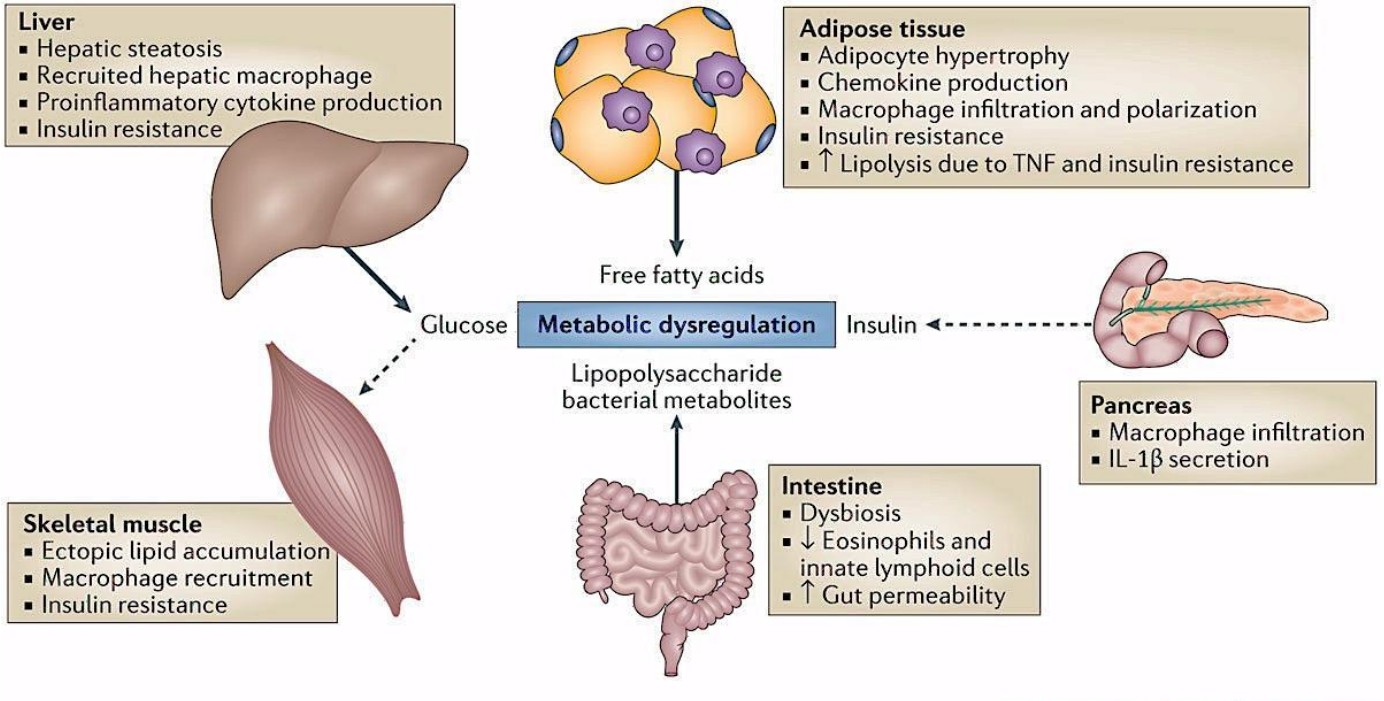
Chromatin accessibility to identify mechanisms of genetic regulation: metabolic traits case study

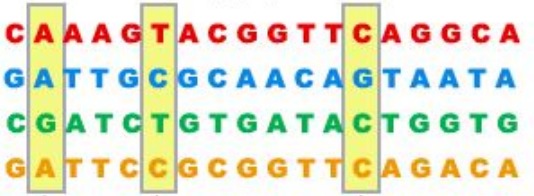
Michael Love
Dept of Genetics
UNC-Chapel Hill

Slides adapted from:
Karen L. Mohlke, UNC Genetics

Multiple tissues and processes regulate metabolic traits

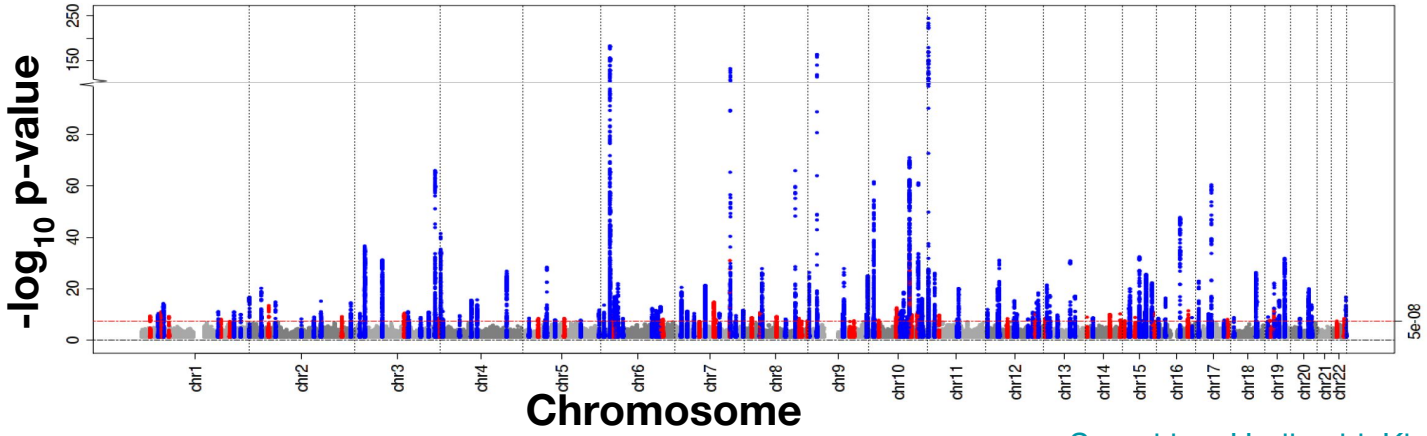


Lackey (2016) *Nat Rev Endocrinol* 12:15



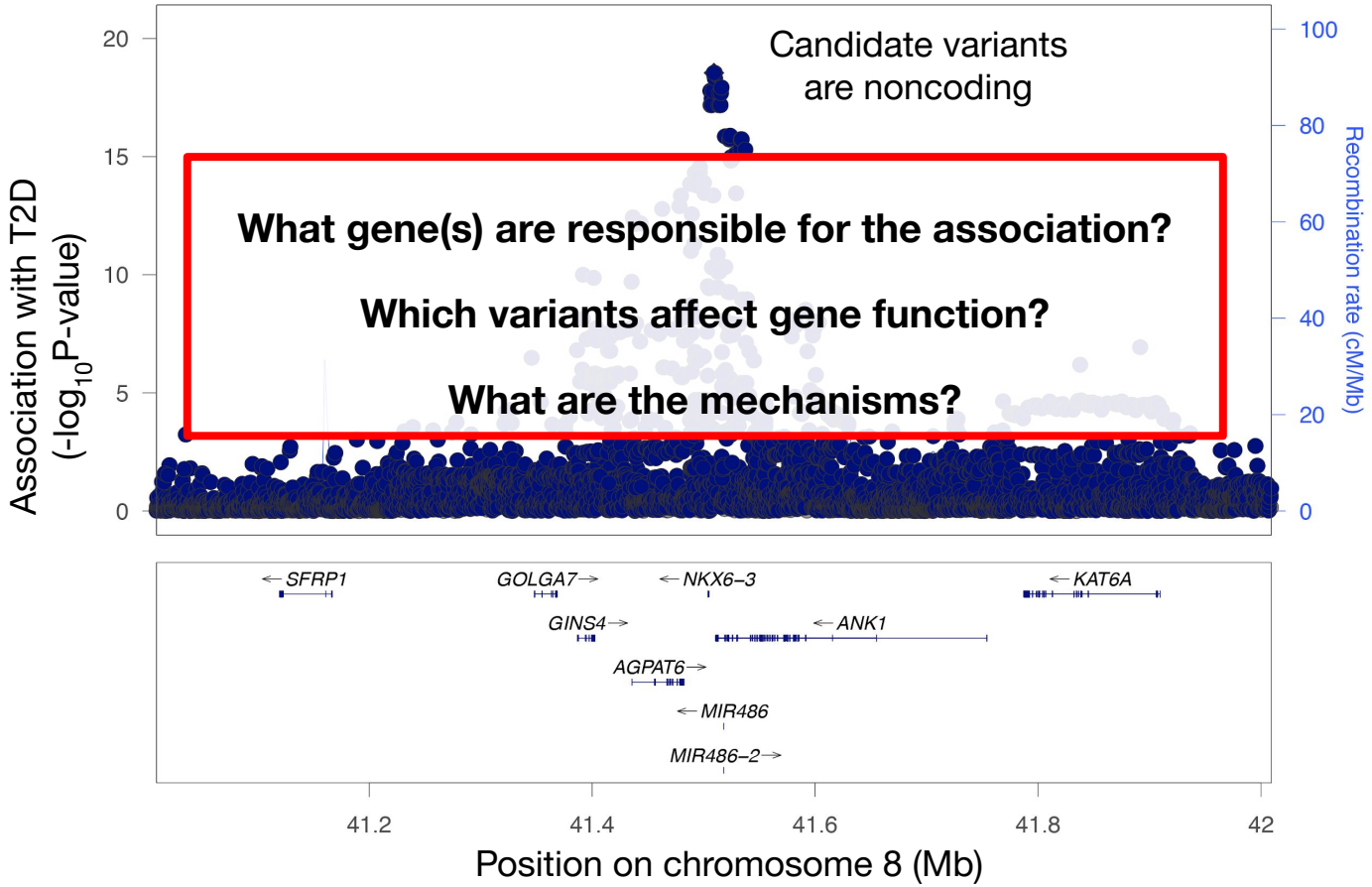
genetic variants ⟷ **type 2 diabetes**

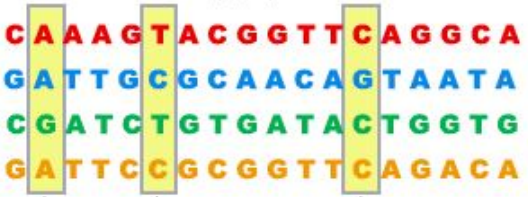
Type 2 diabetes East Asians	~11.7 million variants
77K T2D cases, 356K controls	298 signals at 178 loci



Spracklen, Horikoshi, Kim, Lin et al 2020

Locus associated with type 2 diabetes





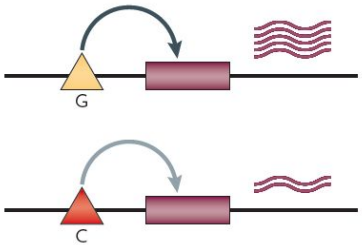
genetic variants ↔ type 2 diabetes

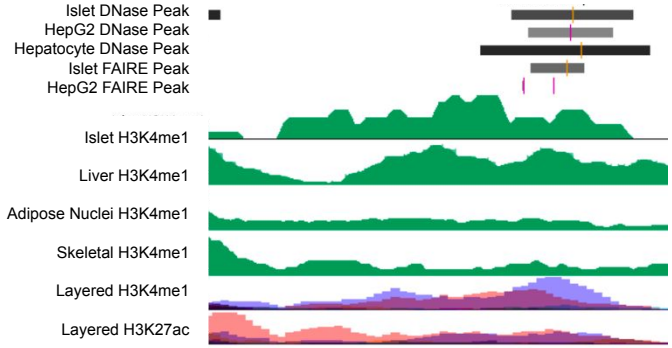
genetic variants ↕ gene expression

genetic variants ↔ chromatin context

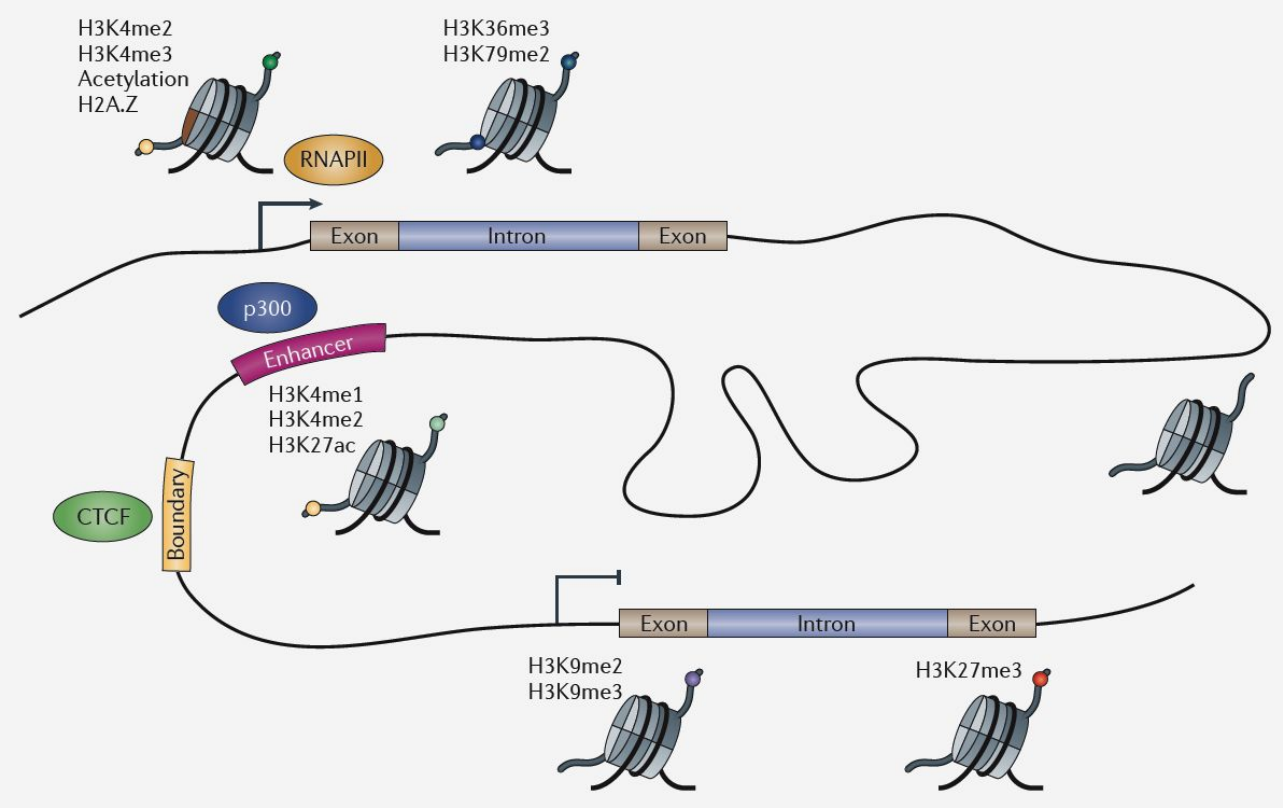
type 2 diabetes ↔ gene expression

type 2 diabetes ↔ chromatin context





Chromatin accessibility and histone marks of functional genomic elements

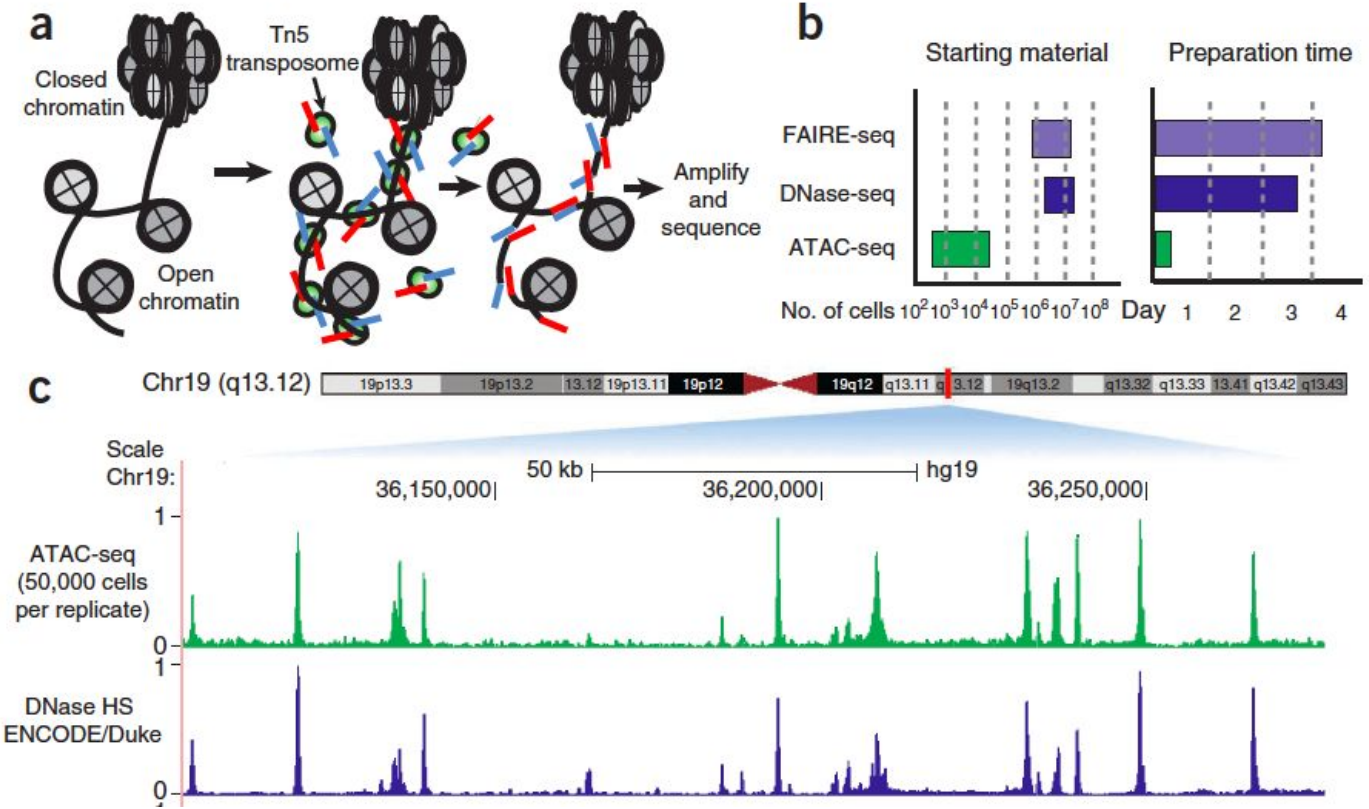


Zhou (2011) *Nat Rev Gen* 12:7

ATAC-seq to identify variants and mechanisms for complex metabolic traits

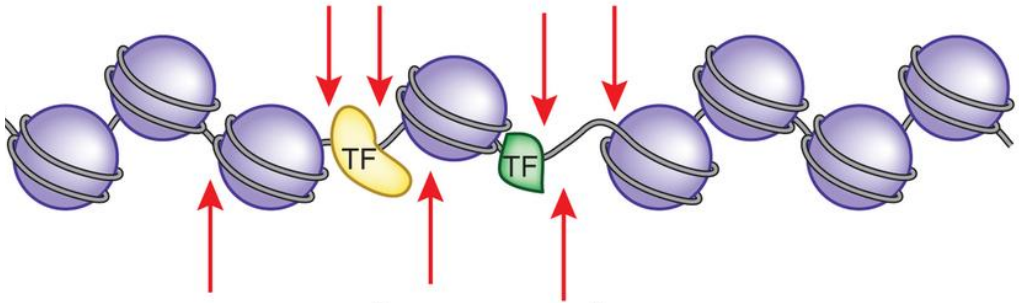
- **ATAC-seq method and quality control**
- **Identify accessible chromatin (regulatory elements)**
- **GWAS candidate variants located in ATAC-seq-defined elements**
- **Cell context/environmental effects on regulatory elements**
- **Variants associated with chromatin accessibility (caQTL)**

Quantify chromatin accessibility in clinical samples ATAC-seq



Buenrostro (2013) *Nat Meth* 10:1213; Corces (2017) *Nat Met*14:959

Isolate nuclei
Integrate Tn5 in open sites



Amplify regions

not known *a priori*

Build library, sequence, align

ATAC-seq read processing

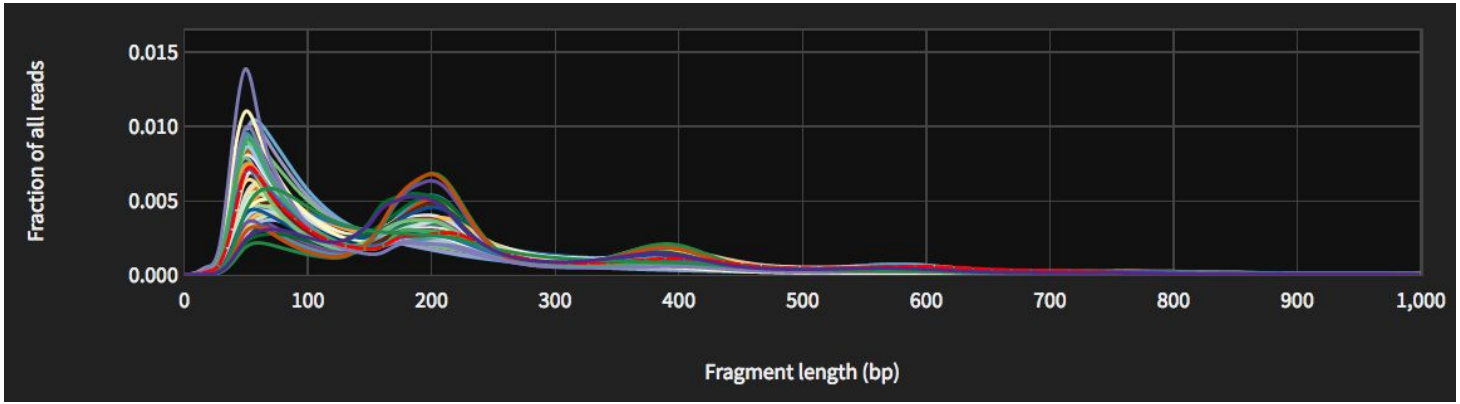
- Remove low quality reads (e.g. fastqc, fastx)
- Trim adapters (e.g. cutadapt, tagdust)
- Map reads (e.g. bwa, bowtie2, GSNAP)
- Remove mitochondrial reads & excluded regions (e.g. Picard tools, Samtools)
- Shift alignments +4/-5 so the 5’ base corresponds to the center of the Tn5 cut site
- Call peaks (e.g. MACS2)

Evaluate ATAC-seq library metrics (ataqv)

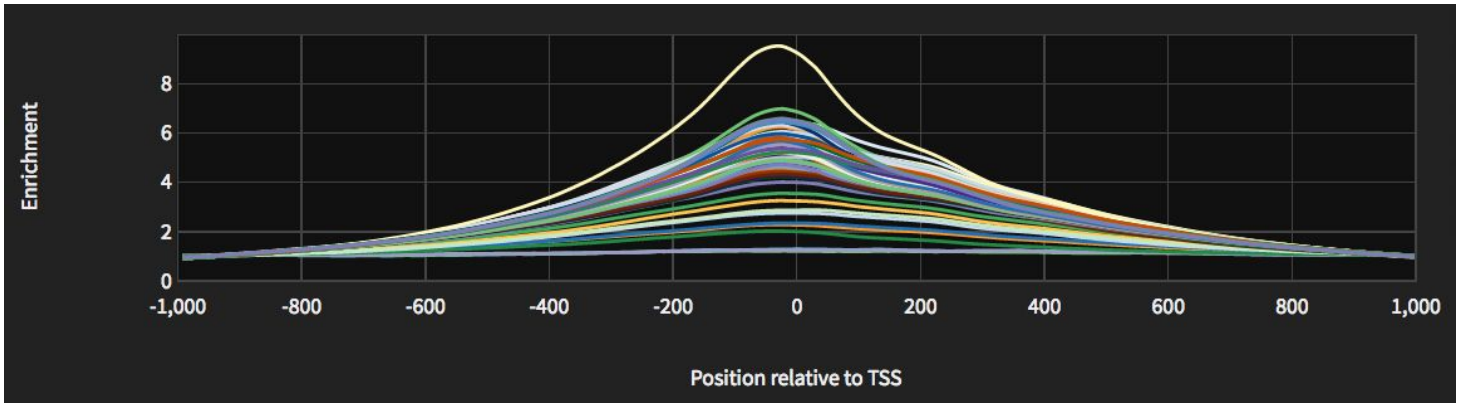
- Reads mapped in proper pairs
- Optical or PCR duplicates
- Reads mapping to autosomal or mitochondrial references
- Ratio of short to mononucleosomal fragment counts
- Mapping quality
- Problematic alignments
- Read coverage of peaks (of provided peak calls)
- Transcription start site enrichment (of provided TSS)

https://github.com/ParkerLab/ataqv/

ataqv: Fragment length vs fraction of all reads (periodic profile is better):

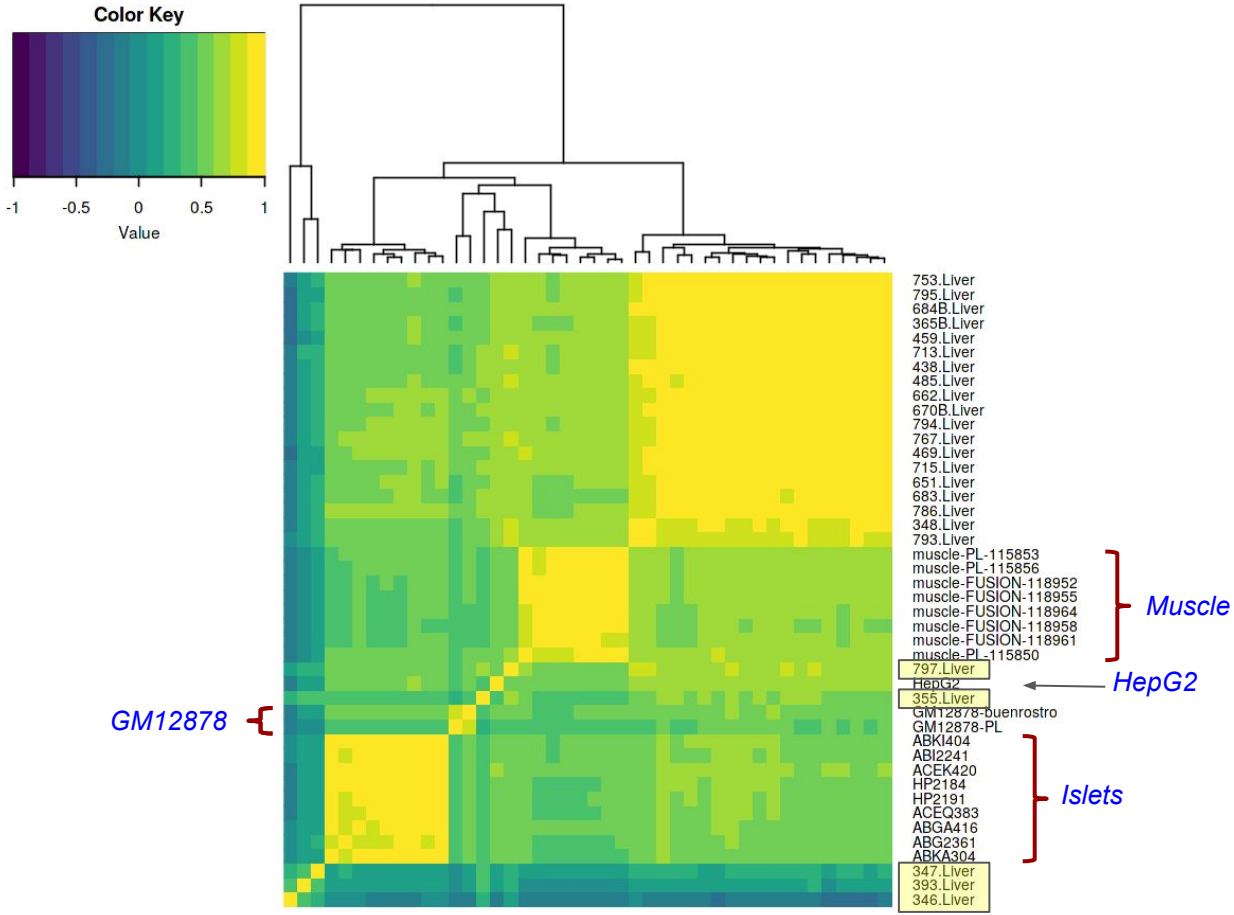


Position relative to TSS vs enrichment (higher enrichment is ~better):



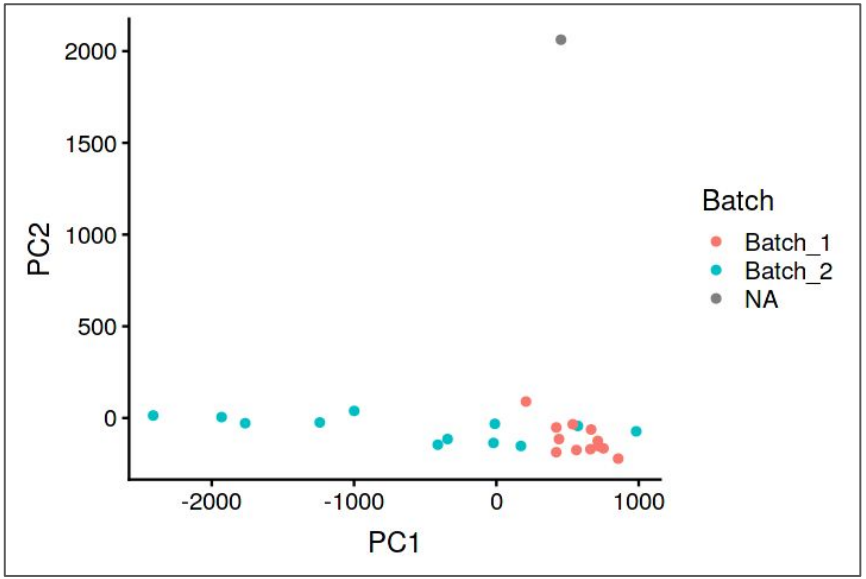
Do samples cluster as expected? BAM correlations

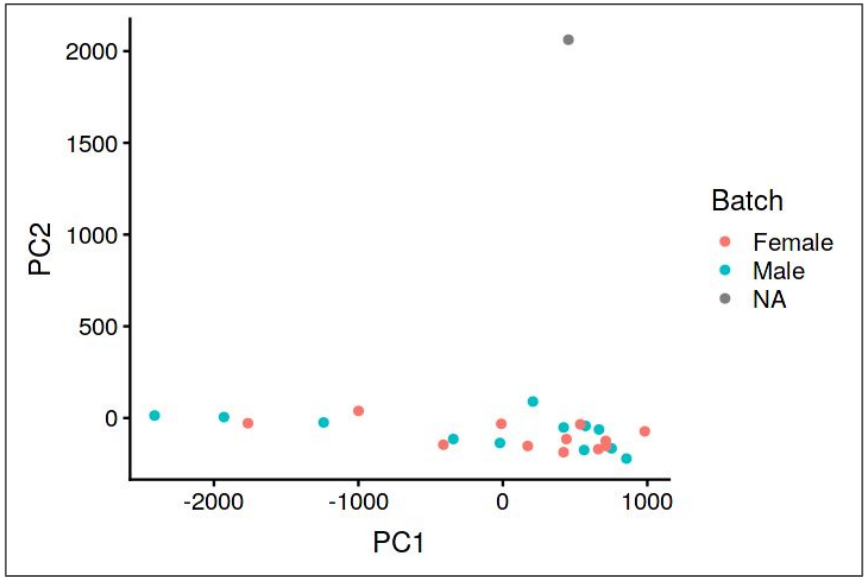
- **Call peaks on each BAM file individually, FDR < 0.05**
- **Create a “pooled” peak file -- Union of peaks from all samples**
- **Count reads overlapping peaks in the pooled file**
- **Compute CPM → Correlation → Hierarchical clustering**


Color Key
-1
-0.5
0
0.5
1
Value
753.Liver
795.Liver
684B.Liver
365B.Liver
459.Liver
713.Liver
438.Liver
485.Liver
662.Liver
670B.Liver
794.Liver
767.Liver
469.Liver
715.Liver
651.Liver
683.Liver
786.Liver
348.Liver
793.Liver
muscle-PL-115853
muscle-PL-115856
muscle-FUSION-118952
muscle-FUSION-118955
muscle-FUSION-118964
muscle-FUSION-118958
muscle-FUSION-118961
muscle-PL-115850
797.Liver
HepG2
355.Liver
GM12878-buenrostro
GM12878-PL
ABKI404
ABI2241
ACEK420
HP2184
HP2191
ACEQ383
ABGA416
ABG2361
ABKA304
347.Liver
393.Liver
346.Liver
Muscle
HepG2
Islets
GM12878

Vivek Rai

Identify potential sources of variation: annotate principal component analysis plot



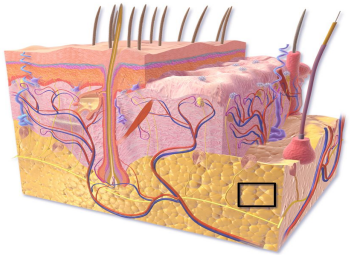


Vivek Rai

ATAC-seq to identify variants and mechanisms for complex metabolic traits

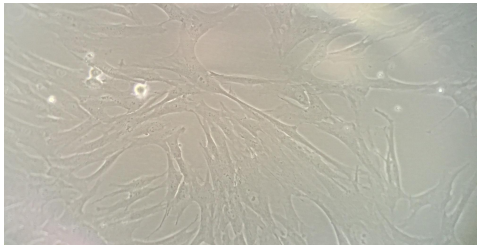
- ATAC-seq method and quality control
- **Identify accessible chromatin (regulatory elements)**
- GWAS candidate variants located in ATAC-seq-defined elements
- Cell context/environmental effects on regulatory elements
- Variants associated with chromatin accessibility (caQTL)

Adipose tissue and adipocyte chromatin accessibility



Adipose Tissue

- Best represents *in vivo* adipose tissue
- Can be assayed across individuals to identify genetic variants that alter chromatin accessibility
- **Tissue extraction site, storage, and handling** can cause differences in chromatin accessibility



SGBS Preadipocytes

- Diploid cell strain allows for consistent genetic background across experiments
- Can be differentiated to mature adipocytes and treated with stimuli
- Cells growing in culture may not fully represent cells within adipose tissue

ATAC-seq in human adipose tissue and cell line (pre)adipocytes

Sample	Total reads	Aligned reads	Percent mitochondrial reads	Nuclear alignments	Remaining reads after blacklist filtering	Remaining reads after duplicates removed	Number of peaks[b]
Tissue 1	129.5	87.4	8.5	80.0	79.0	70.6	58,550
Tissue 2	131.5	83.6	12.8	72.9	71.8	60.6	36,785
Tissue 3	119.3	70.5	11.9	62.2	61.3	57.1	49,962

Average overlap of percent bases in top 25K peaks between tissue samples: 73.8%

Cannon, Currin (2019) *G3* AOP

ATAC-seq in human adipose tissue and cell line (pre)adipocytes

Sample	Total reads	Aligned reads	Percent mitochondrial reads	Nuclear alignments	Remaining reads after blacklist filtering	Remaining reads after duplicates removed	Number of peaks[b]
Tissue 1	129.5	87.4	8.5	80.0	79.0	70.6	58,550
Tissue 2	131.5	83.6	12.8	72.9	71.8	60.6	36,785
Tissue 3	119.3	70.5	11.9	62.2	61.3	57.1	49,962
SGBS adipocytes 1[a]	382.6	275.9	2.1	268.6	267.7	90.4	184,455
SGBS adipocytes 2[a]	245.1	172.9	1.9	168.7	168.1	84.1	172,247
SGBS adipocytes 3[a]	253.7	181.0	1.5	177.2	176.7	87.5	191,141
SGBS preadipocytes 1[a]	97.3	71.8	1.0	70.8	70.7	34.6	171,279
SGBS preadipocytes 2[a]	75.1	54.1	1.1	53.3	53.1	30.5	139,911

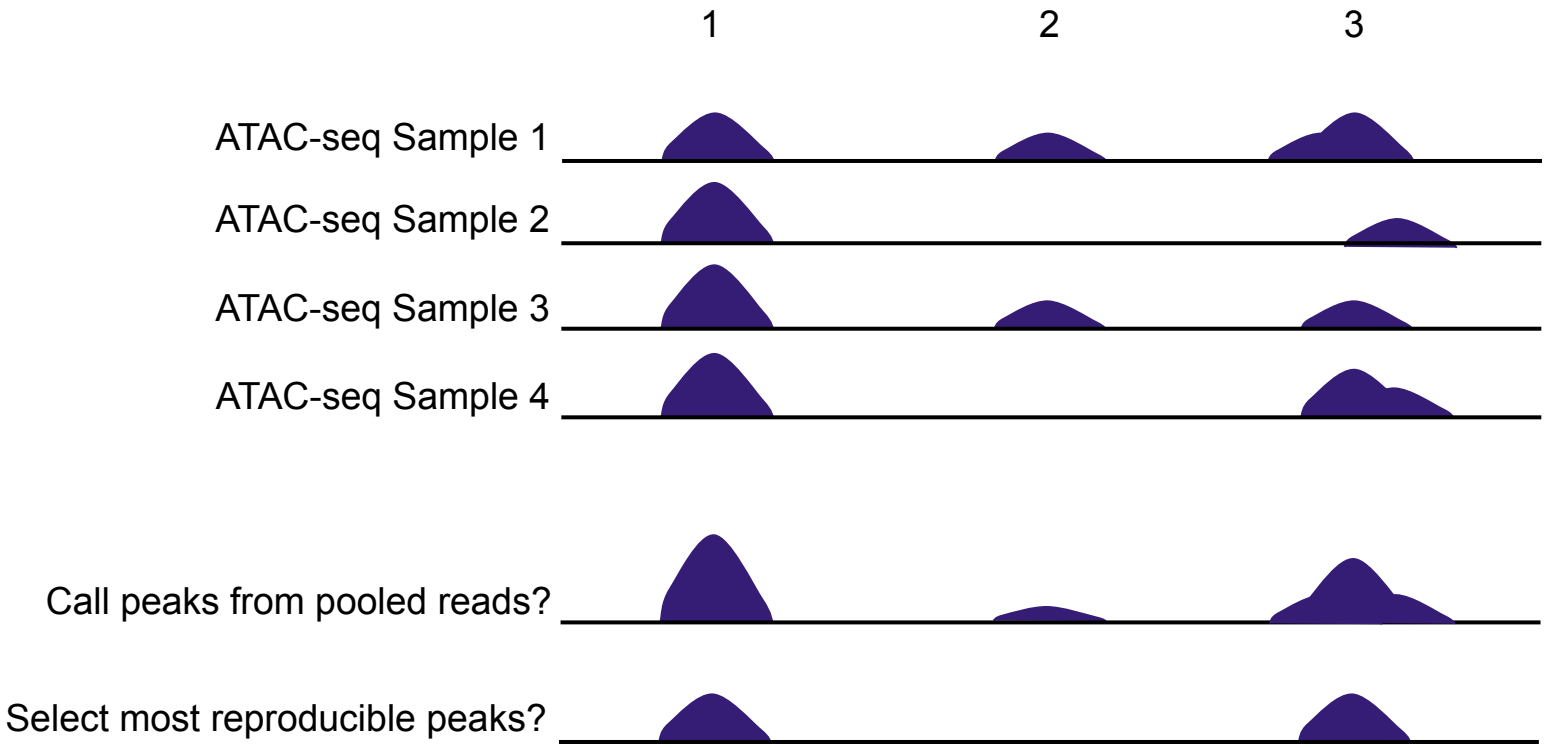
Reads are reported in millions of reads.
[a] Samples were sequenced using paired-end reads, but processed as single-end reads.
[b] We identified 68,571 representative peaks across adipose tissue, 122,924 across SGBS preadipocytes, and 164,252 across SGBS adipocyte samples.

Average overlap of percent bases in top 25K peaks between tissue samples: 73.8%

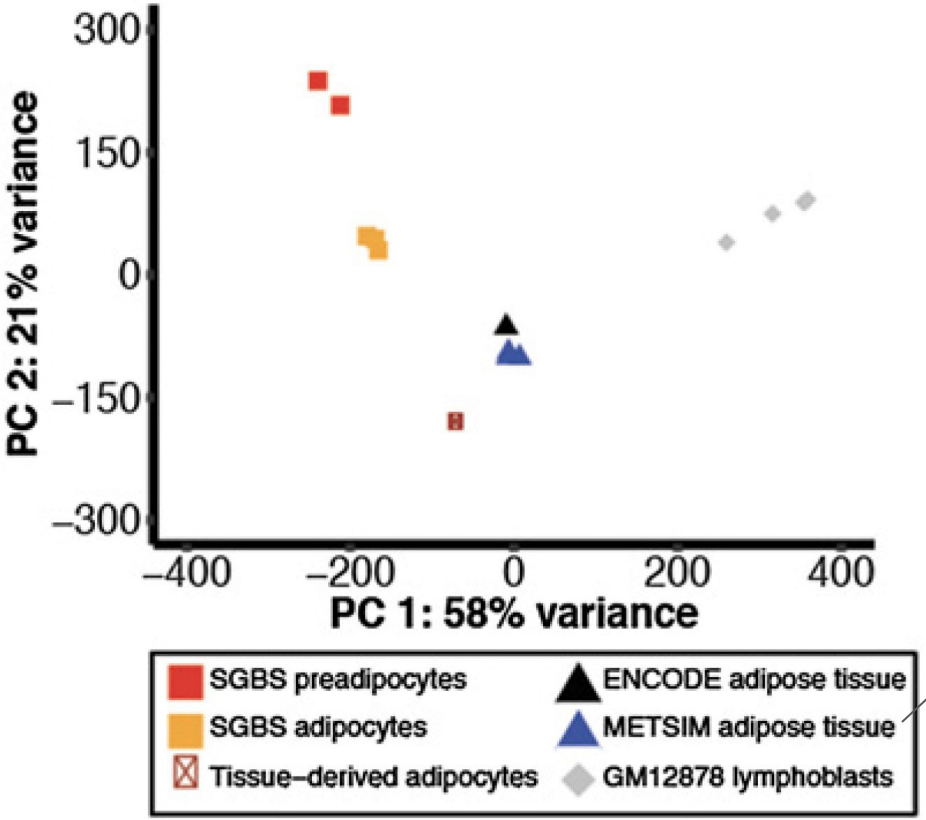
Average overlap of percent bases in top 25K peaks SGBS preadipocytes: 84.8%

Cannon, Currin (2019) *G3* AOP

What are representative peaks?



Principal components analysis of ATAC-seq read counts within representative peaks

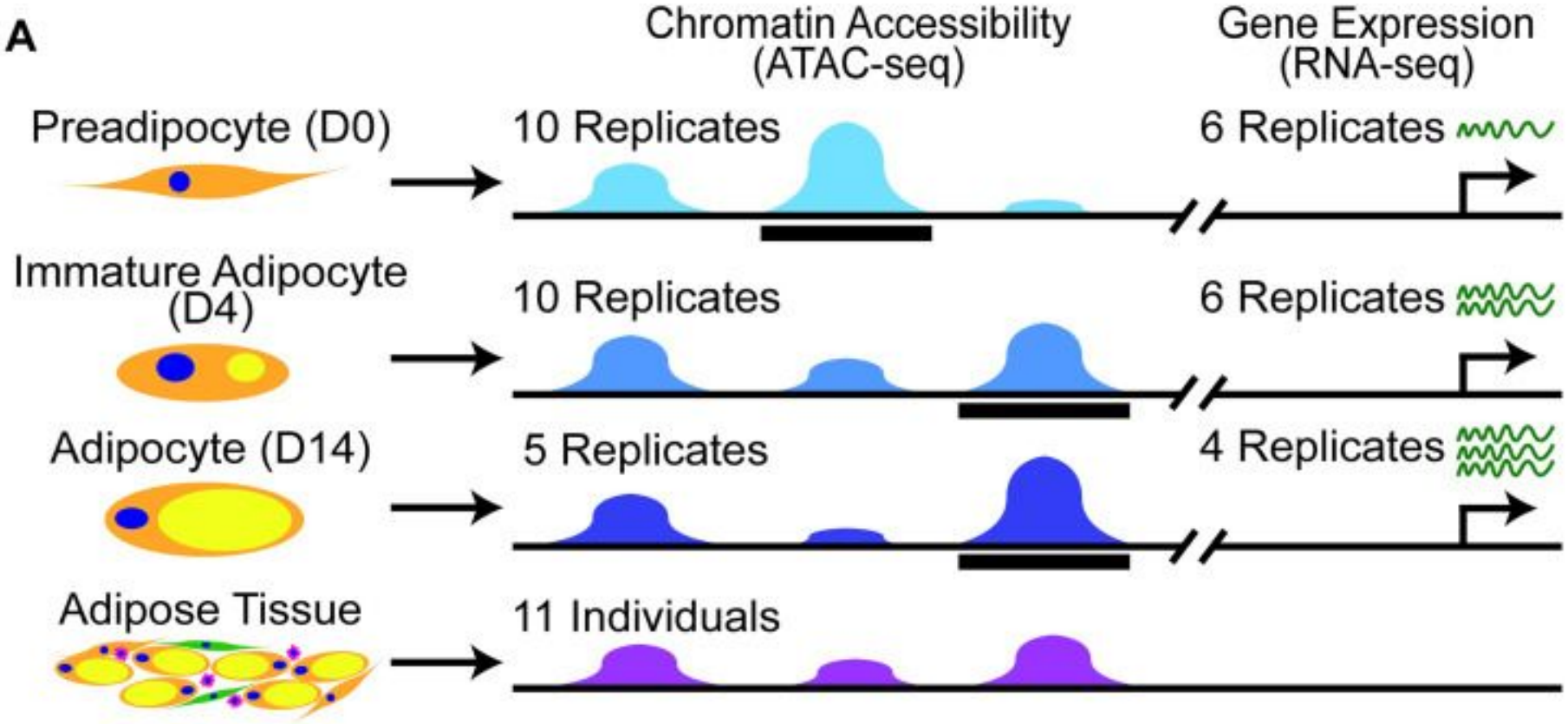


- Counted non-duplicated nuclear reads overlapping the total set of accessible chromatin regions
- Library size scaling and variance stabilization in DESeq2
- plotPCA in DESeq2

Metabolic Syndrome In Men

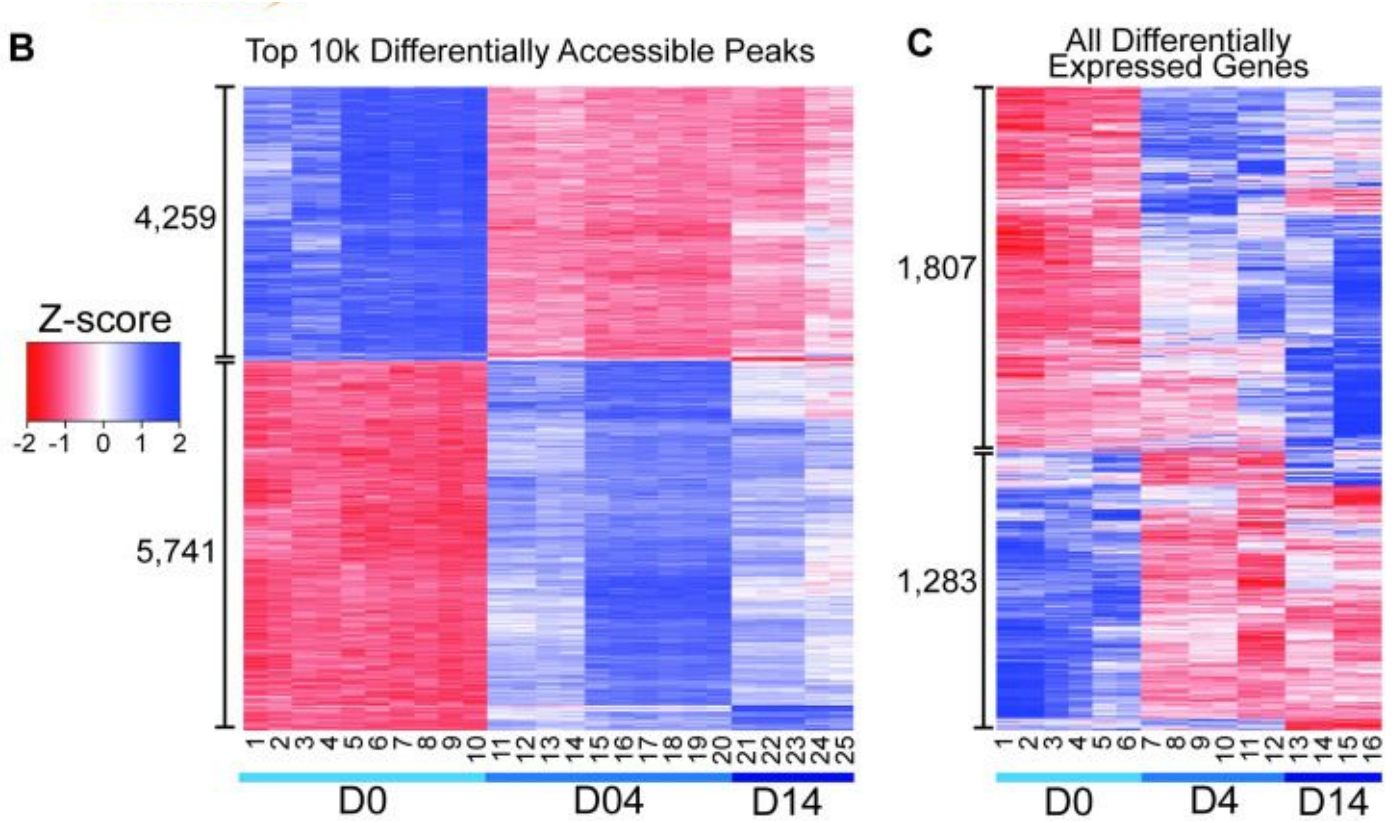
Cannon, Currin (2019) *G3* AOP
Featurecounts: Liao (2014) *Bioinfo* 30:923
DESeq2: Love (2014) *Genome Biol* 15:550
Tissue-derived adipocytes Allum (2015) *Nat Comm* 6:7211

Adipose ATAC-seq marks enhancers and promoters



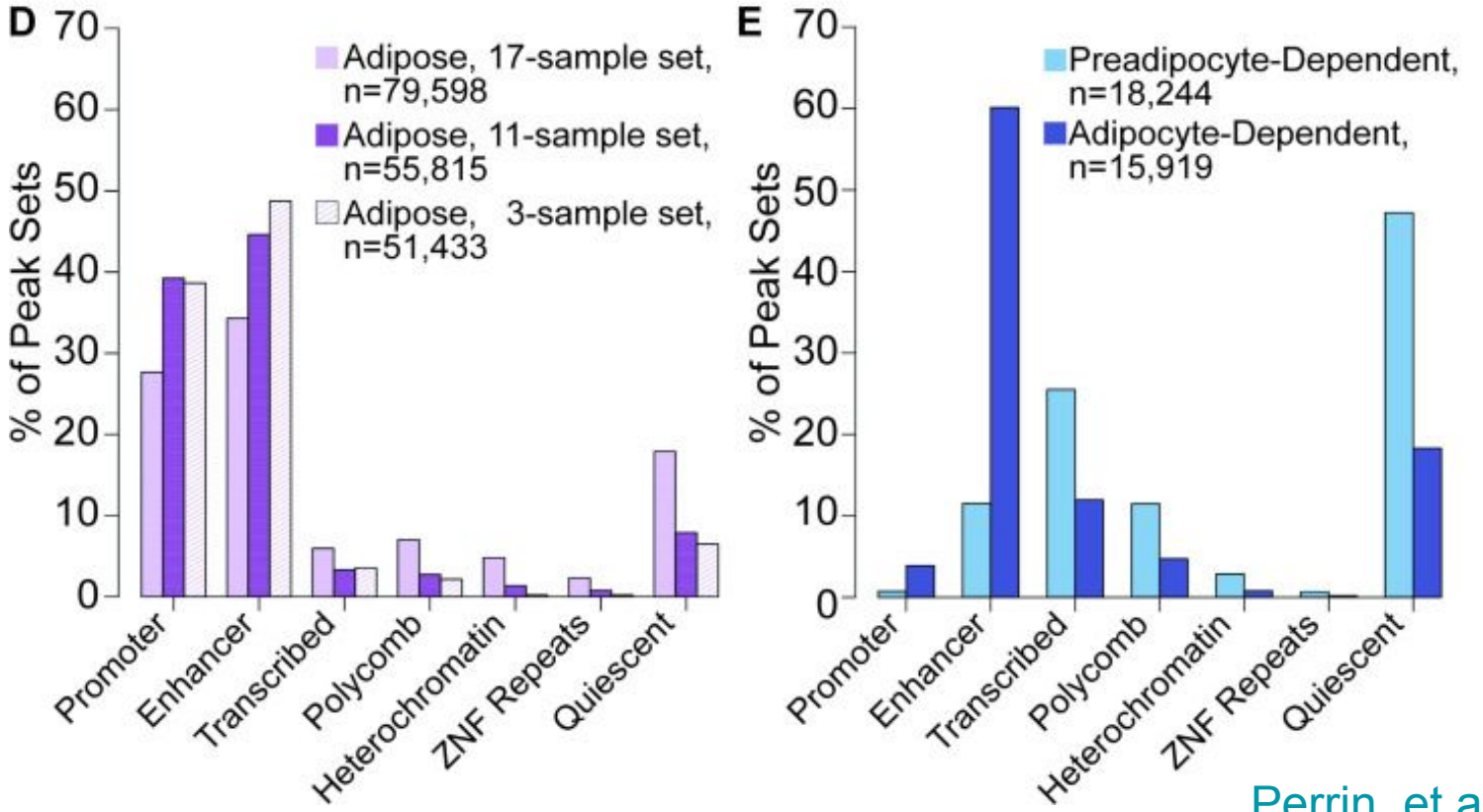
Perrin, et al (2021)

Adipose ATAC-seq marks enhancers and promoters



Perrin, et al (2021)

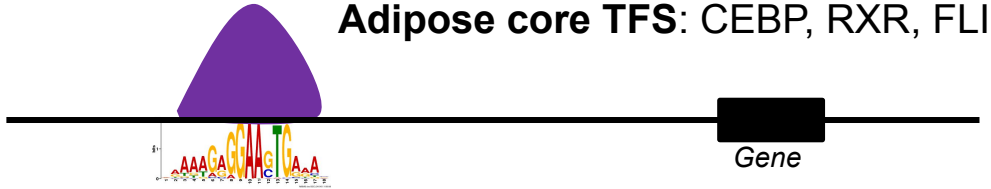
Adipose ATAC-seq marks enhancers and promoters



Perrin, et al (2021)

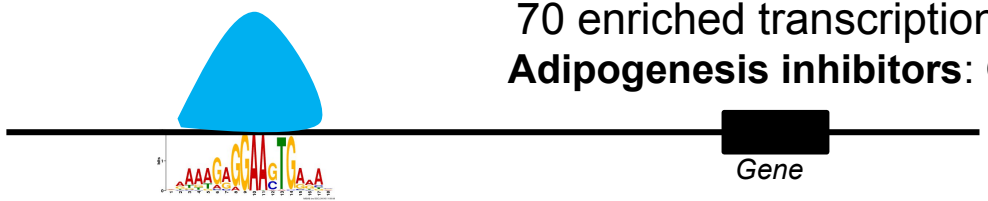
Peaks specific to tissue and preadipocyte cells show different regulatory signatures

METSIM adipose tissue



162 enriched transcription factor motifs
Adipose core TFS: CEBP, RXR, FLI1, ETS1, FOXO1, IRF1

SGBS preadipocytes

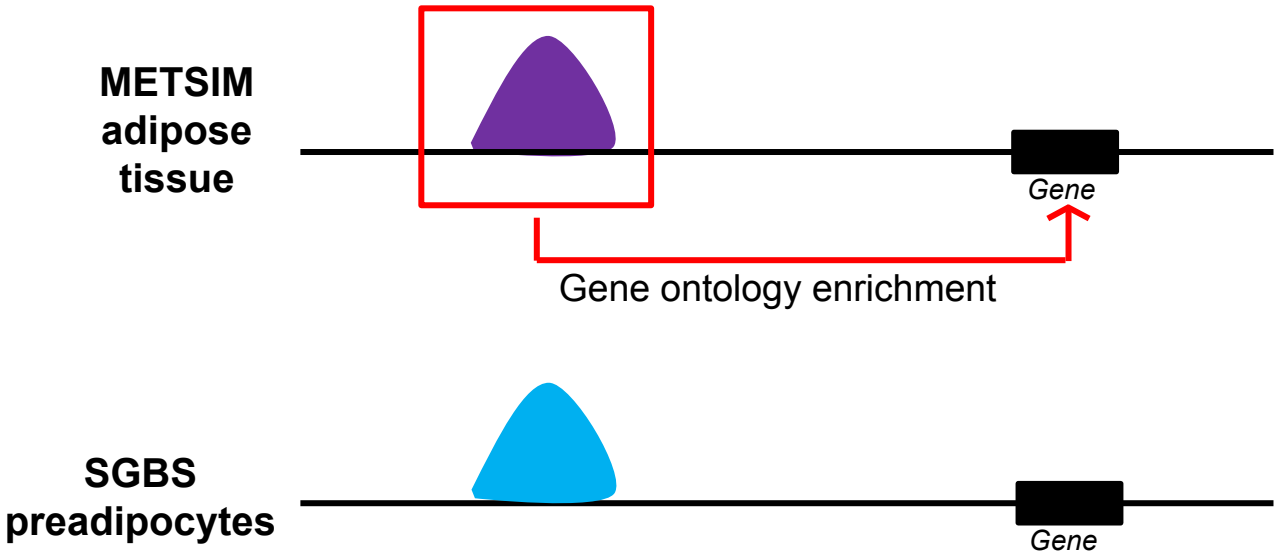


70 enriched transcription factor motifs
Adipogenesis inhibitors: GATA2, GATA3

Enrichment of 519 transcription factor motifs from JASPAR using Motif Enrichment Analysis (AME)
Comparisons made with the top 50,000 representative peaks
Shuffled peak sequences with preserved dinucleotide content as background
Fisher Exact, E<1E-100

Mathelier (2016) *Nuc Acid Res* 44:D110; McLeay *BMC Bioinformatics* 11:165

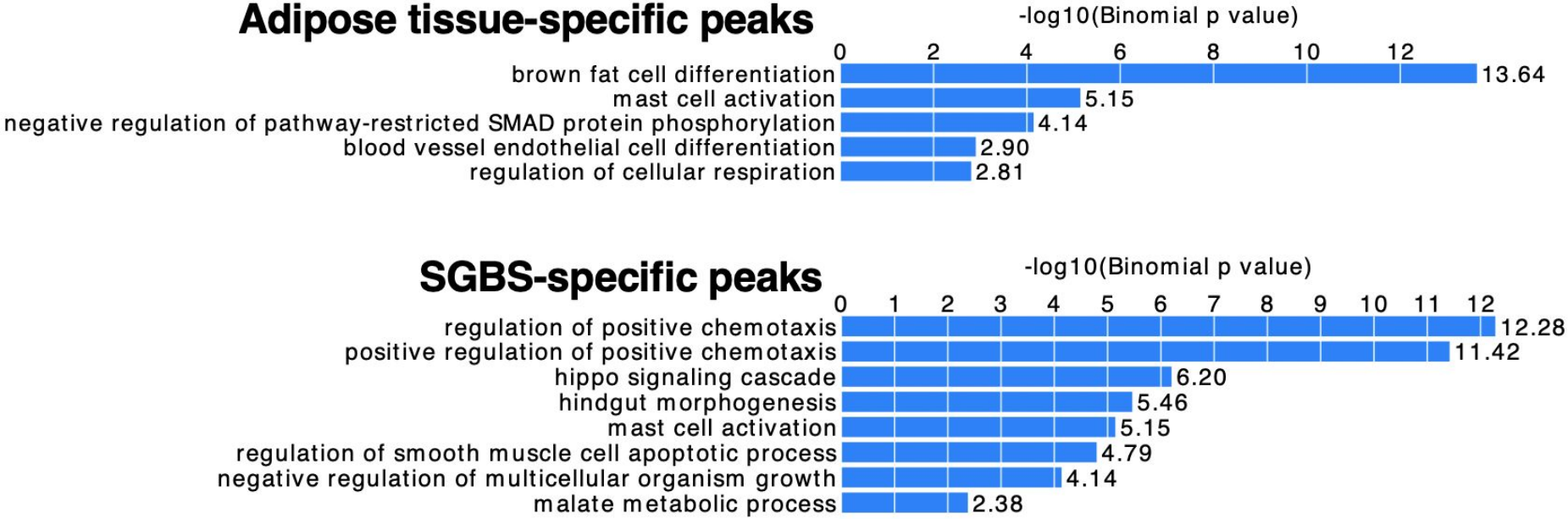
Peaks specific to tissue and preadipocyte cells show different regulatory signatures



Enrichment of GO biological process ontology for genes located near peaks using GREAT
Comparisons made with the top 10,000 representative peaks

McLean (2010) *Nat Biotech 28*: 495

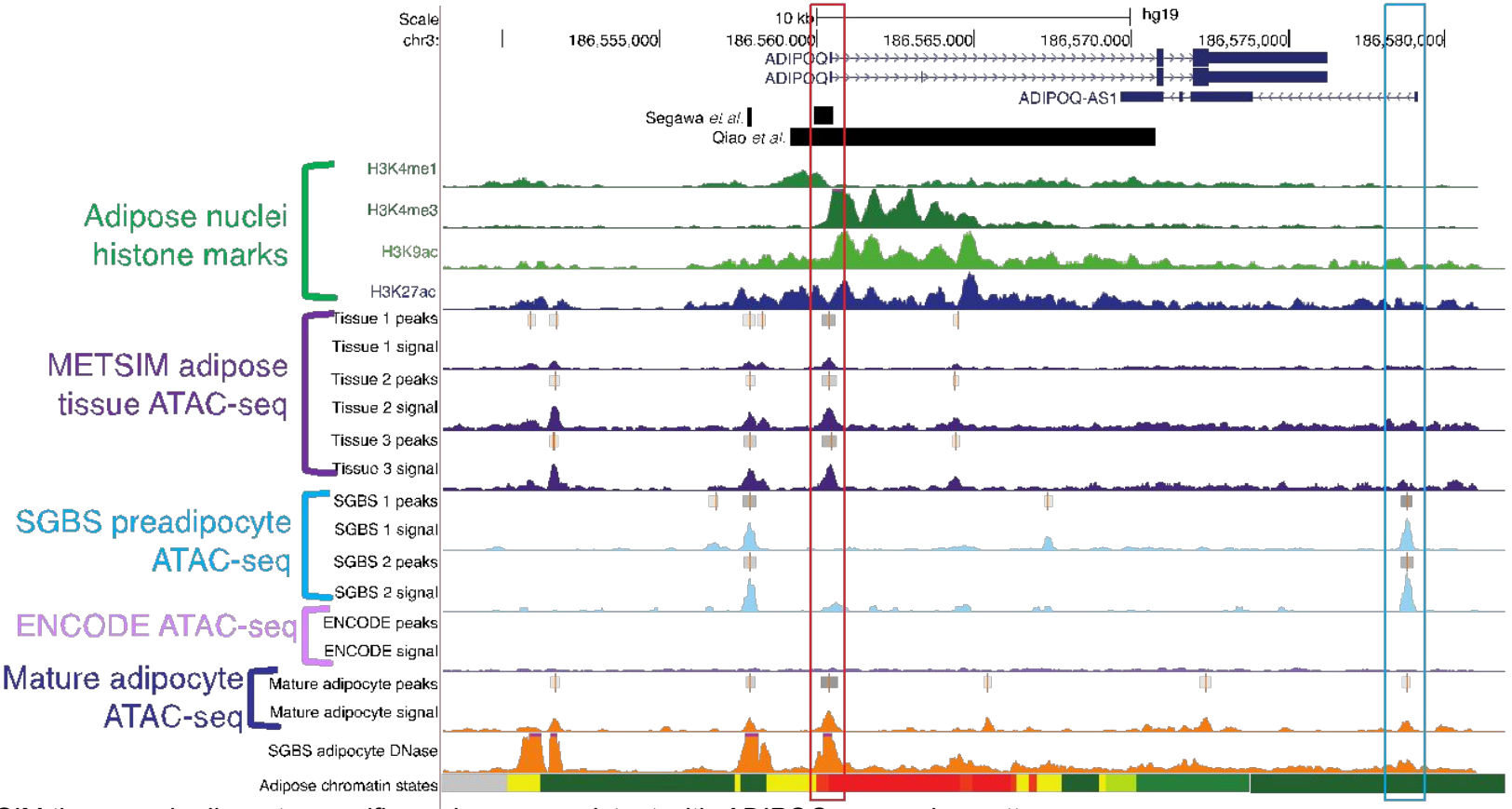
Peaks specific to tissue and preadipocyte cells show different regulatory signatures



Enrichment of GO biological process ontology for genes located near peaks using GREAT
Comparisons made with the top 10,000 representative peaks

McLean (2010) *Nat Biotech 28*: 495

Chromatin accessibility at ADIPOQ



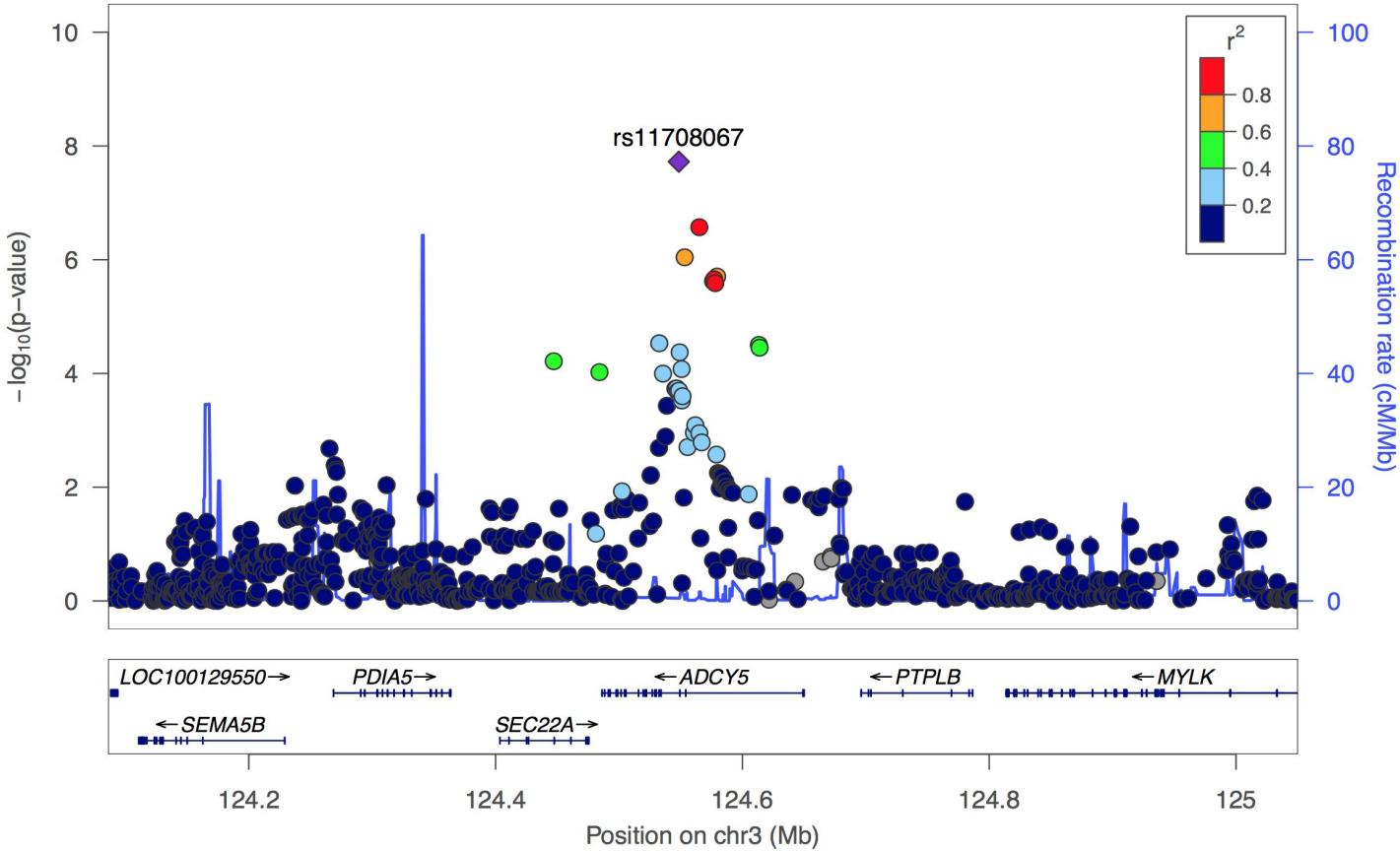
- METSIM tissue and adipocyte-specific peaks are consistent with *ADIPOQ* expression patterns
- Lack of peaks in ENCODE suggest further differences in tissue sampling and/or preparation
- SGBS peaks near *ADIPOQ-AS1* promoter suggest antisense expressed more in preadipocytes

Data from ENCODE, Epigenome Roadmap
Allum 2015 *Nat Comm* 6:7211
Schmidt 2015 *Genome Res* 25:1281

ATAC-seq to identify variants and mechanisms for complex metabolic traits

- ATAC-seq method and quality control
- Identify accessible chromatin (regulatory elements)
- **GWAS candidate variants located in ATAC-seq-defined elements**
- Cell context/environmental effects on regulatory elements
- Variants associated with chromatin accessibility (caQTL)

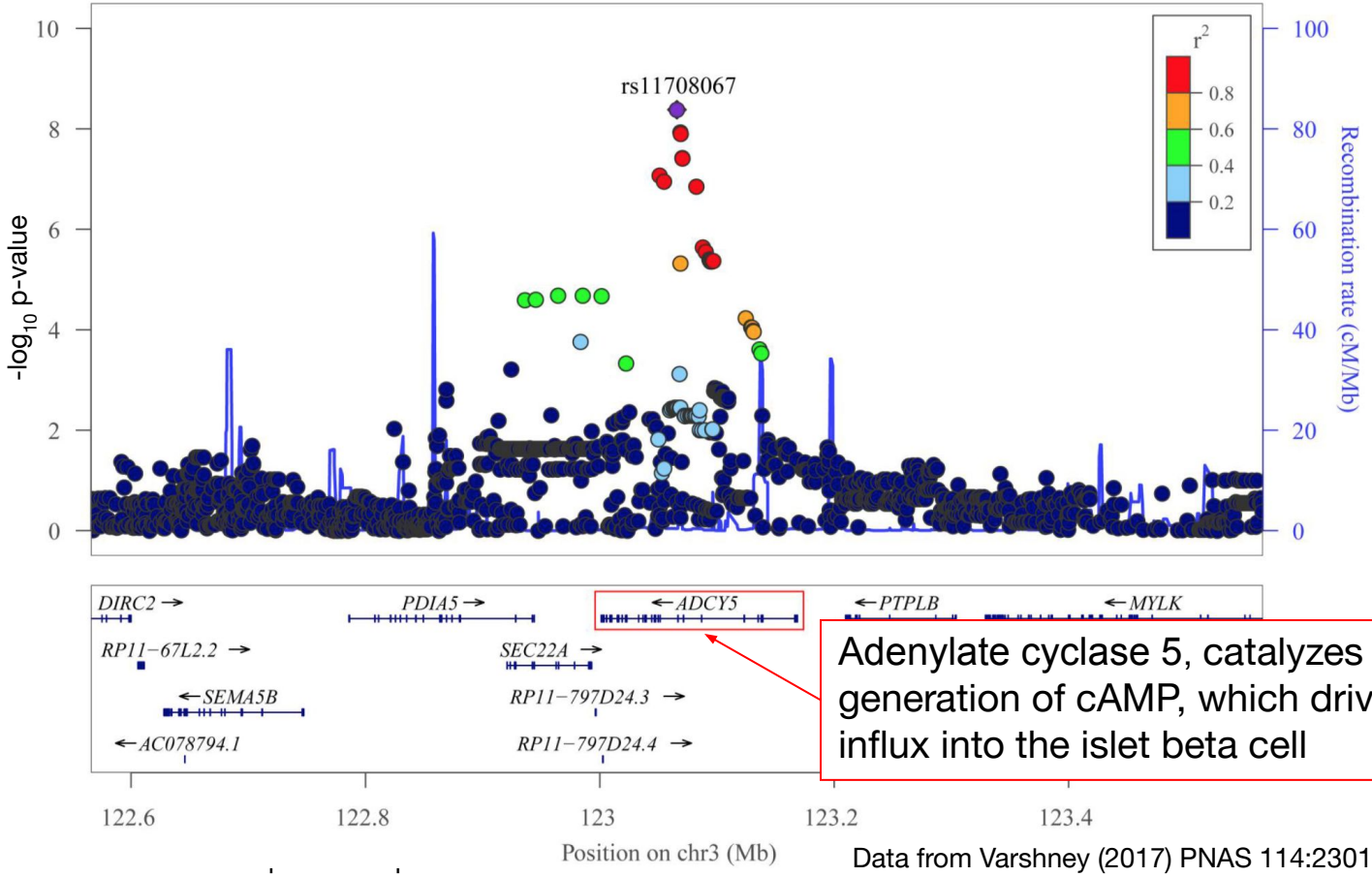
Glucose-associated variants at ADCY5



~58,000 non-diabetic individuals

Data from Manning (2012) Nat Gen 44:659

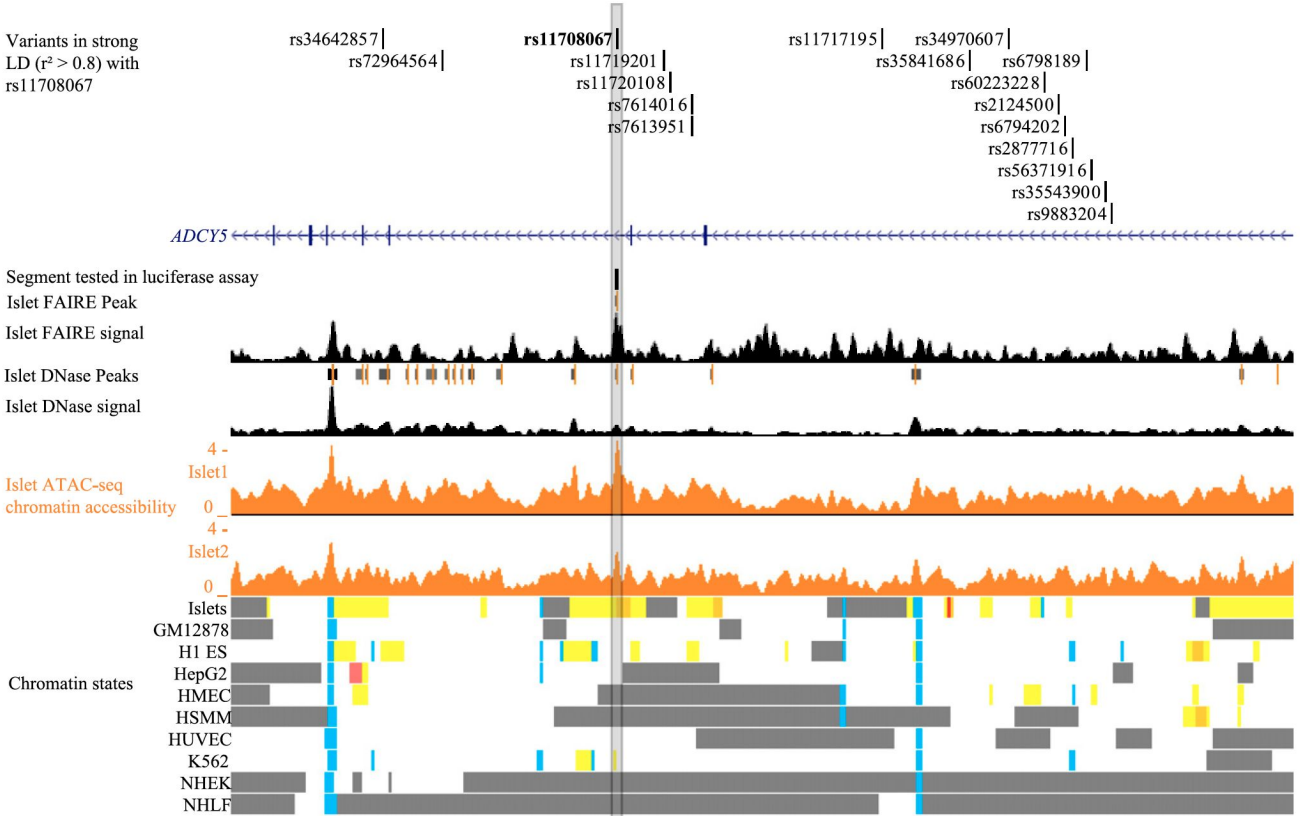
Association with islet ADCY5 expression



Adenylate cyclase 5, catalyzes the generation of cAMP, which drives Ca^{2+} influx into the islet beta cell

Data from Varshney (2017) PNAS 114:2301

Islet regulatory elements overlapping T2D-associated variants at ADCY5

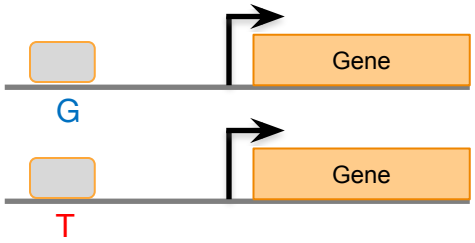


LD, linkage disequilibrium; $r^2 > 0.8$, 1000 Genomes Phase 3 EUR

Roman (2017) Diabetes 66:2521

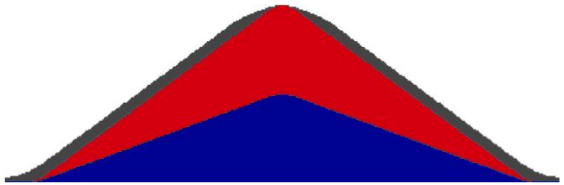
Allelic imbalance suggests allele-specific regulatory activity

Non-specific Protein Binding



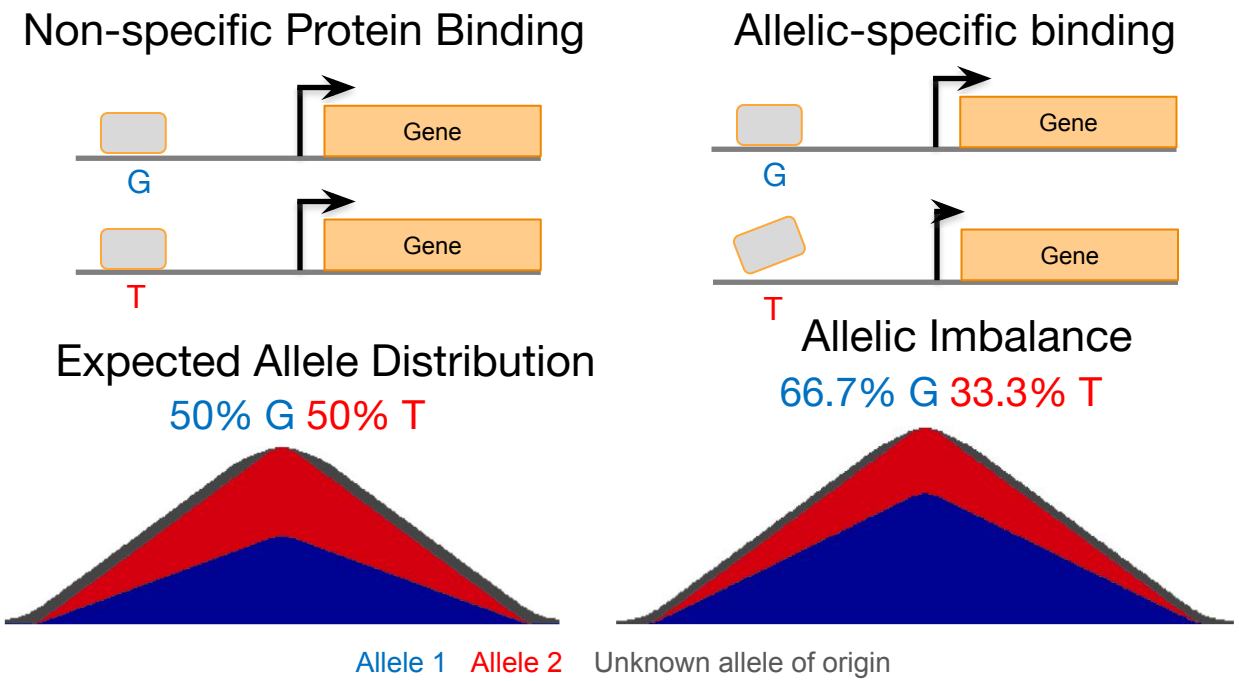
Expected Allele Distribution

50% G 50% T



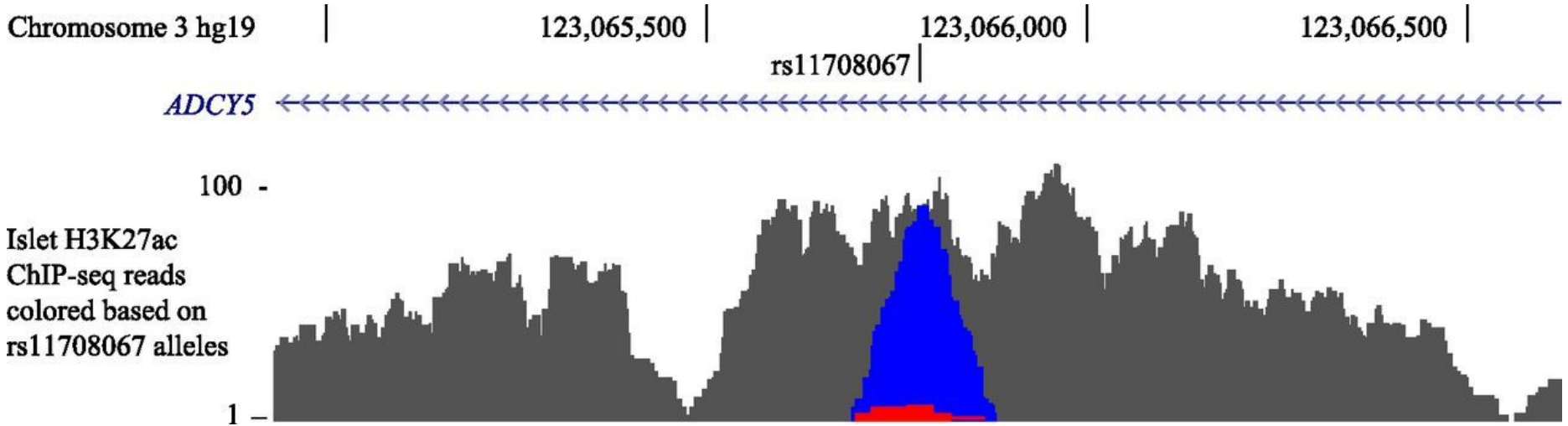
Allele 1 Allele 2 Unknown allele of origin

Allelic imbalance suggests allele-specific regulatory activity



Buchkovich (2015) *BMC Med Gen* 8:43

Islet regulatory elements overlapping T2D-associated variants at ADCY5

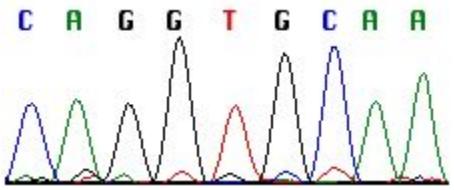


Evidence of allelic imbalance in H3K27ac ChIP-seq reads at rs11708067. H3K27ac ChIP-seq reads in a primary human islet sample heterozygous at rs11708067. Blue indicates reads that contain the G allele of rs11708067, red indicates reads that contain the A allele of rs11708067, and gray indicates reads in the peak that do not overlap rs11708067.

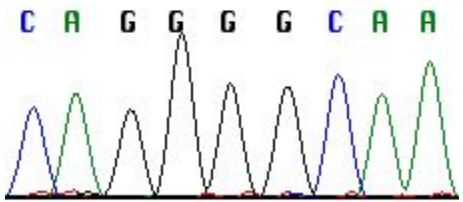
Roman (2017) Diabetes 66:2521

Transcriptional reporter assay

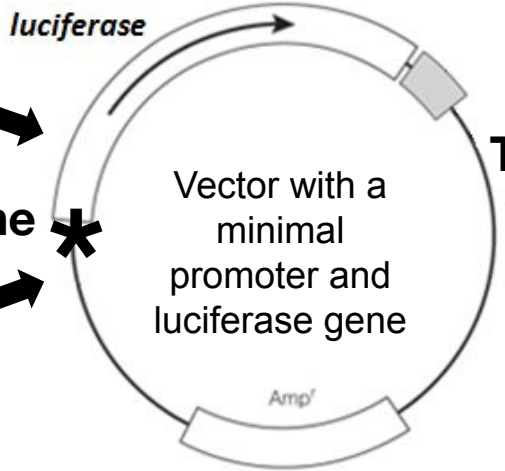
Risk alleles



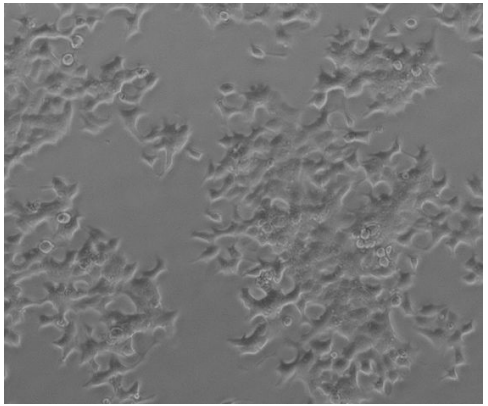
Non-risk alleles



Clone



Transfect

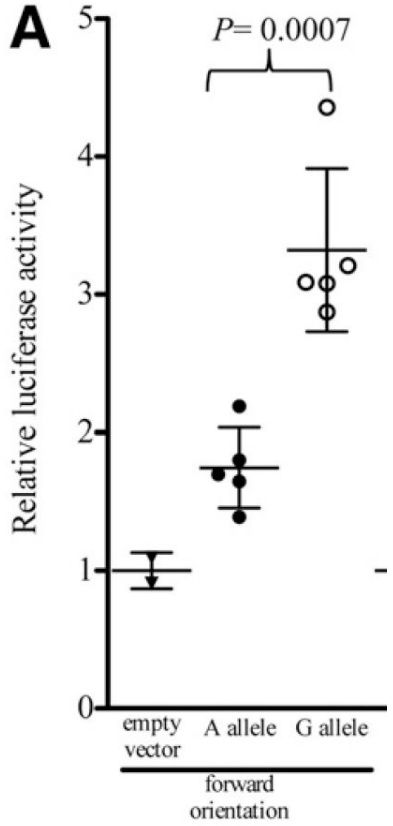


832/13 rat β-cells
MIN6 mouse β-cells

3-5 replicate clones per allele
Forward and reverse orientations

- **Does the cloned element alter transcriptional activity vs empty vector?**
- **Do alleles show significant differences?**

rs11708067 exhibits allelic differences in transcriptional activity



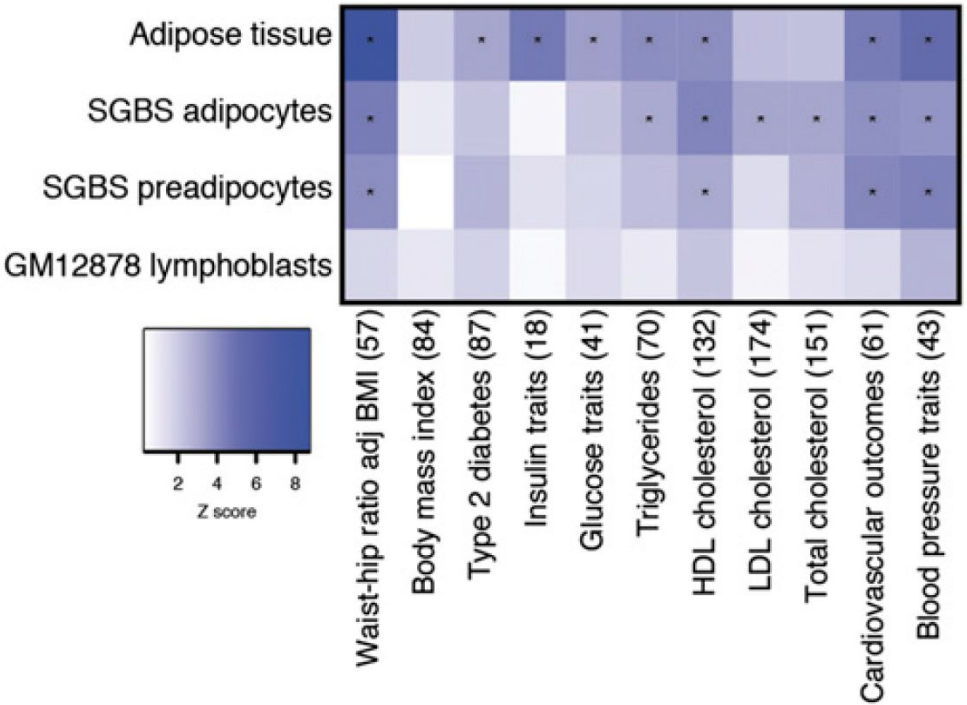
832/13 rat insulinoma cells

231-bp segments containing allele A or G of rs11708067

Direction consistent with islet eQTL

Roman (2017) Diabetes 66:2521

Enrichment of cardiometabolic GWAS variants in ATAC-seq peaks



Top 50,000 representative peaks

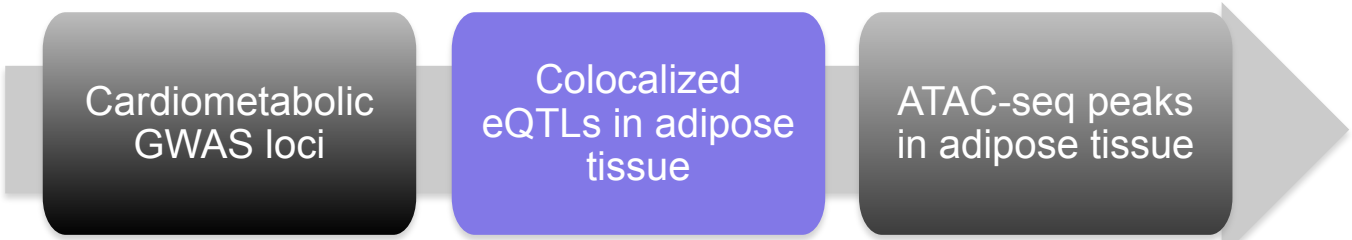
Compare overlap of GWAS variants relative to control variants matched for number of LD proxies, allele frequency, and gene proximity (GREGOR)
Schmidt (2015) *Bioinformatics* 31:2601

Asterisk: enrichment P<.005

Genetic variation in adipose tissue and adipocyte accessible chromatin regions is frequently associated with cardiometabolic traits

Cannon, Currin (2019) *G3* AOP

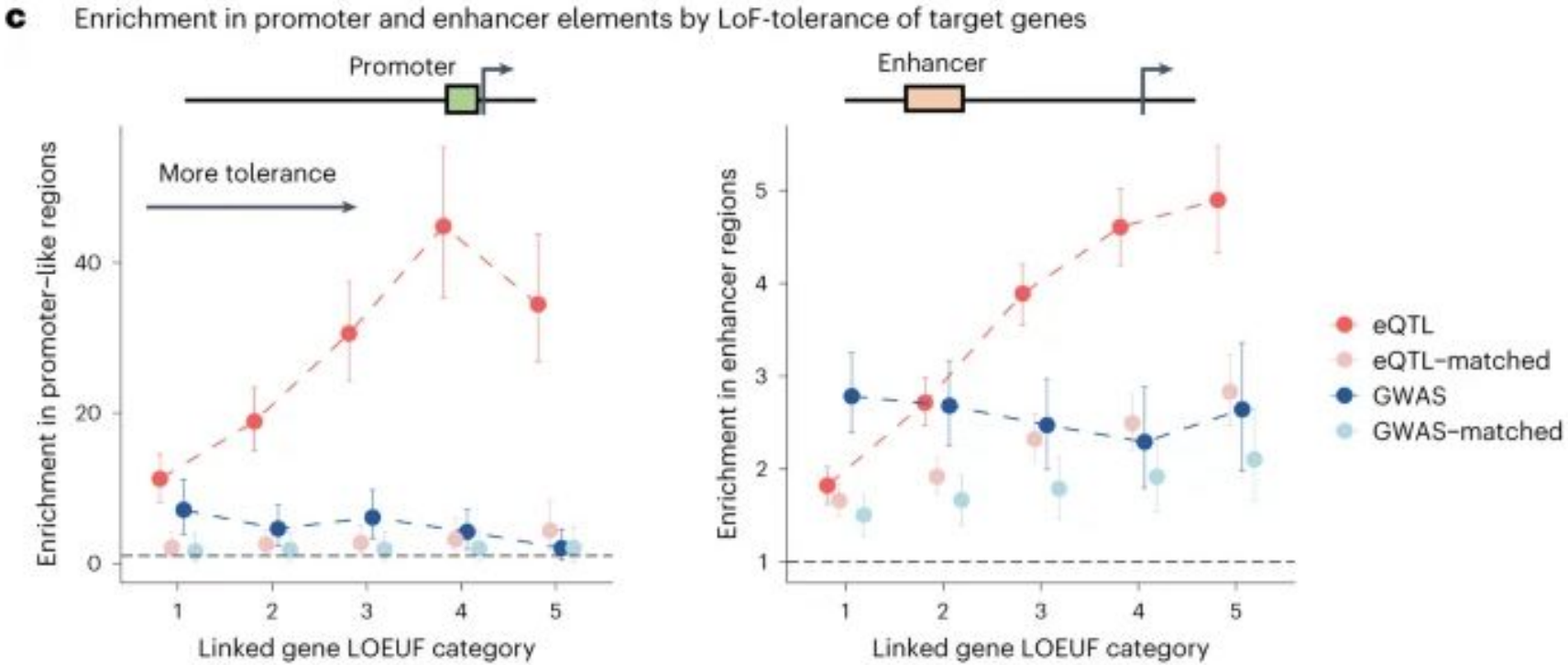
Integrate GWAS, expression, and ATAC-seq data



GWAS and eQTL signals considered colocalized when lead eSNP r^2>0.8 with GWAS SNP and signal was attenuated by reciprocal conditional analysis (Civelek 2017 *AJHG* 100:428)

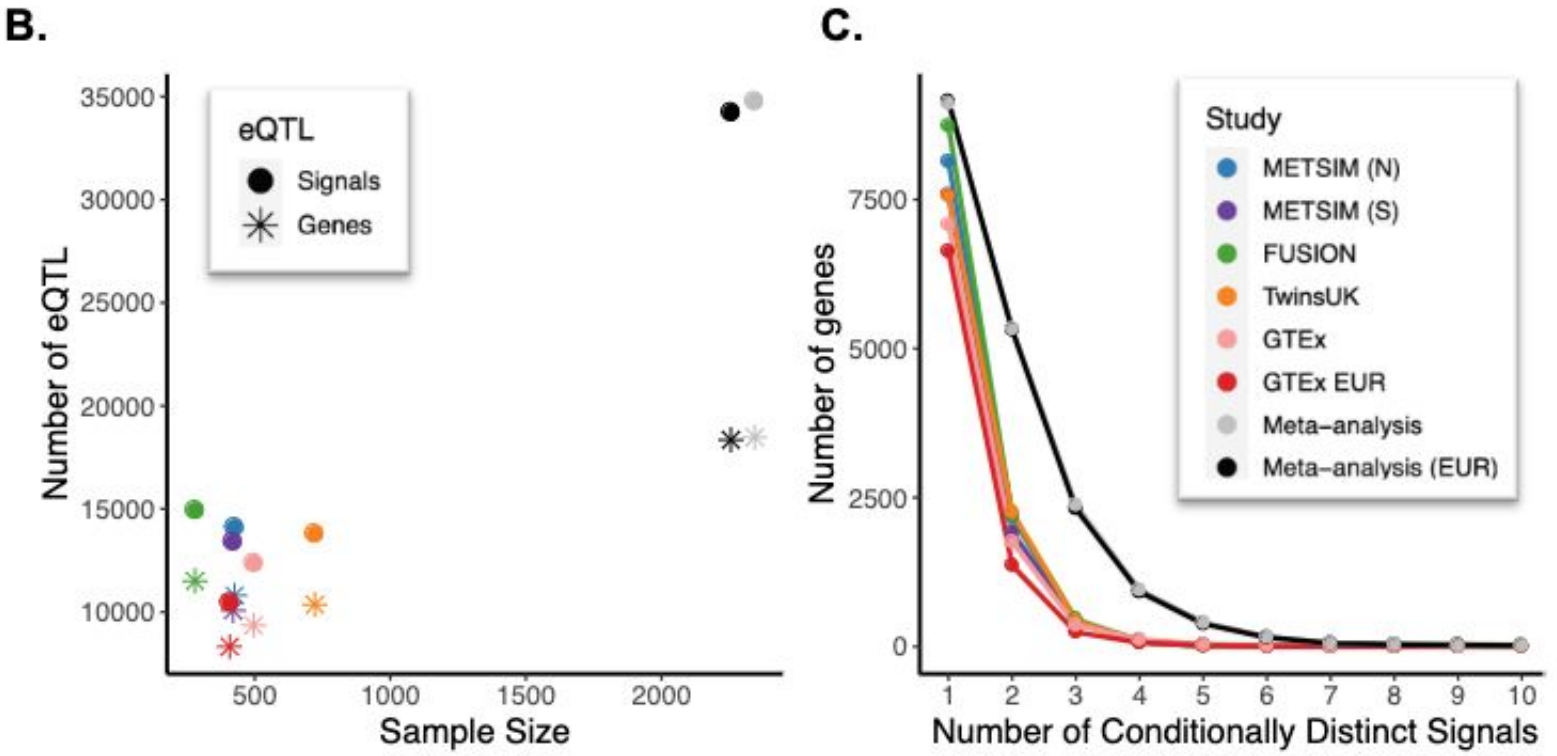
How well does this work? What is the "yield" through the arrow?

Most eQTL studies underpowered to colocalize with GWAS



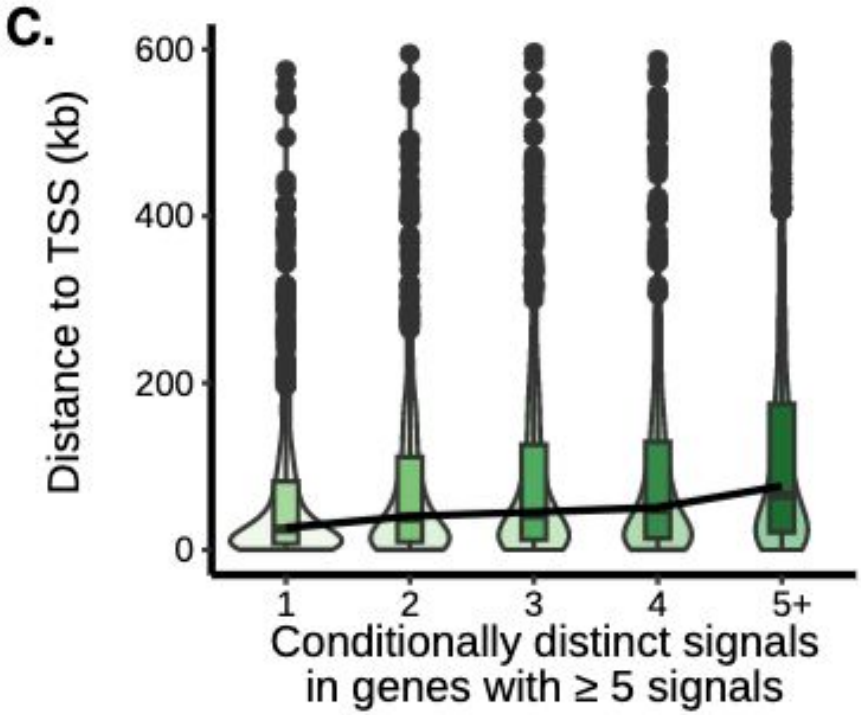
Mostafavi et al (2023)

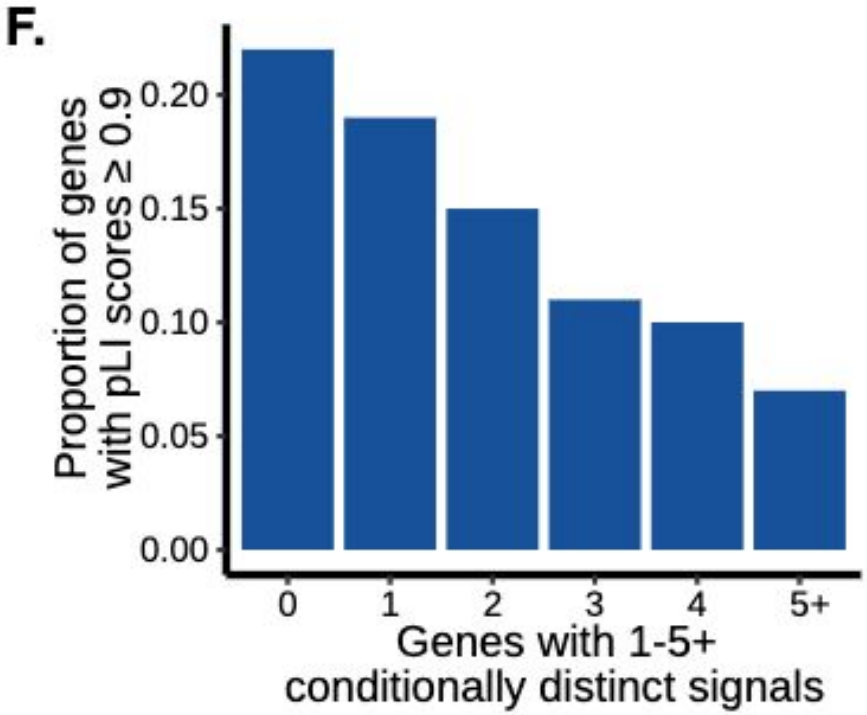
Most eQTL studies underpowered to colocalize with GWAS



Brotman et al (2023)

Most eQTL studies underpowered to colocalize with GWAS



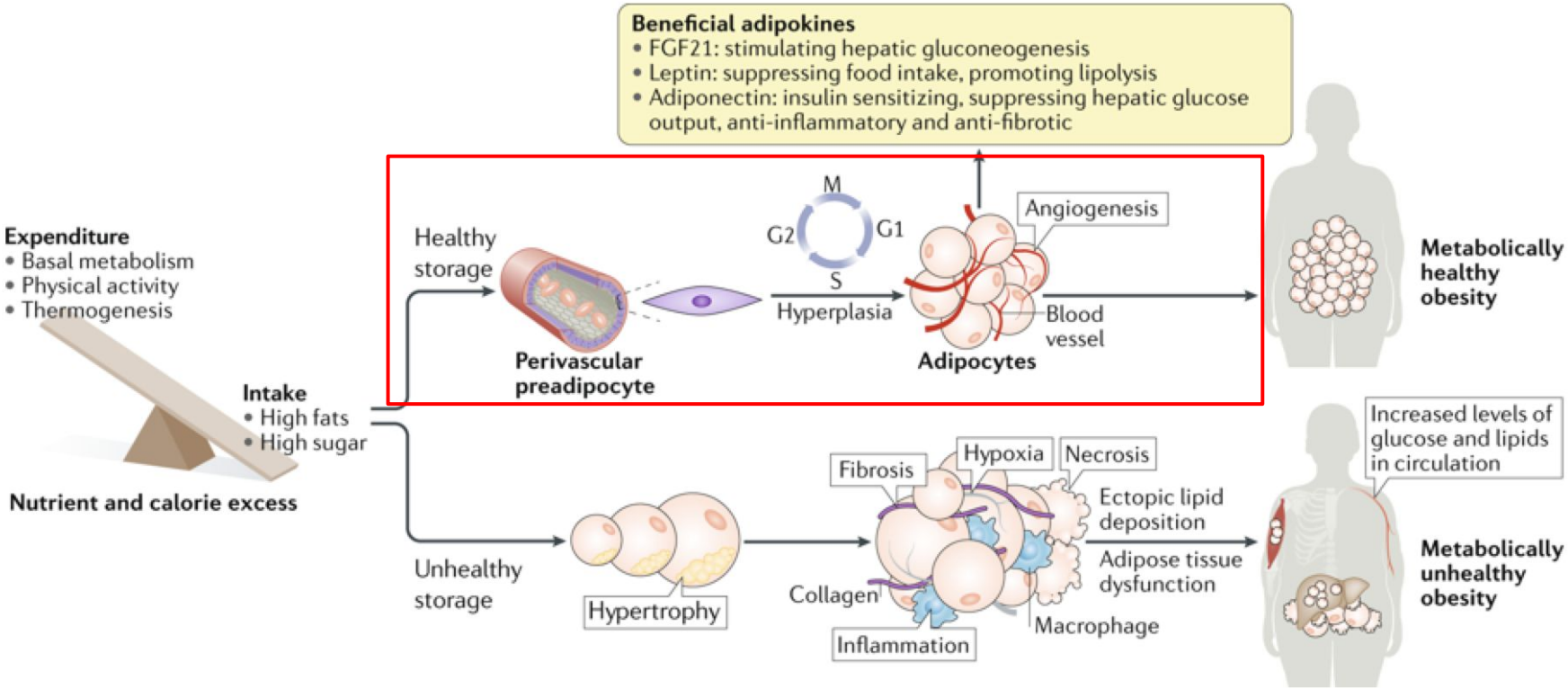


Brotman et al (2023)

ATAC-seq to identify variants and mechanisms for complex metabolic traits

- ATAC-seq method and quality control
- Identify accessible chromatin (regulatory elements)
- GWAS candidate variants located in ATAC-seq-defined elements
- **Cell context/environmental effects on regulatory elements**
- Variants associated with chromatin accessibility (caQTL)

Adipocyte cell context and obesity



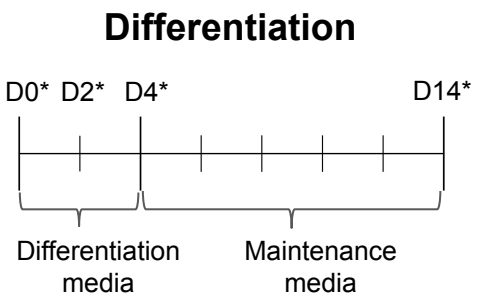
Ghaben et al. Nat Reviews Molecular Cell Bio. 2019.

Identify differential chromatin accessibility regions that affect gene expression due to differentiation

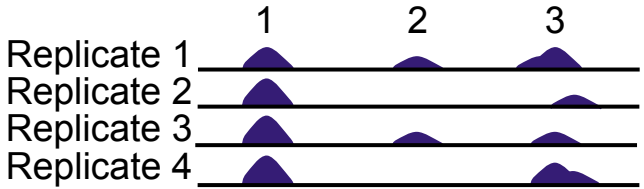
Identify genetic variants in those differentially accessible regions

Strategy to identify differential ATAC-seq peaks

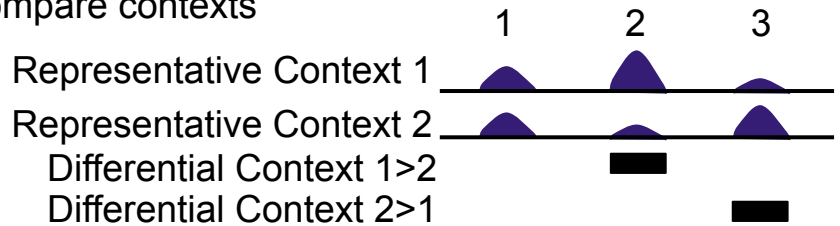
A. Context-specific treatment



B. ATAC-seq mapping for each context

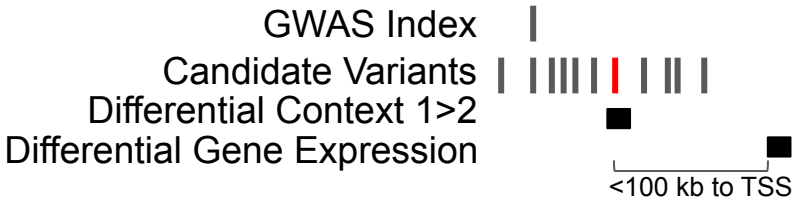


C. Compare contexts



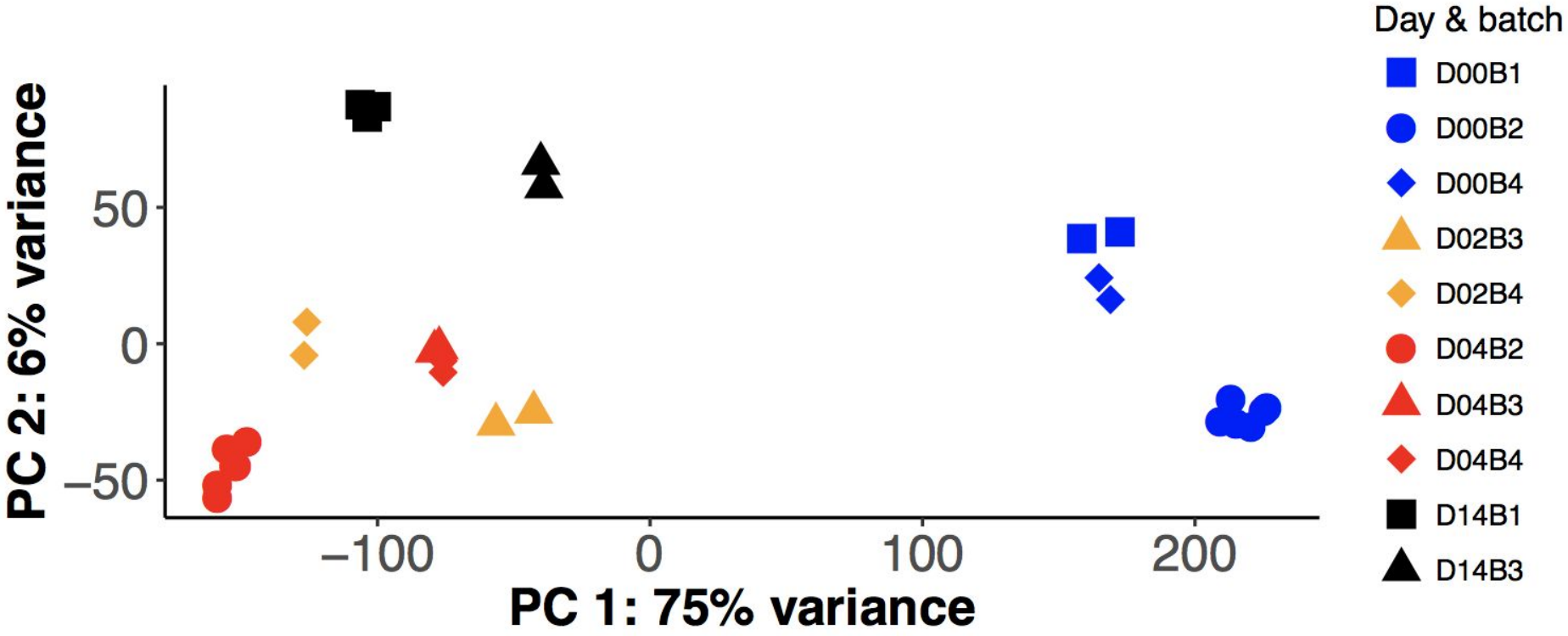
D. Characterize biological process and transcription factor enrichment of representative ATAC-seq and differential gene expression

E. Data Integration to Identify candidate regulatory elements



F. Test alleles for context-dependent effects on transcriptional activity

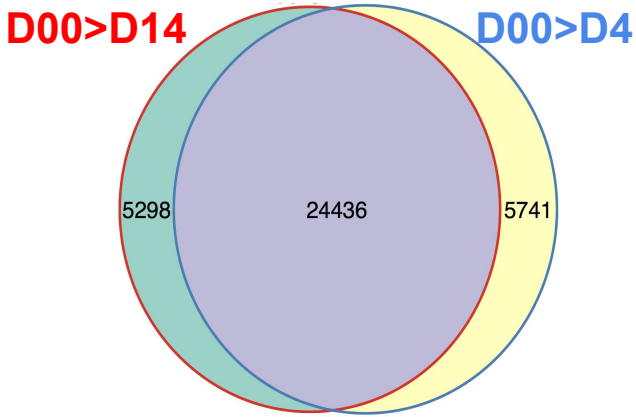
PCA of ATAC-seq data



Counts variance-stabilized & adjusted for library size with DESeq2 before PCA analysis
Big batch effects, and days 2 and 4 do not cluster well
ATAC-seq PC1 accounts for majority of variation, separates by differentiation stage

Regions of context-specific accessible chromatin

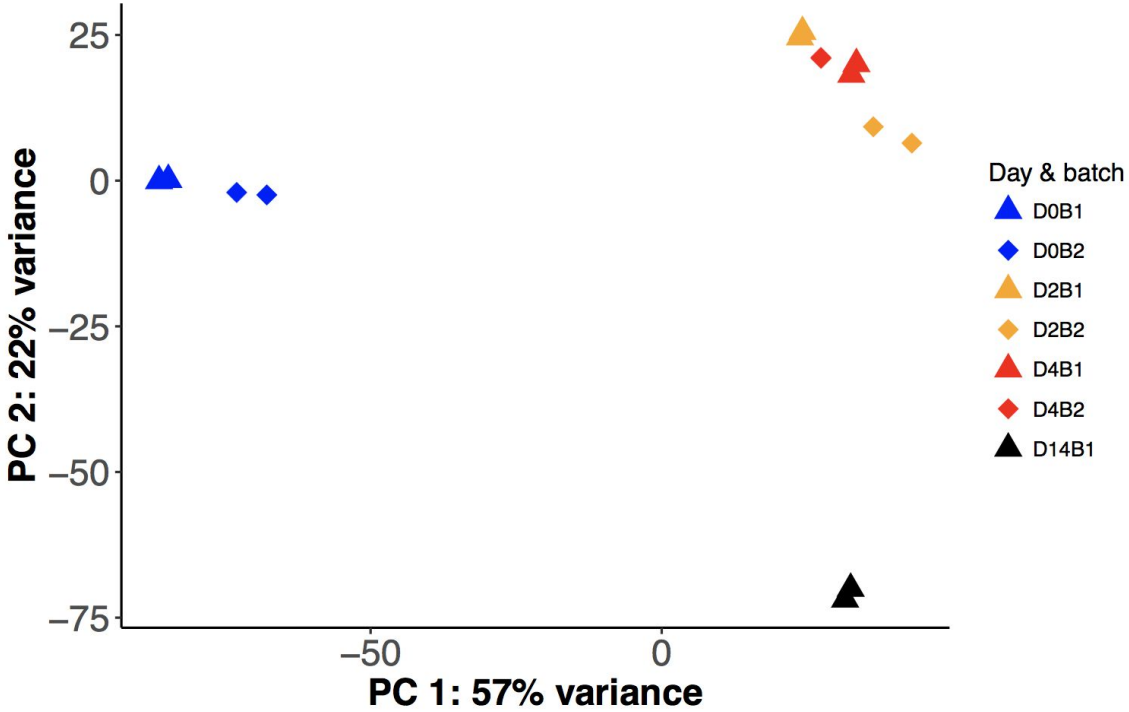
More accessible timepoint:	Compared to:	# Regions
D0	D4	30,177
D0	D14	29,734
D4	D0	30,038
D4	D14	2,661
D14	D0	20,151
D14	D4	480



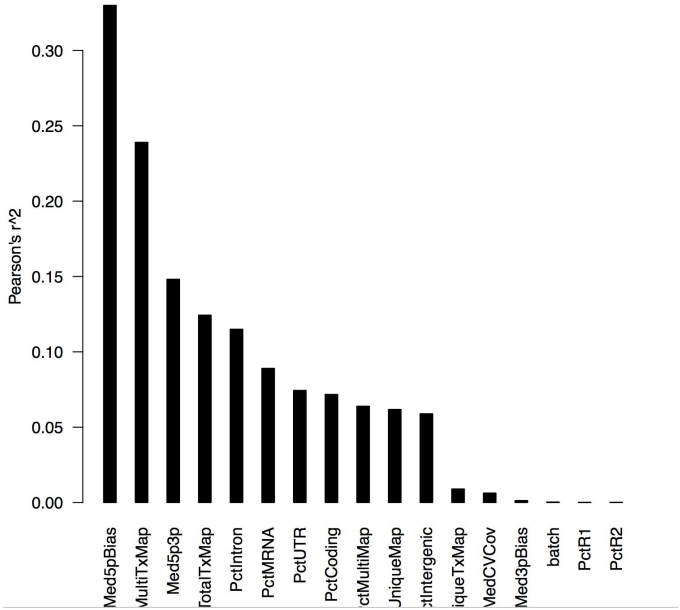
~80% of CA differences between preadipocytes (D0) and different timepoints of differentiated adipocytes (D4 and D14) are shared

DESeq2, Log Fold Change >1, False Discovery Rate <5%

PCA of RNA-seq data



PC1 also correlates with RNA-seq technical factors

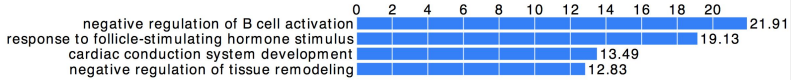


PC1 separates preadipocytes from any days of differentiation
PC2 separates 2 and 4 days from 14 days of differentiation

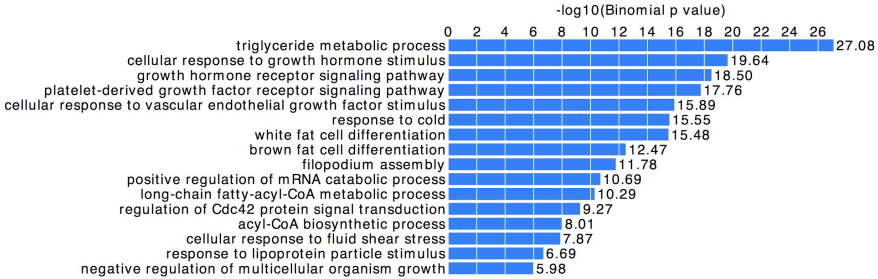
Context-specific ATAC-seq and gene expression are associated with expected biological processes

Chromatin Accessibility

Enriched in D0 (vsD14)



Enriched in D14 (vsD0)



Gene Expression

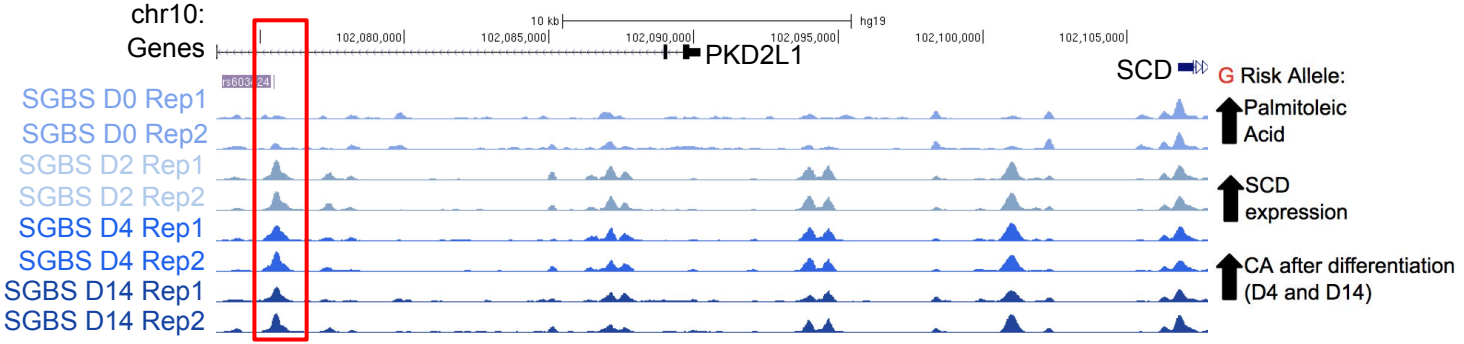
Enriched in D0 (vsD14)

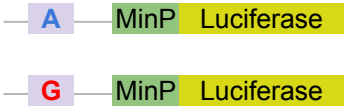
mitotic cell cycle (GO:0000278)	2.4E-16
mitotic cell cycle process (GO:1903047)	2.4E-16
cell cycle (GO:0007049)	3.5E-15
mitotic nuclear division (GO:0140014)	3.7E-15
cellular process (GO:0009987)	3.2E-10
regulation of cellular process (GO:0050794)	9.5E-09

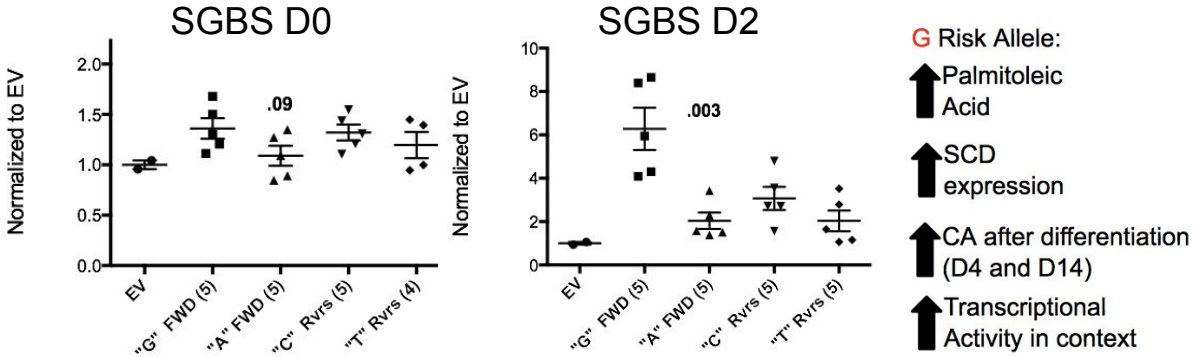
Enriched in D14 (vsD0)

fatty acid metabolic process (GO:0006631)	1.3E-08
lipid metabolic process (GO:0006629)	6.2E-08
response to stimulus (GO:0050896)	3.3E-07
signal transduction (GO:0007165)	1.7E-05
cellular response to insulin stimulus (GO:0032869)	2.1E-05

Context-specific CA to Identify Regulatory Element that Affects Transcriptional Activity

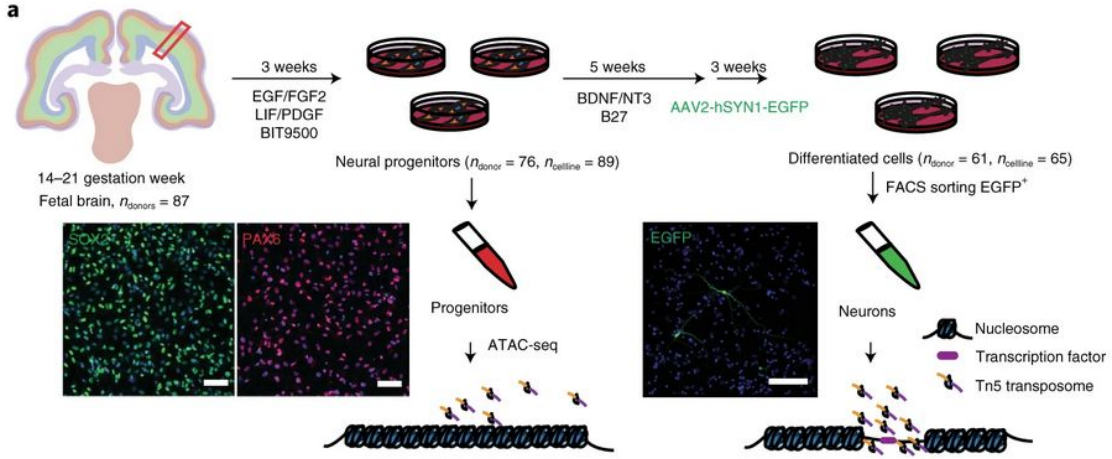


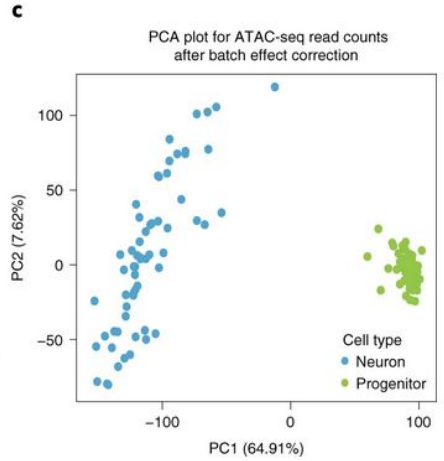


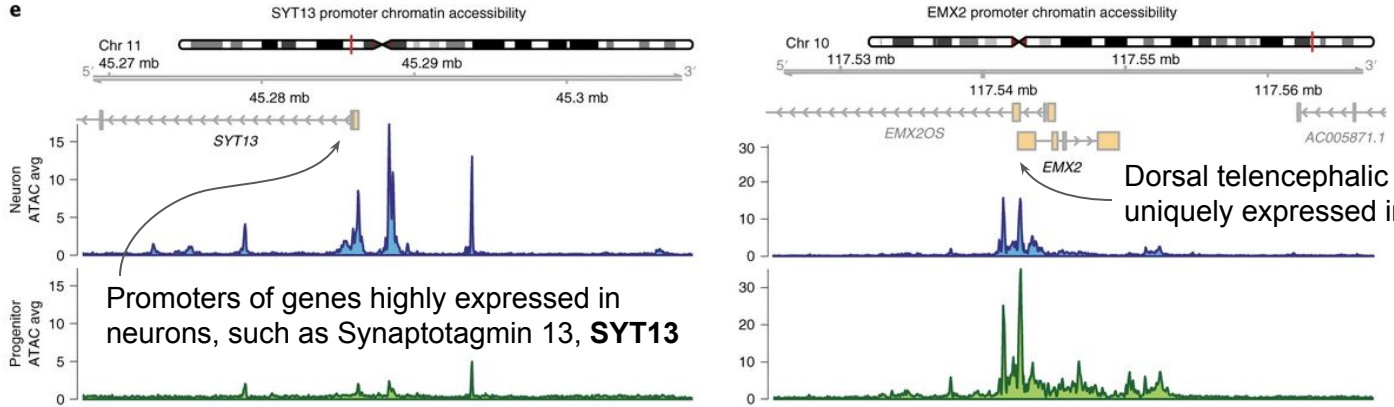


Bars represent 4 independent clones ± SD
pGL4.23 vector, min promoter, 249 bp insert surrounding rs603424
Luciferase assay performed by Swarooparani Vadlamudi

Same approach: brain cell types





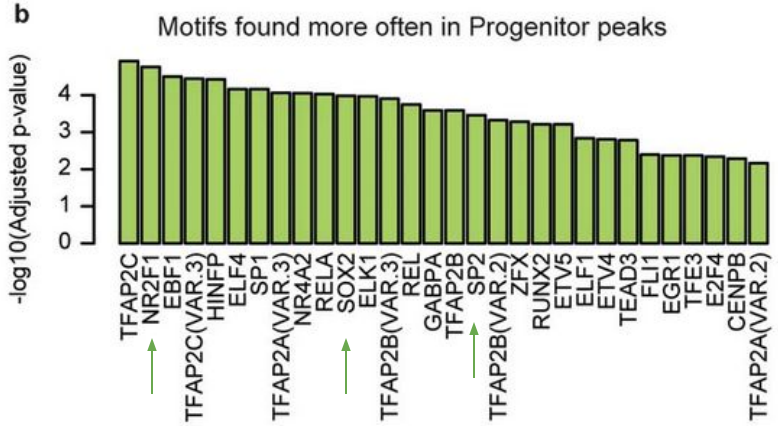


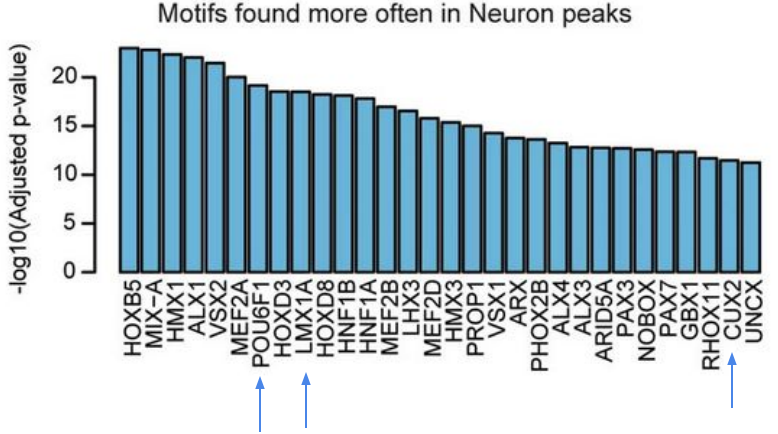
Promoters of genes highly expressed in neurons, such as Synaptotagmin 13, **SYT13**

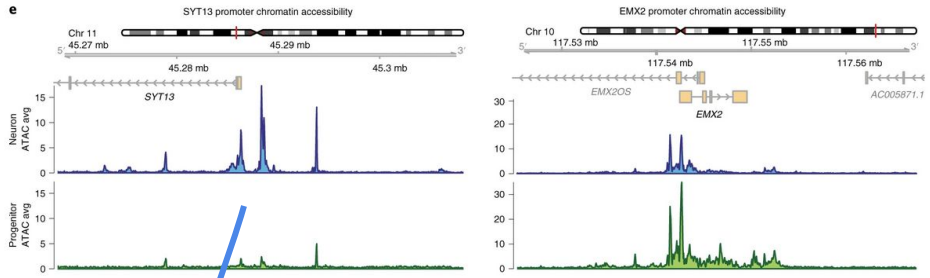
Dorsal telencephalic marker **EMX2**, uniquely expressed in progenitors

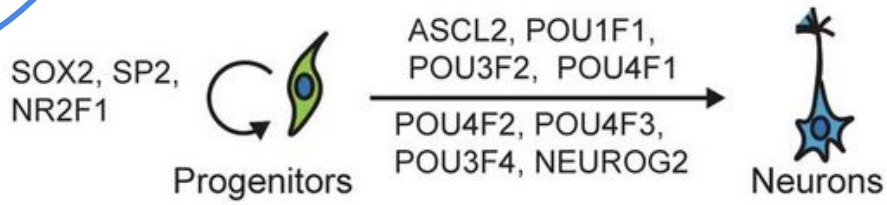
[Liang et al (2021)]

Same approach: brain cell types



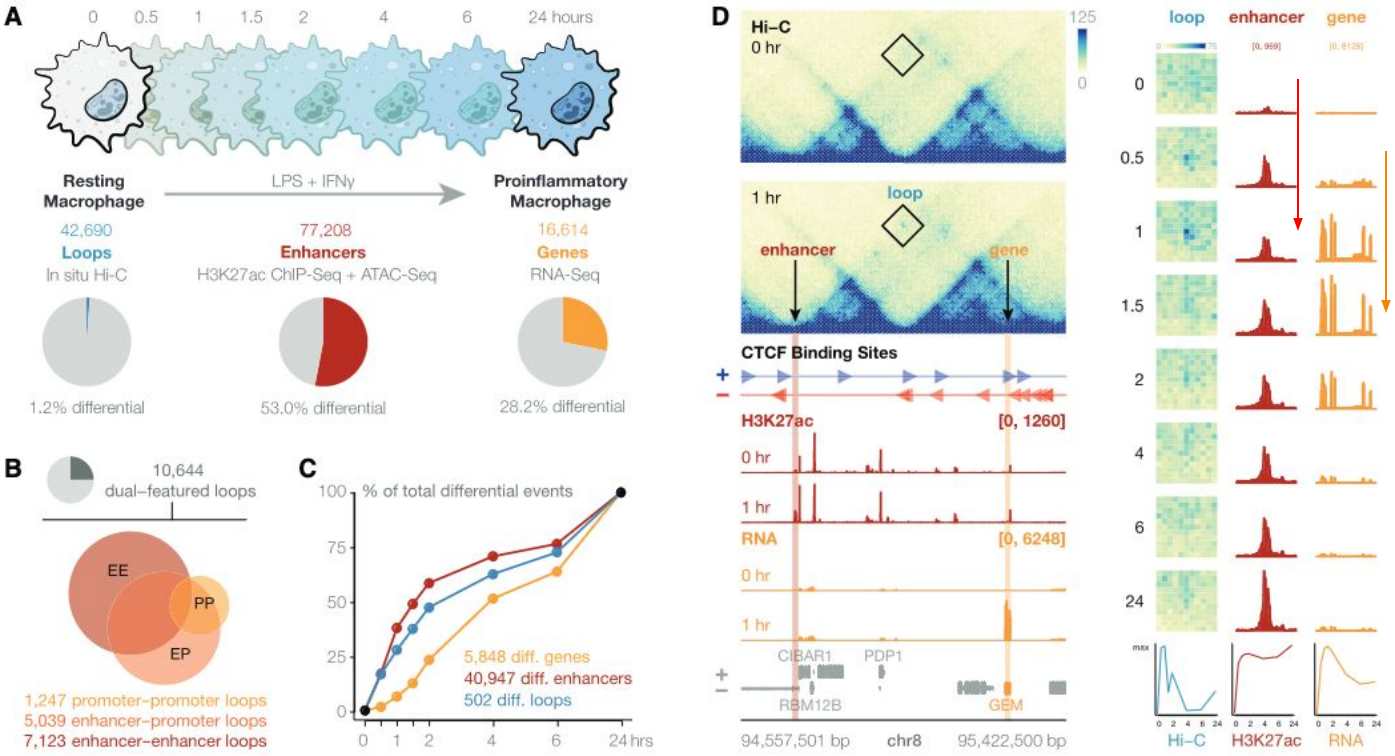






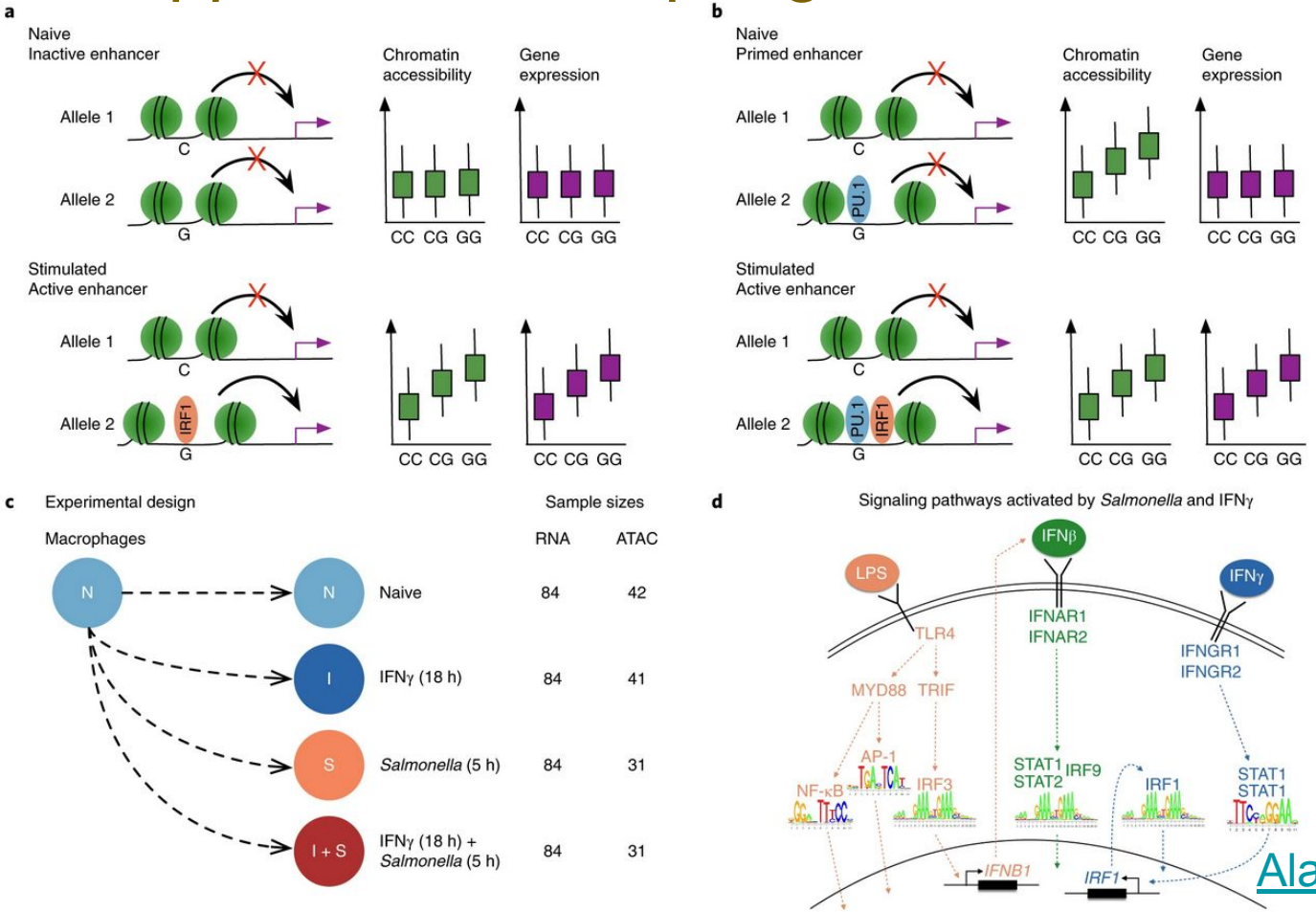
[Liang et al (2021)]

Same approach: macrophage stimulation I



Reed et al (2022)

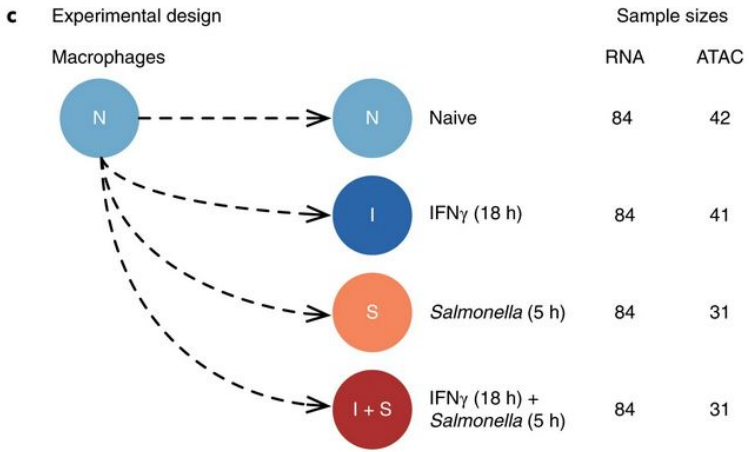
Same approach: macrophage stimulation II



[Alasoo et al (2018)]

Same approach: macrophage stimulation II

Data can be explored in the *fluentGenomics* package, showing use cases for *plyranges* and *tximeta*:



F1000Research
F1000Research 2020, 9:109 Last updated: 27 NOV 2023

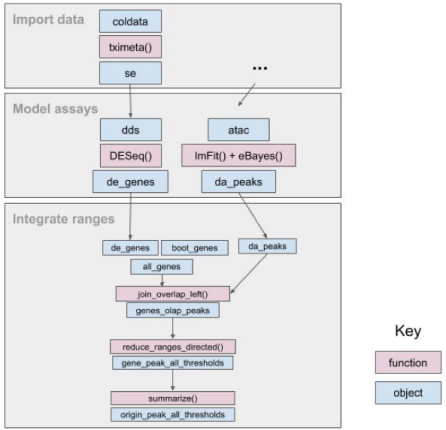
Check for updates

METHOD ARTICLE

Fluent genomics with *plyranges* and *tximeta* [version 1; peer review: 1 approved, 2 approved with reservations]

Stuart Lee[1,2], Michael Lawrence[3], Michael I. Love[4,5]

[1]Epigenetics and Development Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia
[2]Econometrics and Business Statistics, Monash University, Clayton, Victoria, 3800, Australia
[3]Bioinformatics and Computational Biology, Genentech Inc, South San Fransisco, California, 94080, USA
[4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27516, USA
[5]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA



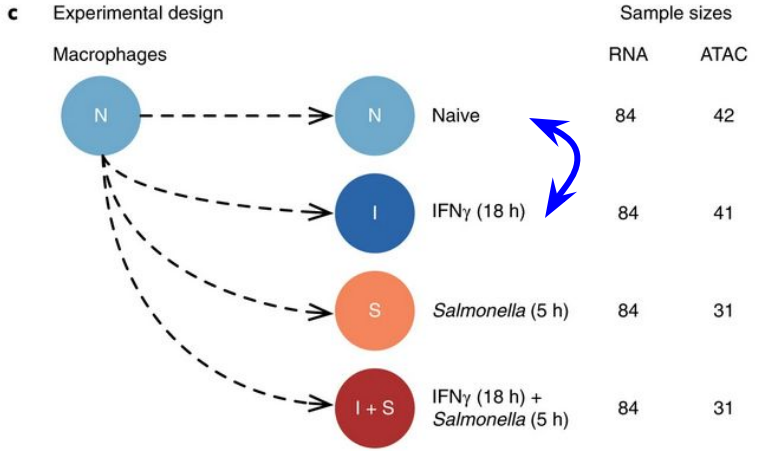
Alasoo et al (2018)

Same approach: macrophage stimulation II

Integration of RNA-seq and ATAC-seq differential results

```
da_peaks <- peaks |>
  filter(da_padj < .01, abs(da_log2FC) > .5)
tss_by_de <- all_genes |>
  mutate(de_sig =
           case_when(
             de_padj <= .01 ~ "de",
             TRUE ~ "non-de"
         )) |>
  filter(!dplyr::between(de_padj, .01, .99)) |>
  anchor_5p() |>
  mutate(width=1)
dist_res <- tss_by_de |>
  add_nearest_distance(da_peaks)
dist_res_clean <- dist_res |>
  as_tibble() |>
  tidyr::drop_na()
dist_res_clean |>
  group_by(de_sig) |>
  summarize(mean = mean(distance), sd = sd(distance))
```

```
# A tibble: 2 × 3
  de_sig  mean     sd
  <chr>  <dbl>  <dbl>
1 de     3173.  5927.
2 non-de 7705. 13266.
```

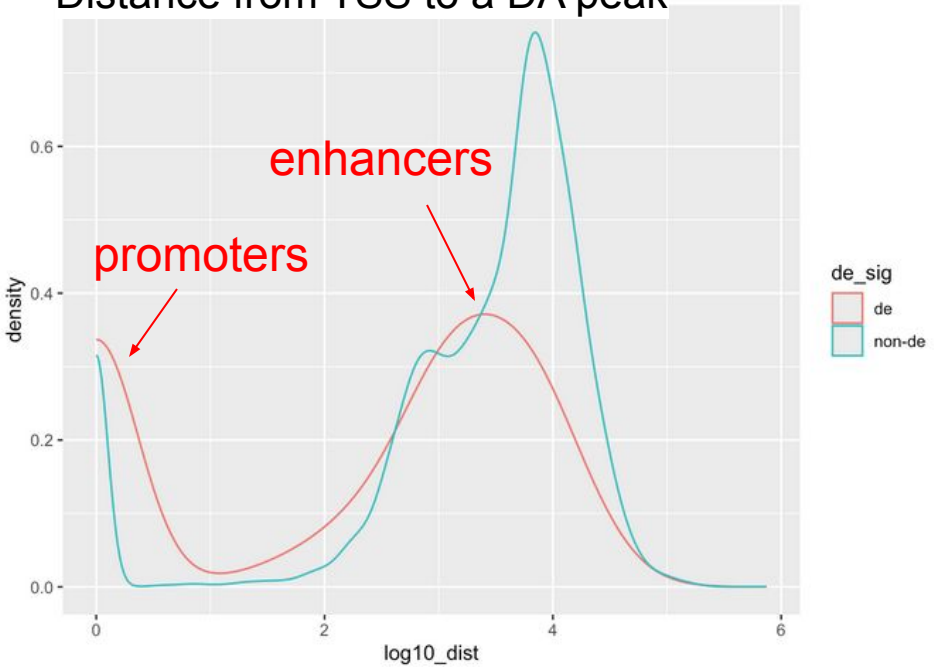


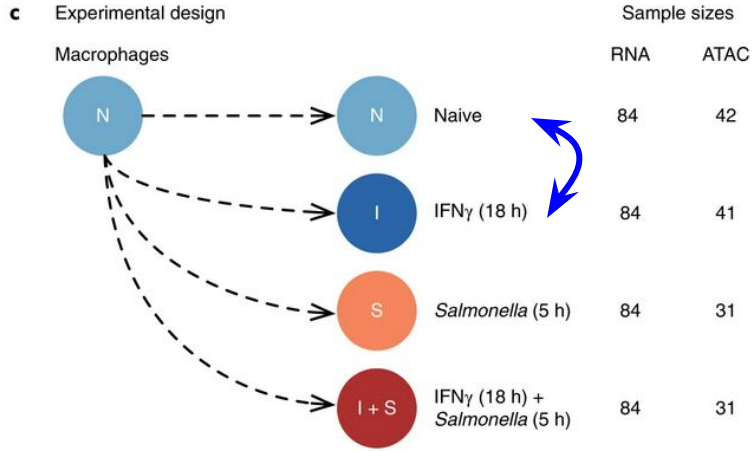
Alasoo et al (2018)

Same approach: macrophage stimulation II

```
library(ggplot2)
dist_res_clean |>
  filter(distance < 1e6) |>
  mutate(log10_dist = log10(distance + 1)) |>
  ggplot(aes(log10_dist, color=de_sig)) +
  geom_density()
```

Distance from TSS to a DA peak





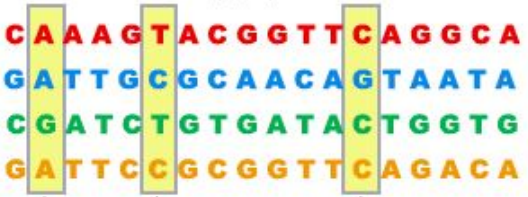
More workflow in the CSAMA GitHub:

`lab/3-thu+fri/lab-98-atac-seq/`

Alasoo et al (2018)

ATAC-seq to identify variants and mechanisms for complex metabolic traits

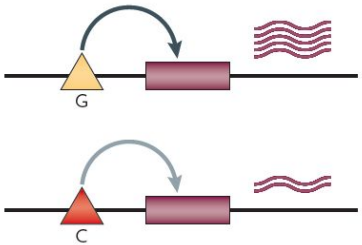
- ATAC-seq method and quality control
- Identify accessible chromatin (regulatory elements)
- GWAS candidate variants located in ATAC-seq-defined elements
- Cell context/environmental effects on regulatory elements
- **Variants associated with chromatin accessibility (caQTL)**

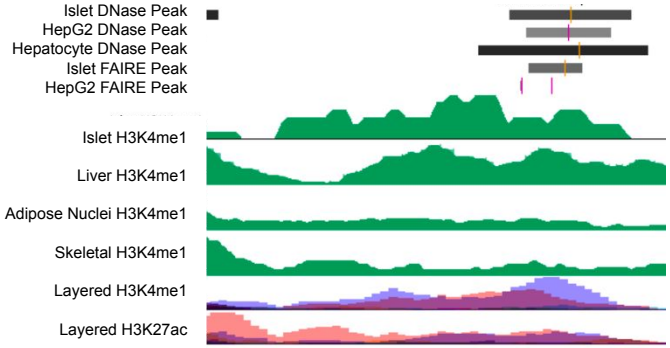


genetic variants ↔ type 2 diabetes

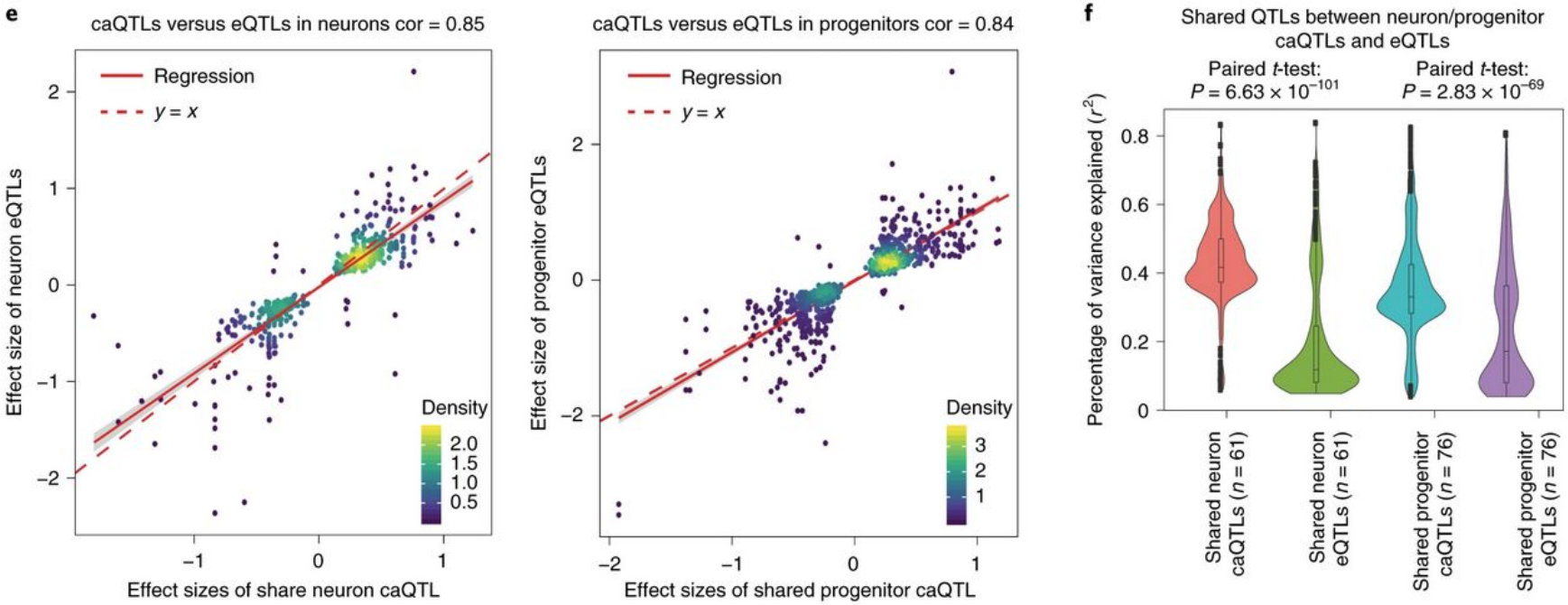
genetic variants ↕ gene expression

genetic variants ↘ chromatin context



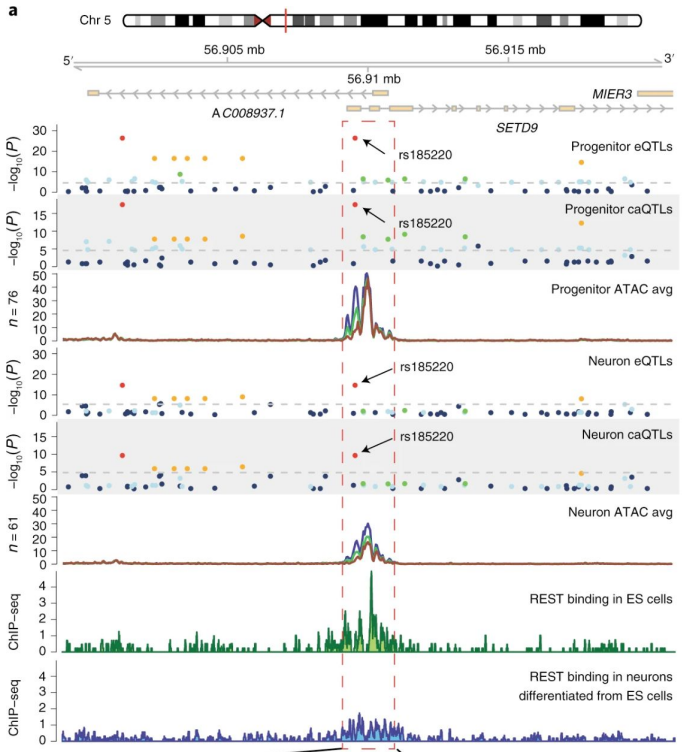


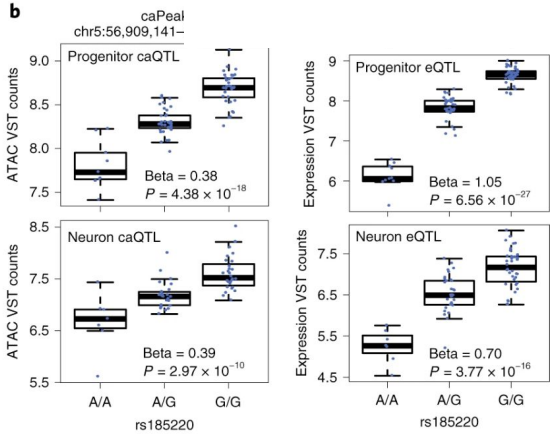
Chromatin QTL in brain cell types

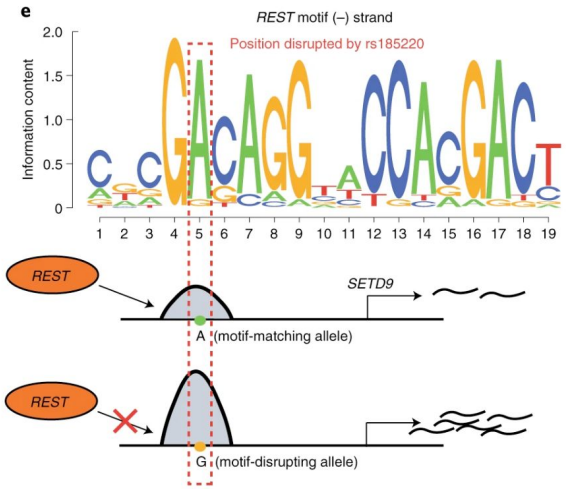


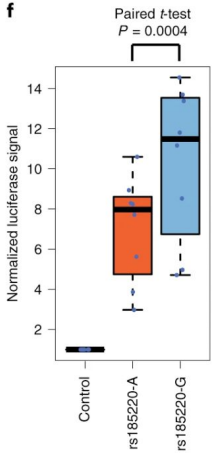
Liang et al (2021)

Chromatin QTL in brain cell types



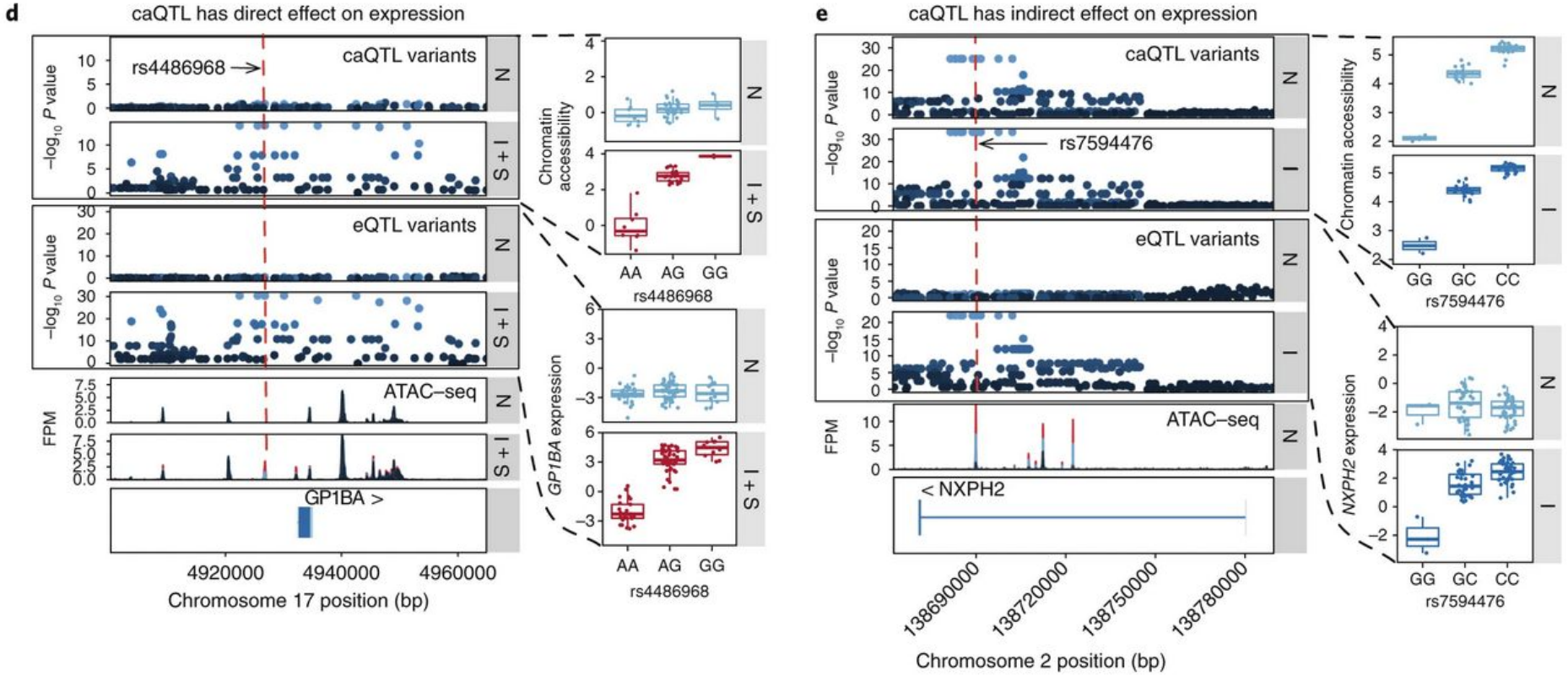






[Liang et al (2021)]

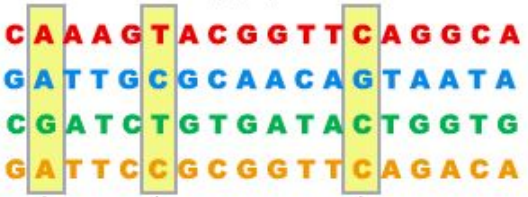
Chromatin QTL in macrophage stimulation II



Alasoo et al (2018)

ATAC-seq to identify variants and mechanisms for complex metabolic traits

- **ATAC-seq method and quality control**
- **Identify accessible chromatin (regulatory elements)**
- **GWAS candidate variants located in ATAC-seq-defined elements**
- **Cell context/environmental effects on regulatory elements**
- **Variants associated with chromatin accessibility (caQTL)**



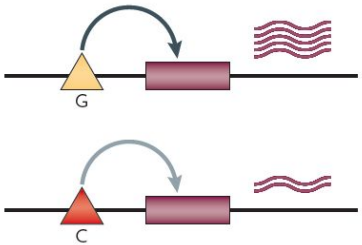
genetic variants ↔ type 2 diabetes

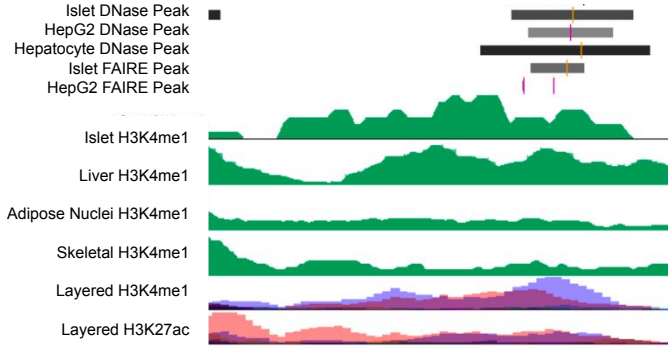
genetic variants ↕ gene expression

genetic variants ↔ chromatin context

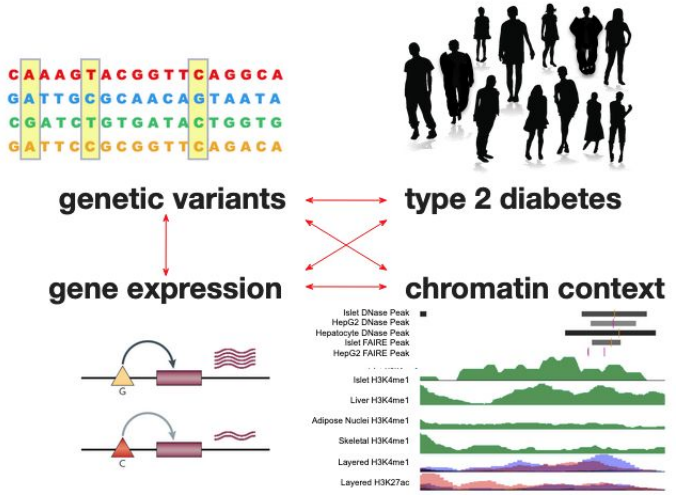
type 2 diabetes ↔ gene expression

type 2 diabetes ↔ chromatin context





Numbers



- What % of GWA signals follow this "canonical pattern"?
  - Variant → TF binding → total expression change
  - Splicing/UTR another non-coding pathway (e.g. causal variant is not missense)
  - Chromatin conformation, CTCF sites
- Can't know with under-powered molecular QTL studies
  - Need high power to detect distal, small effect QTLs, among the right cell types / context

Acknowledgements (metabolic)

UNC Chapel Hill
Martin Buchkovich
Maren Cannon
Kevin Currin
James Davis
Apoorva Iyengar
Hannah Perrin
Chelsea Raulerson
Tamara Roman
Cassandra Spracklen
Swarooparani Vadlamudi
Ying Wu
Terry Furey
Yun Li
Michael Love
Jason Stein

University of Michigan
Stephen Parker
Peter Orchard
Vivek Rai
Arushi Varshney
Michael Boehnke
Christian Fuchsberger
Anne Jackson
Laura Scott
Heather Stringham
Ryan Welch

Duke University
Greg Crawford
Alexias Safi
Lingyun Song

University of Eastern Finland
Markku Laakso
Henna Cederberg
Johanna Kuusisto
Alena Stančáková

NHGRI/NIH
Francis Collins
Lori Bonnycastle
Michael Erdos
Narisu Narisu

University Virginia
Mete Civelek

UCLA
A. Jake Lusis
Paivi Pajukanta