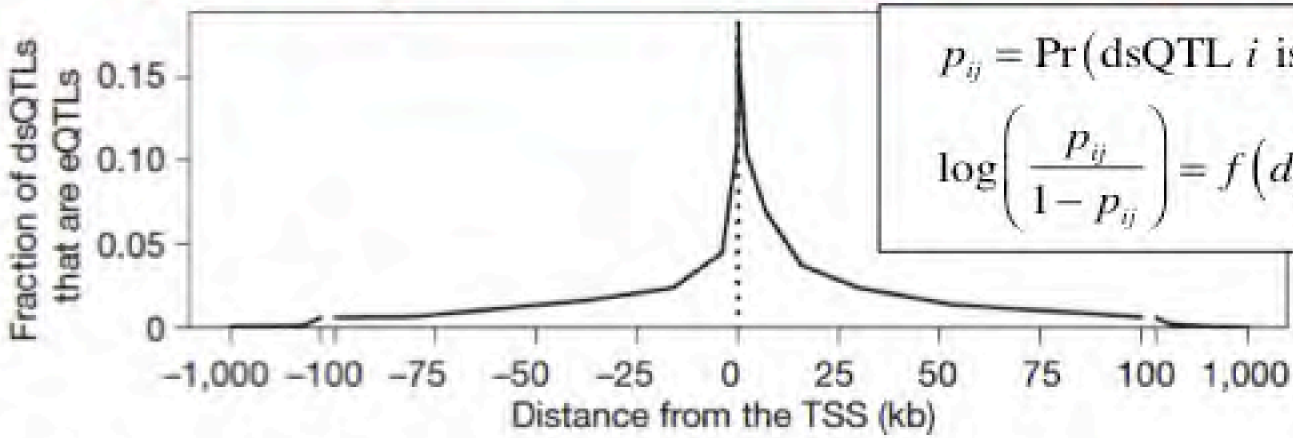
- Dip indicates footprints caused by protection of bound proteins
- SNPs in CENTIPEDE footprints are more likely to be dsQTLs

Exploring the *cis*-regulatory architecture



dsQTLs that are eQTLs are enriched next to the transcription start site.

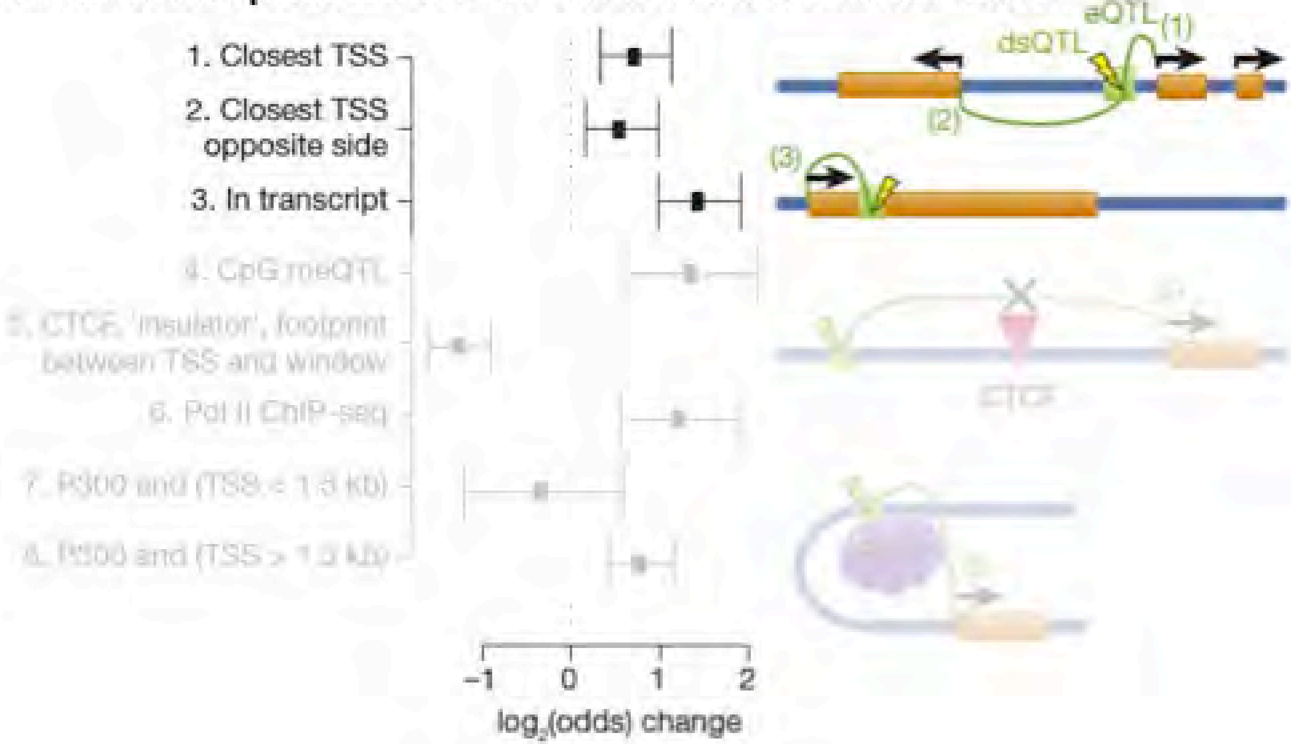
Exploring the *cis*-regulatory architecture



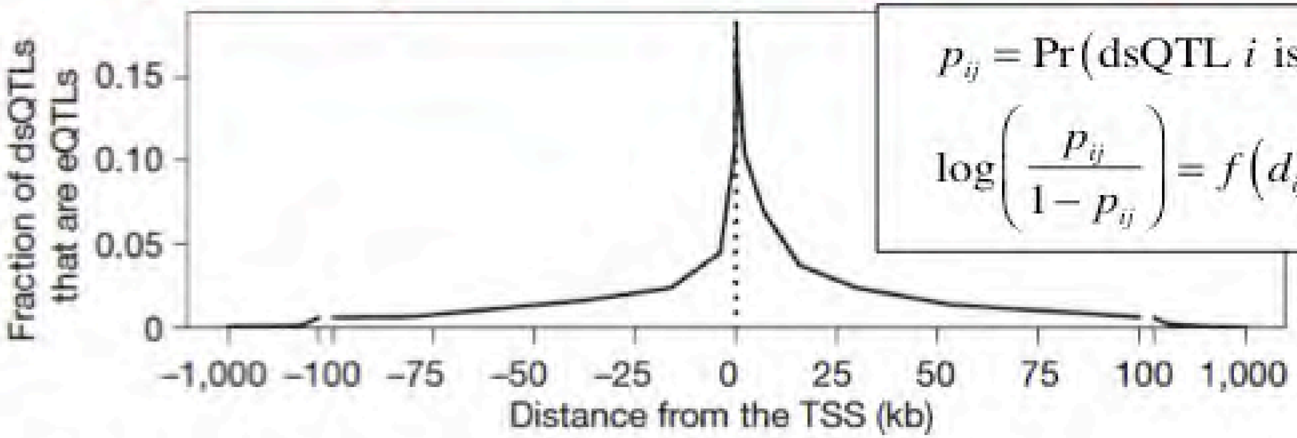
$$p_{ij} = \Pr(\text{dsQTL } i \text{ is an eQTL for gene } j)$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = f(d_{ij}) + \sum_k \beta_k A_{ij}^{(k)}$$

Annotations predictive of whether a dsQTL is an eQTL



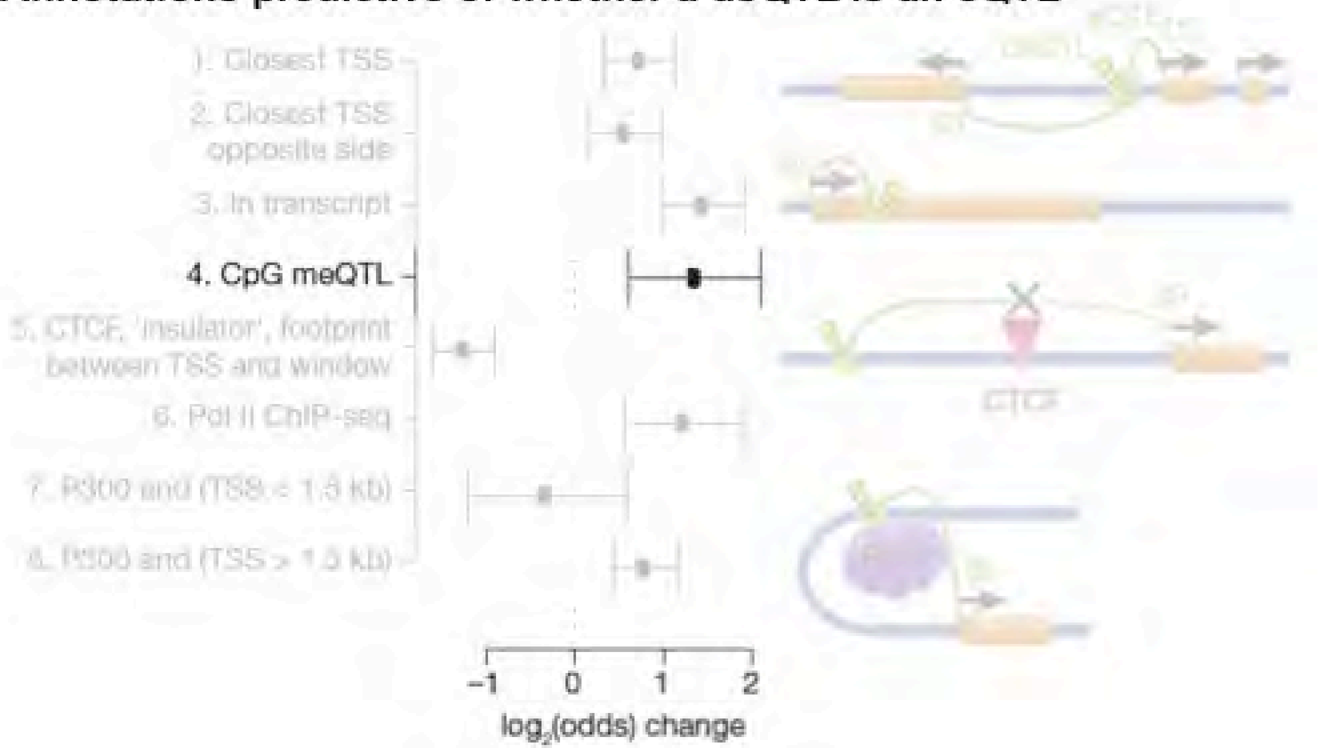
Exploring the *cis*-regulatory architecture



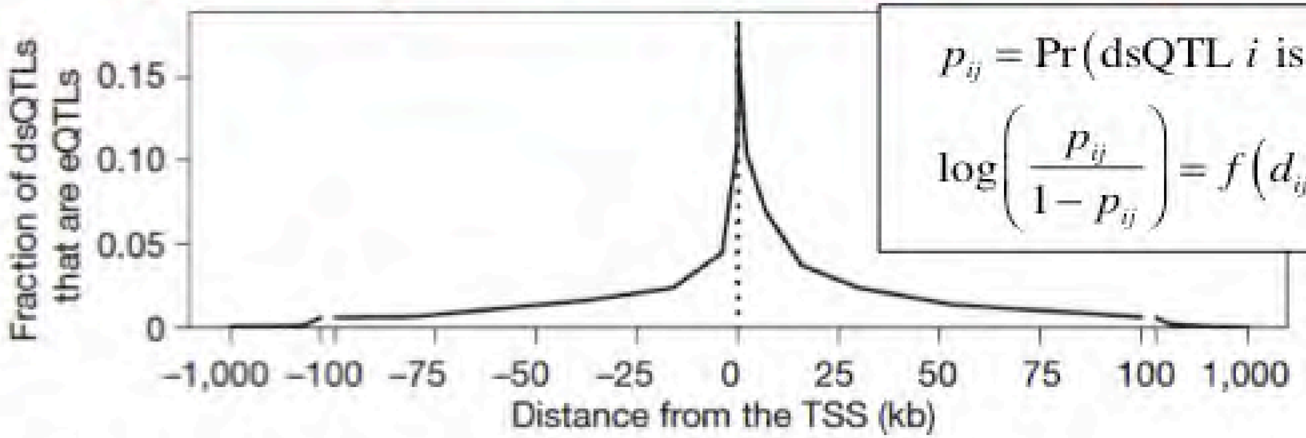
$$p_{ij} = \Pr\left(\text{dsQTL } i \text{ is an eQTL for gene } j\right)$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = f\left(d_{ij}\right) + \sum\nolimits_k \beta_k A_{ij}^{(k)}$$

Annotations predictive of whether a dsQTL is an eQTL



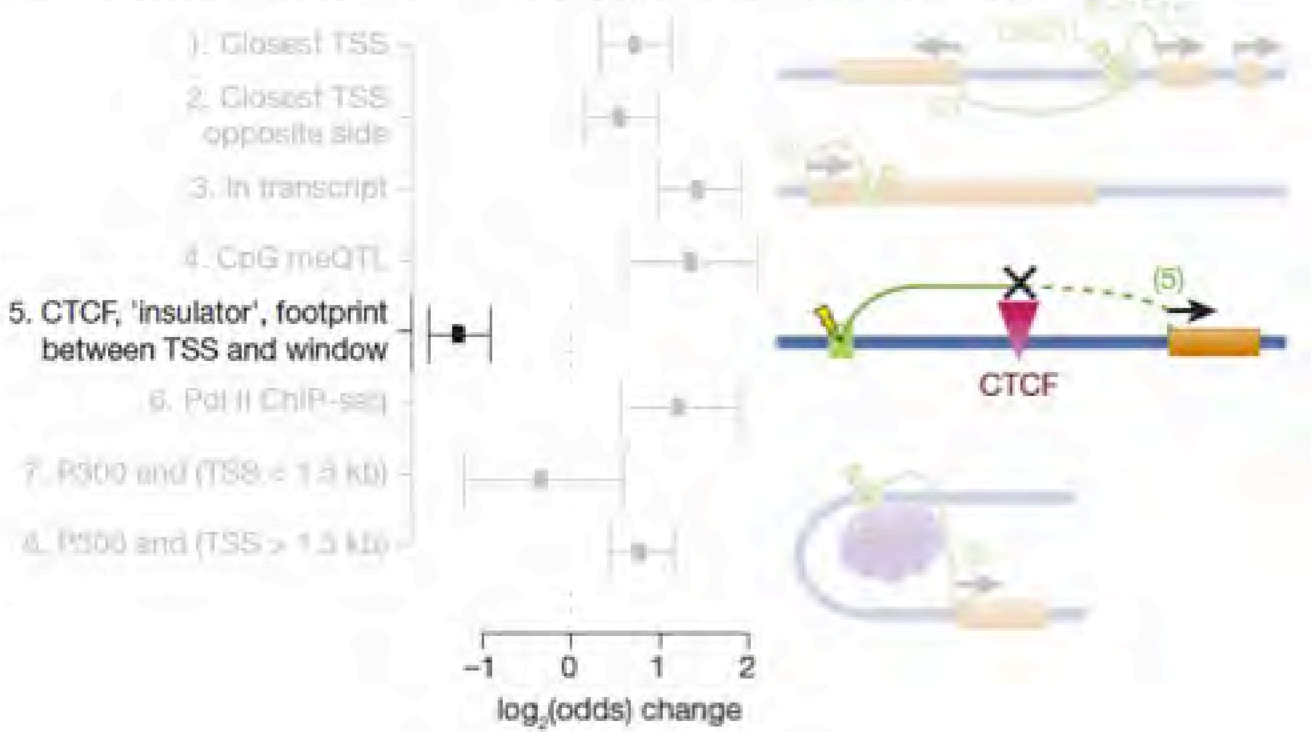
Exploring the *cis*-regulatory architecture



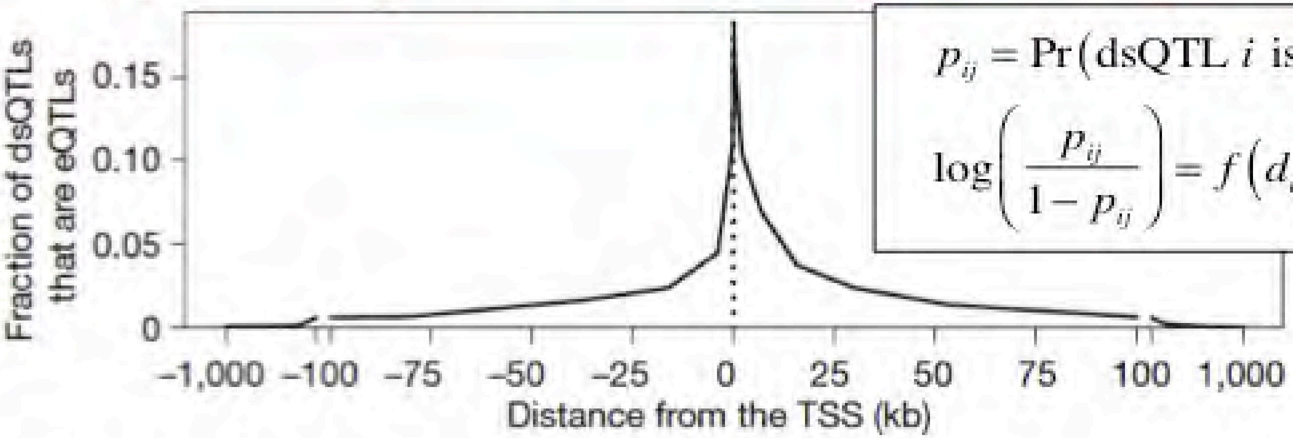
$$p_{ij} = \Pr(\text{dsQTL } i \text{ is an eQTL for gene } j)$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = f(d_{ij}) + \sum_k \beta_k A_{ij}^{(k)}$$

Annotations predictive of whether a dsQTL is an eQTL



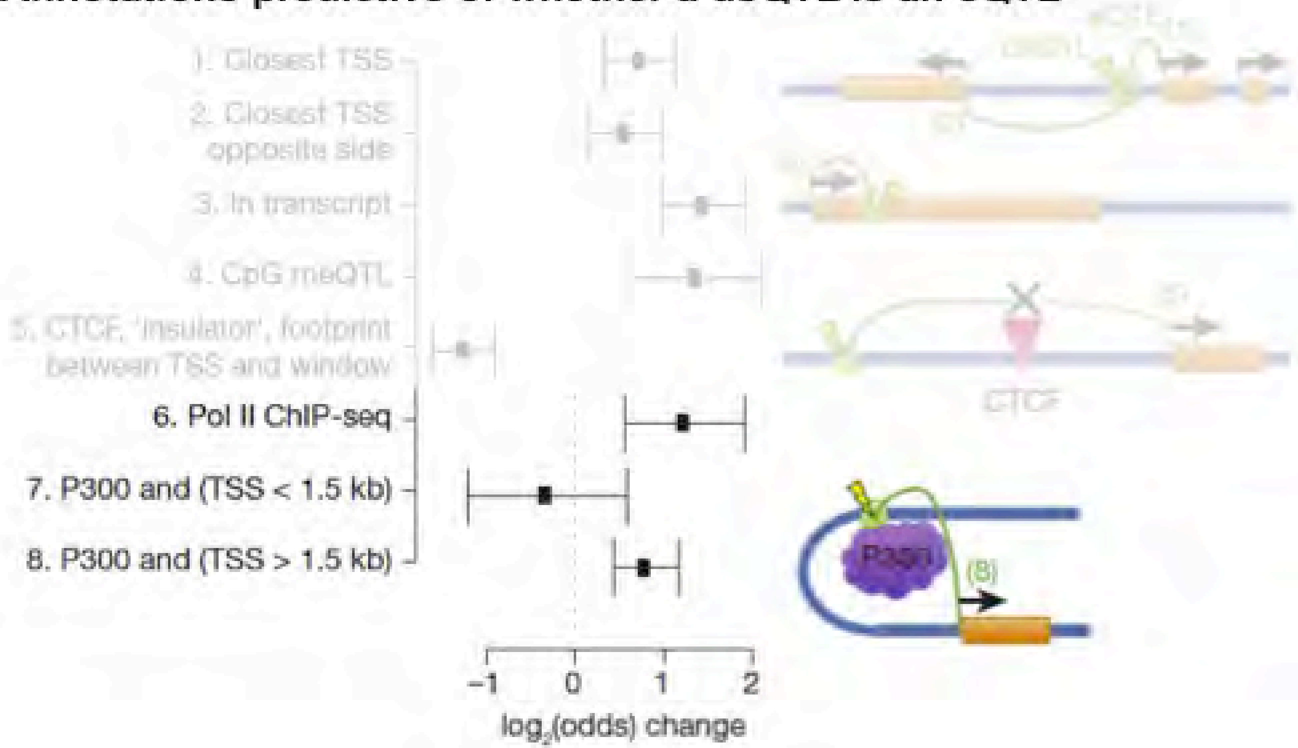
Exploring the *cis*-regulatory architecture



$$p_{ij} = \Pr(\text{dsQTL } i \text{ is an eQTL for gene } j)$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = f(d_{ij}) + \sum_k \beta_k A_{ij}^{(k)}$$

Annotations predictive of whether a dsQTL is an eQTL


1. Closest TSS
2. Closest TSS opposite side
3. In transcript
4. CpG meQTL
5. CTCF, 'insulator', footprint between TSS and window
6. Pol II ChIP-seq
7. P300 and (TSS < 1.5 kb)
8. P300 and (TSS > 1.5 kb)
log2(odds) change
dsQTL
eQTL
(1)
(2)
(3)
(5)
CTCF
P300
(8)