

Genetic Variation and Transcriptional Regulation

**BIOM262 - Quantitative Methods in
Genetics and Genomics**

Graham McVicker – Salk Institute

Human phenotypic diversity



Alzheimer's Disease

Ankylosing Spondylitis

Multiple Sclerosis

Leukemia

Psoriasis

Breast Cancer

Schizophrenia

Rheumatoid Arthritis

Coronary Heart Disease

Celiac Disease

Autism

Crohn's Disease

Parkinson's Disease

Type I Diabetes

Systemic Lupus Erythematosus

ATCCTTCCAGAGAAAATAGAAAATAGAACTTTGAGGTTGAATCTCTTTAATGTAATG
TTTTCTCGAATCCAAGTGTTCACACTATAACAATAGGAGTAGAAATTGTCACCACTC
TGTGGCCAAA^YTCACTTTCTTCTTTTATT^YACATTAAAAAAATTACTT
TAAGTTCCAGGATA^CATGTGCAGGATGTGCAGGTTGTTACATAGTAAATGTTTATT
TAAATTAAATTAAACACTTTTAT^YTTAAGTCATACAAC^TCTCATGCCAGTAGTTAAT
ATTACCTTG^MAAGTTGGTATGGTTGATGAATTGCATCCTGTTAATAATTGCTACAGATT
TTTGAATAATTGCAGACCAGTTGATGR^TCCTGGGTTGGCATAAGTACATGAAGAGATTAC
TTTCTGTGAGCTTCTTGGGATGAAGAAATTAGT^{GT}TTTTTTAATTAAAGAA
AT^WTTTATTATTTTACATGATTATTCCC^ACTGAAAAAATAAATCCCACCGGG^YATAA
AGTGT^ATTTTTAAGTCACAGAGTAACCC^ACTGAAAGCTAGTTTCAGACTTAGGCA
GTTCATGCTGTAAGCCCGAGATCTCATGGTCACCC^CTGCAAGAGAAATATCTAATTGAAA
AAAAATATGAAGAGTATTAAATTG^WTAGTGCTAAAATGACATAAGGGATCTCACTGGG
CTTGAGATATTAAGTATTAAAATTGTT^RAGGTTAAATTGTTAGTAAC^TGTTATTGCA
TAGAAAATGTGCCAAATGTCAGTAAATAAAAAACTTTT^TAAAATAAAAATTACAGA
ARAATTATGACR^AACTACAAAGAGGTTCTGTACAAACCC^CCTCCCAGTTCTCTTACTAT
TAACATCTTAAATTAGTATGTTACATTGTCACAATTAGTGAACCAATTGATA^CATTA
GTACKAACTAAAGTCAGTGT^CCTTTACTGGAGAATGGTGTAGAAACTAAGGTCTGGG
CACTGTGGTATGGTGGTGCTATTGAGATGT^YGTTATTAGGTTCTTCTCAGCTGAC
AGAGCAAAGAAATATATGTGTATATTAAACCTATGTGTACACATACTATGATTATT
TCR^AATATGTAACAT^STGTATCTTATTAAAGCTAAATATGAGTCATATGGTGTCTCAAT
TCTAATCAATTACTGTATAGATTATTCTAGCCTCTTGTCTTAC^TGTAAC^TCCT
ATTCAAACCGTAAAAATCTGTCTTCCACCACCTACTATCTGCTTACCTAATTCT^YAT
TTCCAGTTA^KGTATACAGTGGCTTCAGAATTATTACATATA^RCCCTGTGGGATACAAC^T
TTGTCAACTAGAGTGGTGCTTATGTAAGTTCTATCTTAC^TGACTCTACTC

Which genetic variants affect human traits?

ATCCTTCCCTAGAGAAAAATAGAAAATAGAACCTTGAGGTTGAATCTCTTTAAATGTAATC
TTTTCTCGAATCCAAGTGTACACTATACAATAGGAGTAGAAATTGTCACCACTG
TGTGCCAAAYTCACTTTCTTCTTTTATTTYACATTAAAAAAAAAATTGTTACTT
TAAGTTCAGGATACATGTCAGGATGTCAGGGTTACATAGGTAATGTTTATT
TAAATTAACTTAACTTTATXTTAAGTCATAACCTCTCATAGGCTAGTTAA
ATTACCTTGMAGTTGGTATGGTATGTAATTGTCATAGGCTACAGATT
TTTGAATAATTGCAGACCAGTTGATGRCTCTGGGTTGGCATAAGTACATGAAGATTAC
TTTTCTGTCAGCTTCTGGGATGAAGAAATTAGTGTGTGTGTGTGTGTGT
ATWTTATTATTTTACATGATTATTCCCAGTGAAGAAATTCCACCGGGYATAA
AGTGTATTTTTAAGTCACAGAGTAACCCAACTTGAAGCTAGTTTCAGACTTAGGC
GTTCATGCTGTAAGCCCCGAGATCTCATGGTCACCCCTTGCAAGAGAAATCTAATTGAA
AAAAATATGAAGAGATTAAATTGWTAGTGTCTAAATGACATAAAGGGATCTCACTGG
CTTGAGATATTAAAGTATTAAATTGTTAAGGTTAAATTGTTAGTAACTTGTATTGCA
TAGAAAATGTGCCAAATGTCAGTAAATAAAAAACTTTTTTAAATTTAAATTACAGA
ARAATTATGACRATACTACAAAGAGGTTCTGTACAACCCCTCCAGTTCTCTTACTAT
TAACATCTAAATTAGTATGTTACATTGTCAAAATTAGTGAACCAATATTG
GTACKAACTAAAGTCAGTGTCTTTACTGGAGAATGGTGTAGAAACTAA
CACTGTGGTATGGGGTGTCTTACTGGAGATGTYGTTATTAGGTTCTTCT
AGAGCAAAAGAAATATATGTCGTATTAACCTATGTCACACATACATCA
TCRATATGTAACATSTGTTATTTTAAAGCTAAATGAGTTCATATGGT
TCTAATCAAACTACTGTATAGATTATTCTAGCCTCTCCTGCTTATCTGTAATTCT
ATTCAAACCGTAAAAATCTGCTTCCACCACTACTATCTGCTTACCTAAATTCTYAT
TTCCAGTTAKGTATACAGTGGCTTCAGAATTATTACATATARCCTGTGGGATACA
TTGTCAACTAGAGGGTGTATGTAAGTTCTCTATCTTAGTTACTGACTCTACT
ATTTCAAAGTGTCTTAGTCAGAACACTTCACACTCTCCCTAGTGAAGTTGTT
ATAYGTTAGTAAACACAGATTCTTTYTGCACTGTCGATTCCATTAGGTTCTCT
CTCCAACTCTCTAAATTATTATTAAATTCTATACATCAACAGGTTATTCTGTG
TGTAARGTTCTATAGGTTTGACAAATACAAAGGTCRTGACCCATCATTACA
TACRGAATGTTCACTGCYCTAAAAAATATCCCTGCTTGCCTATTCAACCCCTCC
CTCCTTCCCAAACTCTGGCAACCACTGATCTGTTATCGTGAGCTGTCCTTCC
GAATGCATATAATTGAAATCATACAAATATGTAAGACTTTCACCTGGCTTATTGTT
CAATATGCTTAAACATTCCATGCCATGCTCTATGTCAGTTAGTCACTTTACT
GCTGGTAGTATTCTRCACTAGAAATGRCACACRNTGTTATTCCATTYGCTGATTGA
GTATATCAATATACCTGGAACATGACTGCTAGATAGYATAGTAAGACTATATTAKCT
TGCAAGAAAATGCCAAACTGTTAAAGTGGCTGTACCCATTGTGCCACCGAAC
TGCCACTGATCCAGTATTGTCAGTTTGGATTTAKCCATTCTAAAGGTGAGTGATG
GTATCTATTGTCGTTTAATTGTAATACTCTAATGACAAATGATGGTGGATTTCTT
CATATGYTTGTTTCCCATTTGATATCTCTTAAAGTGTCTGTTGGATTTGCT
TACTTTTTAAACTGGGTTGATTGTTCTTTCTTTCTTTCTTTCTTTGAGAYG
GAGTCTYGCTCTTAGGCCAGGCTGGAGTCAGTGGCCCATCTCGGTCACTGCAAGCYC
TGCCCTCGGGTTCAAGTGTACCTCGTACCTCAGCCTCCGAGTAGCTGGGACTACAGG
GCCGCCACACACCTGGCTAATTGTTGTTGGAGACGAGGTTCACTT
TCGGTCAGGCTGGCTTAAWCCTCTGACCATAGATGATCTGCTGCTTGGCCTCCAAA



Topics

- Genome-wide association studies
- Molecular quantitative trait loci (QTLs)
- Gene expression QTLs
- Chromatin QTLs
- DNA methylation QTLs
- Intersection of molecular QTLs and GWAS

Testing for genetic association

TTGTTAGAGGTTAAATT

G G



TTGTTAGAGGTTAAATT

G A



TTGTTAGAGGTTAAATT

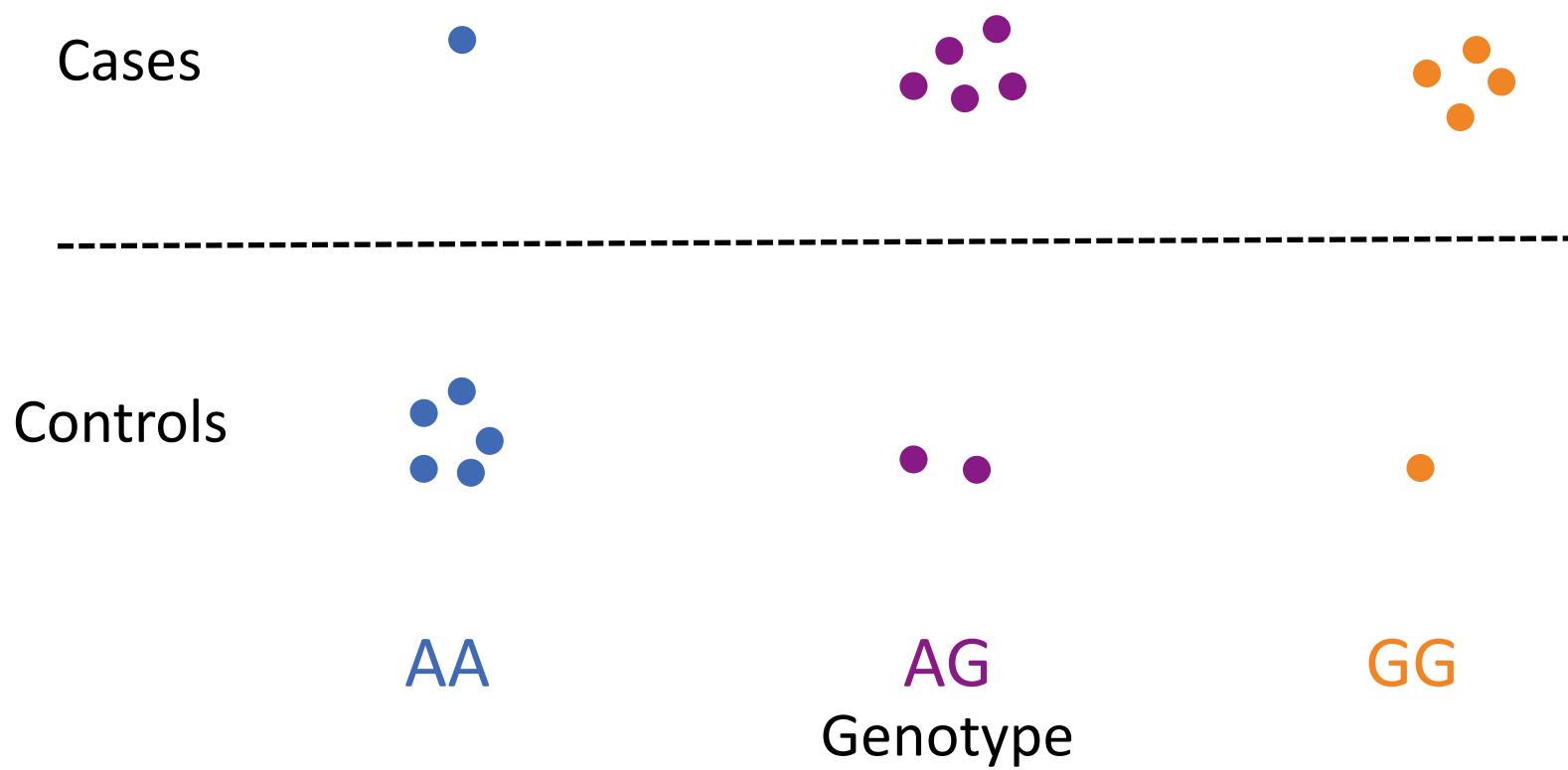
A A



TTGTTAAAGGTTAAATT

TTGTTAAAGGTTAAATT

Case-Control Association Study

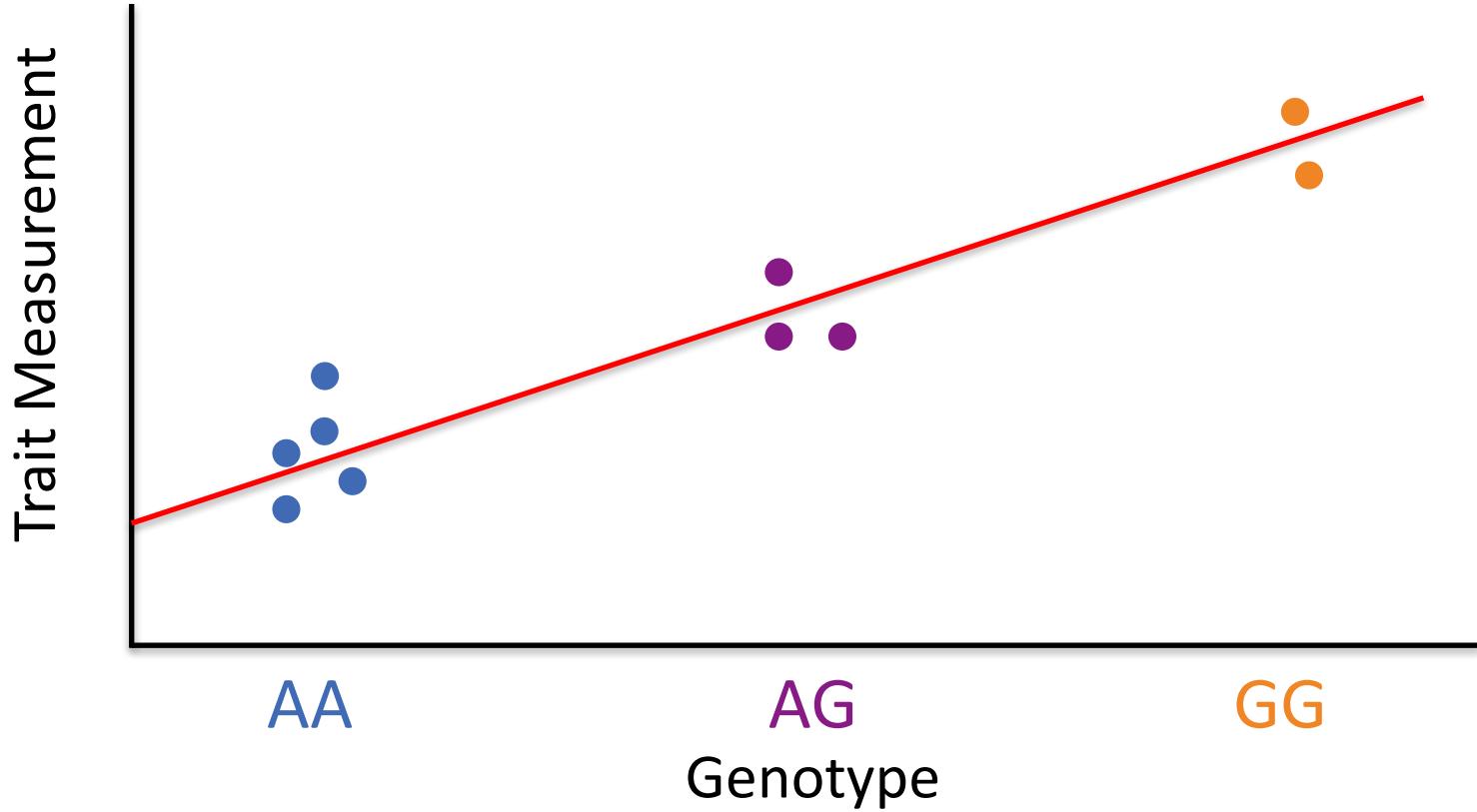


Case-Control Association Study

	# A alleles	# B alleles
Cases	750	250
Controls	800	200

χ^2 p-value = 0.009

Quantitative Trait Association Study



Discussion

- What are examples of human traits?
 - Quantitative Traits?
 - Binary (Case/Control) Traits?

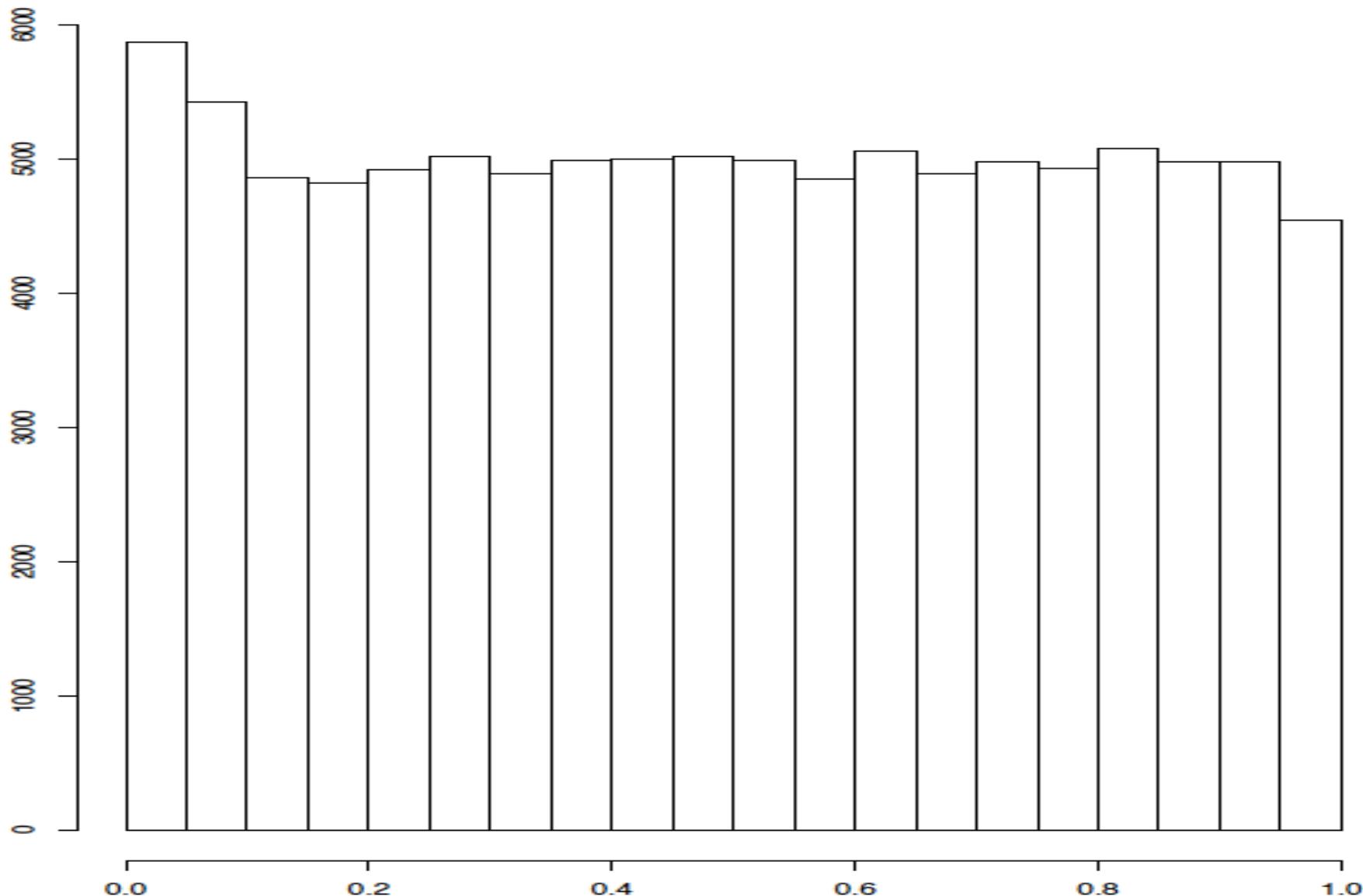
In class exercise

- GWAS study results from Rheumatoid Arthritis (Okada et al. 2014)
- Meta-analysis of European and Asian individuals
- 29,880 RA cases and 73,758 controls
- ~10 million SNPs tested
- **Datafile:** RA_GWAS.txt
(reduced to random set of 100k SNPs)

Exercise

- Read RA_GWAS.txt table into R
- Make a histogram of the p-values

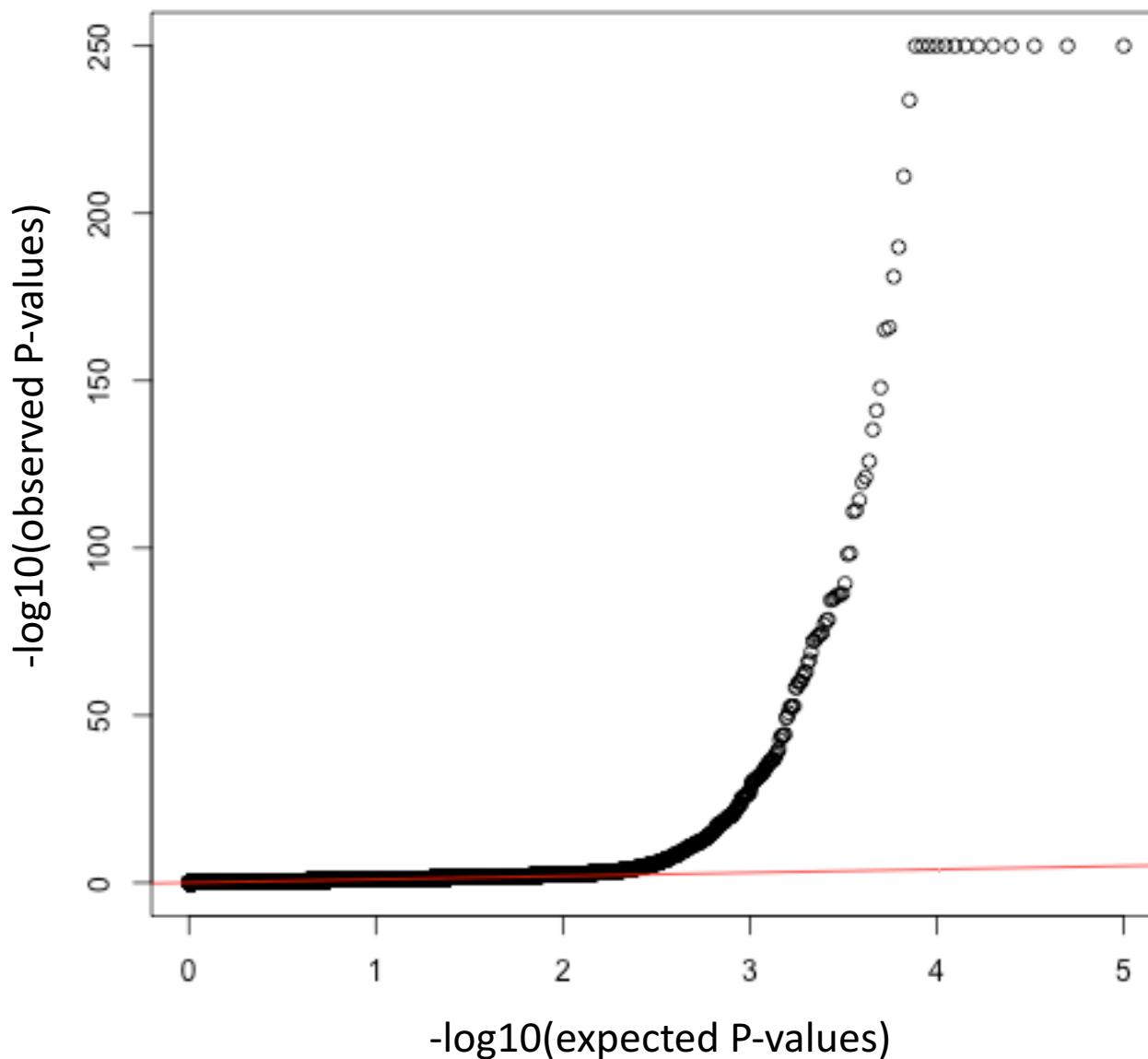
P-value distribution



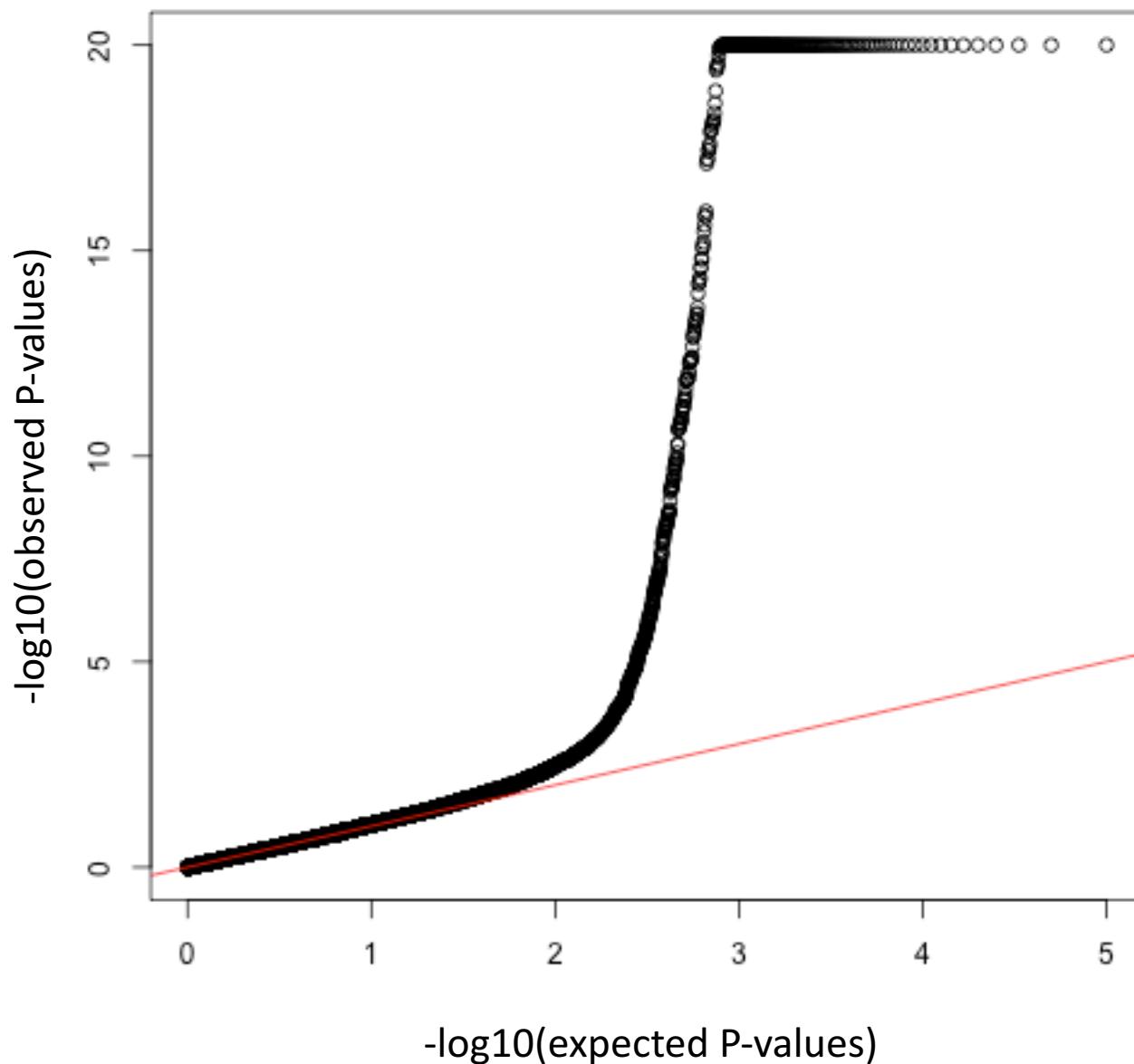
Exercise

- Make a quantile-quantile plot of $-\log_{10}$ expected vs. observed p-values
- Expected p-values are uniformly distributed

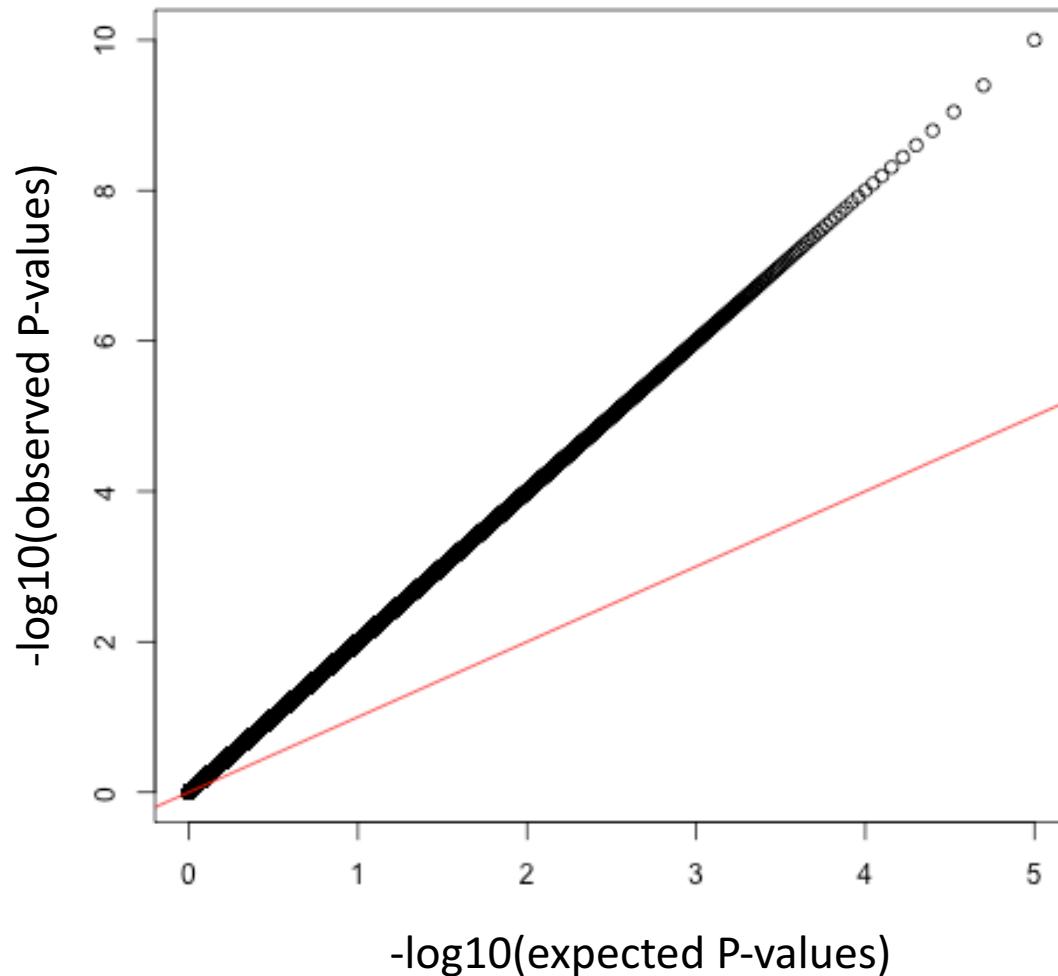
P-value QQPlot



truncated -log10 P-value QQPlot



example of deflated p-values



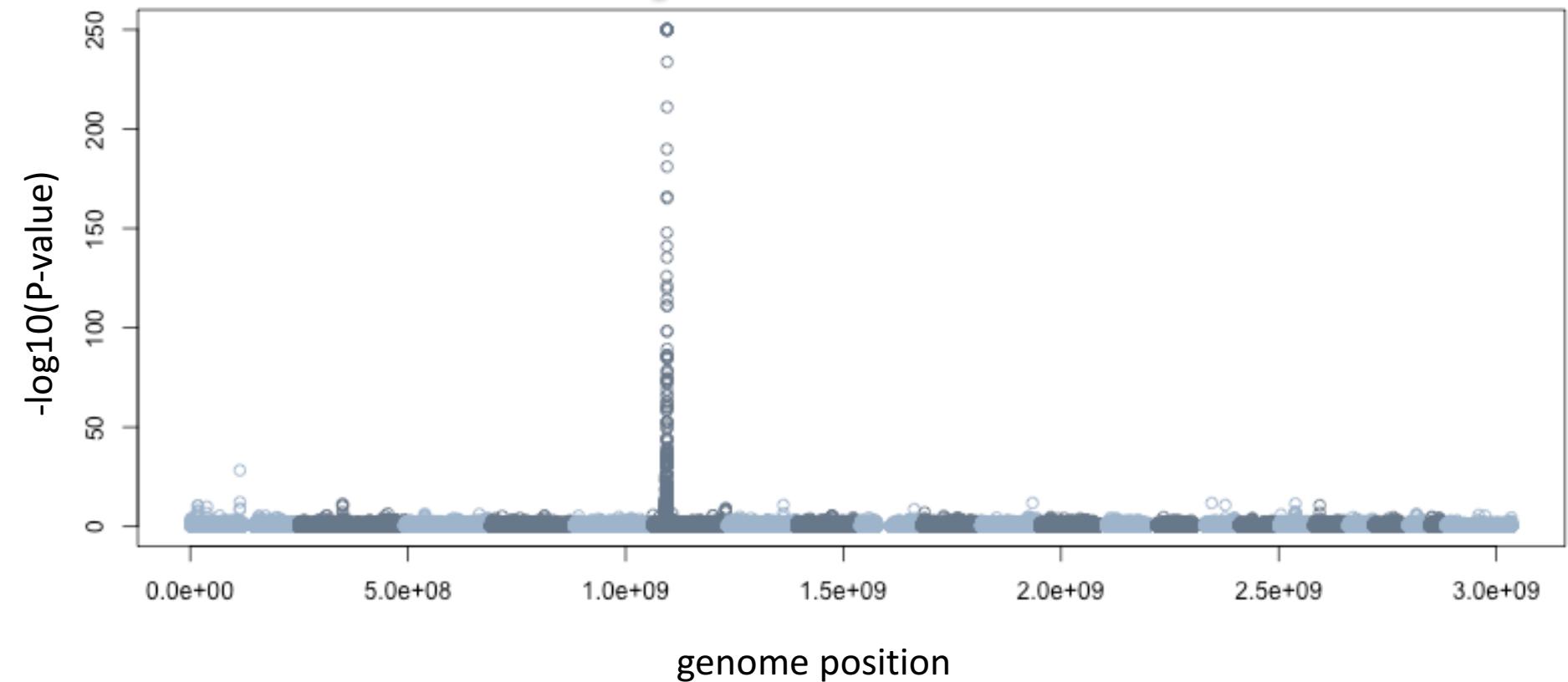
Discussion

- What can cause "deflated" p-values?
- Inflated p-values?
- What other information can QQPlot of – log10 P-values provide?

Exercise

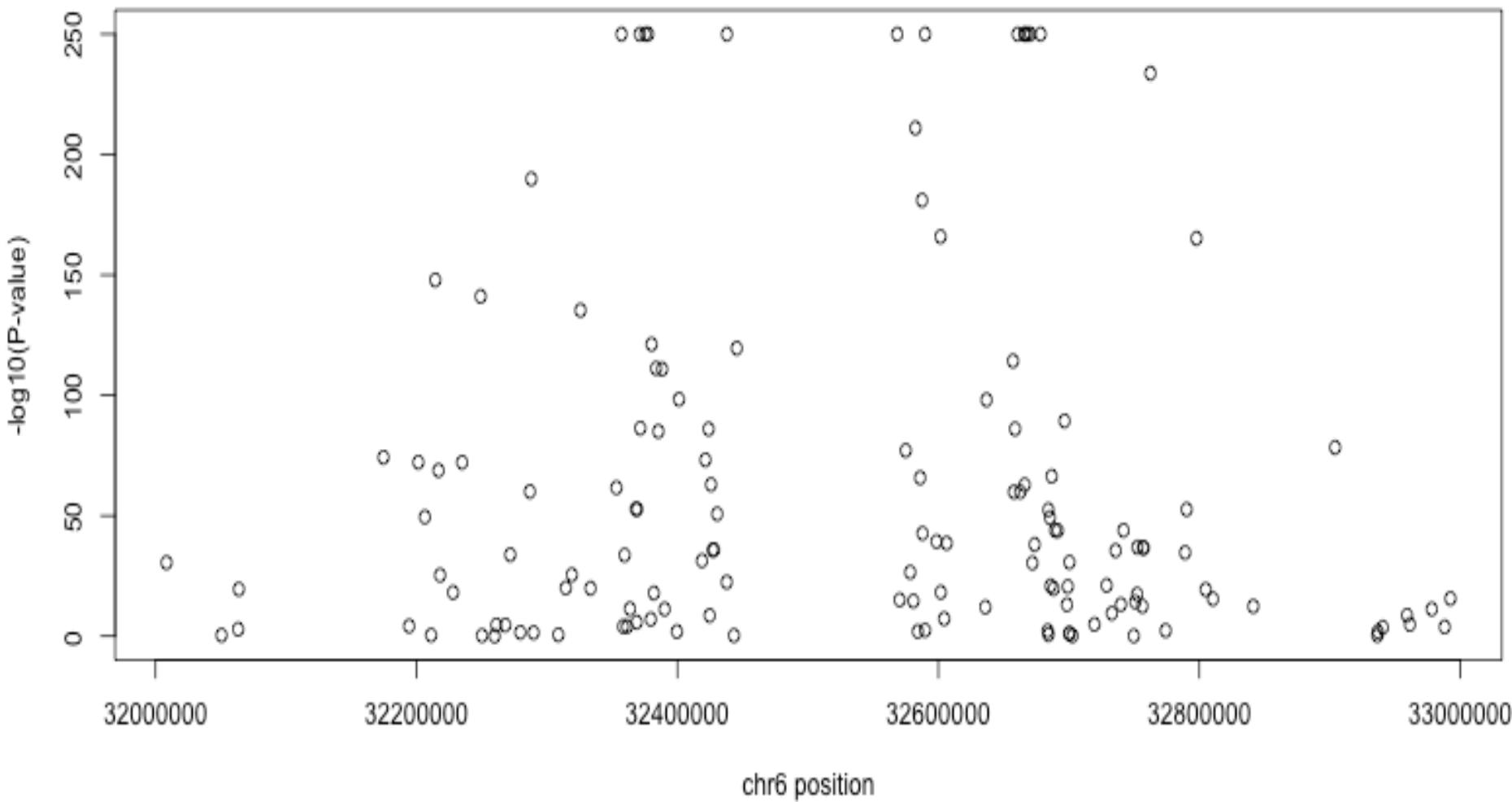
- Make a “manhattan” plot with SNP genome position on X-axis, $-\log_{10}$ p-values on Y-axis
- Try to make neighboring chromosomes different colors

What region is this signal from?

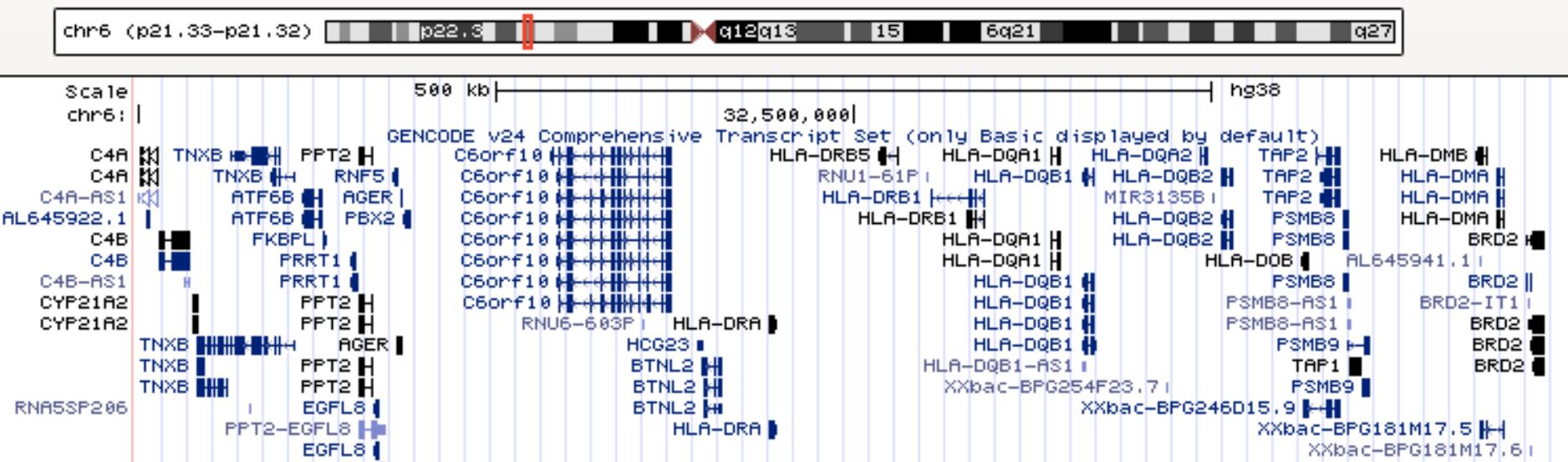


Exercise

- What SNP is giving the lowest p-value?
- What gene(s) are in this region (look up in genome browser)

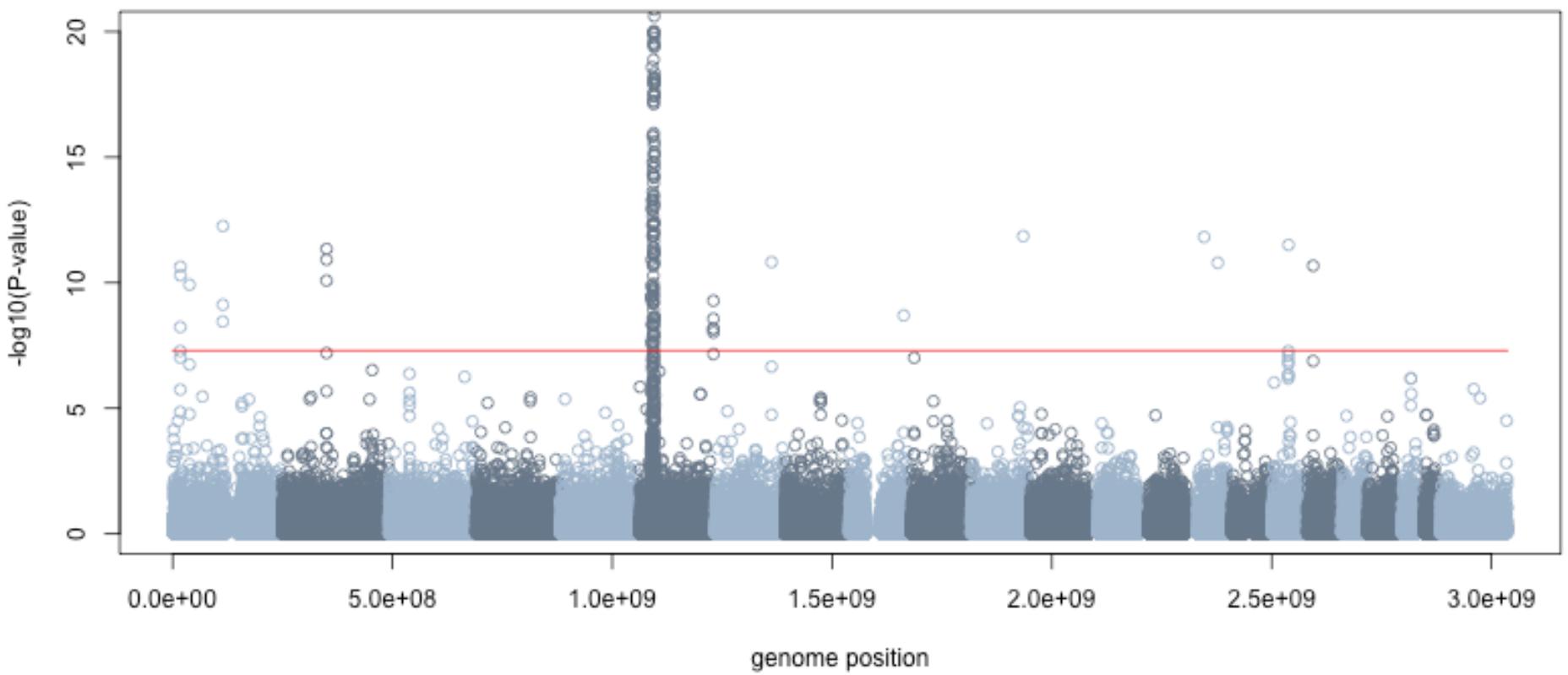


MHC Class 2



Exercise

- What about outside of the MHC region?
- remake manhattan plot, but threshold low p-values to $1e-20$
- Draw a line indicating genome-wide significance at $p=5e-8$
- Roughly how many significant hits are there?

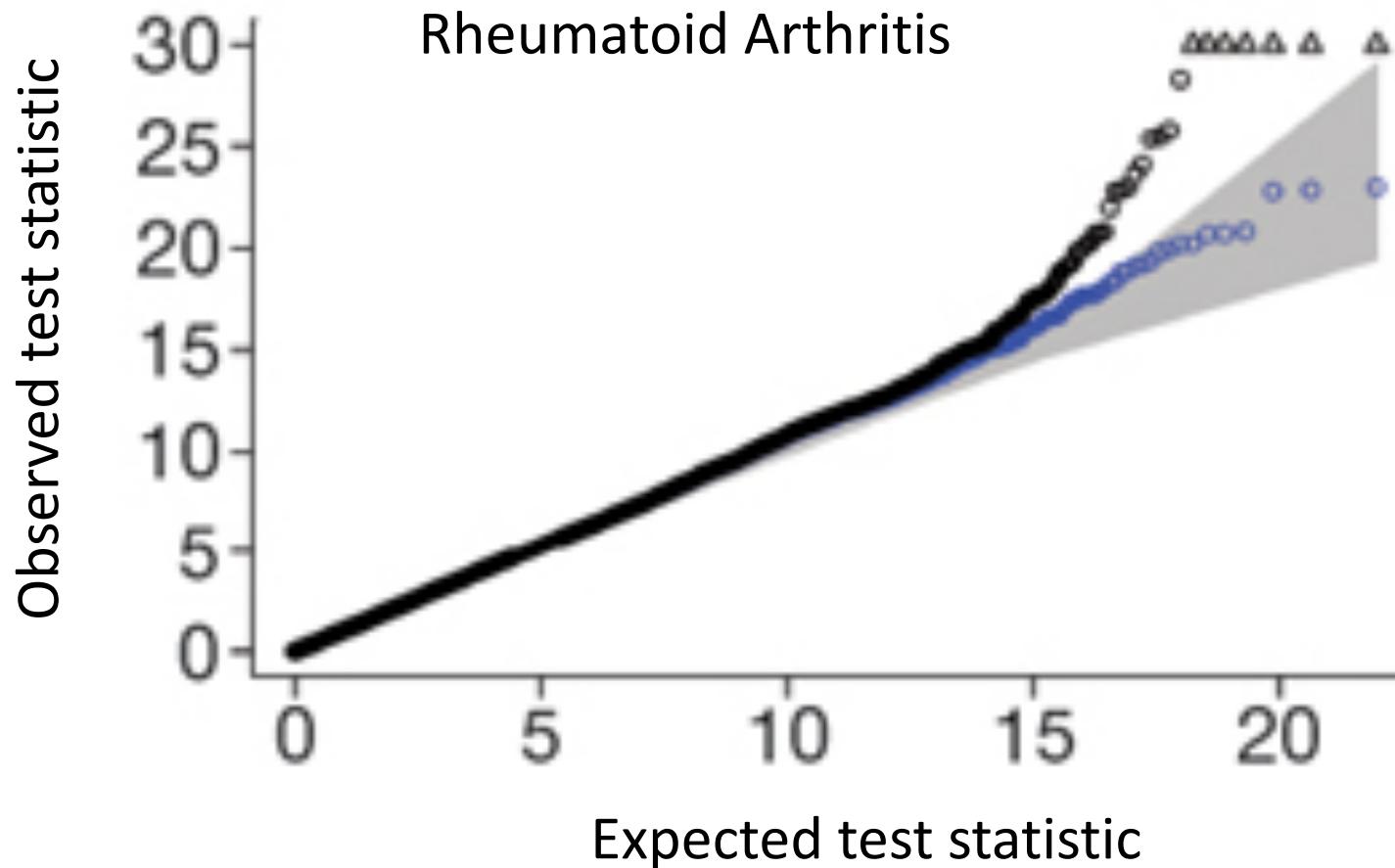


Discussion

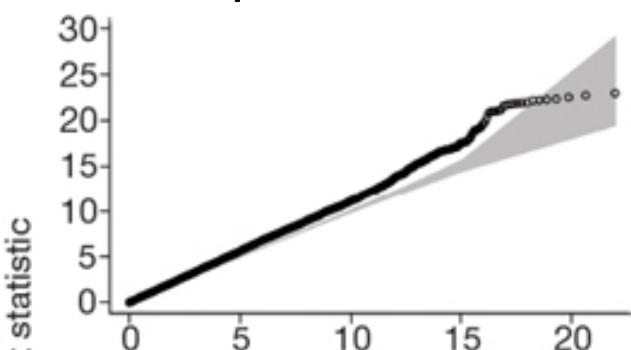
- Why use $5e-8$ as a threshold for genome-wide significance?
- Why not simply correct for the number of SNPs tested?

Wellcome Trust Case Control Study

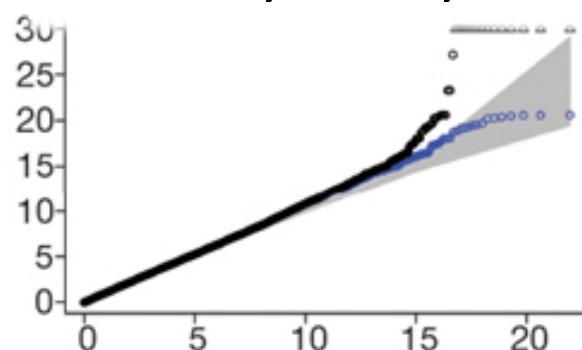
- 14,000 cases for 7 diseases
- 3,000 shared controls



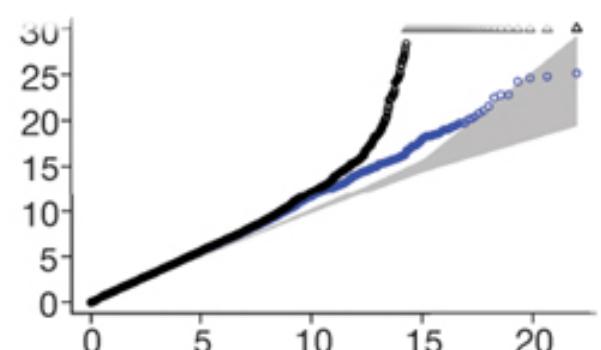
Bipolar Disorder



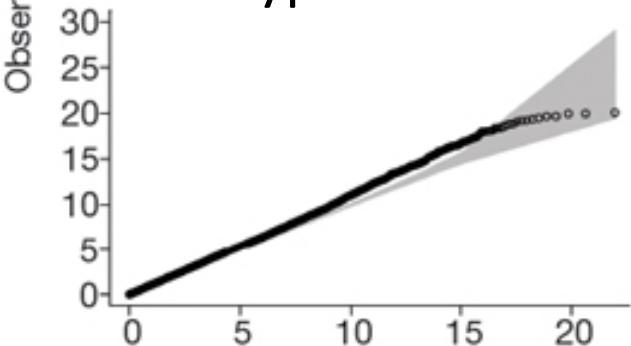
Coronary Artery Disease



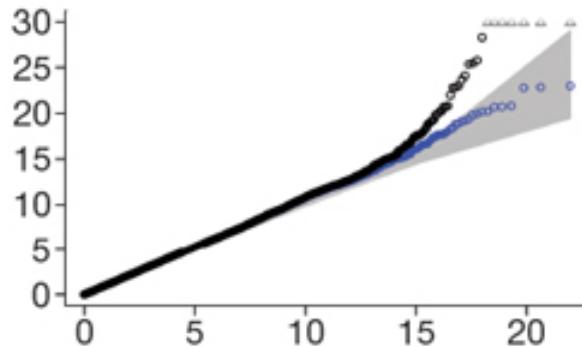
Crohn's Disease



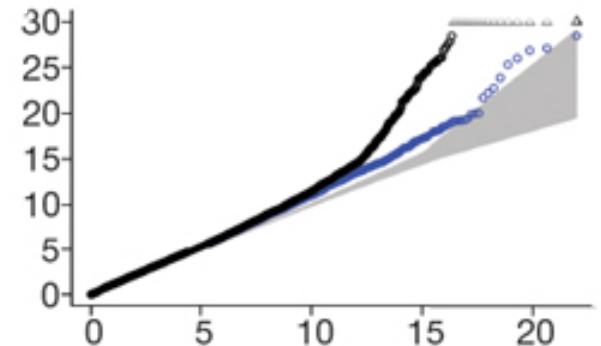
Hypertension



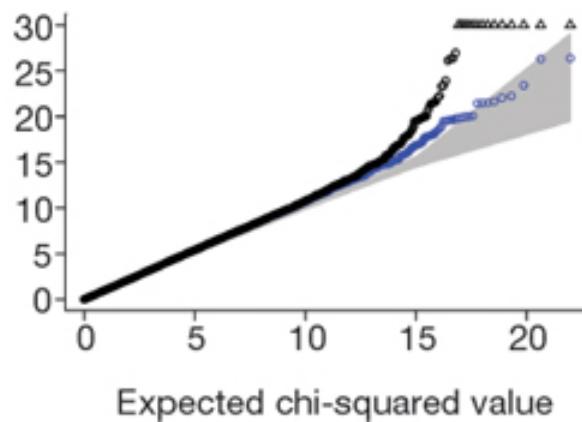
Rheumatoid Arthritis



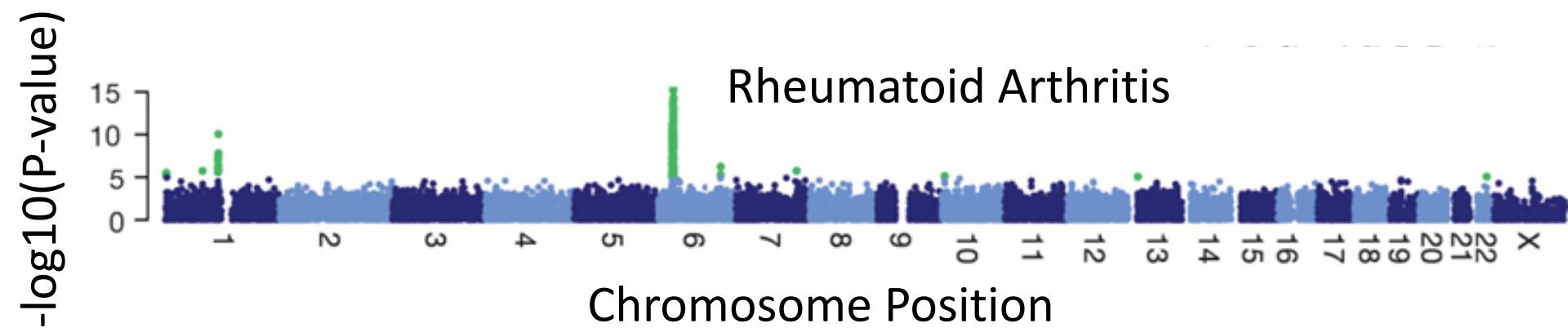
Type 1 Diabetes

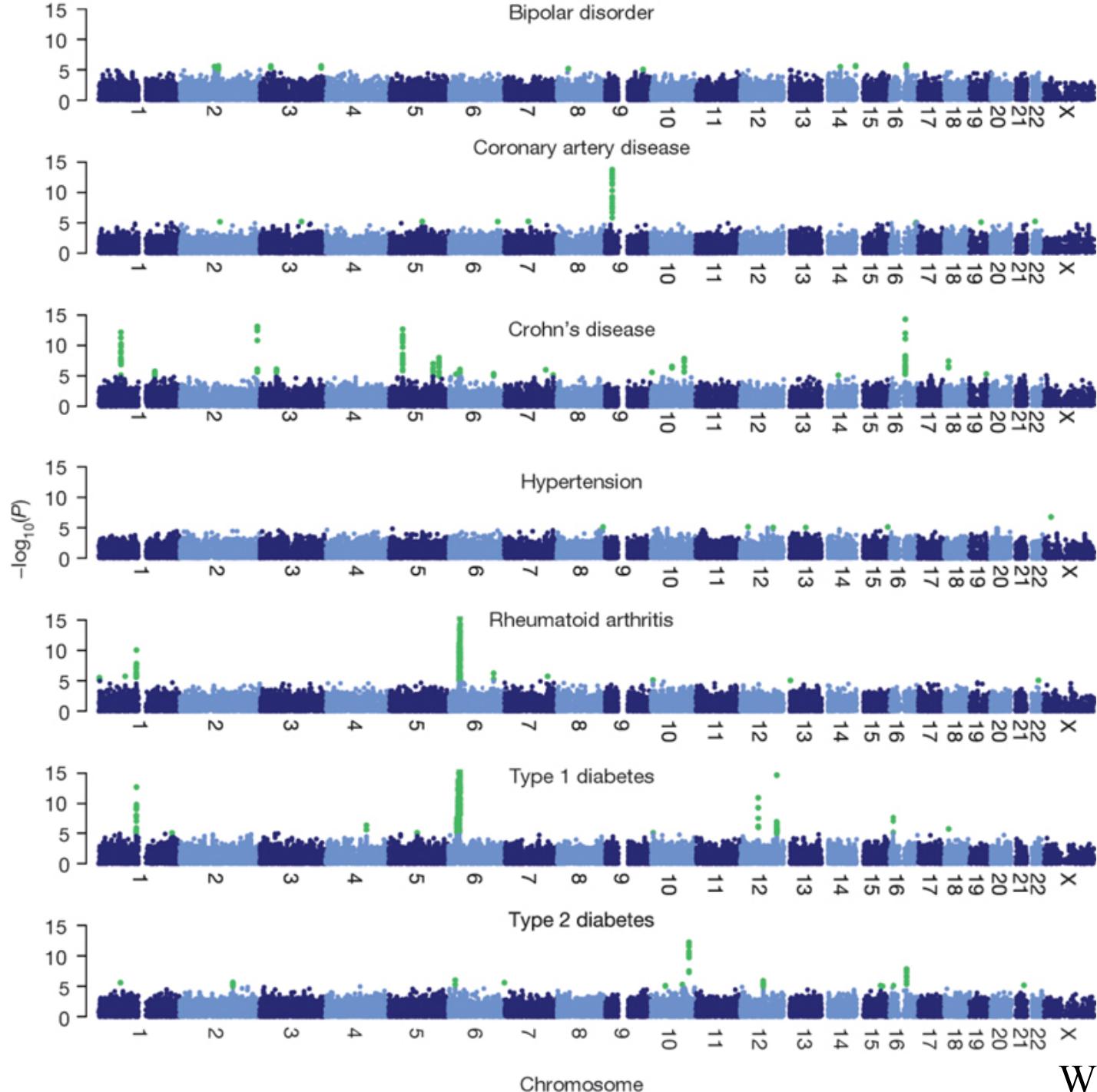


Type 2 Diabetes

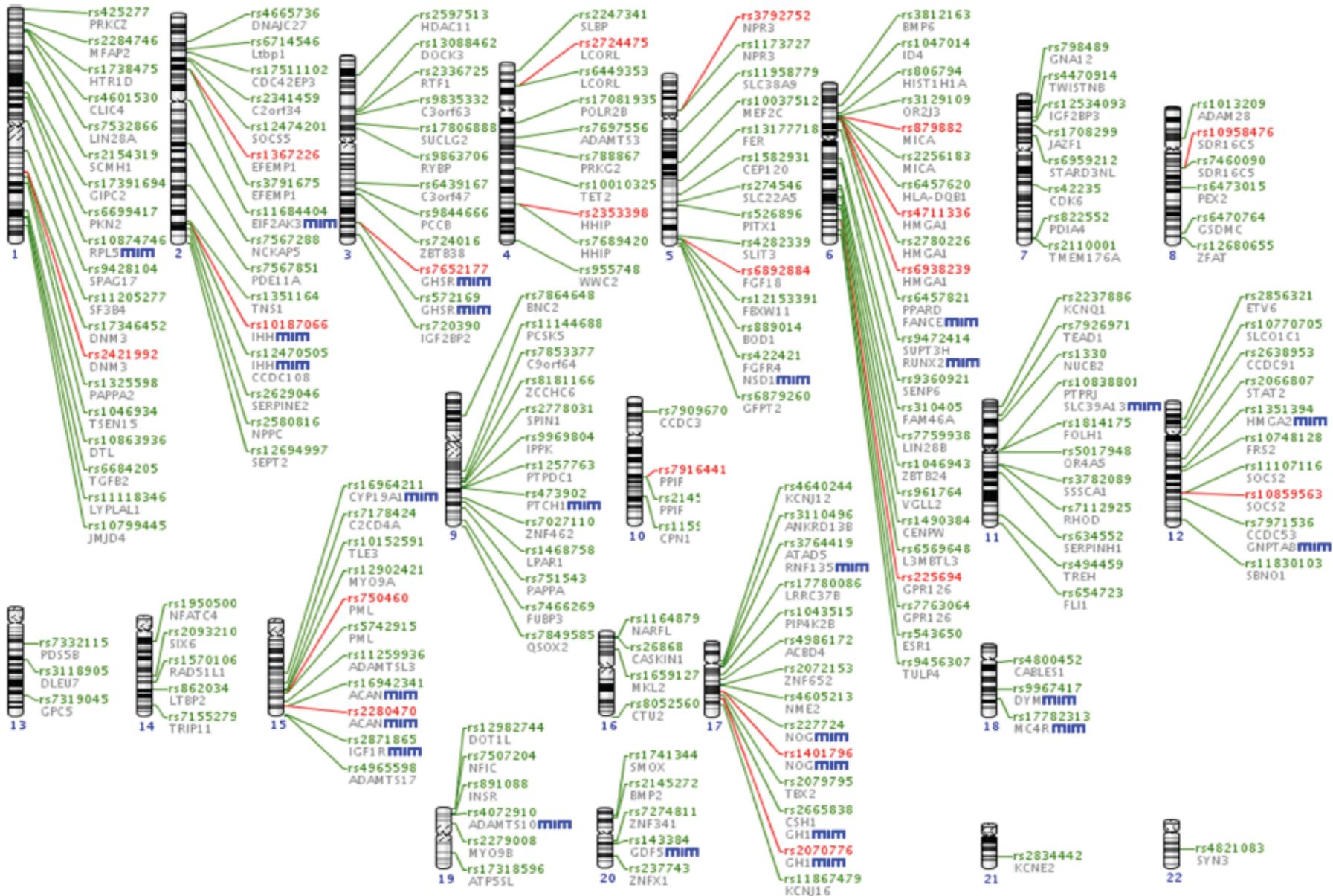


Manhattan Plot of Association Signals

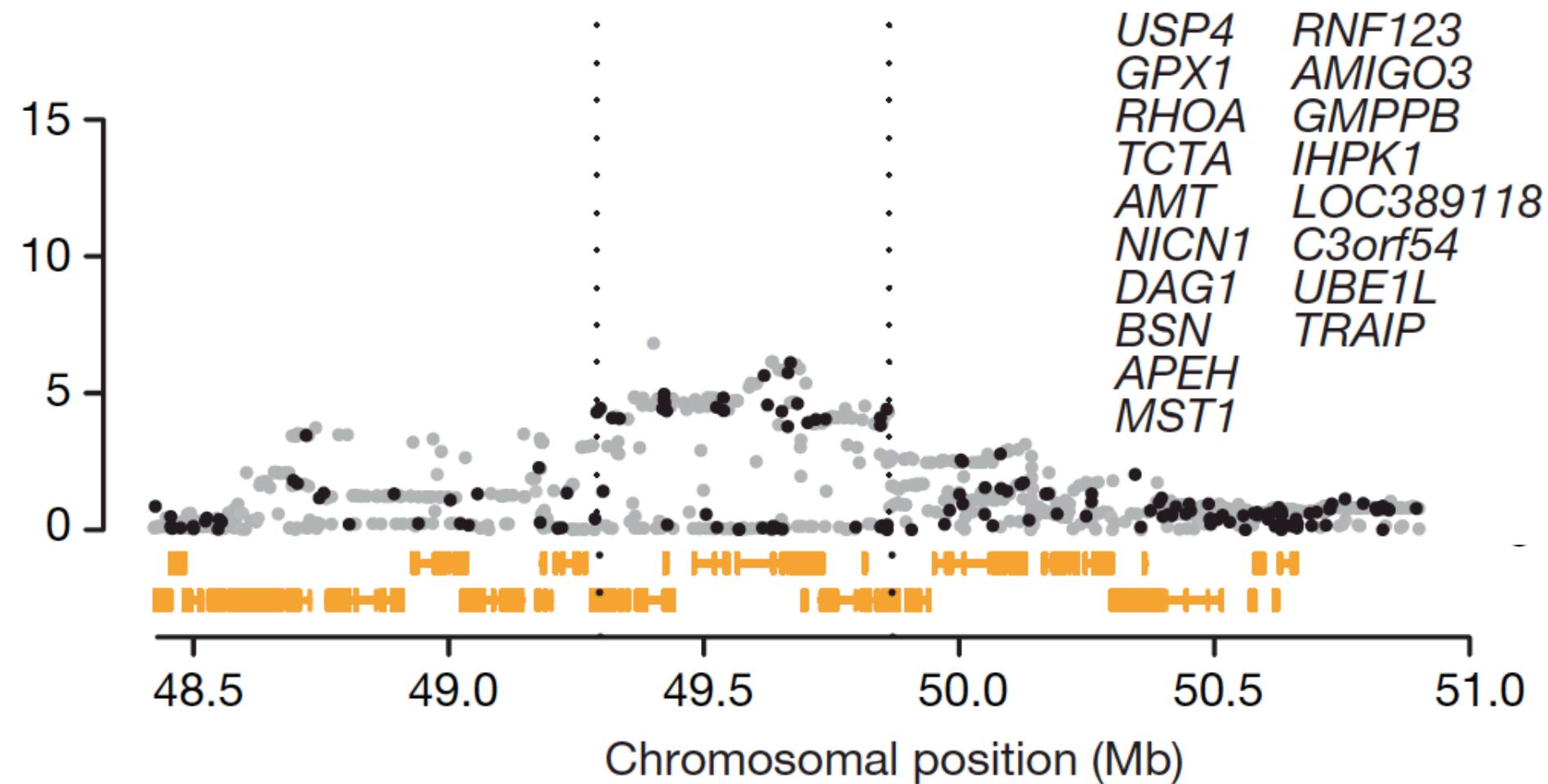




QTLs for human height



Crohn's Disease Hit Region, Chromosome 3



Discussion

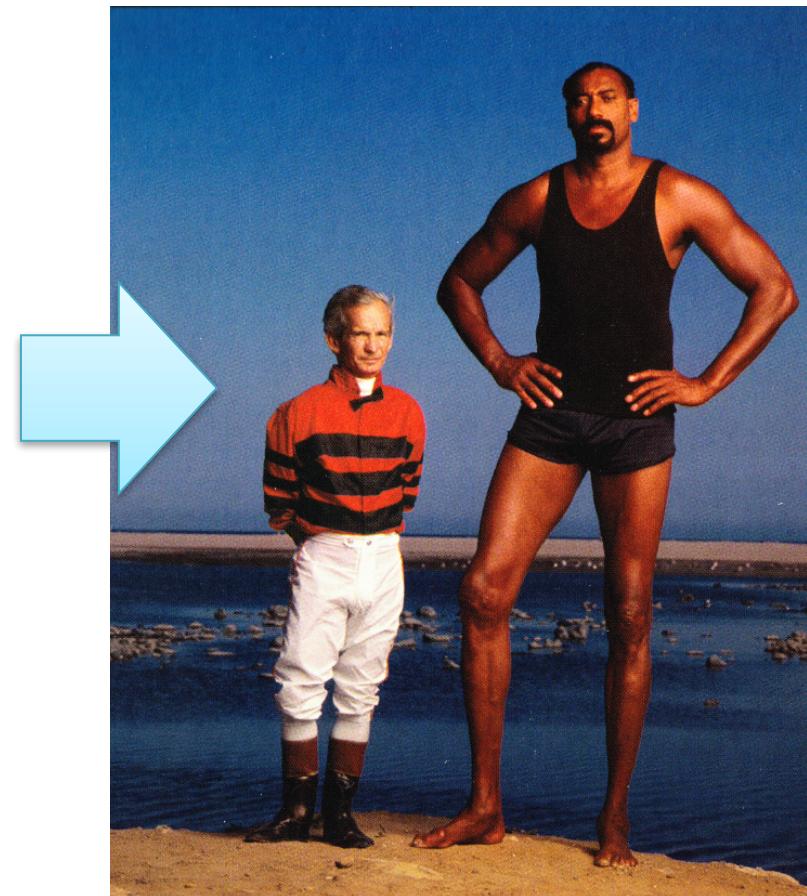
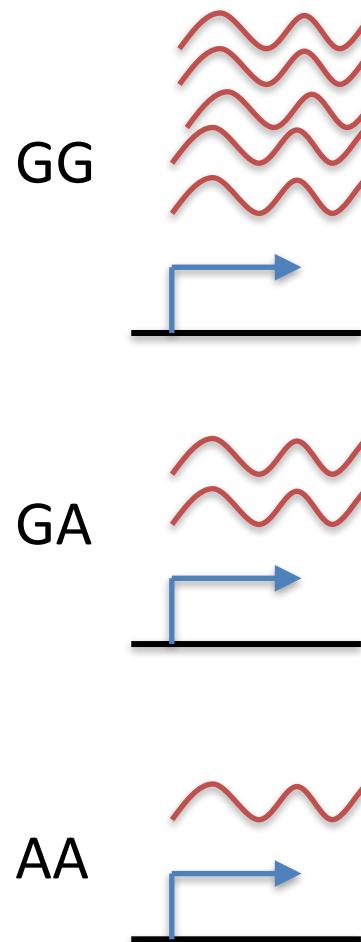
- What are advantages of GWAS over candidate gene studies?
- What are some disadvantages / limitations of GWAS?

Difficulty in interpreting GWAS hits

- Most associated variants (hits) are far from known genes
- Which gene(s) do they effect?
- What is their function / mechanism?
- What cell types are the variants active in?
- Which variants are causal?

Molecular traits as an intermediate phenotype

AAATAGAACCTTGAGGTGAATCTCTTTAATGTAATG
TTACACTATAACAATAGGAGTAGAAATTGTCACCACTC
TTCTTTTTTATTTYACATTAaaaaaaaATTTACTT
GGATGTGCAGGTTGTCATAGGTAATGTTTATTT
ATYTTAACATACAACCTCATAGCCAGTACTTAAAT
GTTGATGAATTGCACTCTGTTAATAATTGCTACAGATT
TGATGRCTCGGGTTGGCATAAGTACATGAAGATTAC
GATGAAGAAATTAGTGTTTTTAATTAAAGAA
TTTATTTCCCACTGAAAATAATCCCACCGGGATAAA
AGTAACCCAACCTGAAGCTAGTTTCAGACTTAGGCA
CTCATGGTCACCCCTGCAAGAGAAATATCTAATTGAAA
TTGWTAGTGCTAAATGACATAAAGGGATCTCACTGGG
TTGTTARAGGTTAAATGTTAGTAACTGTTATTGCA
TAAATAAAAAAACTTTTTTAAAGATAAAATTACAGA
GAGGTTCTGACAAACCCCTCCAGTTCTTACTAT
ACATTGTCACAATTAGTGAACCAATTGATACATTA
CTTTACTGGAGAATGGTTAGAAACTAAGGTC
TTGAGATGTGTATTTTAGGTTCTTCAGC
TATATTAACCTATGTGACACATACATCTATGAT
TTATTAAGCTAAATATGAGTTCATATGGTGTCTT
TATTCTAGCCTCTCTTGCTTATCTGTAACCT
TCTTCCACACCACTACTATCTGCTTACCTAATTTC
CTTCAGAAATTACATACATACCTGGGATACAACT
TGTAAGTTCTCTATCTTACTGACTCTACTC
GAACATTCACTCATACTCCCTAGTGAAGTTTC
TTTYTGCAGTCTGCATCCATTAGGGTCCCT
TTTTAAATTCAATACATCAAGTTATTCTTGTGC
CAAATACAAAGTGTGTACCCATCATTACAATGTCA
AAAAAATCCCTGTCTTGCCTATTCAACCCTTCCC
ACCACTGATCTGTTATCGTGGAGCTGTCTCTTCCA
ACAATATGTAGACTTTCACCCCTGGCTTATTGTTAG
CATGTCCTTATGTGCTGTAGTTCATTAACCTTACT
AAATGTRCCACARTTGTATCCATTYGCAGATTGAA
ATGACTGCTAGATAGYATAGTAAGACTATATTCA
TTTTAAAGTGGCTGTACCATGTGCCACCAGCAACTCC
GTTTTTGAGTTTAKCATTCTAAAGGTGAGTGATG
TGTAATACTCTAAATGACAATGATGGTGGATTCTT
ATATCTTCTTACTGATGTGCTGTTGGATGTTGCT
TTGAGTTCTTCTTCTTCTTCTTCTTGTGAGAYG
TGGAGTGCACTGGGCCATCTGGCTCACTGCAAGC
TTCGTACCTCAGCTCCGAGTAGCTGGGACTACAGGC
TTTTTTGTATTTGGAGACGAGGTTTACCATG
CCTGACCATAGATGATCTGCTGTCTGGCCTCCAAA



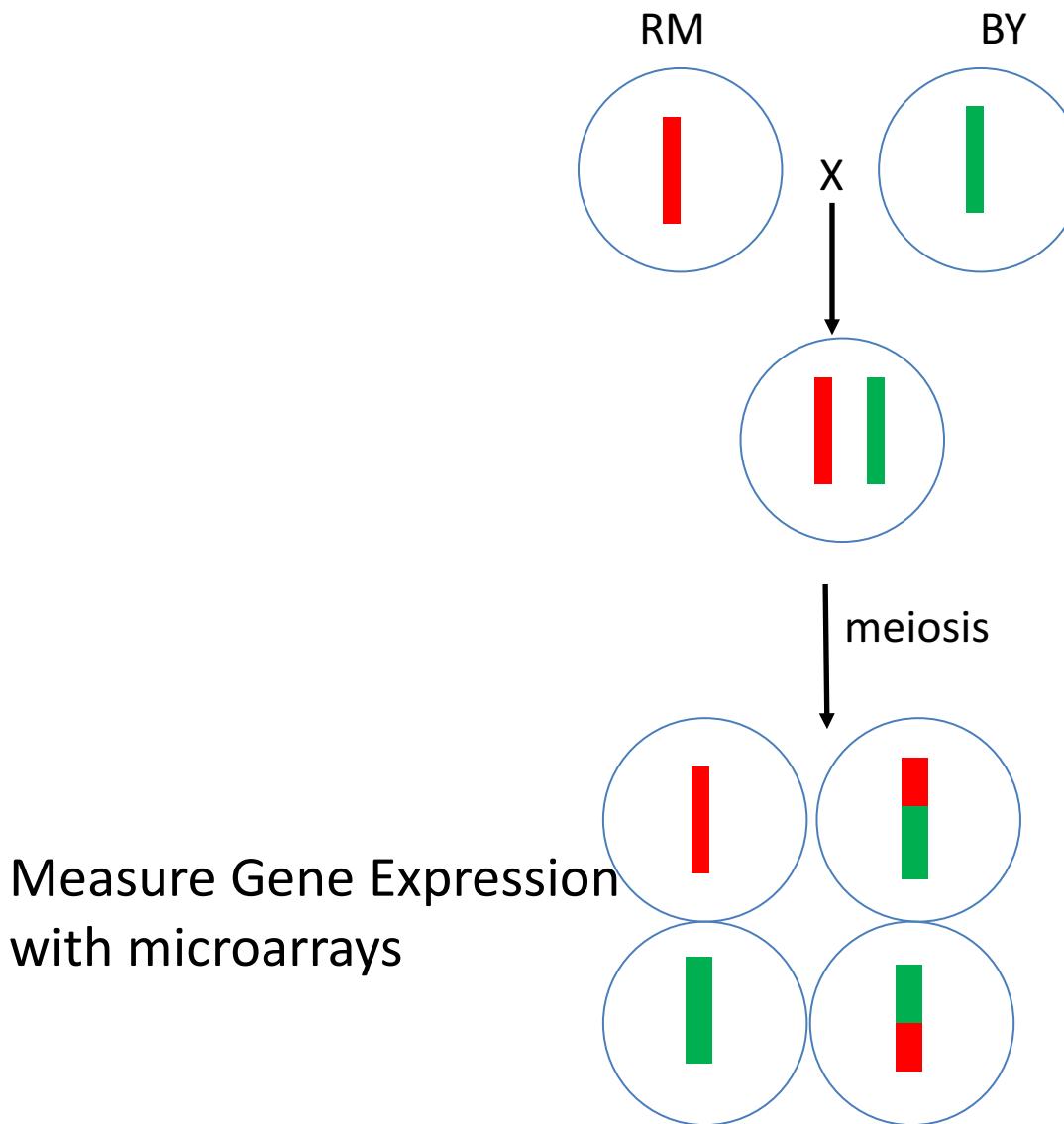
Why use molecular traits?

- Close to underlying genetics
 - affected by small number of polymorphisms
 - require smaller sample sizes
- Can measure 1000s of traits in single experiment (e.g. RNA-seq)
- Reveal molecular basis of organismal traits
 - implicate specific cell types

Discussion

- What else could be used as molecular or cellular traits?
- How would you measure these?

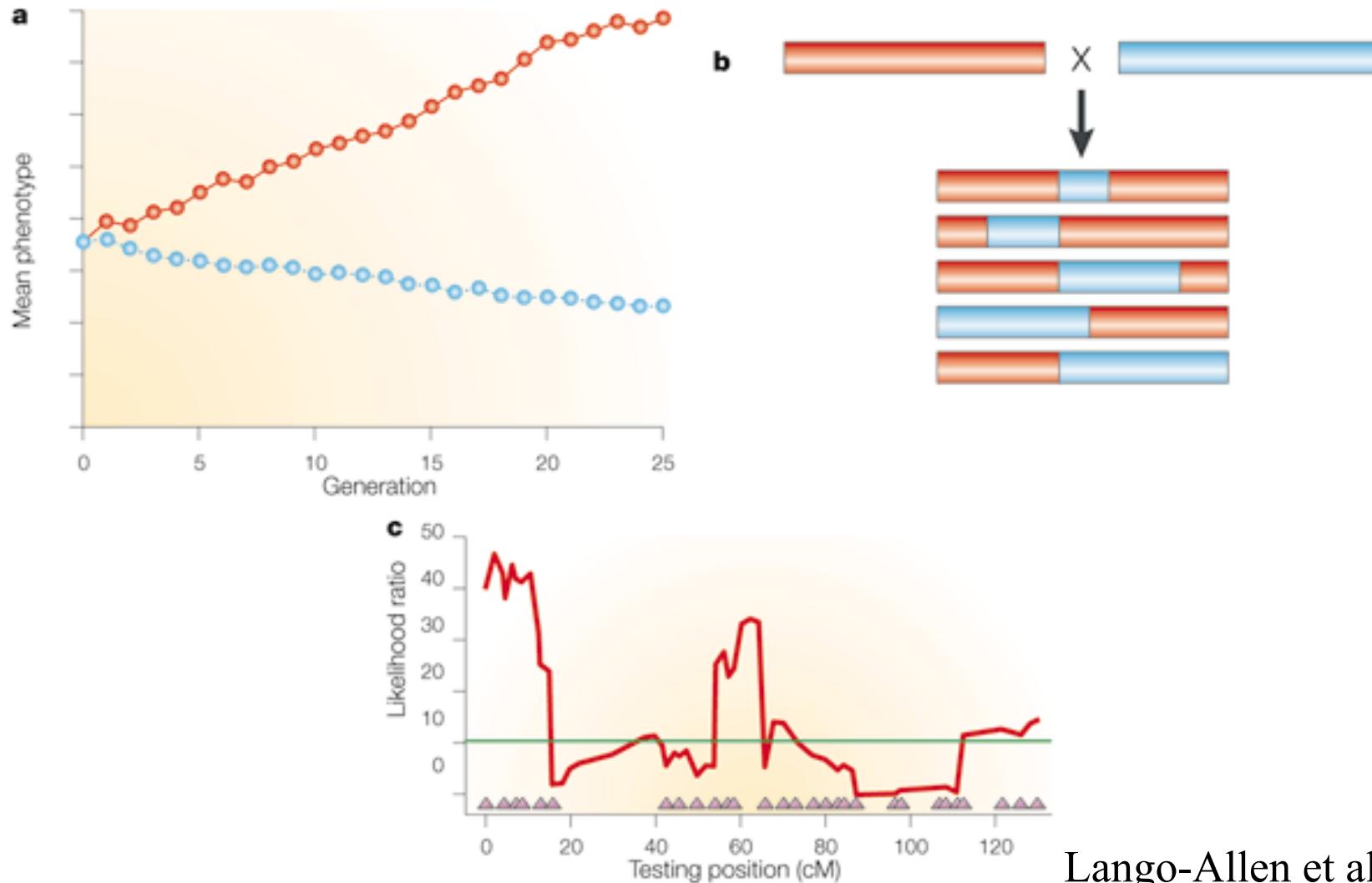
The first gene expression QTL mapping study



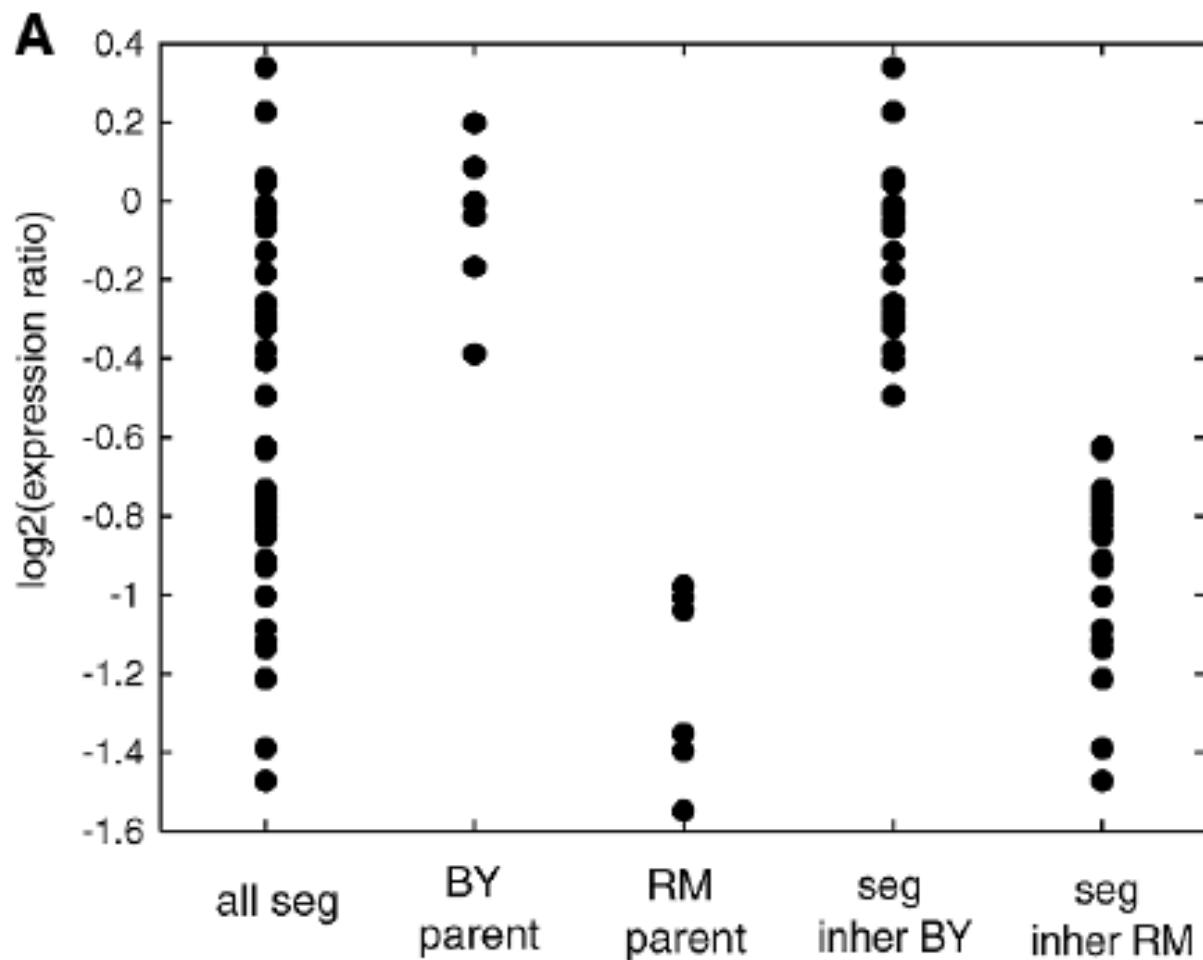
Cross two diverged yeast strains “RM” and “BY”

Genotype “segregants” to identify parental genome segments

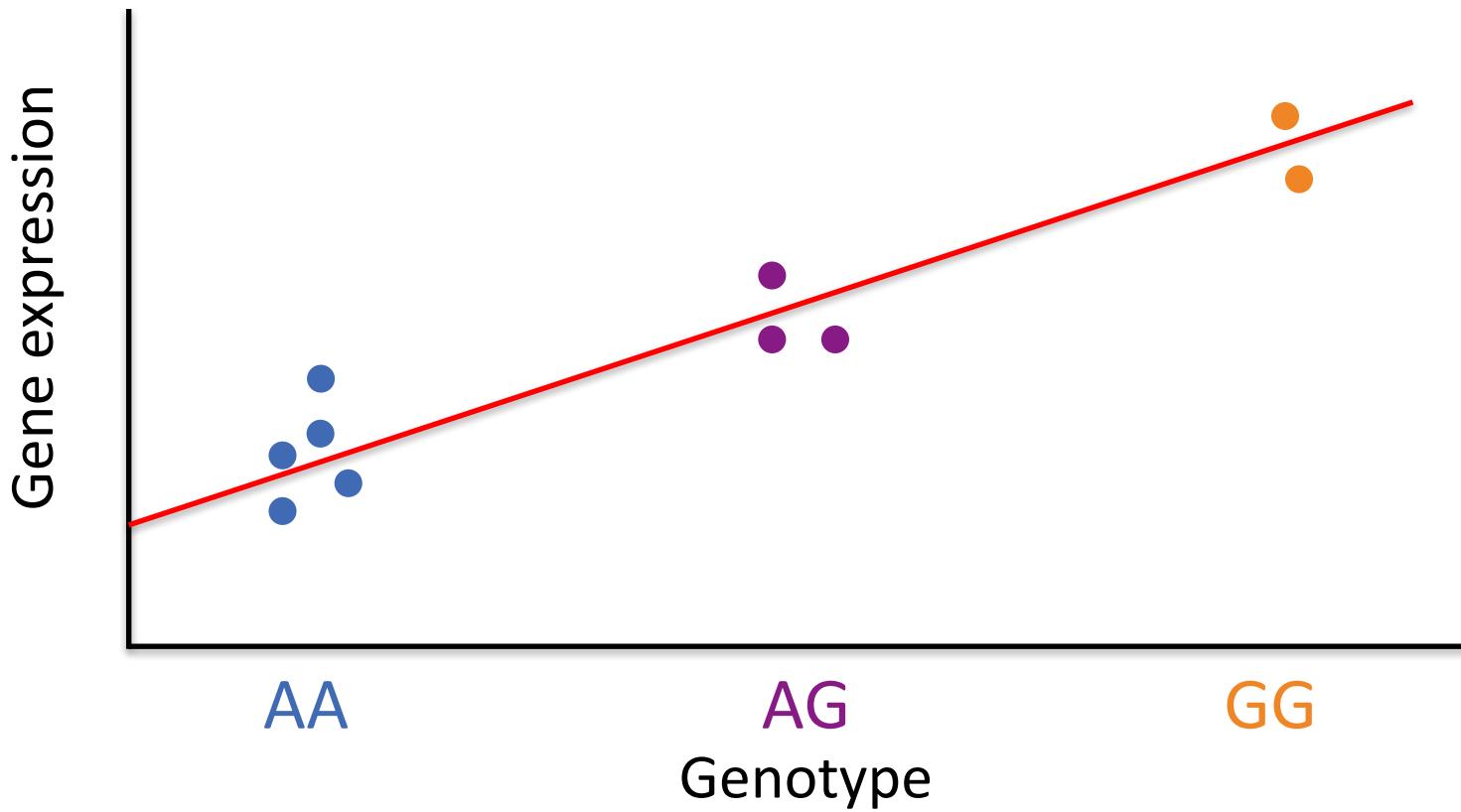
Quantitative Trait Locus Mapping with recombinant inbred lines



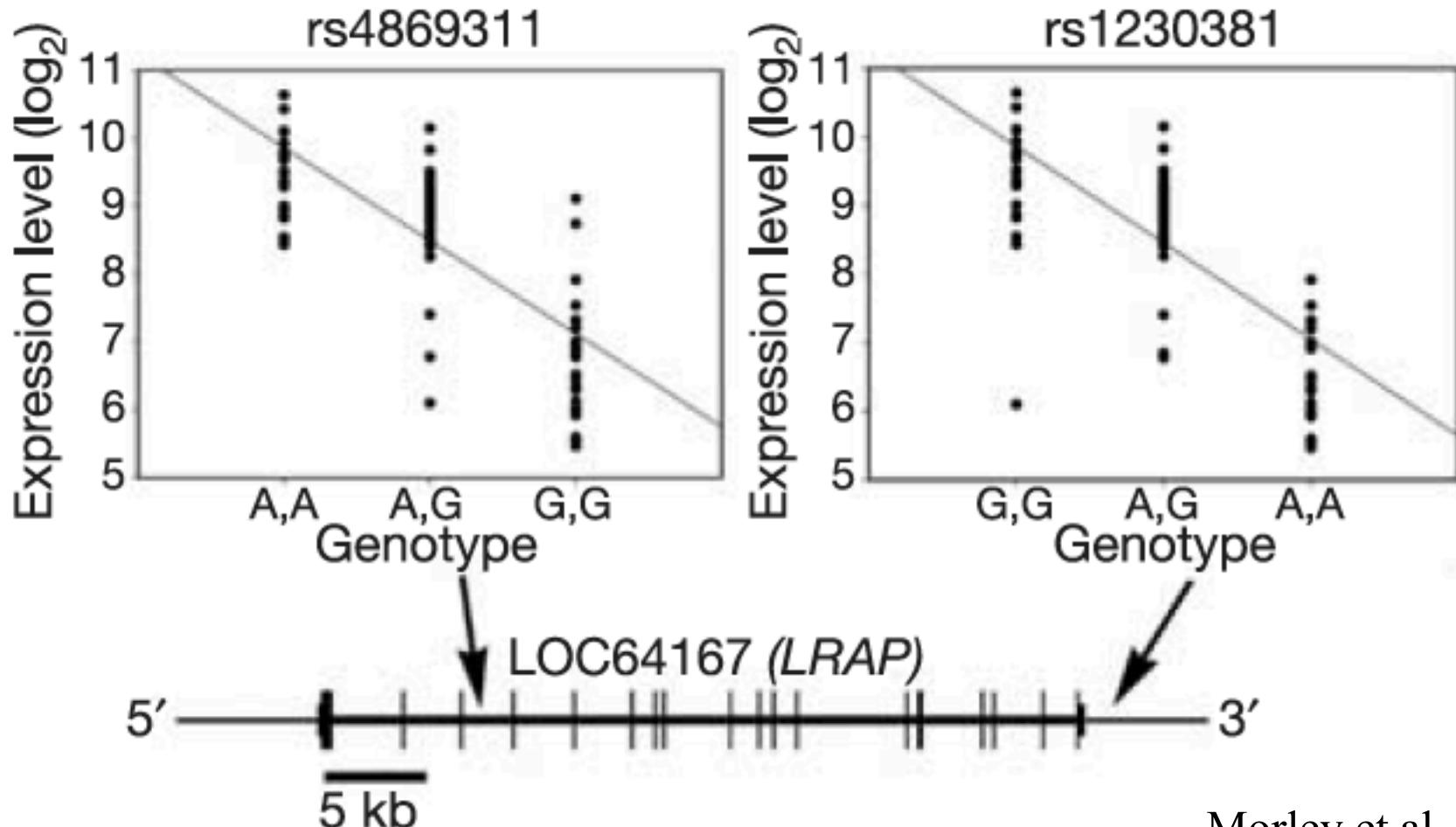
A yeast eQTL



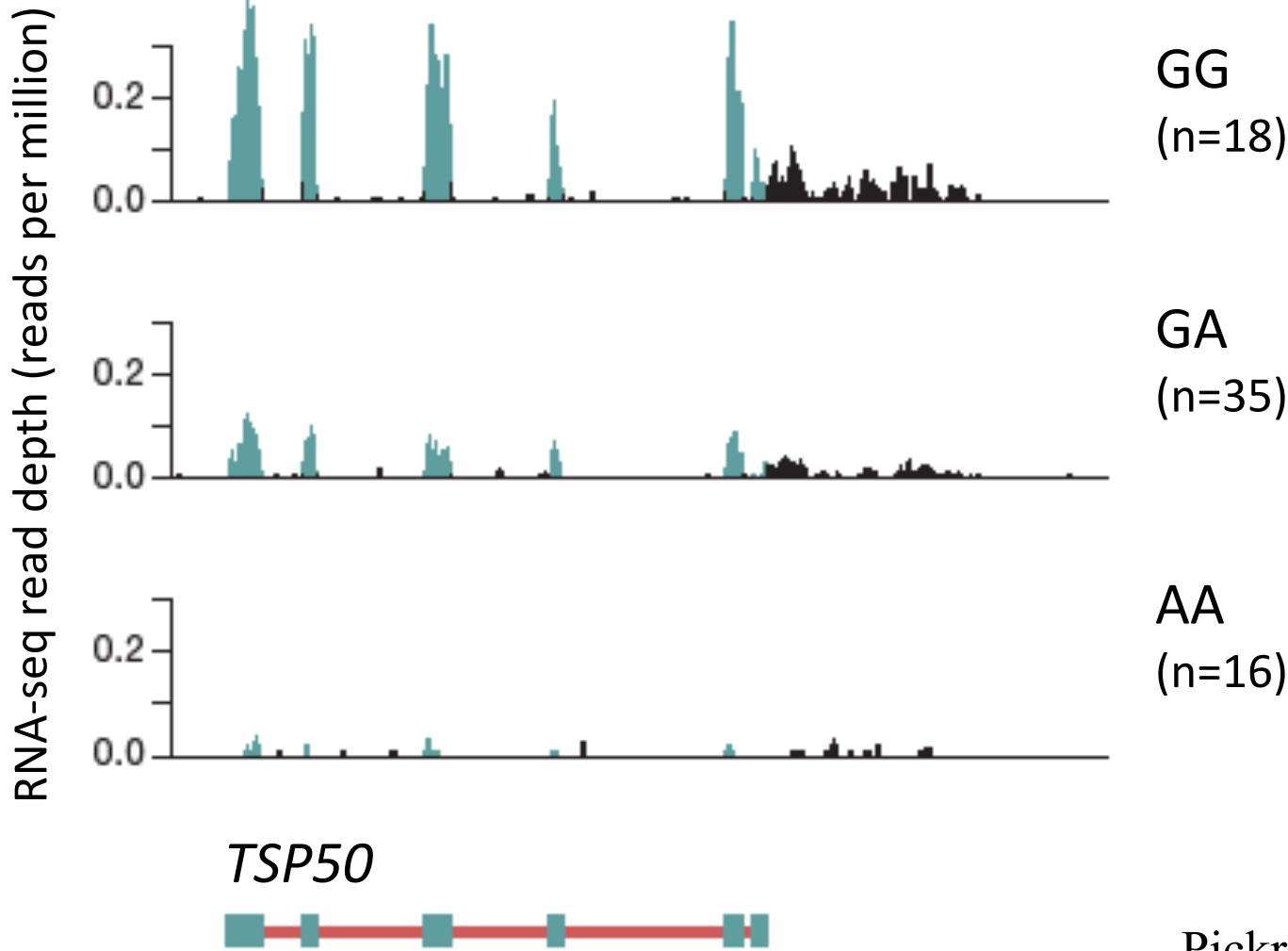
eQTL mapping in humans



Pedigree-based mapping of eQTLs in human B cell lines

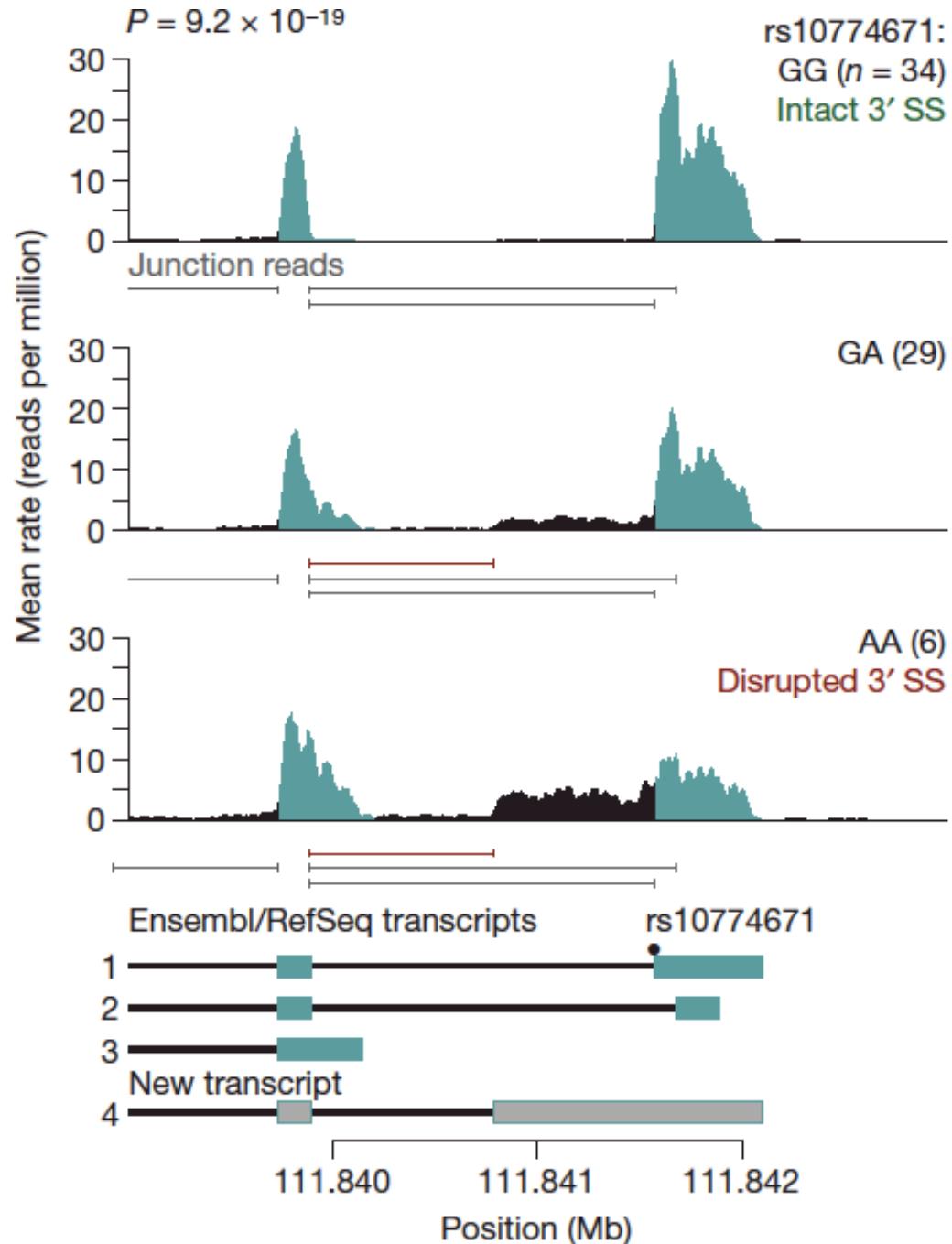


eQTL mapping with RNA-seq

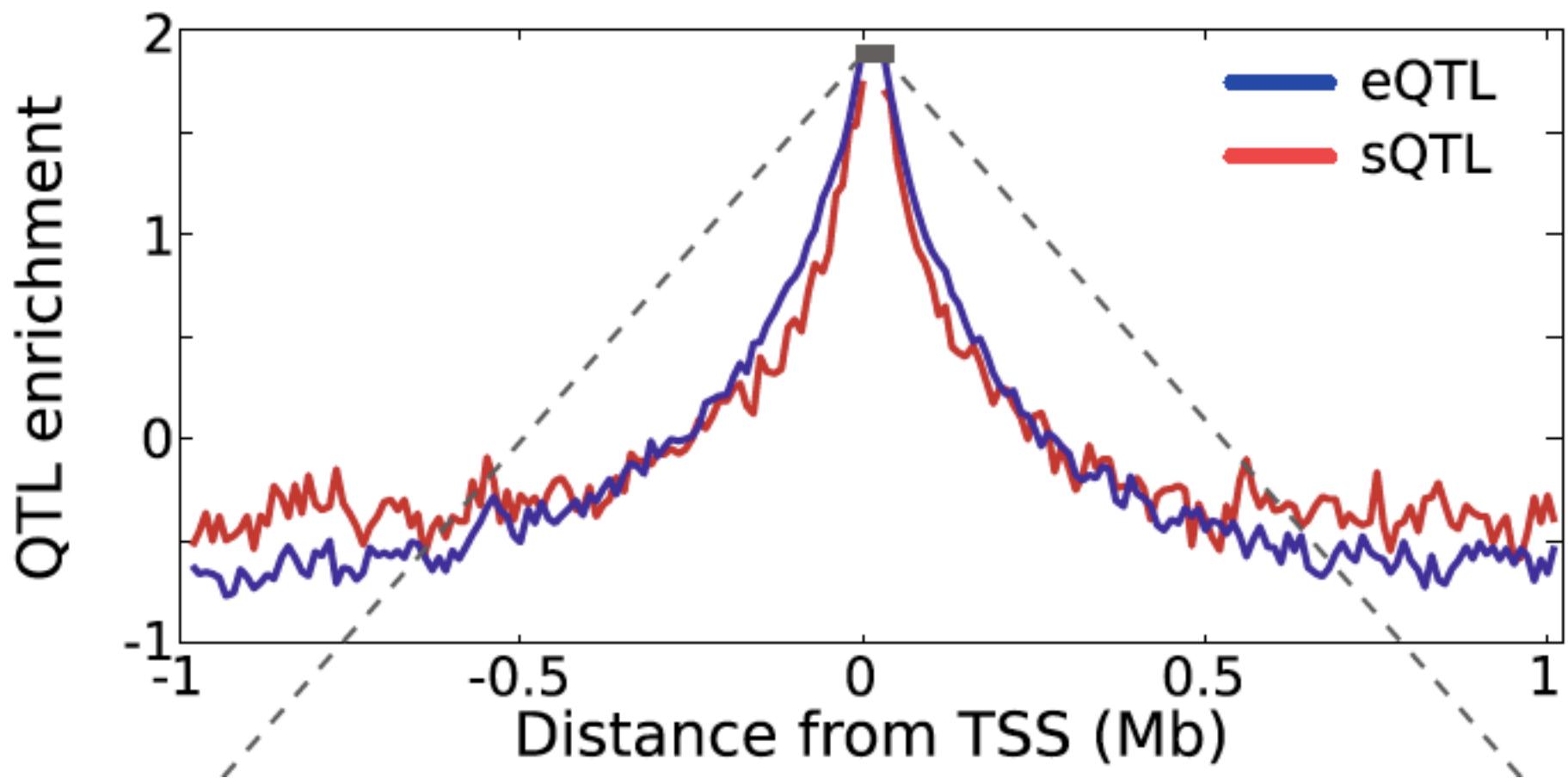


Pickrell et al. 2010

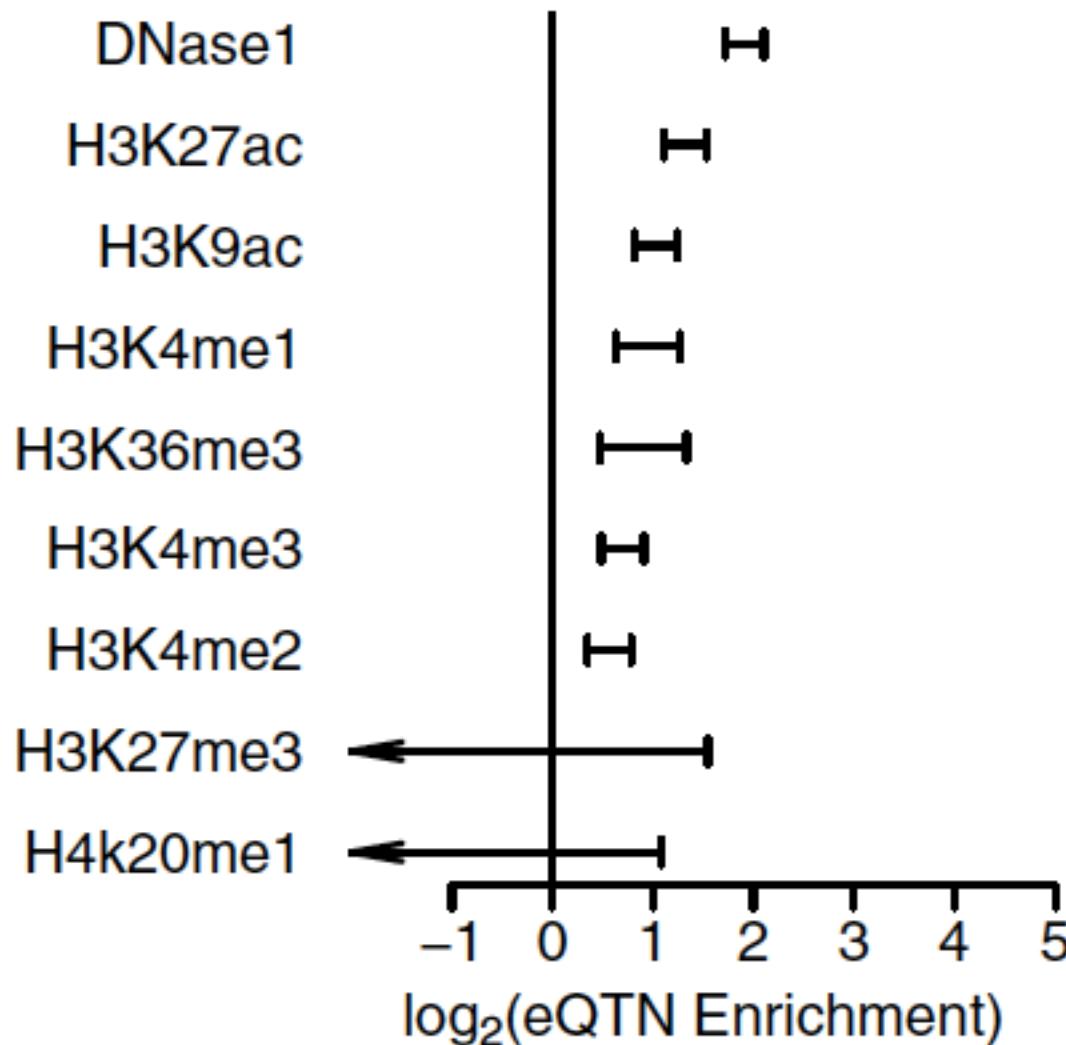
Splicing QTLs (sQTLs)



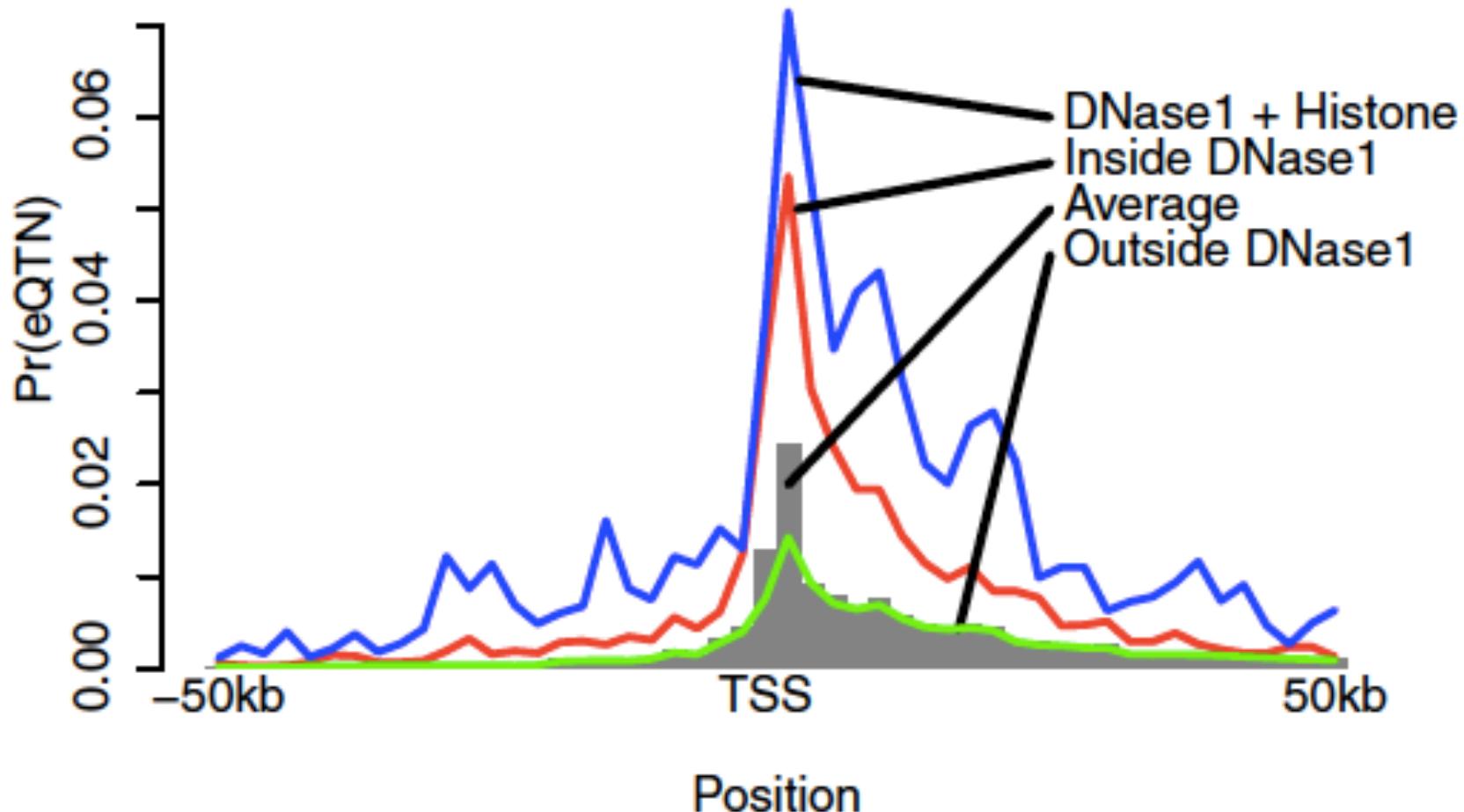
eQTLs are enriched near transcription start sites



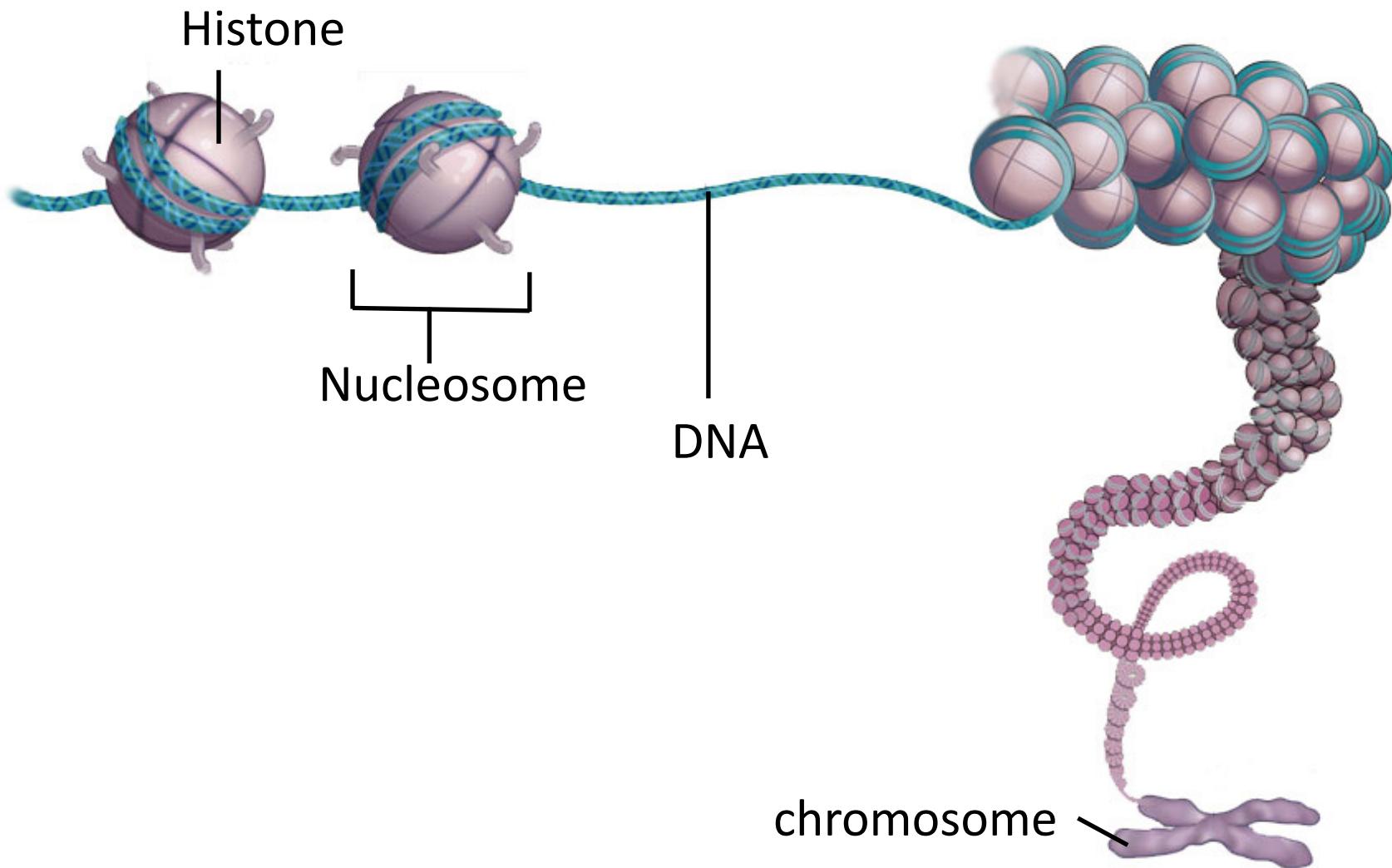
eQTLs are enriched in regions with open/active chromatin



Distance from TSS and DNase sensitivity are predictive of eQTLs

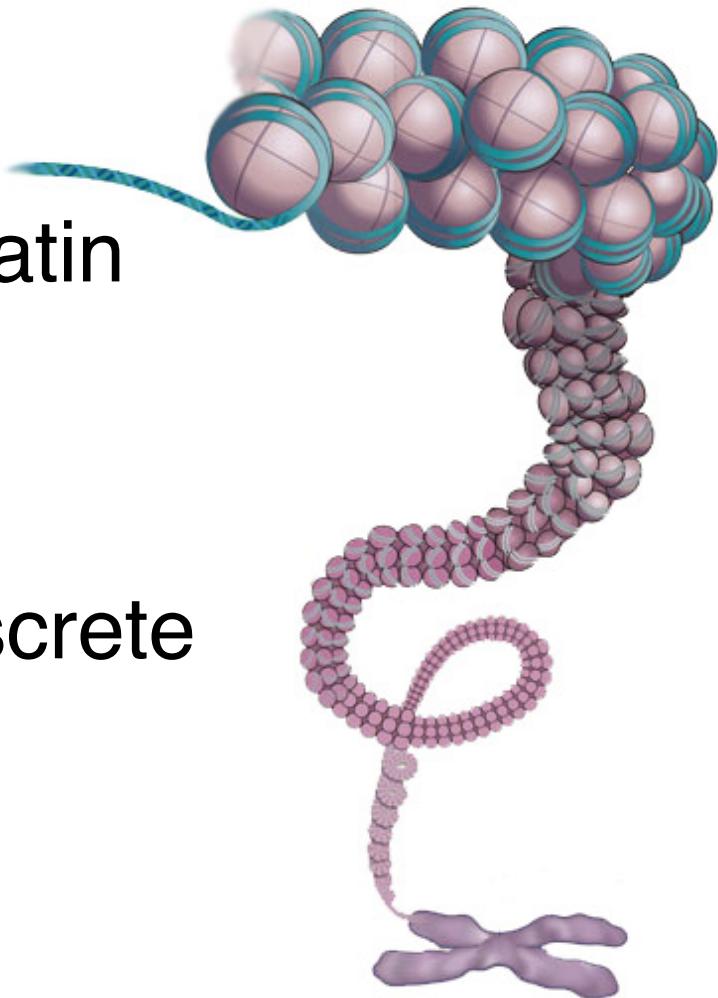


Chromatin as a molecular trait

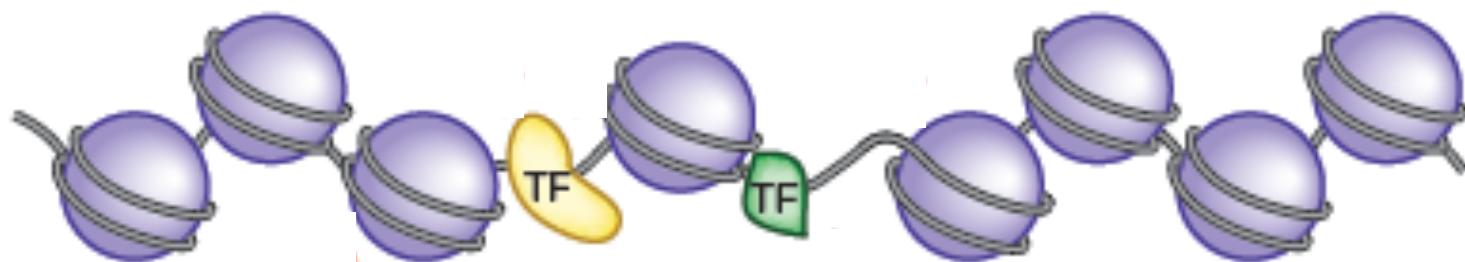


Discussion

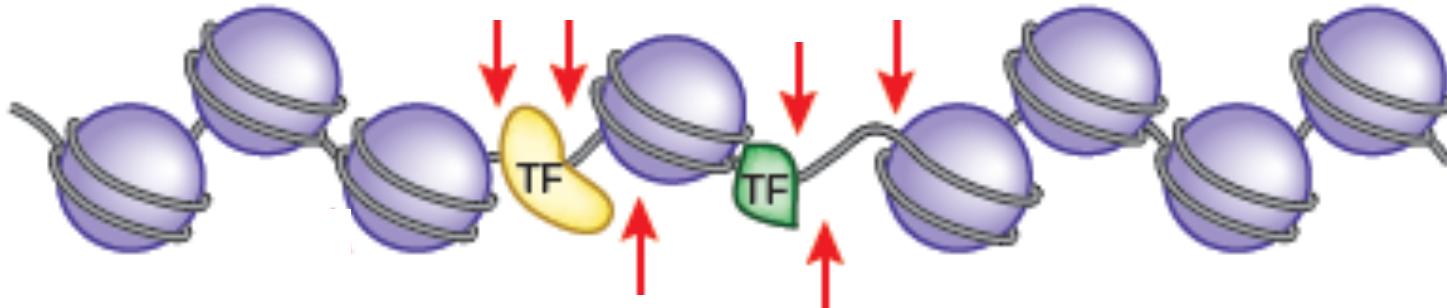
- What aspects of chromatin could be treated as a molecular trait?
- Do these traits have discrete or continuous values?



Nucleosomes are depleted in regulatory regions

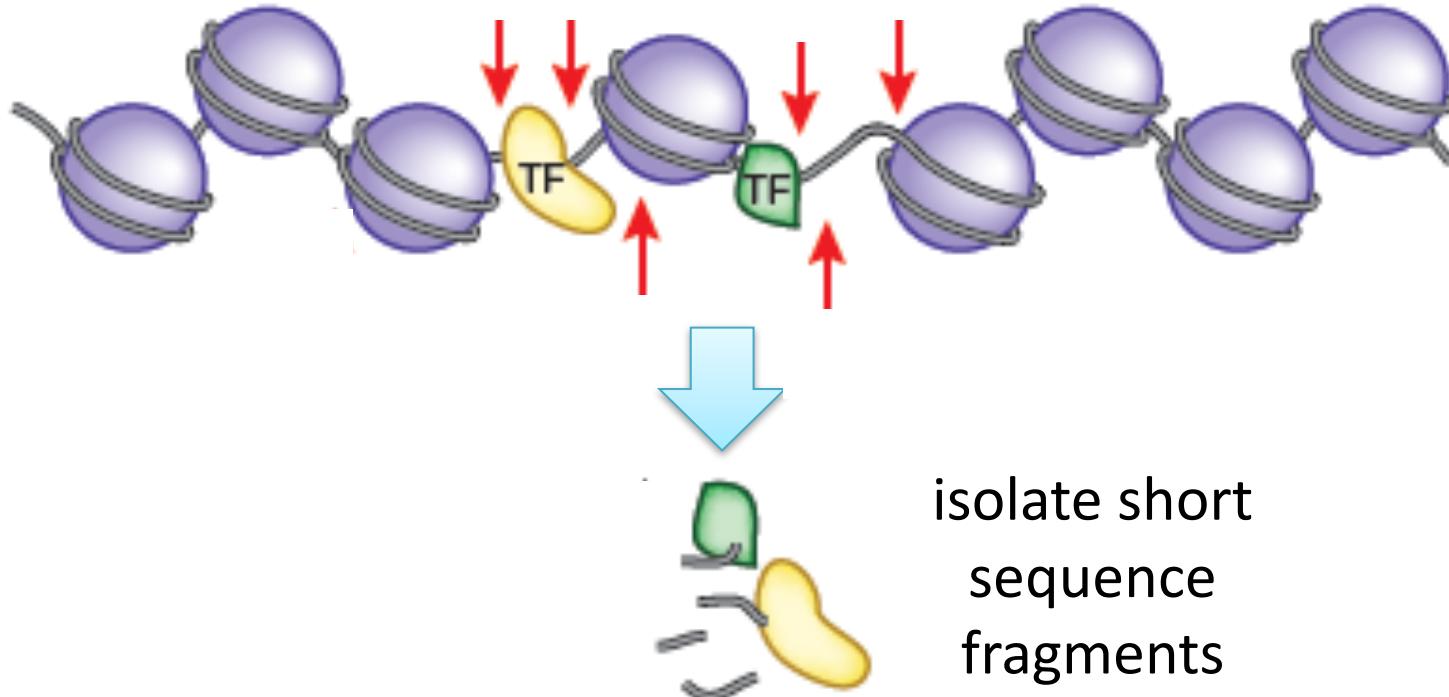


Measuring chromatin accessibility with DNase-seq

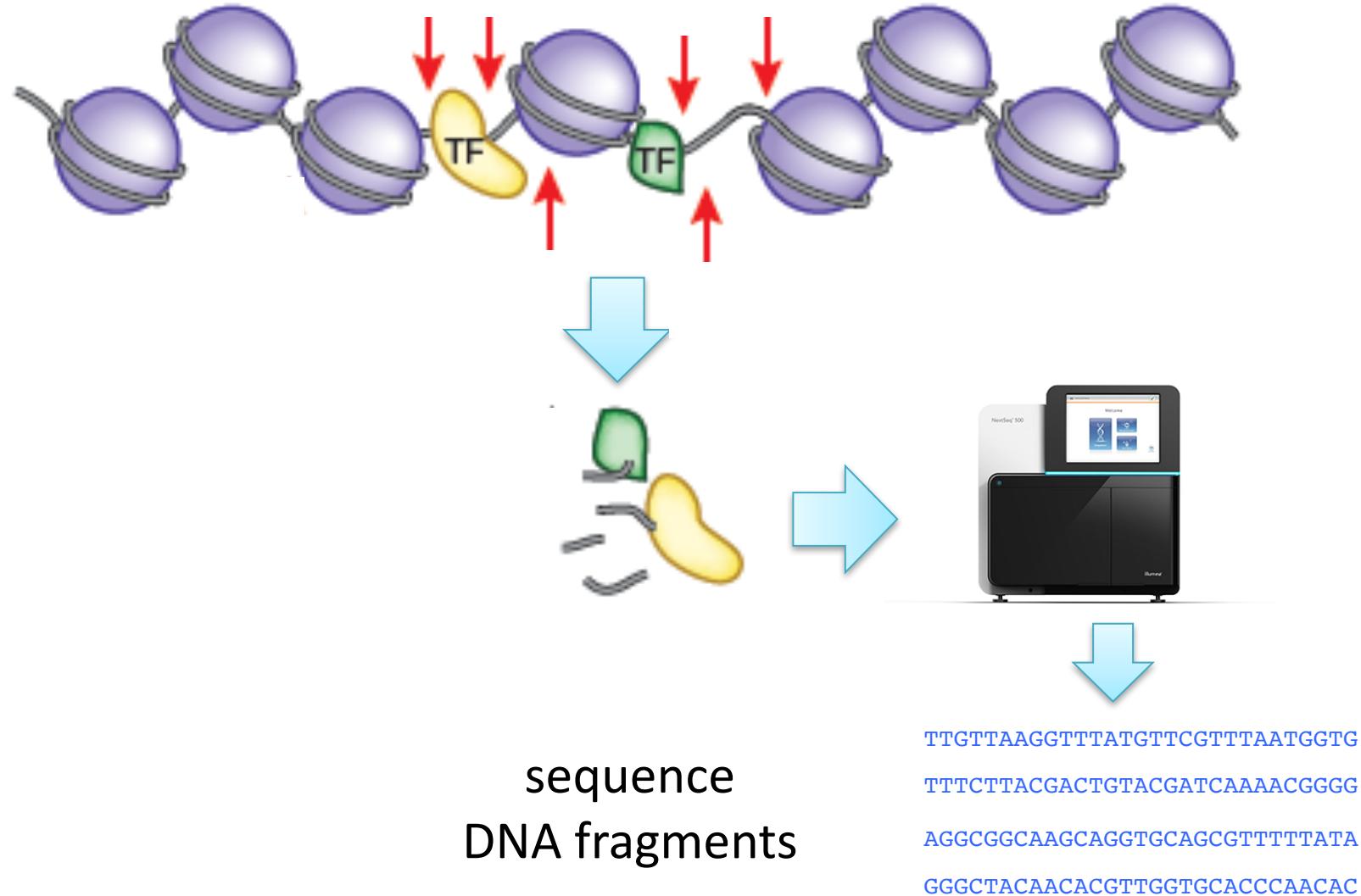


digest chromatin with Deoxyribonuclease 1 (DNase1)

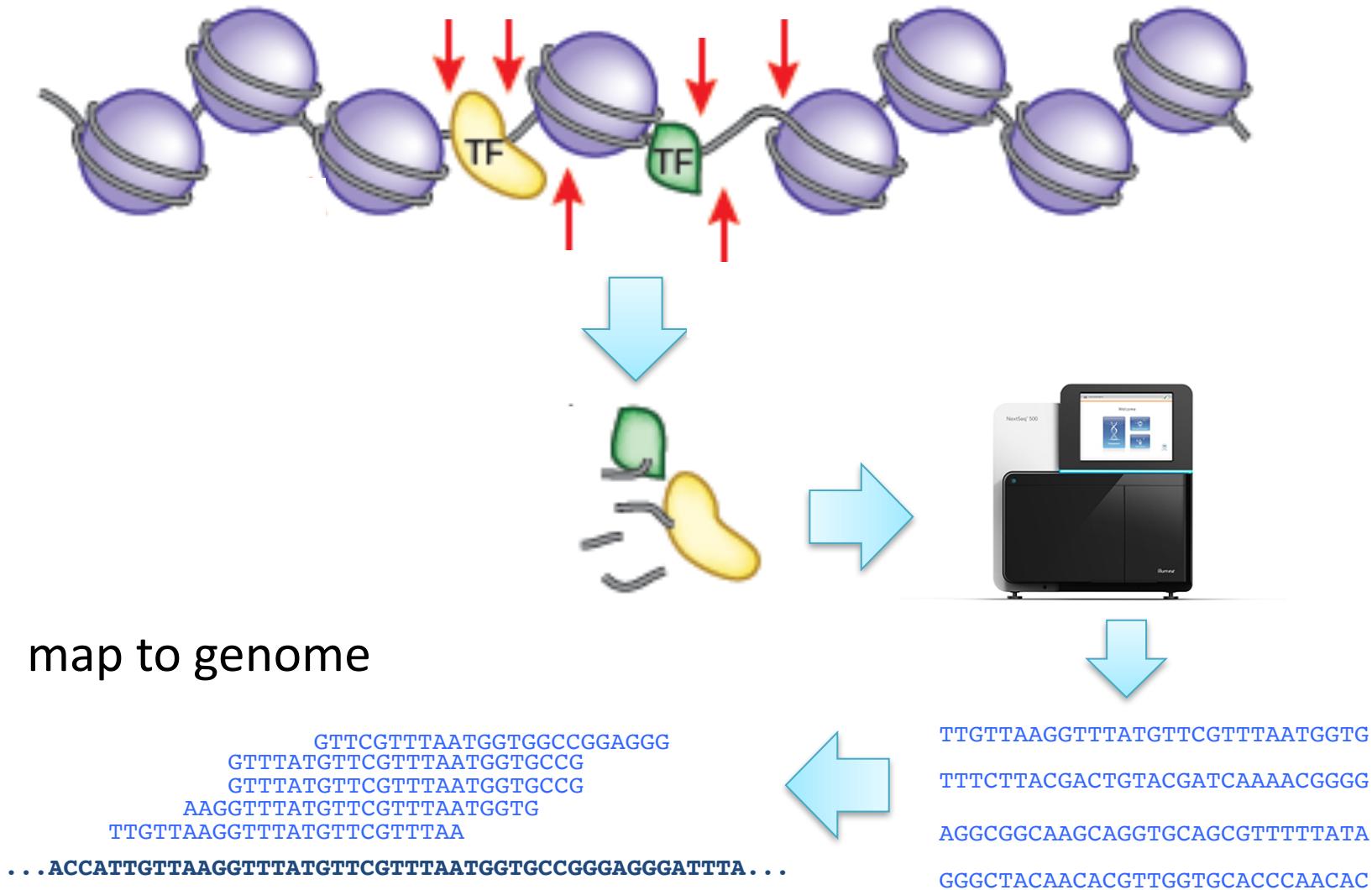
Measuring chromatin accessibility with DNase-seq



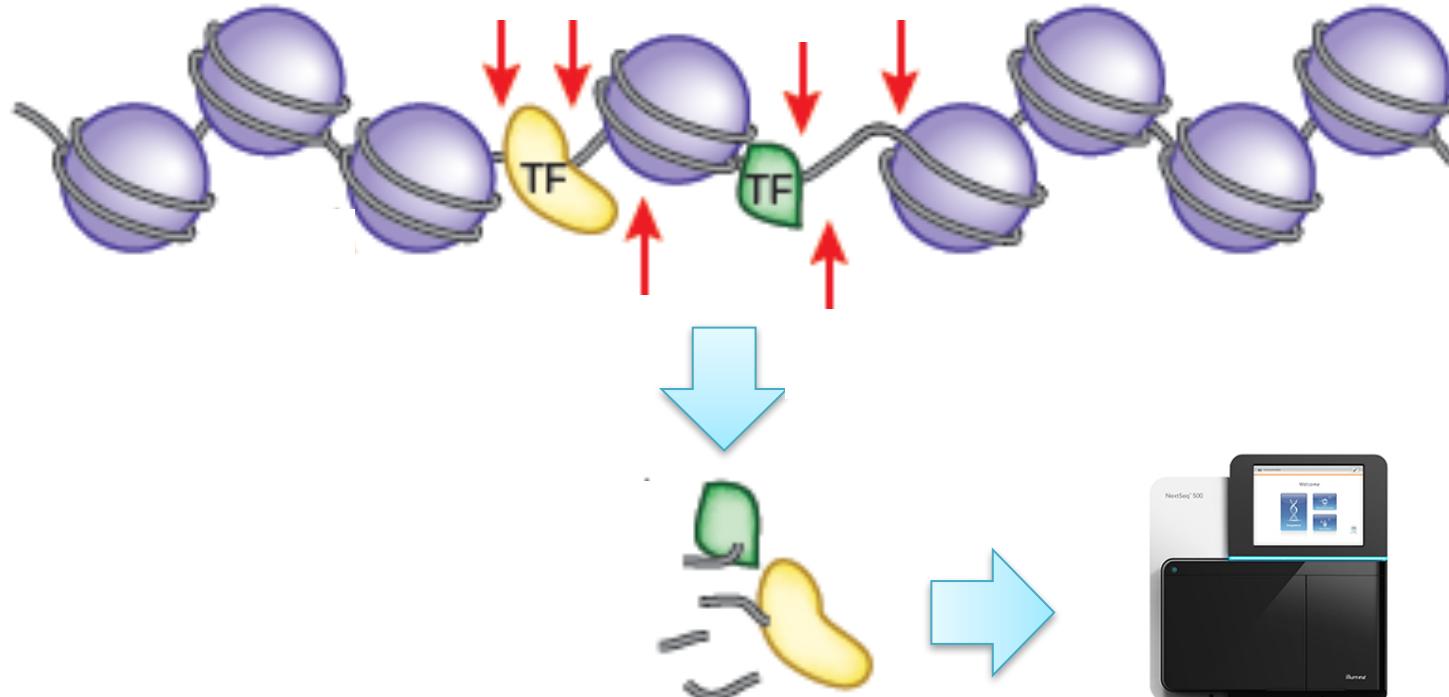
Measuring chromatin accessibility with DNase-seq



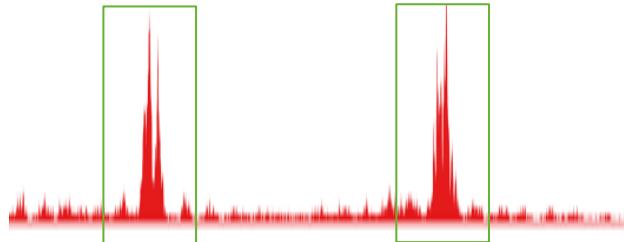
Measuring chromatin accessibility with DNase-seq



Measuring chromatin accessibility with DNase-seq

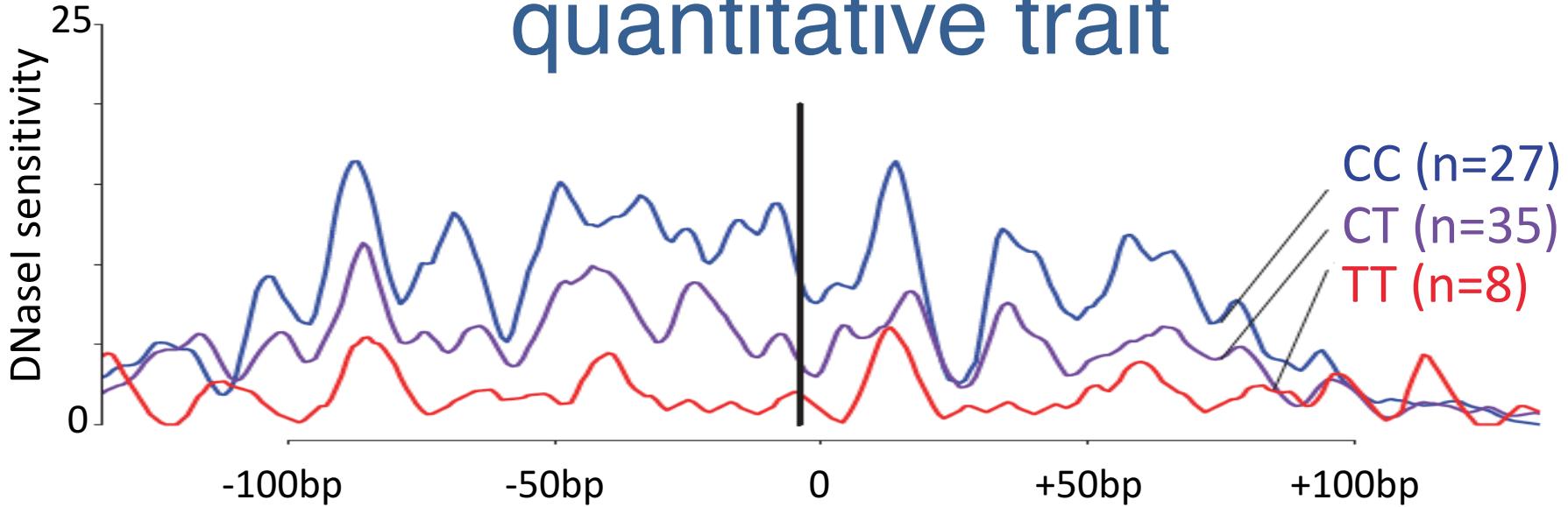


DNase hypersensitive sites



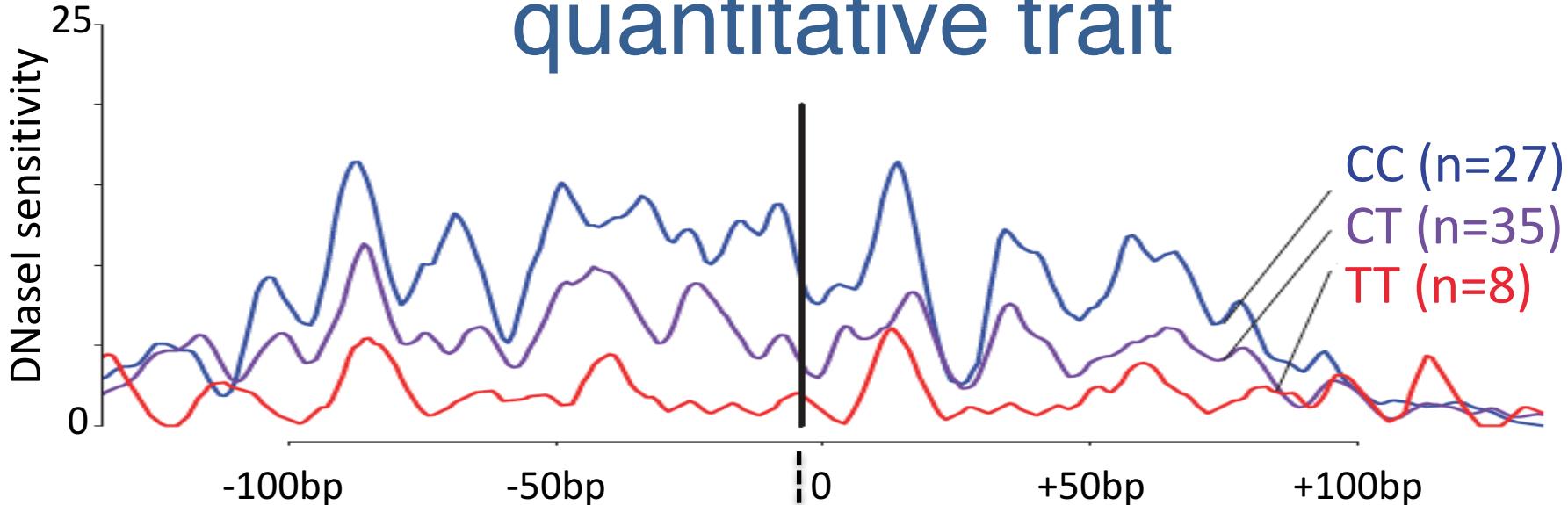
TTGTTAACGGTTATGTTCGTTAATGGTG
TTTCTTACGACTGTACGATCAAACGGGG
AGGCAGCAAGCAGGTGCAGCGTTTTATA
GGGCTACAAACACGTTGGTGCACCCAACAC

DNase sensitivity as a quantitative trait

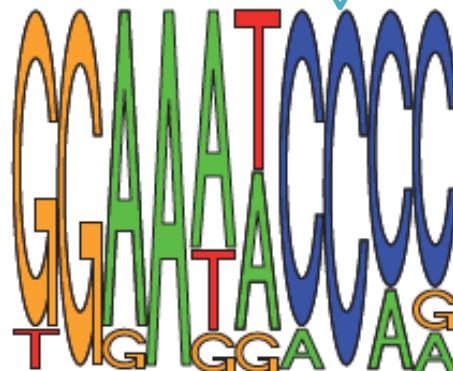


distance from center of DNase peak

DNase sensitivity as a quantitative trait

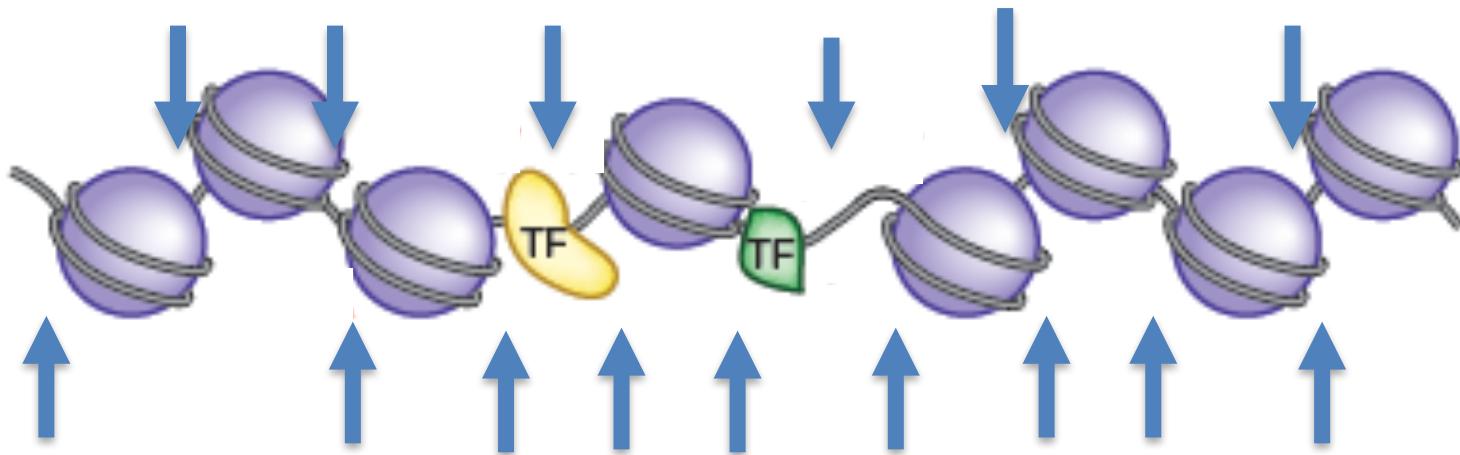


T/C



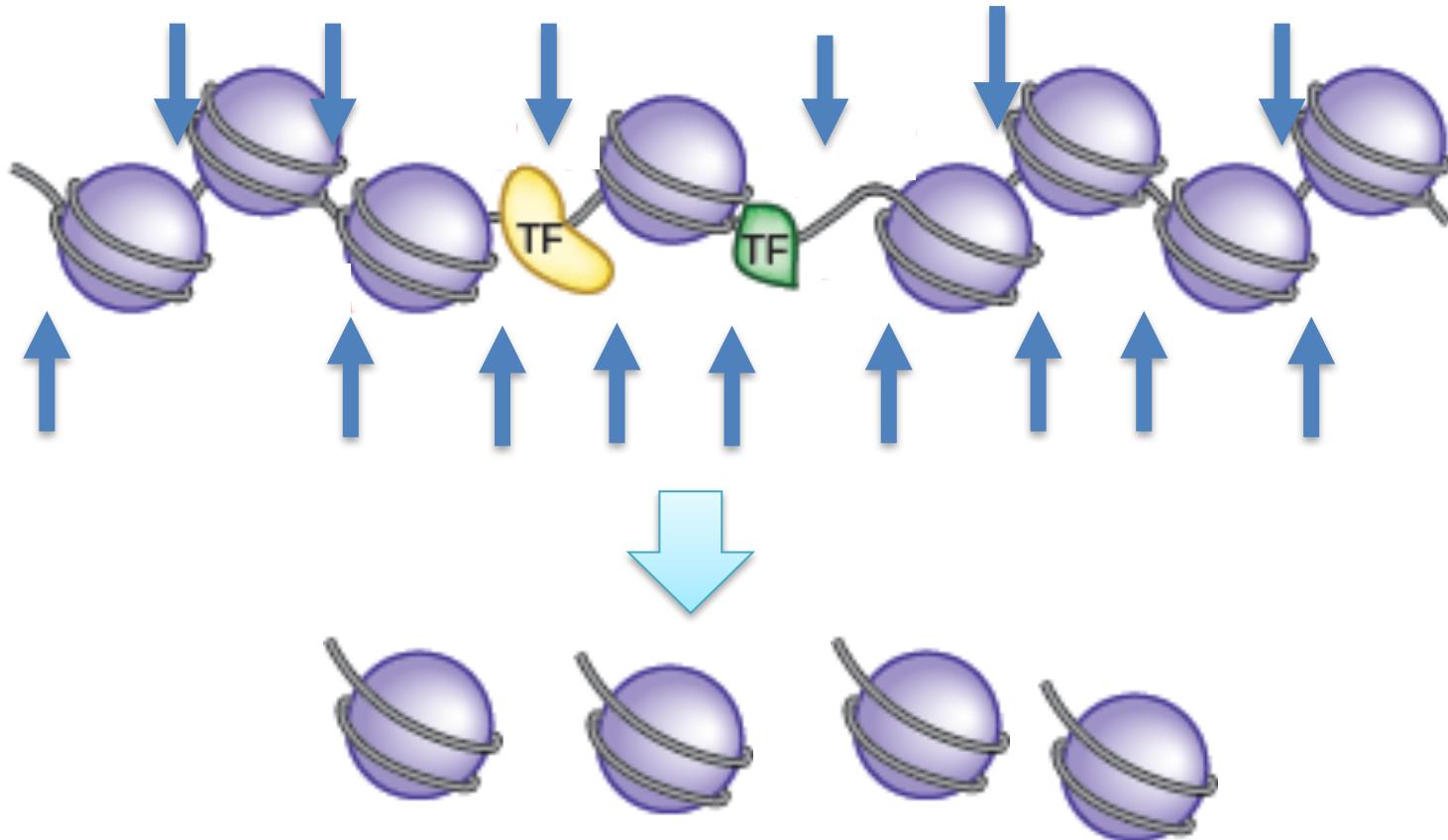
NF- κ B motif

Determining nucleosome positions with MNase-seq



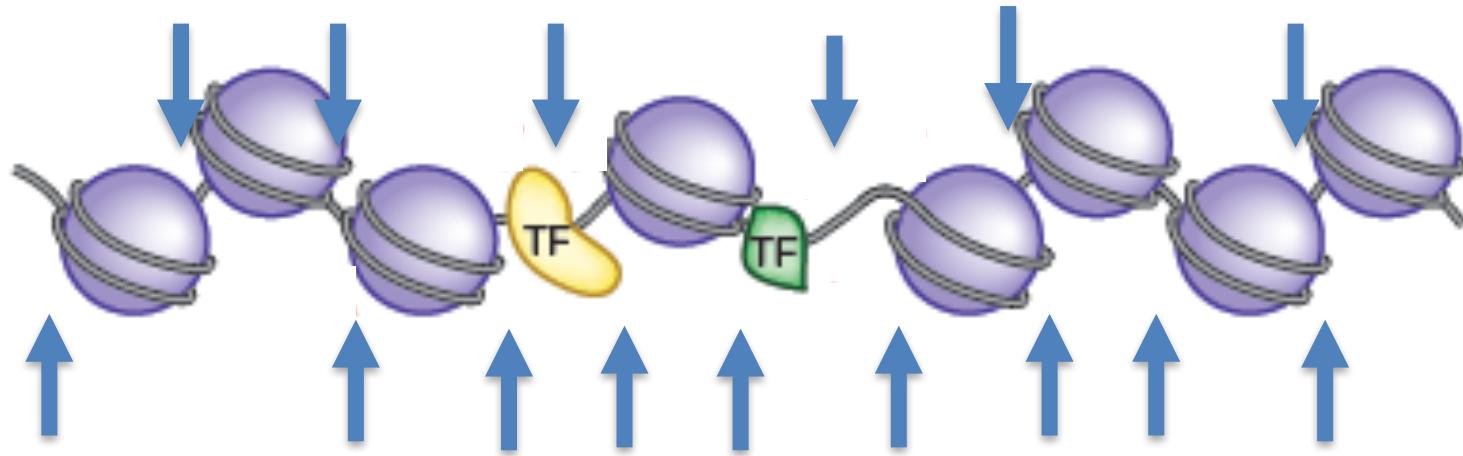
digest chromatin with micrococcal nuclease (MNase)

Determining nucleosome positions with MNase-seq



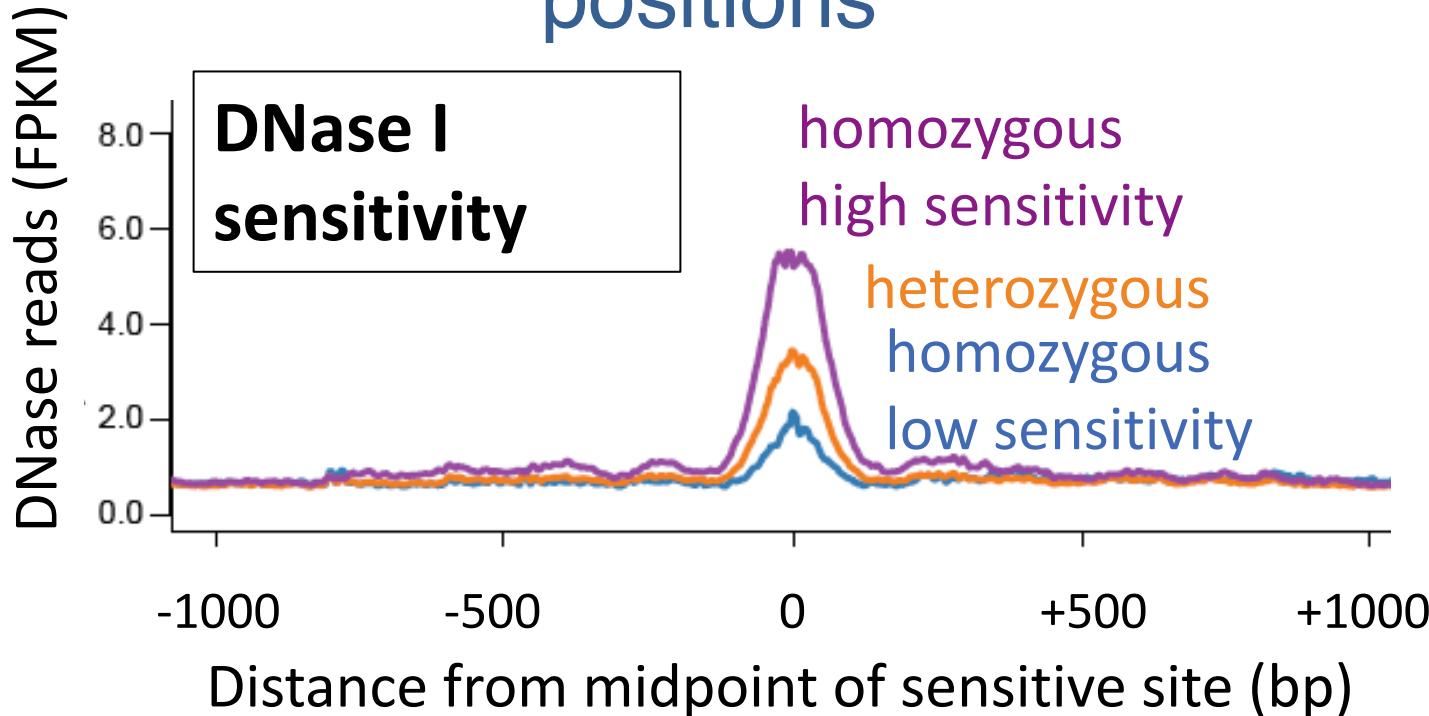
isolate nucleosome-sized fragments and sequence ends

Determining nucleosome positions with MNase-seq

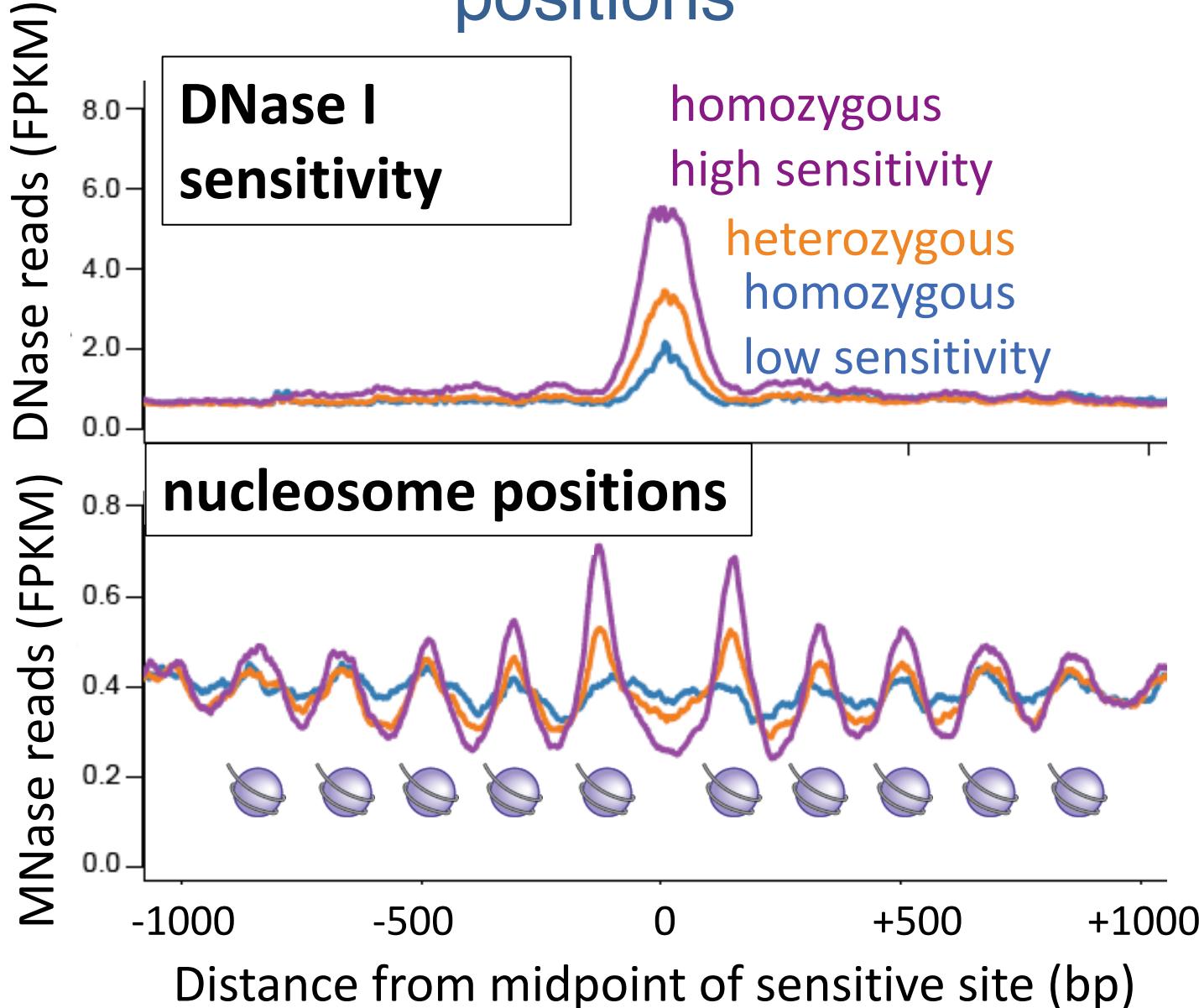


peaks are
consistently-
positioned
nucleosomes

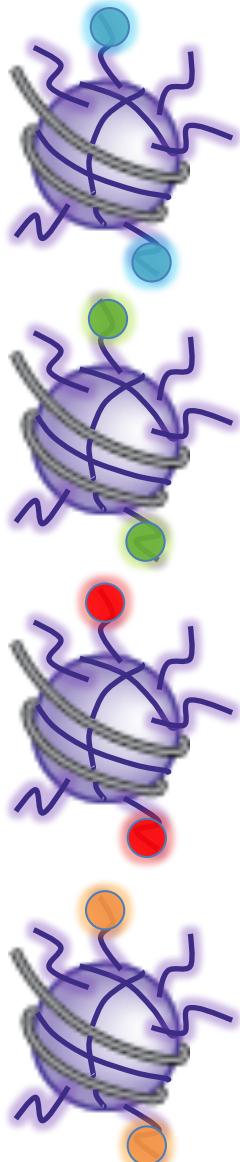
dsQTLs are associated with nucleosome positions



dsQTLs are associated with nucleosome positions



Histone modifications



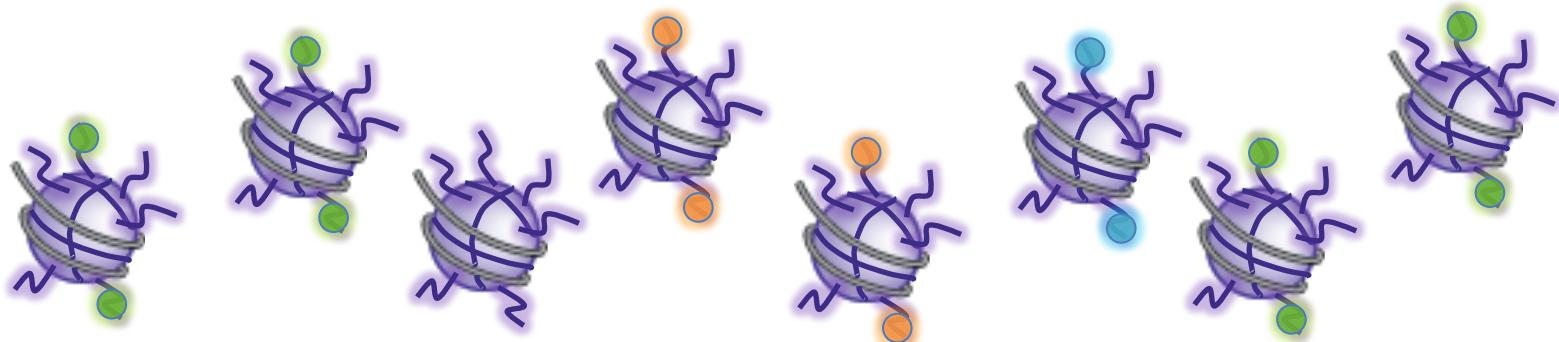
H3K4me1: active/open chromatin outside of promoters

H3K4me3: active promoters

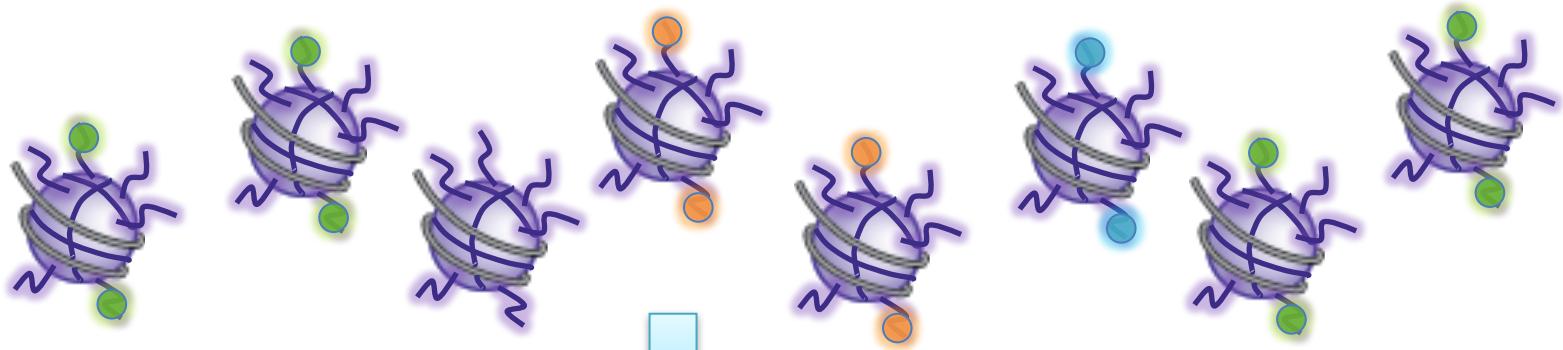
H3K27ac: active promoters & enhancers

H3K27me3: silenced genes

Measuring histone marks with ChIP-seq



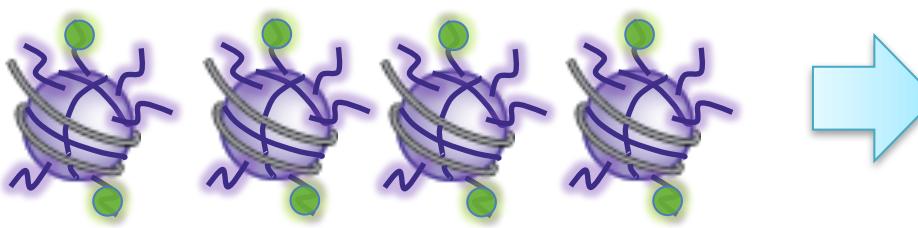
Measuring histone marks with ChIP-seq



Chromatin
Immunoprecipitation
(ChIP)

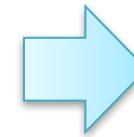
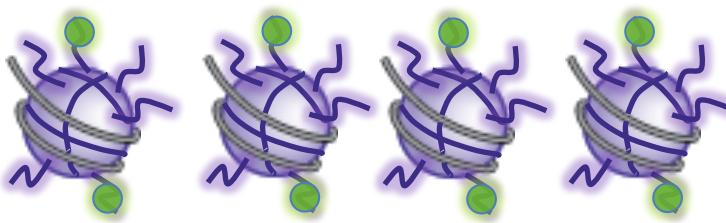
Measuring histone marks with ChIP-seq

High-throughput
DNA sequencing



TTGTTAACGGTTATGTCGTTAACGGTG
TTTCTTACGACTGTACGATCAAAACGGGG
AGGCAGCAAGCAGGTGCAGCGTTTTATA
GGGCTACAACACGTTGGTCACCCAACAC

Measuring histone marks with ChIP-seq



Map reads to
genome

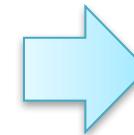
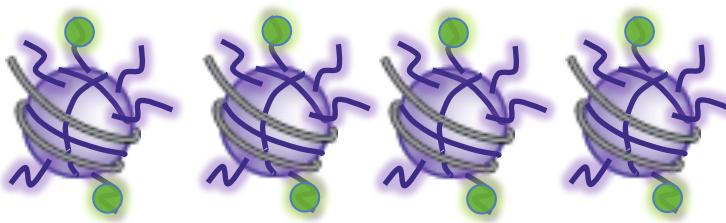
GTTCGTTAACGGTGGCCGGAGGG
GTTTATGTTCGTTAACGGTGCGC
GTTTATGTTCGTTAACGGTGCGC
AAGGTTATGTTCGTTAACGGTG
TTGTTAACGGTTATGTTCGTTAA

...ACCATTGTTAACGGTTATGTTCGTTAACGGTGCCCCGAGGGATTAA...



TTGTTAACGGTTATGTTCGTTAACGGTG
TTTCTTACGACTGTACGATCAAAACGGGG
AGGCAGGCAAGCAGGTGCAGCGTTTTATA
GGGCTACAAACACGTTGGTGCACCCAACAC

Measuring histone marks with ChIP-seq

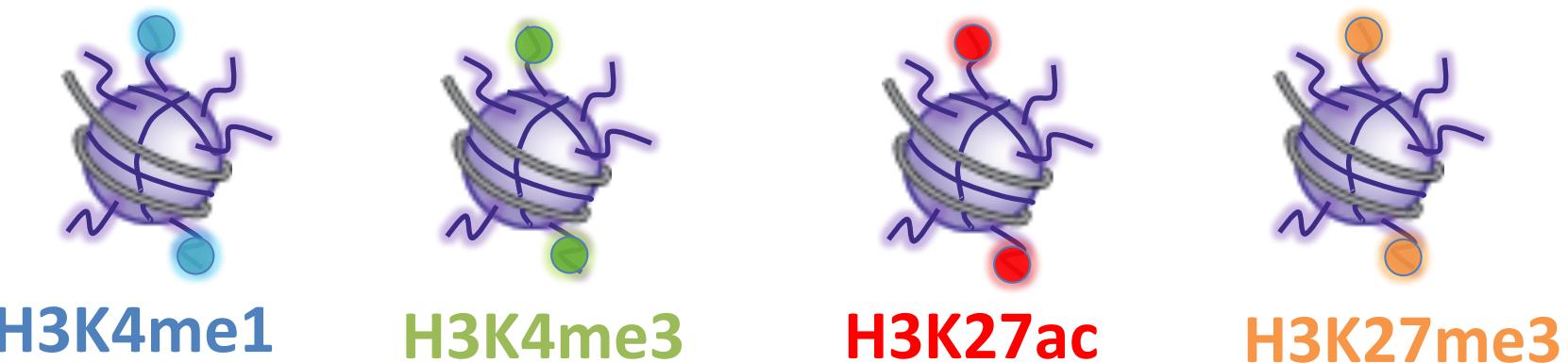


Identify ChIP-seq peaks

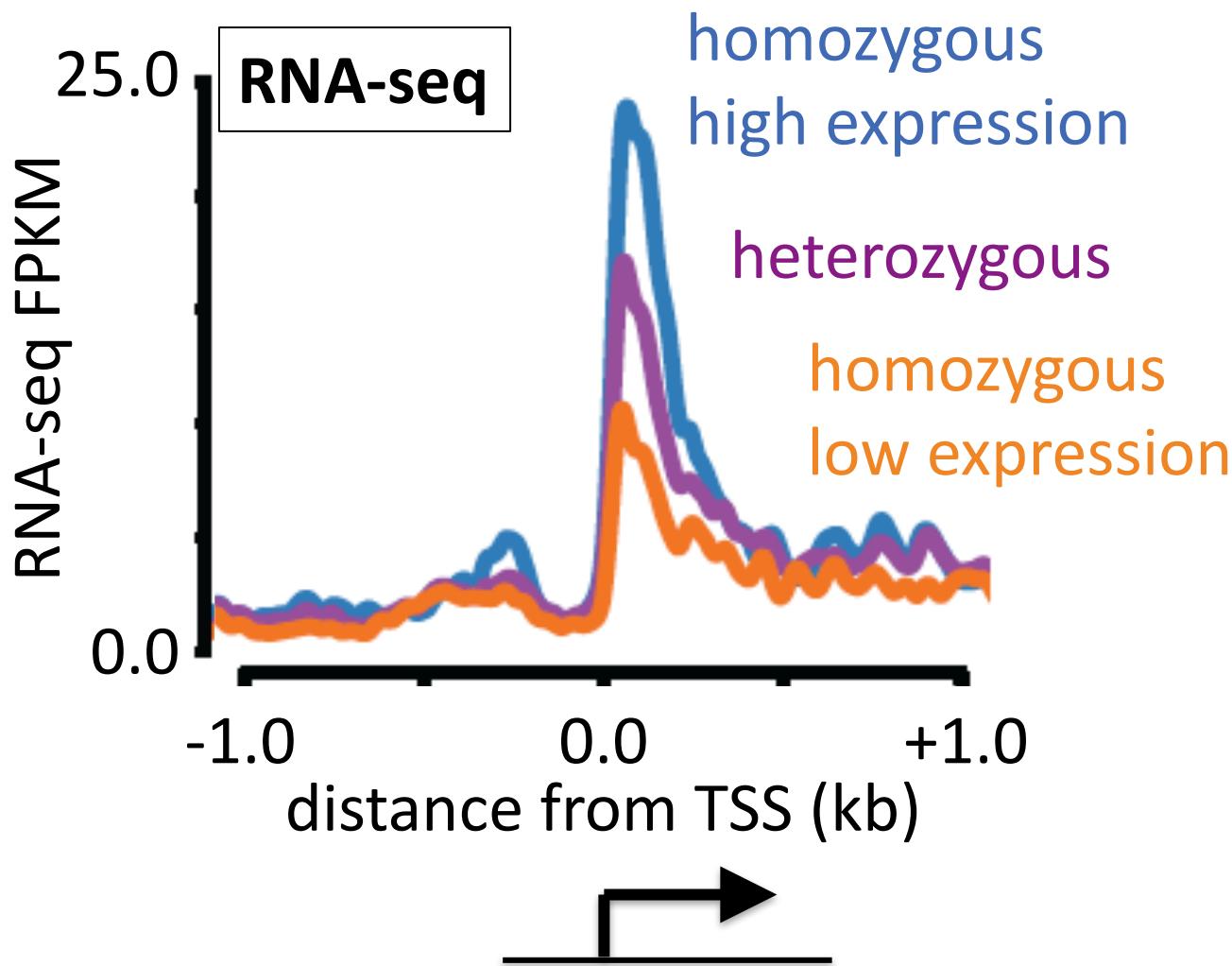


TTGTTAACGGTTATGTCGTTAATGGTG
TTTCTTACGACTGTACGATCAAAACGGGG
AGGCAGGCAAGCAGGTGCAGCGTTTTATA
GGGCTACAAACACGTTGGTGCACCCAACAC

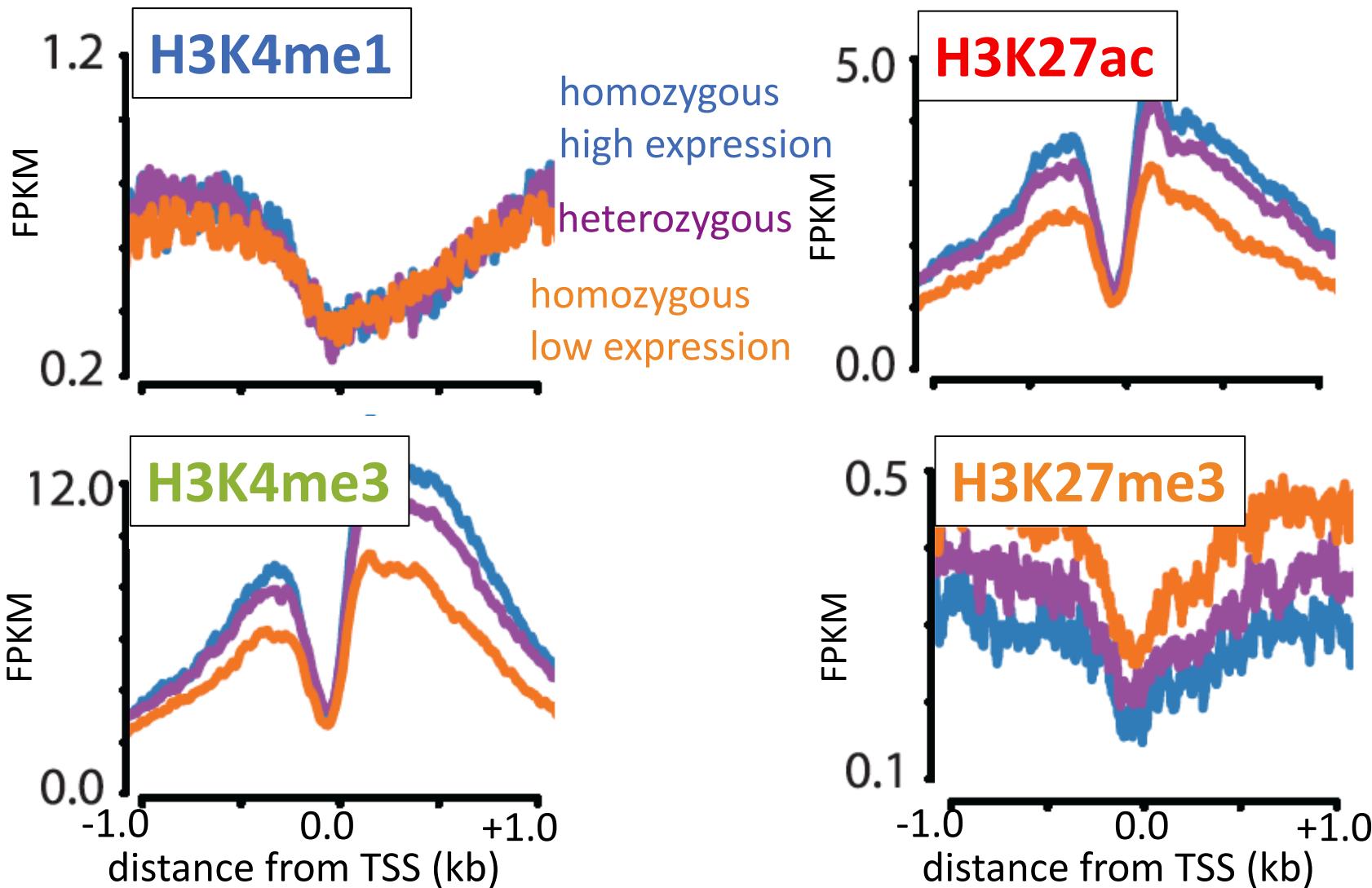
ChIP-seq data from 10 individuals



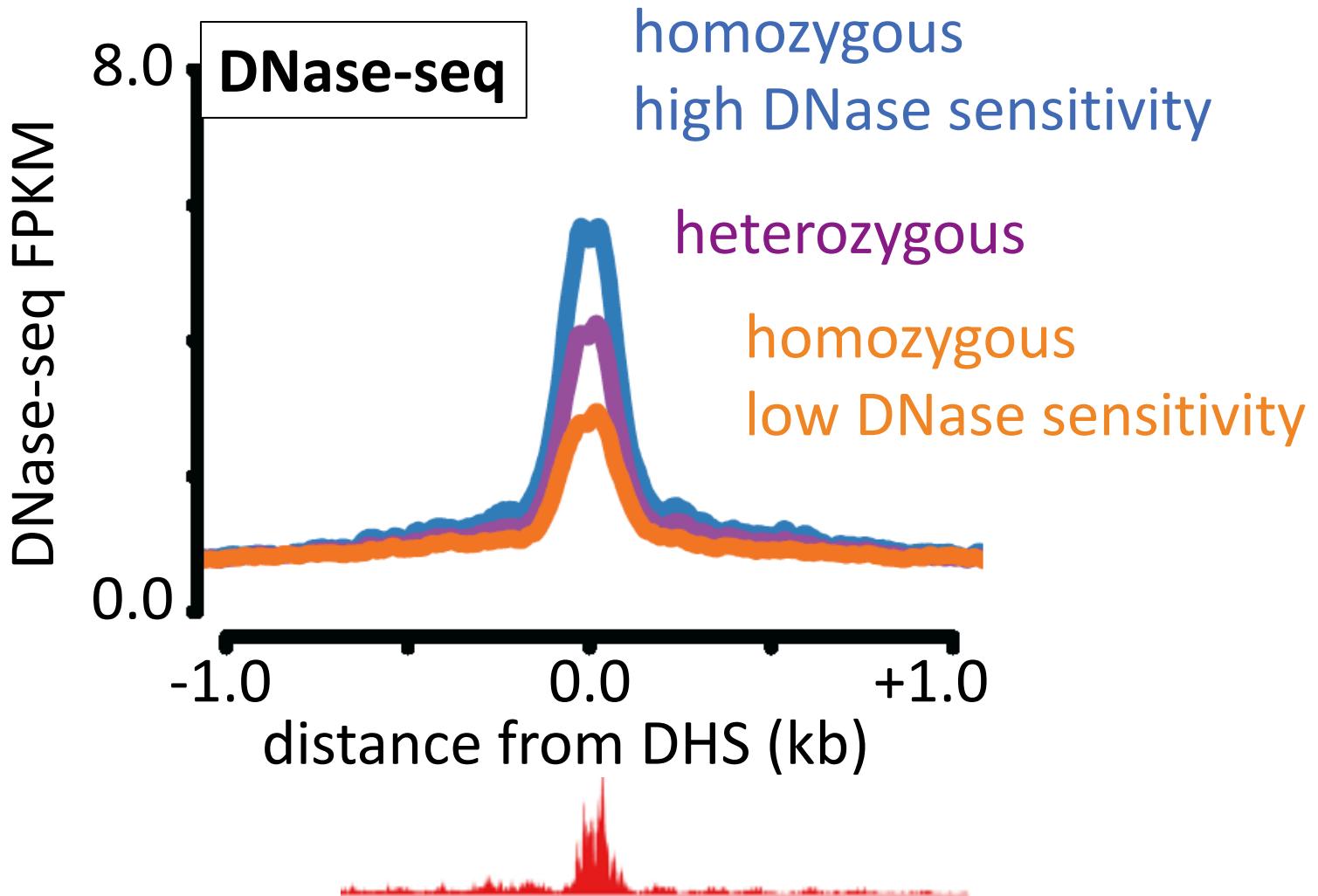
Are eQTLs also associated with histone modifications?



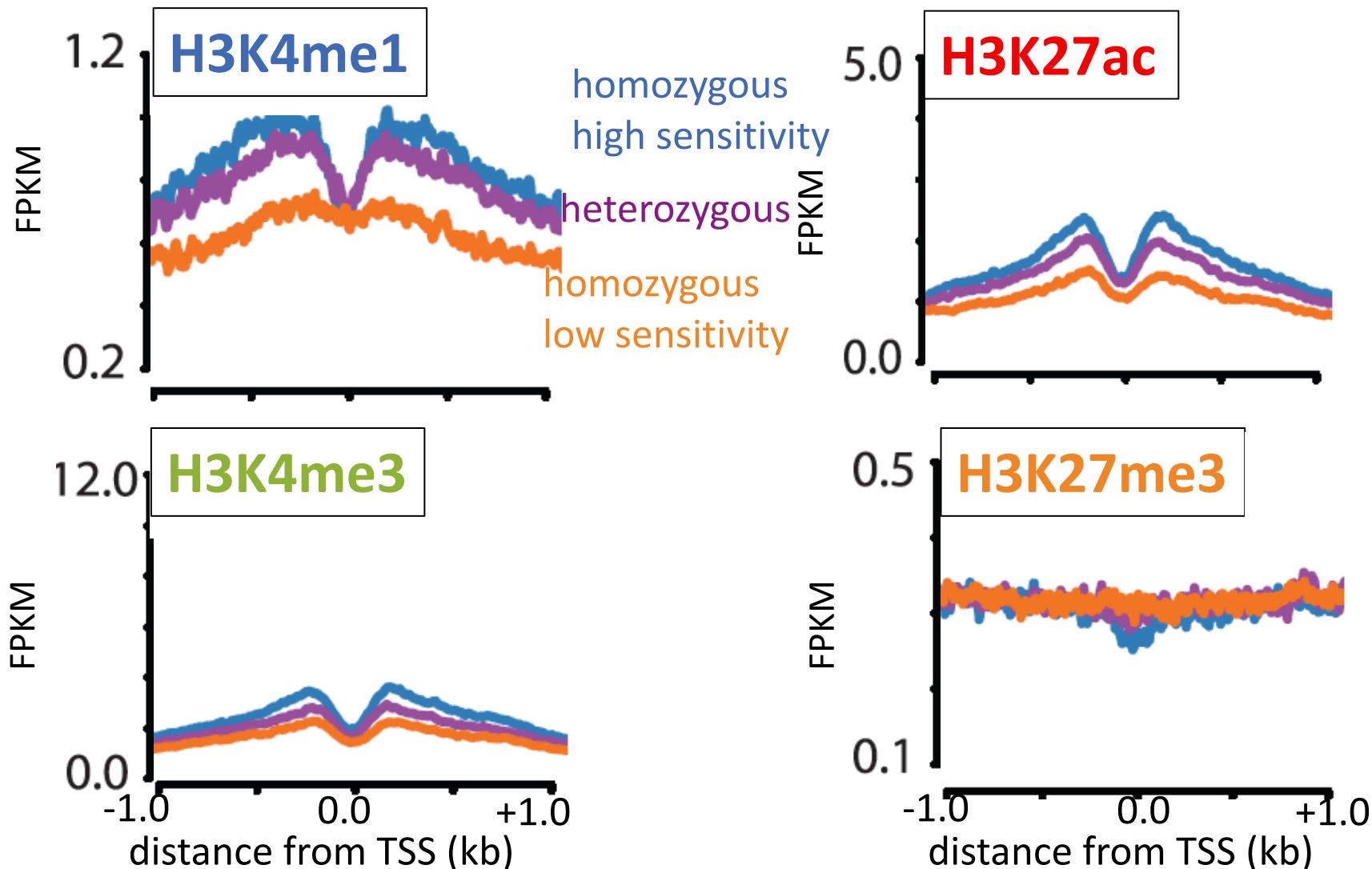
Are eQTLs also associated with histone modifications?



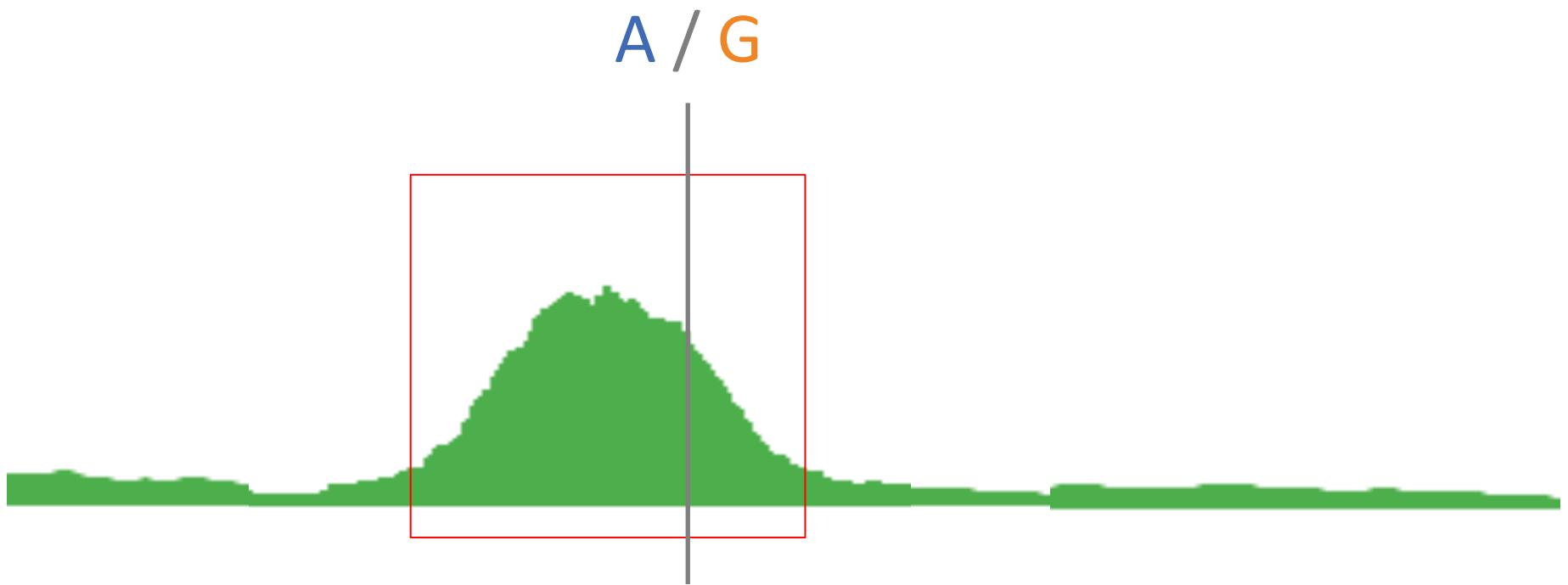
Are dsQTLs also associated with histone modifications?



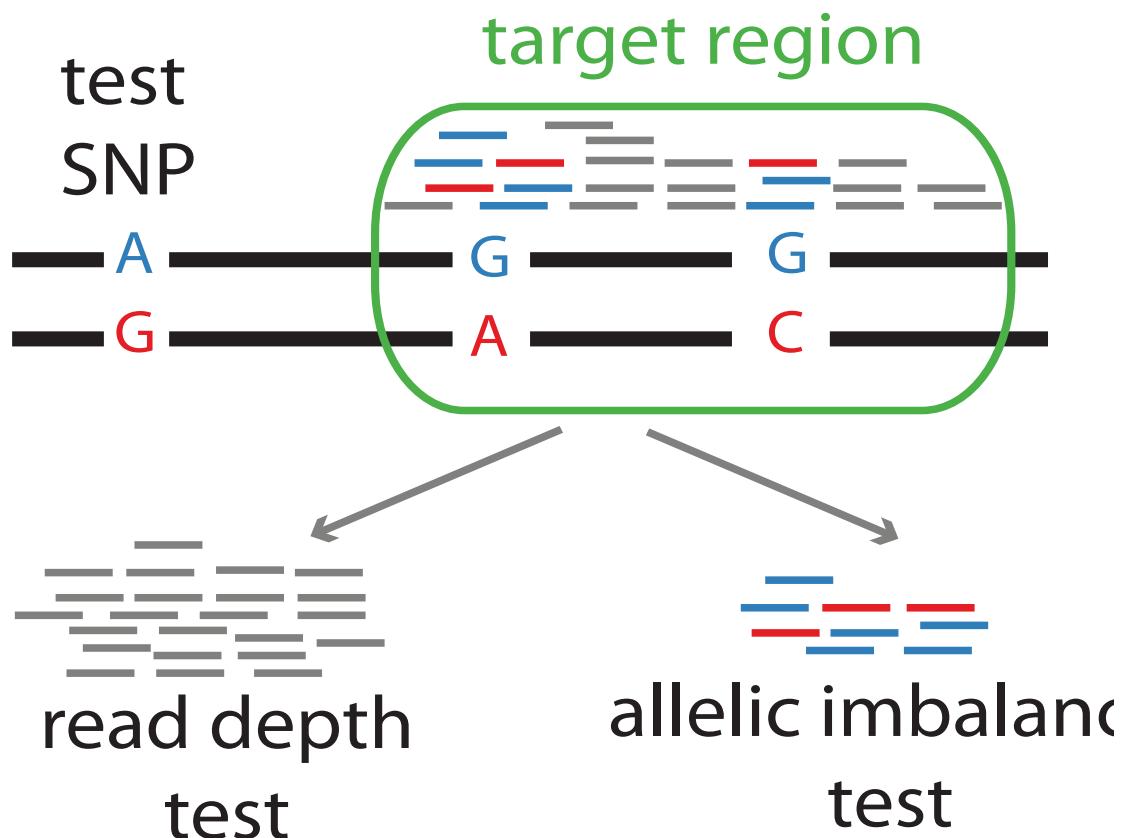
Are dsQTLs also associated with histone modifications?



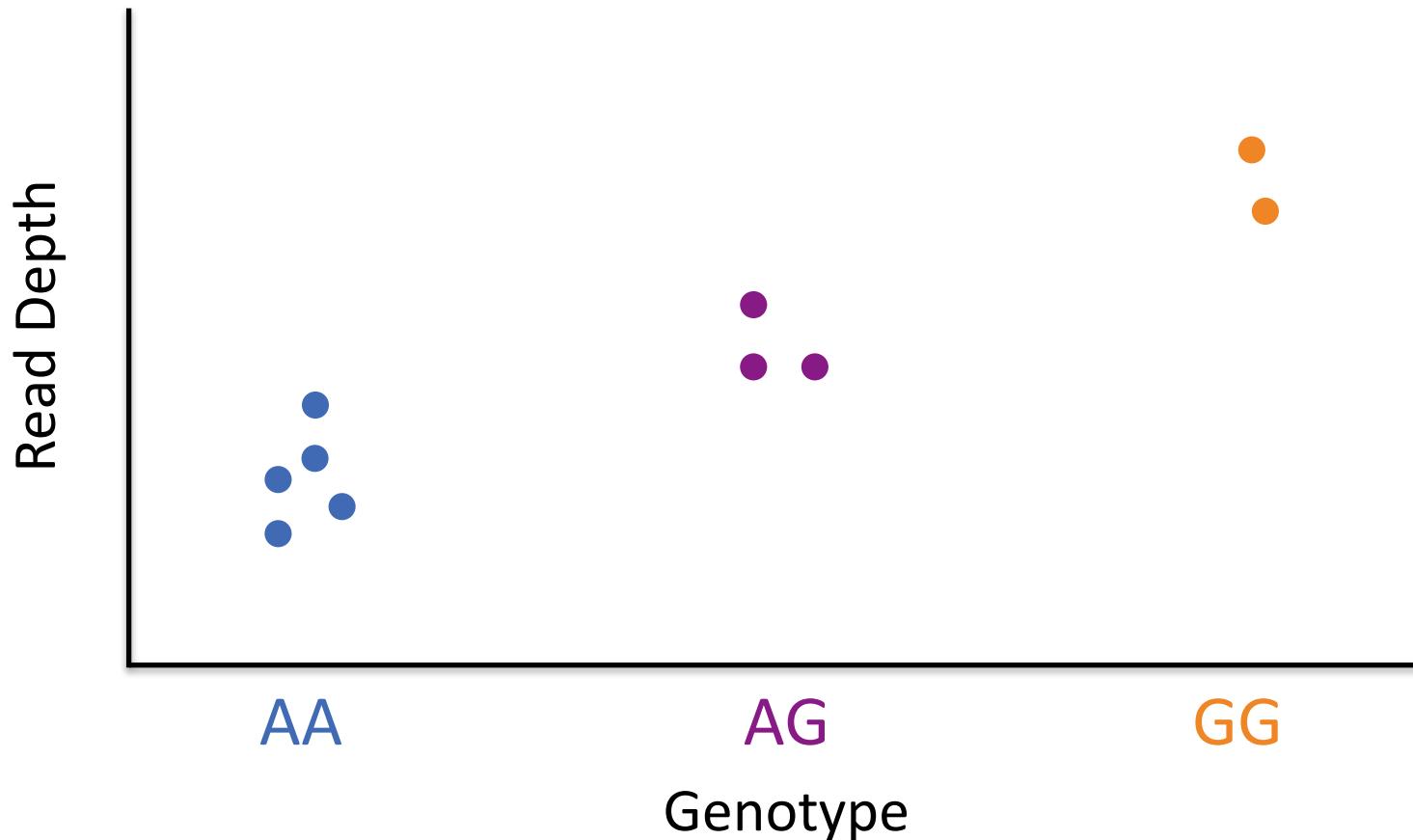
Can we identify Histone Mark QTLs?



Combined Haplotype Test



Read Depth Association Test



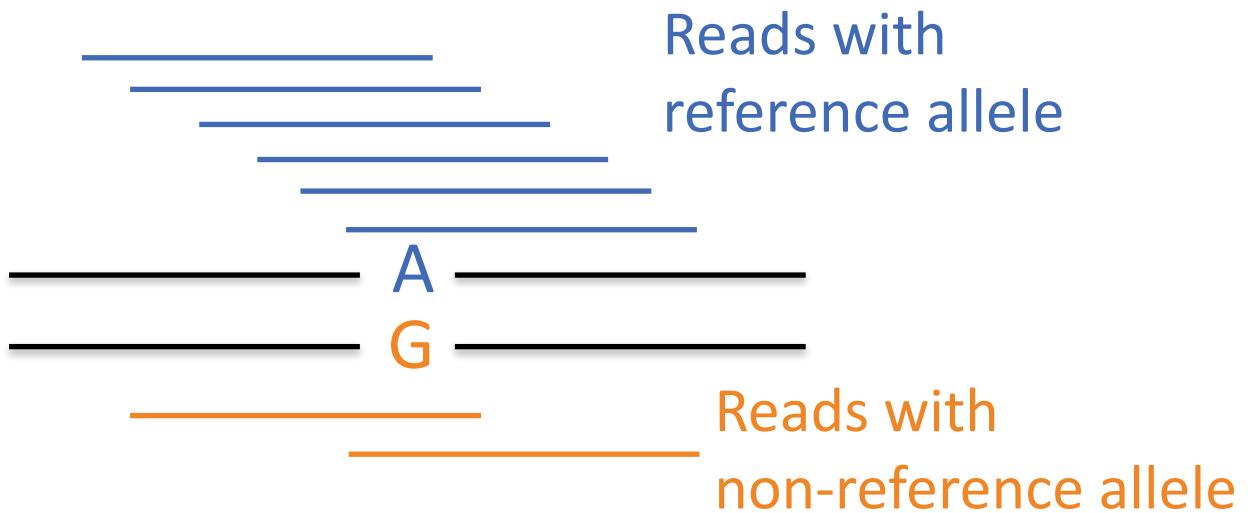
Model read counts with Poisson distribution:

$$\lambda = 2\alpha$$

$$\alpha + \beta$$

$$2\beta$$

Allelic Imbalance Test



Model reference proportion with binomial distribution:

$$p = \frac{\alpha}{\alpha + \beta}$$

Haplotype Imbalance Test



- Phase SNPs
- Test allelic imbalance across entire haplotype

Combined Haplotype Test

Read Depth Association Test

$$\lambda = \begin{cases} 2\alpha & \text{if homozygous AA} \\ \alpha+\beta & \text{if heterozygous AB} \\ 2\beta & \text{if homozygous BB} \end{cases}$$

+

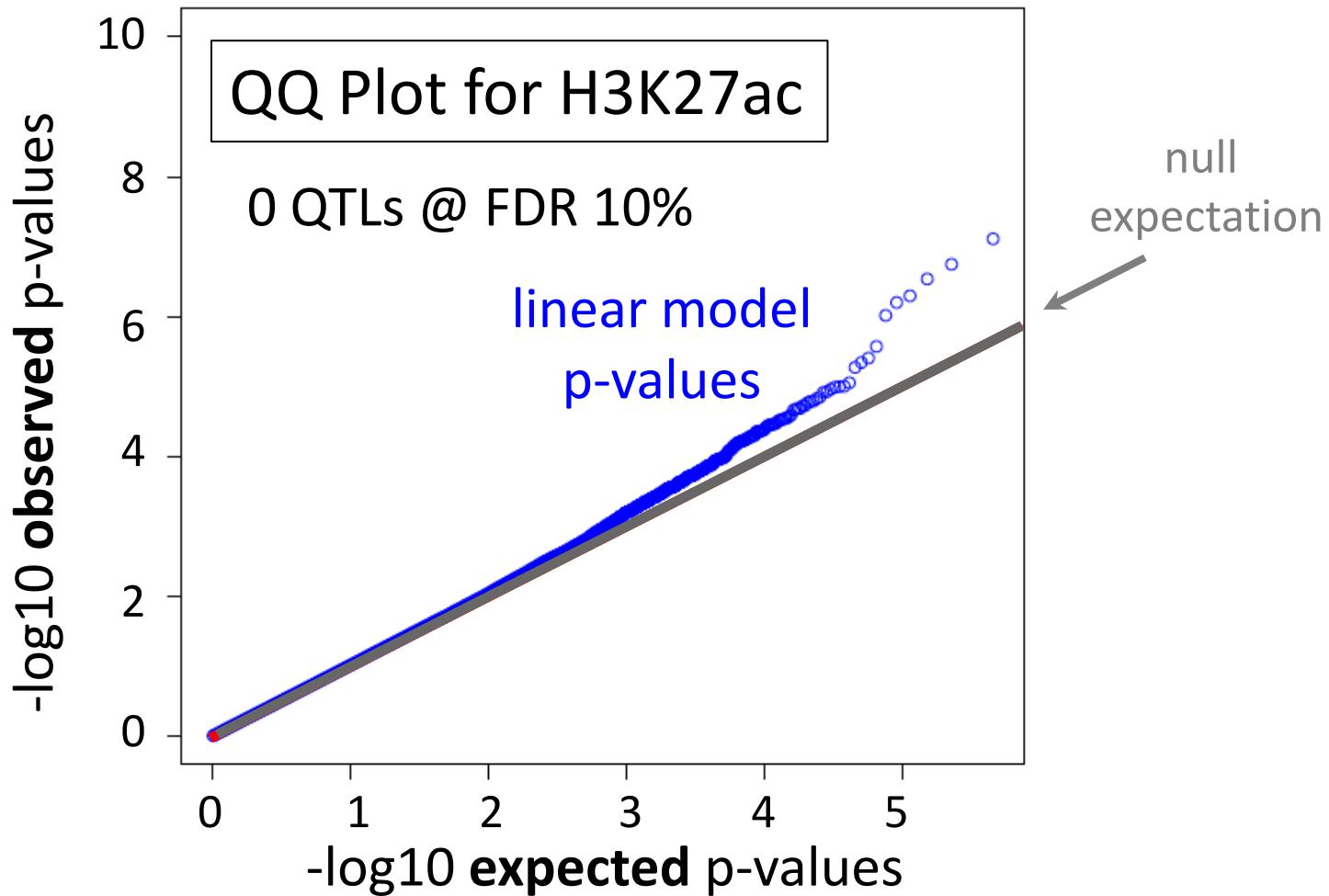
Haplotype
Imbalance Test

$$p = \frac{\alpha}{\alpha+\beta}$$

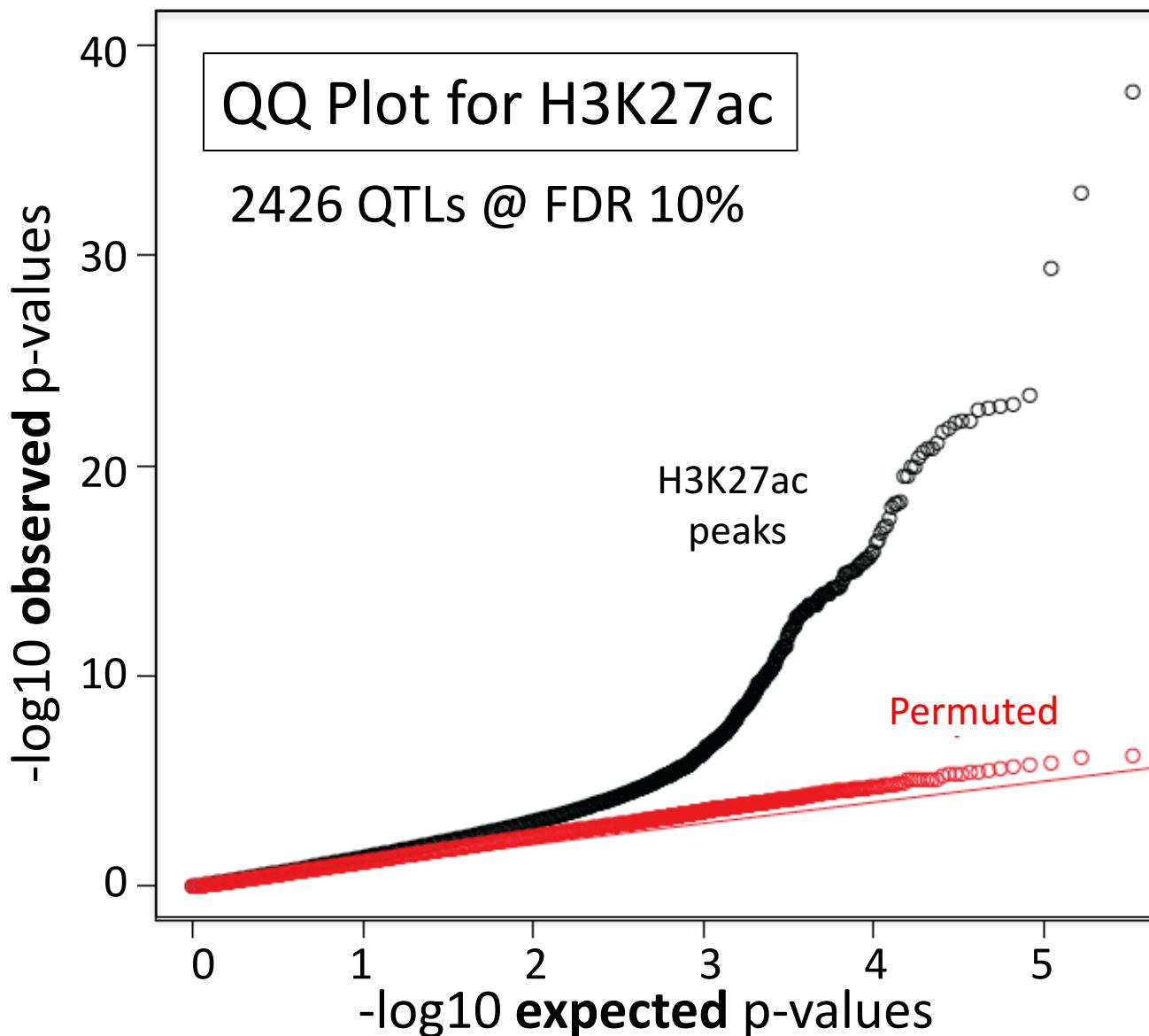
null hypothesis, $H_0: \alpha = \beta$

alternative hypothesis, $H_1: \alpha \neq \beta$

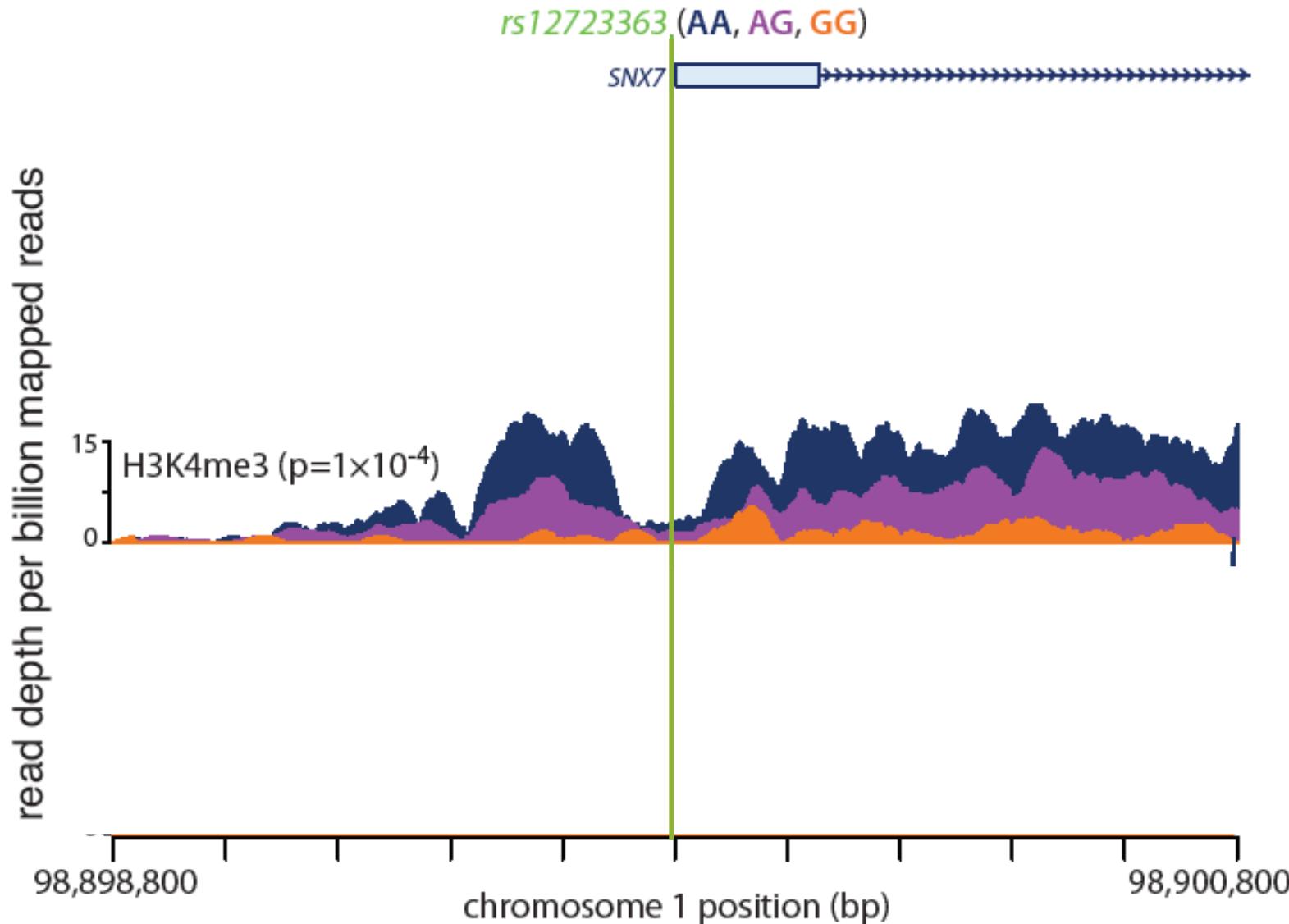
Standard mapping of H3K27ac QTLs with 10 Individuals



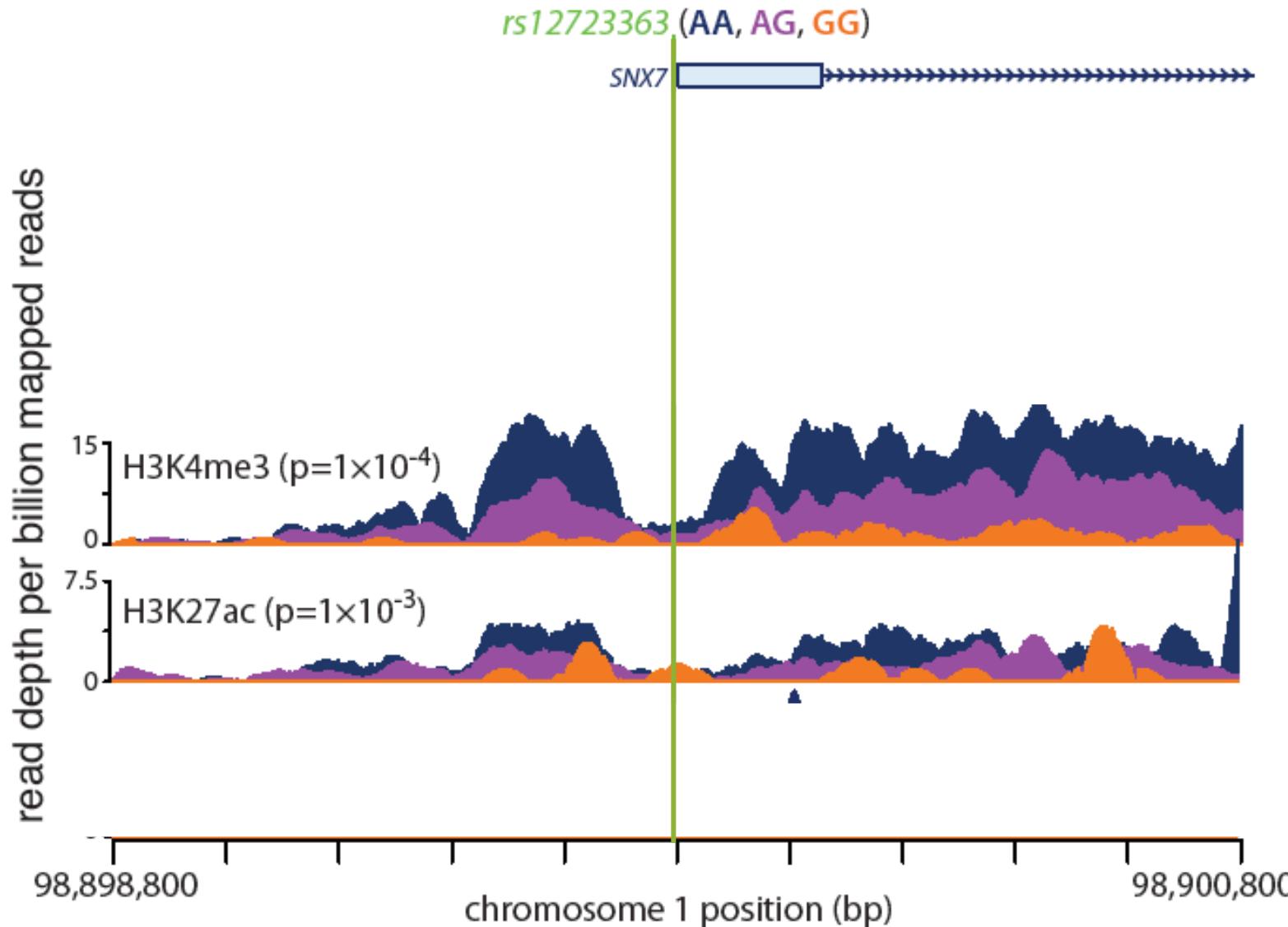
Combined Haplotype Test Mapping of H3K27ac QTLs



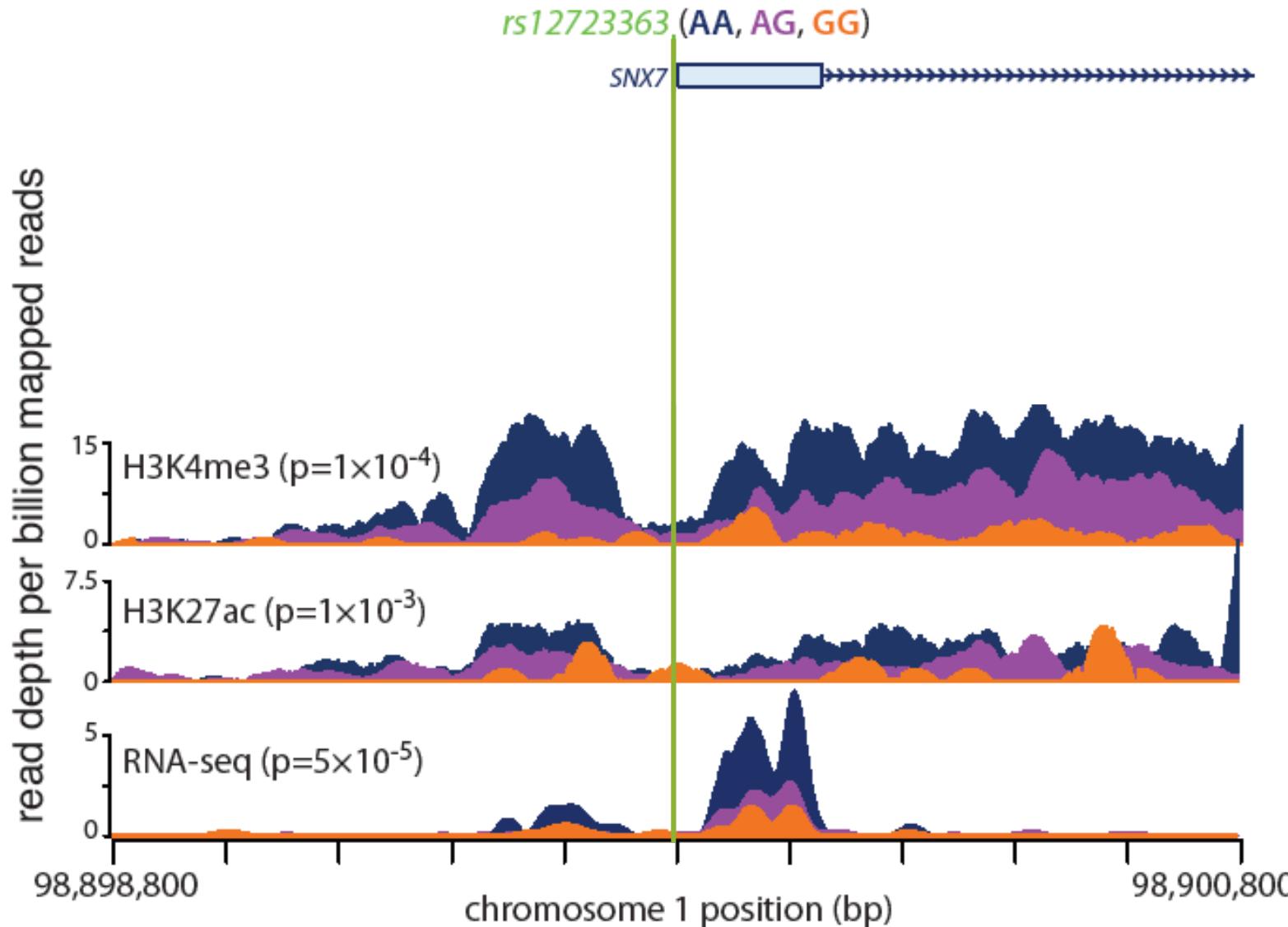
Example of a histone mark QTL



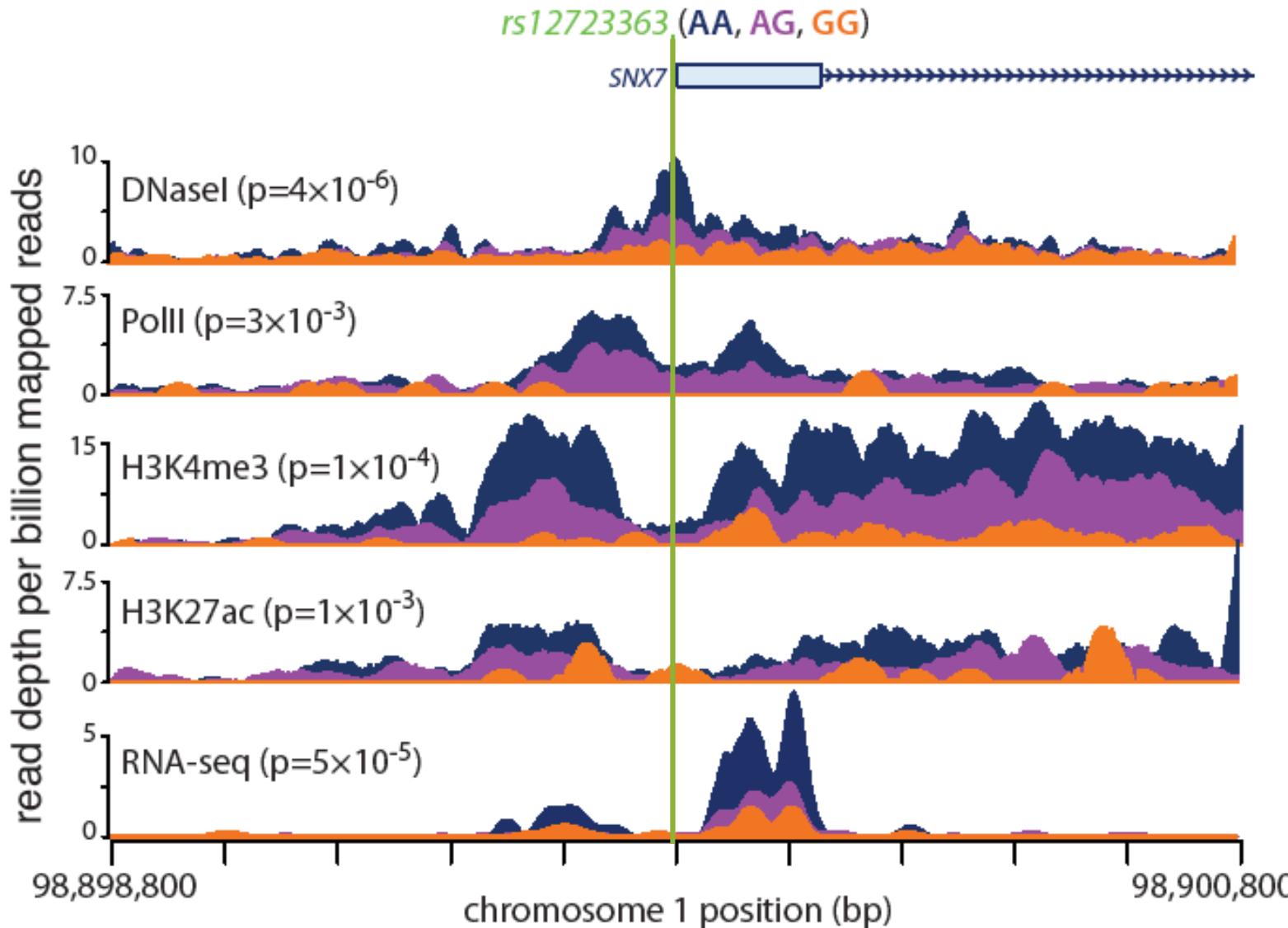
Example of a histone mark QTL



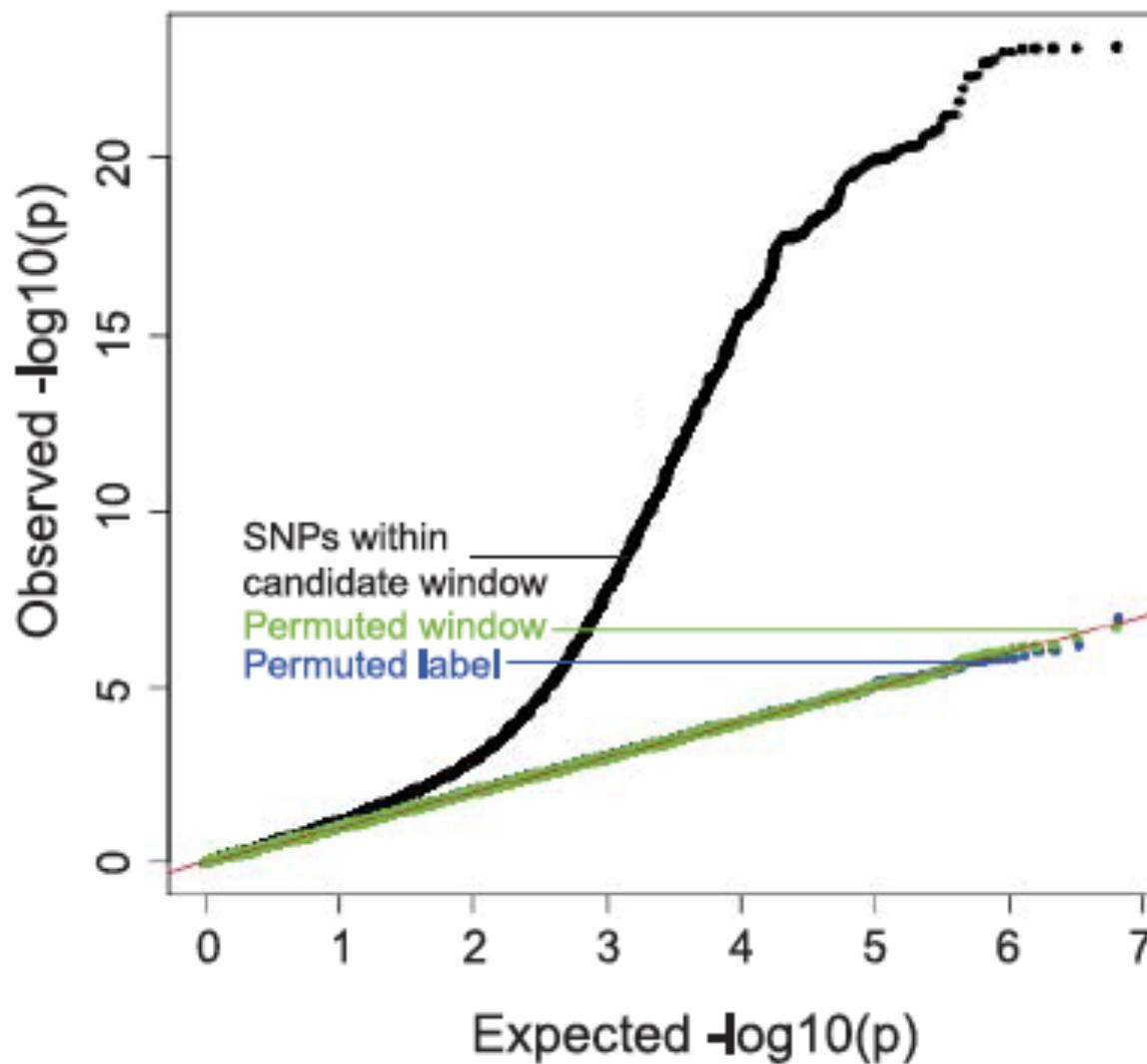
Example of a histone mark QTL



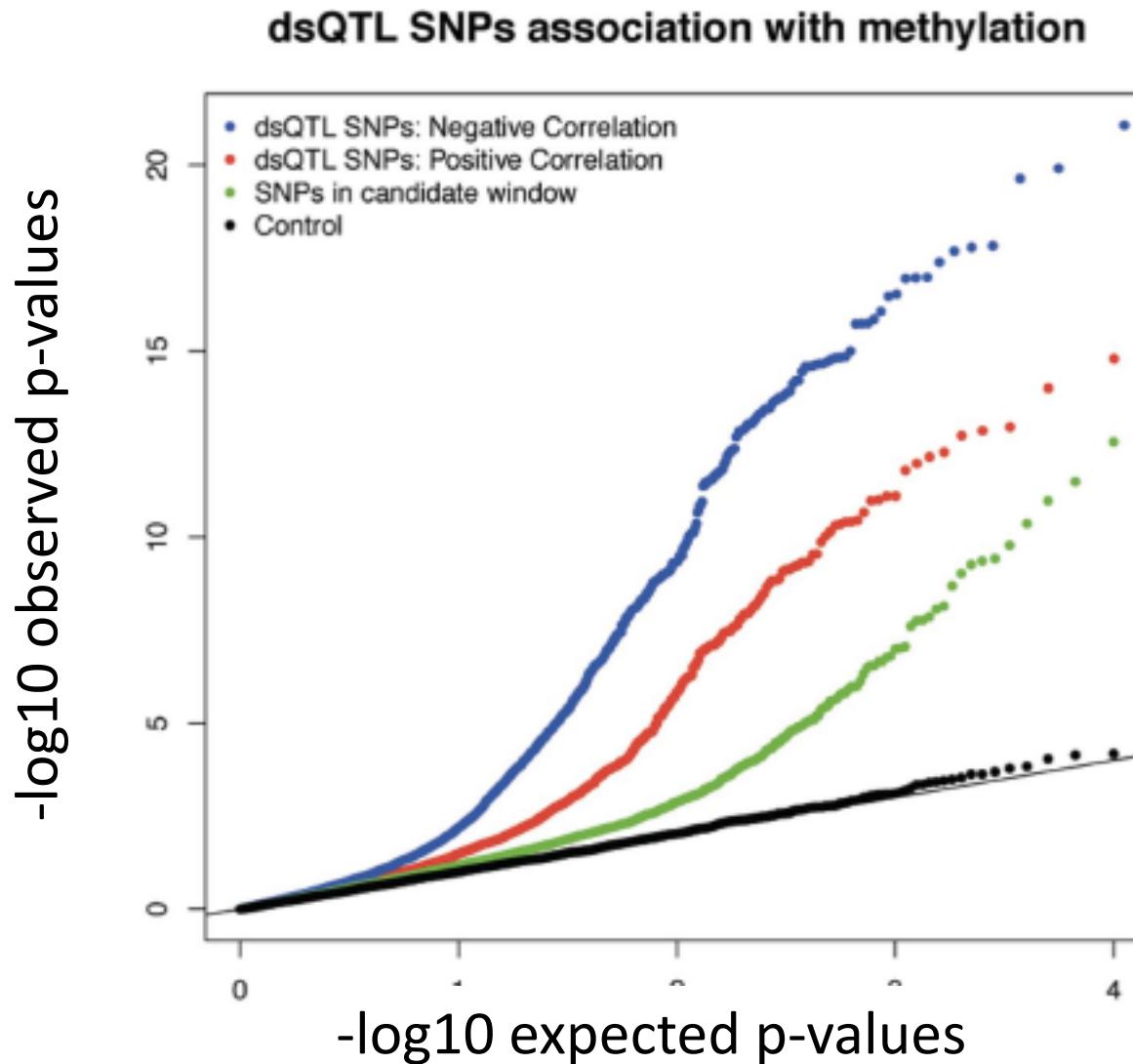
Example of a histone mark QTL



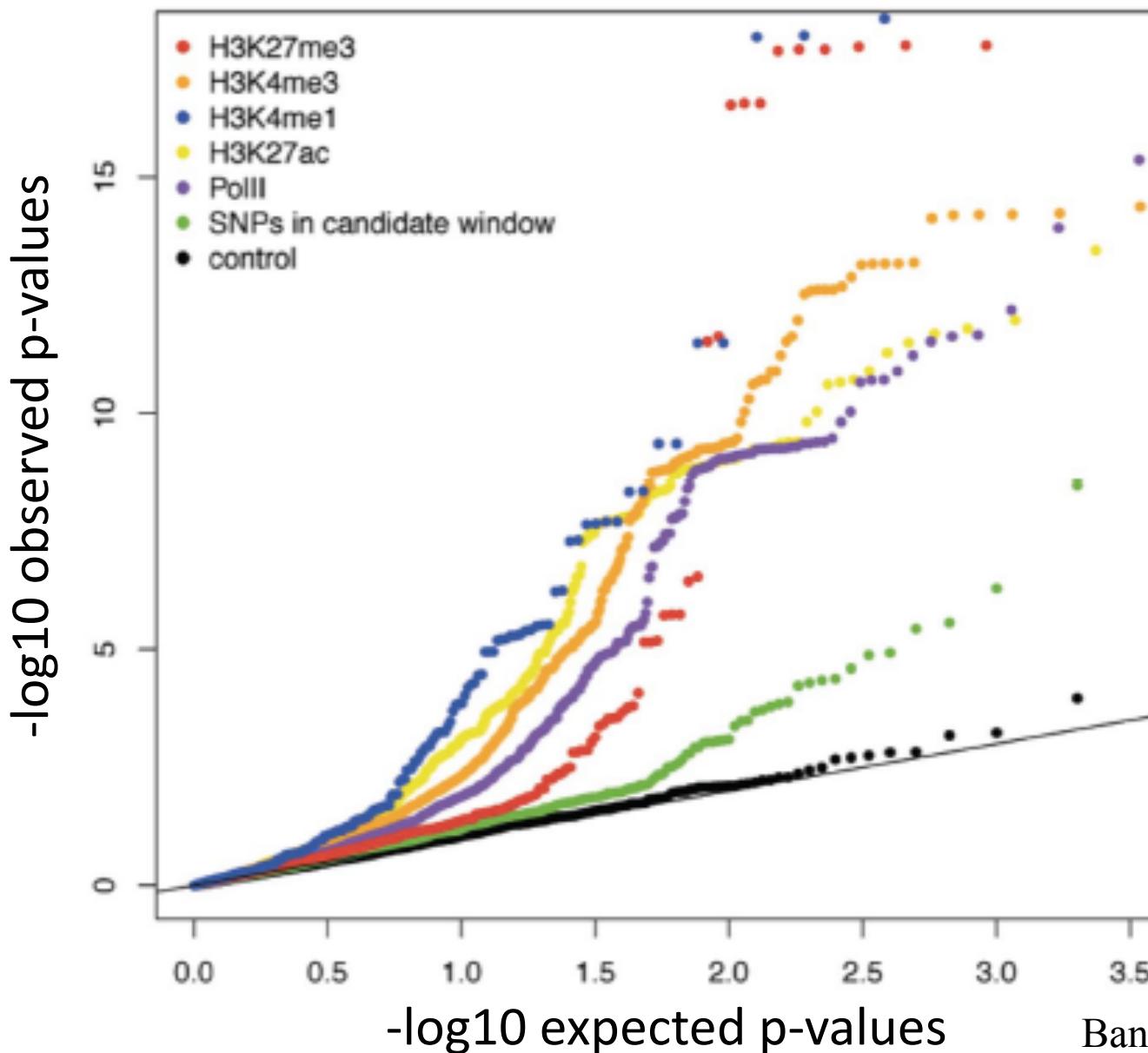
DNA methylation QTLs



DNA Methylation QTLs are often also DNase sensitivity QTLs

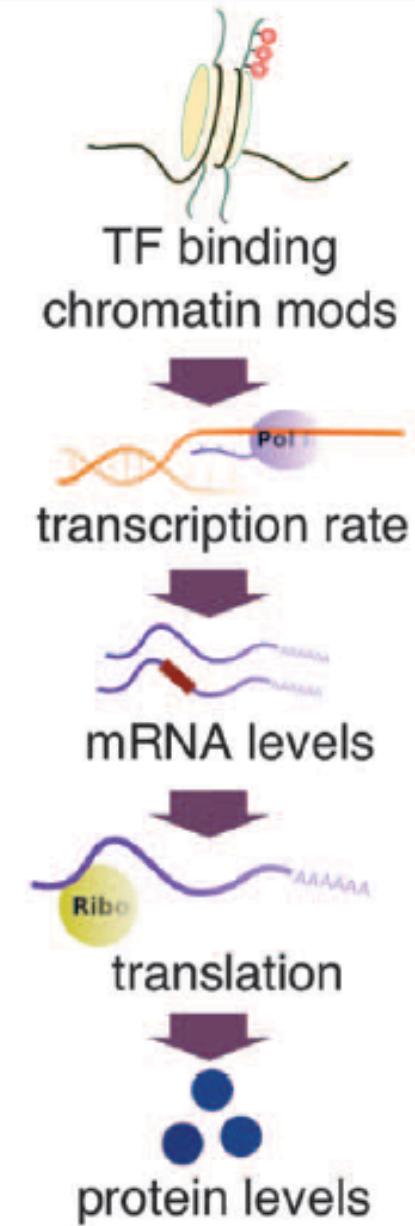


Histone mark QTLs are often DNA methylation QTLs



A Cell phenotypes

Regulatory cascade



Quantification

H3K27ac ChIP-seq,
DNA methylation,
DNase-seq
(n=59, 64, 67)

4sU-seq
(n=65,64;30m,60m)

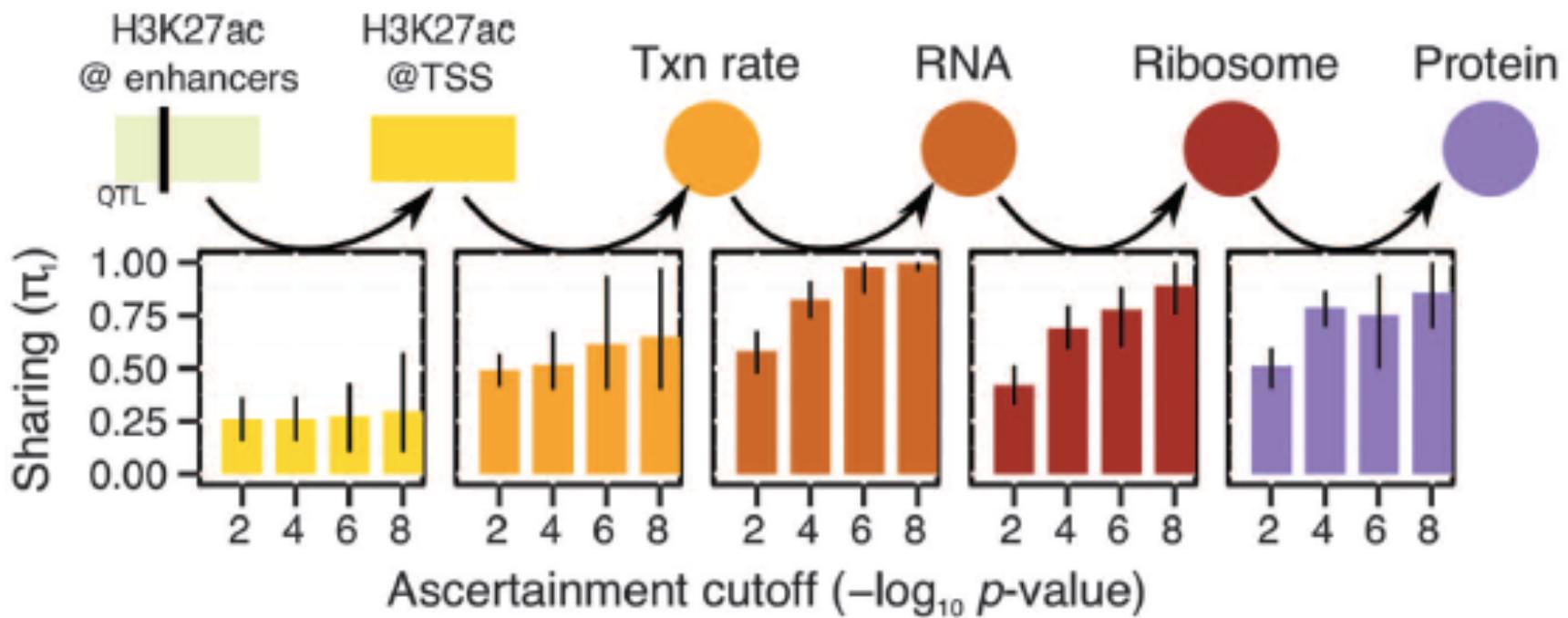
RNA-seq
(n=86,69;G,P)

RNA decay
(n=70)

Ribo-seq
(n=70)

Mass Spec.
(n=62)

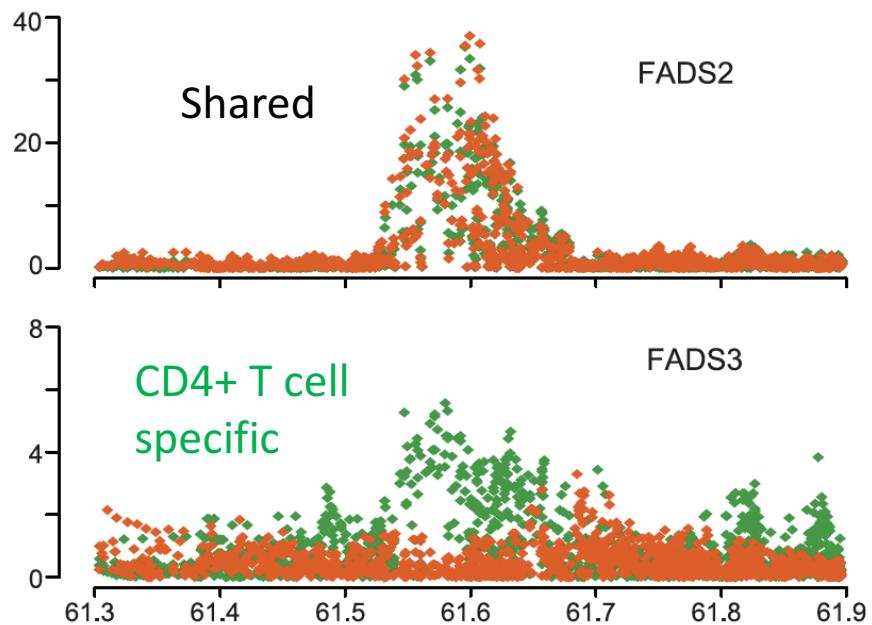
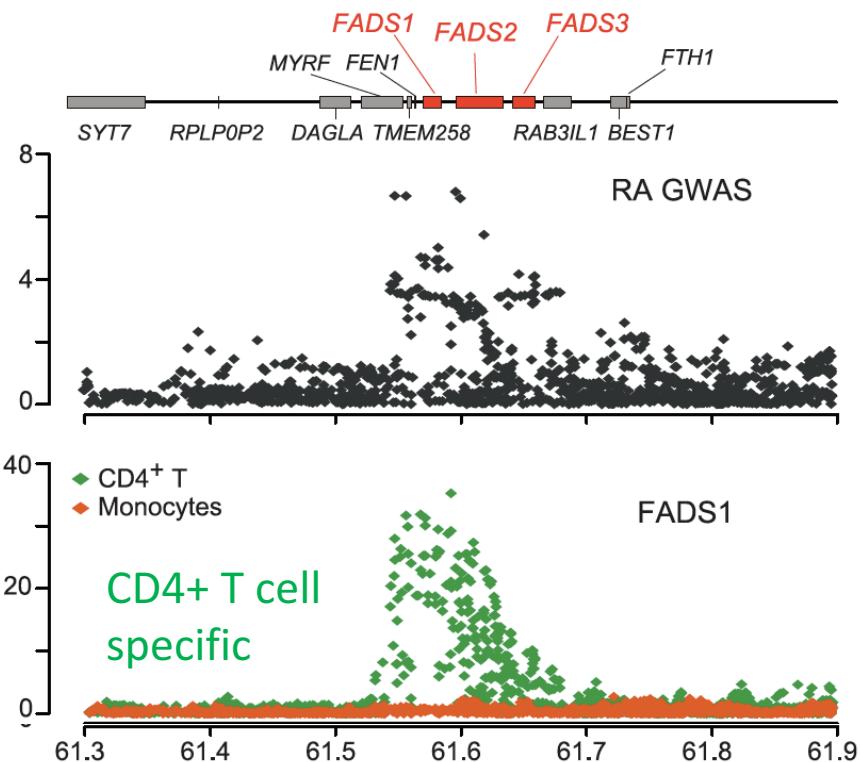
Sharing of regulatory QTLs



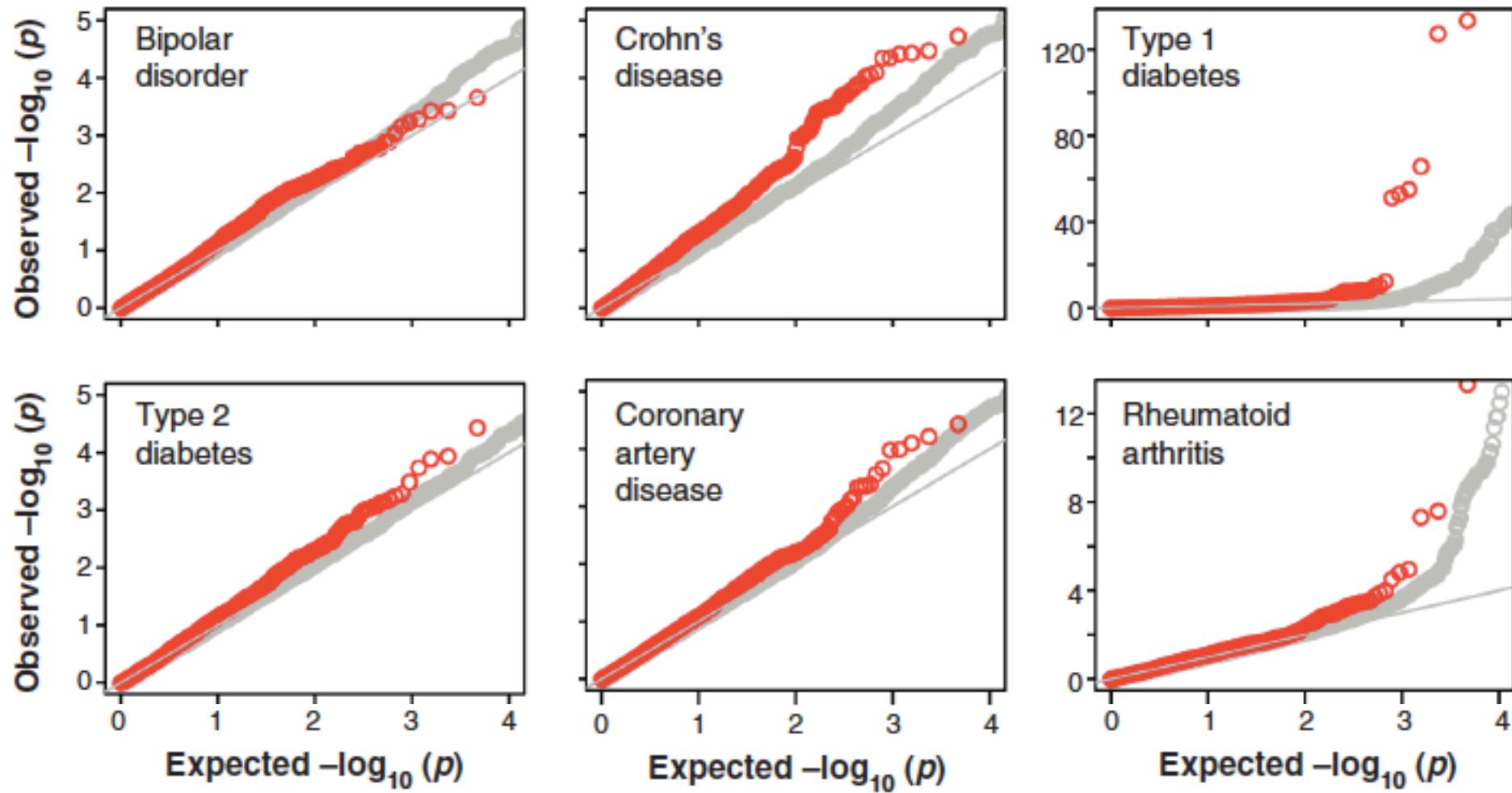
Discussion

- How can molecular traits be connected to organismal traits?
- What can we learn from molecular traits?

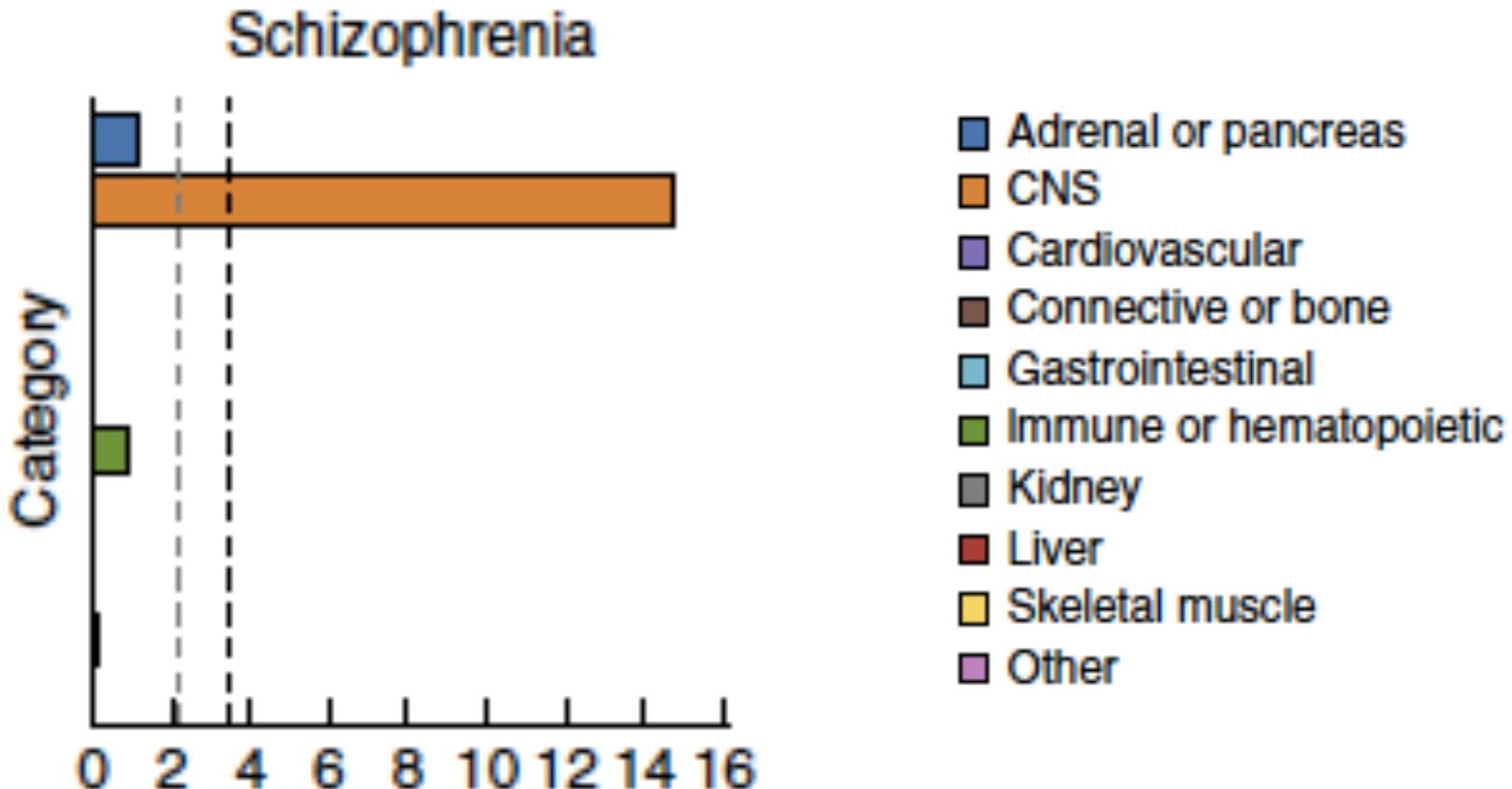
Intersecting eQTLs with GWAS

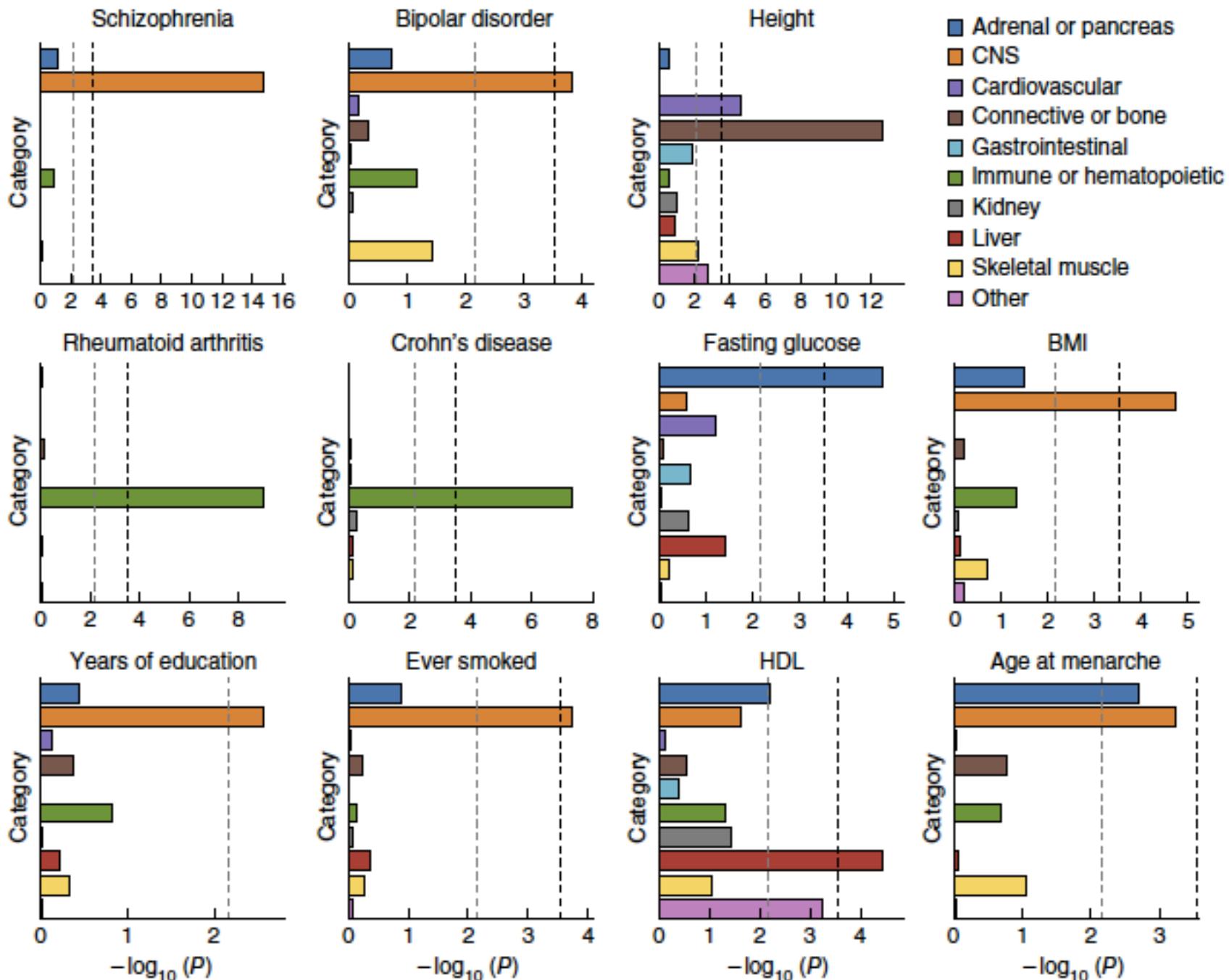


Whole Blood eQTLs are enriched for Crohn's Disease and T1D GWAS Hits



GWAS enrichment in cell-type specific annotations





Summary

- Many molecular traits can be associated with genetic variants
- Molecular QTLs can reveal mechanism underlying organismal traits
- Smaller samples are needed to map molecular QTLs that organismal traits
- Genetic associations are challenging to interpret