

2019

Investigating Genomic Regulatory Elements in Adipocytes via Computational Analysis

Andrea Bejar
abejar@wellesley.edu

Follow this and additional works at: <https://repository.wellesley.edu/thesiscollection>

Recommended Citation

Bejar, Andrea, "Investigating Genomic Regulatory Elements in Adipocytes via Computational Analysis" (2019). *Honors Thesis Collection*. 627.
<https://repository.wellesley.edu/thesiscollection/627>

This Dissertation/Thesis is brought to you for free and open access by Wellesley College Digital Scholarship and Archive. It has been accepted for inclusion in Honors Thesis Collection by an authorized administrator of Wellesley College Digital Scholarship and Archive. For more information, please contact ir@wellesley.edu.

Investigating Genomic Regulatory Elements in Adipocytes via Computational Analysis

Andrea Mia Bejar

Submitted in Partial Fulfillment
of the
Prerequisite for Honors
in Biological Sciences.

Wellesley College Advisor: Melissa Beers
Beth Israel Deaconess Medical Center Advisor: Linus Tsai

May 2019

© 2019 Andrea Mia Bejar

Acknowledgments

I would like to thank Dr. Linus Tsai and Dr. Evan Rosen for allowing me the opportunity to work in their labs, and for all the support their incredible mentorship has provided me throughout the past two years. You were my first view of what my future could be. I am grateful of the opportunities you have provided me and for giving me the chance to work under such amazing scientists.

I want to thank Rachael Ivison and Christopher Jacobs for tolerating my questions and for always being there to debrief me after difficult lab meetings. Thank you, Rachael, my editor-in-chief, for editing more drafts than I wrote. And thank you Chris, for sharing your knowledge of obscure topics only you would know about. I truly appreciate both of your help throughout this process.

Thank you to Professor Beers for all the patience, expertise and guidance you provided me through three years of research at Wellesley. Thank you for always steering me back on track when I got too buried in the details. Your encouragement kept me going when I was close to giving up.

Thank you to Professor Biller, Professor Klepac-Ceraj, and Professor Carrico-Moniz for serving on my committee. Your feedback was invaluable to me. Your questions were not only challenging but enriched my own understanding of this project.

I would lastly like to thank my incredible mother for teaching me the value and excitement of a STEM education, and for always being the one I call when I need a friend. Your lab coat will always hang proudly in my closet.

The extent of my research would not have been possible without a generous gift from the Samuel and Hilda Levitt Fellowship.

Abstract

Obesity prevalence has more than doubled since the 1980s. One possible consequence of obesity is insulin resistance (IR), a condition underlying type 2 diabetes mellitus (T2DM). So far, genome wide association studies (GWAS) have attributed 18% of heritable risk for T2DM to genetic variants, but one shortcoming of GWAS is knowing which genes are affected by identified variants. This study aimed to confront this weakness and investigate how epigenetic regulation affects metabolic phenotype. Three histone marks (H3K27ac, H3K4me1, H3K4me3) were targeted by chromatin immunoprecipitation for *in vivo* samples collected from human subcutaneous adipocytes and metabolically-relevant tissue samples curated from the ENCODE database. We developed the Extremity analysis method to identify enhancers and promoters enriched in histone ChIP-Seq data. Additionally, a full suite of well-established bioinformatics methods were employed, including differential enrichment analysis (DEA) and motif enrichment analysis (MEA), and existing obesity related GWAS were incorporated. Close correlation was found between Extremity and DEA, MEA provided enriched motifs that bind known adipogenic TFs, and the GWAS showed variants in T2DM and WHRadjBMI overlapping H3K4me3 have less heritability in adipocytes than the other two marks. This study provides a new method and potential targets for further understanding epigenetic variation and its effect on metabolic phenotype.

Table of Contents

Introduction	5
<i>Insulin Function and Insulin Resistance</i>	<i>6</i>
<i>Overview of Adipocyte Biology and the Pathogenesis of Insulin Resistance</i>	<i>8</i>
<i>Adipose Distribution and Metabolic Risk.....</i>	<i>10</i>
<i>Genome Wide Association Studies (GWAS)</i>	<i>11</i>
<i>Identifying Genetic Regulatory Elements using Histone Modification ChIP-Seq</i>	<i>13</i>
<i>Experimental Aims</i>	<i>15</i>
Methods.....	18
<i>Sample Collection and Selection</i>	<i>18</i>
<i>Human Adipocyte Sample Preparation</i>	<i>19</i>
<i>ENCODE Samples.....</i>	<i>19</i>
<i>Adipocyte Peak Calling and Filtering</i>	<i>19</i>
<i>Differential Enrichment Analysis</i>	<i>20</i>
<i>Extremity Analysis.....</i>	<i>21</i>
<i>Motif Enrichment Analysis</i>	<i>21</i>
<i>GWAS: Linkage Disequilibrium Score Regression and Partitioning Heritability.....</i>	<i>22</i>
<i>Computational Analysis</i>	<i>23</i>
Results	23
<i>Promoter and Enhancer Enrichment do not Vary Largely Between Adipocyte Samples</i>	<i>24</i>
<i>Differential Enrichment Analysis between Adipocytes and Non-adipocytes</i>	<i>28</i>
<i>Adipocyte Extremity Shows Correlation with Fold-Change from DEA</i>	<i>29</i>
<i>Comparison of Adipocyte vs Non-adipocyte DEA and IR vs IS DEA.....</i>	<i>31</i>
<i>Motif Enrichment Analysis of Peaks Selected from Adipocyte Extremity.....</i>	<i>33</i>
<i>Heritability for GWAS Traits</i>	<i>37</i>
Discussion.....	39
Supplemental Data	44
References	105

Introduction

The prevalence of obesity has increased dramatically within the last three decades. Roughly 39.8% of adults over 20 in the U.S. are estimated to be obese¹, up from about 15% in 1980 (Figure 1)². Obesity is linked to the development of ailments including metabolic syndromes, coronary artery disease, impotency and infertility, and cancers³⁻⁴. Type 2 diabetes mellitus (T2DM) is often discussed alongside obesity because 87.5% of new diabetics are overweight or obese⁵. T2DM is the seventh leading cause of death in the United States, and 1.6 million new patients are diagnosed each year⁵. Thus, the need to understand the underlying mechanisms of obesity and its associated conditions is increasingly pertinent.

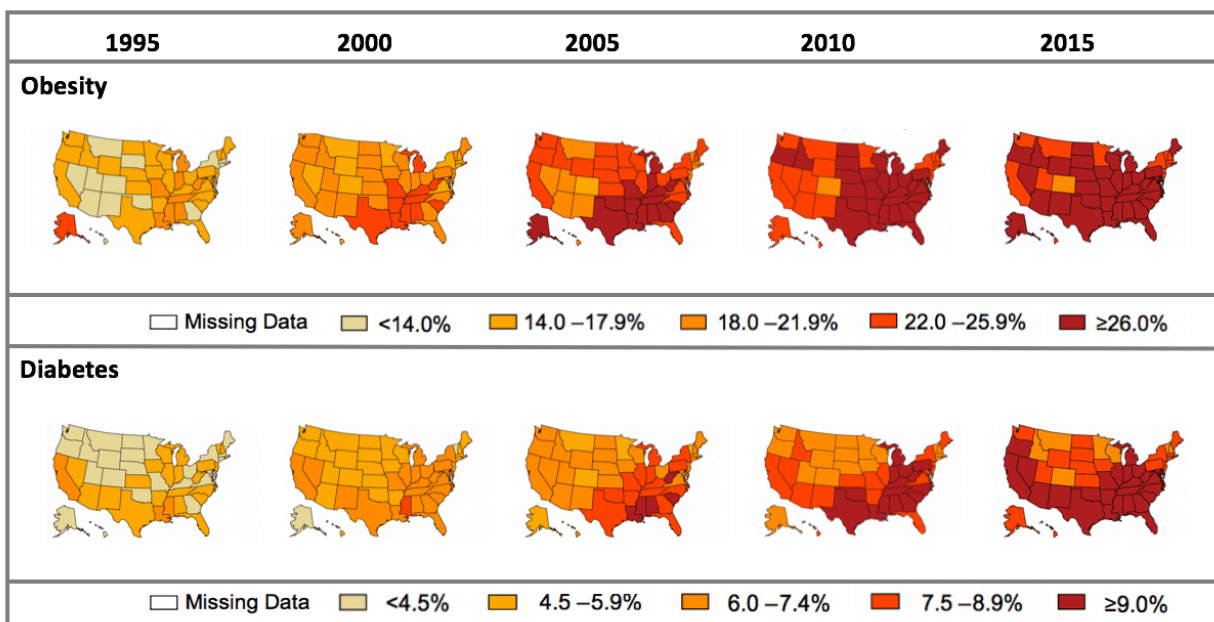


Figure 1: Prevalence of obesity and diabetes in the U.S. from 1995 to 2015⁶. Diabetes prevalence has increased with the rise in obesity.

Overnutrition and increasingly sedentary lifestyles are largely blamed for the rise in global rates of obesity. The first line of treatment for obesity is diet and exercise in order to lose weight. It is clear, however, that an overwhelming majority of diets and weight loss programs

fail to promote sustainable weight loss. Diet failure along with stigma and pressure from society and medical professionals can lead to psychological trauma and eating disorders in patients, exacerbating the associated health risks of obesity⁷⁻⁹.

Identifying other possible causes for the severe increase in obesity rates has been a focus of many studies. While much of the blame is attributed to high-calorie Western diets and inactive lifestyles, it is well known that obesity and metabolic disease are heritable¹⁰⁻¹². With great advancements in the field of genetics, researchers have investigated genetic variants in obese and diabetic patients. A study of genetic variants between T2DM patients and a non-diabetic control group found that 243 regions of the genome were associated with T2DM risk¹³. Yet, these variants account for only 18% of heritable T2DM risk, so this increase in obesity and T2DM cannot be explained by genetics alone¹³. Since the current rising trend in obesity and T2DM has been observed within the last four decades, there has been hardly enough time for microevolution, or changes in allele frequency within a species, which can take approximately 2-100 generations to occur¹⁴. Genetic studies thus suggest a potential epigenetic regulatory factor contributing to the heritability of metabolic disease.

Insulin Function and Insulin Resistance

Insulin is a peptide hormone that is responsible for the regulation of carbohydrate and lipid metabolism, and protein synthesis (Figure 2)¹⁵. This study will focus on the role of glucose uptake and the control of circulating free fatty acid (FFA) concentrations. Insulin is vital for the regulation of blood glucose levels. Blood glucose concentration increases after meals, leading to the secretion of insulin into the blood by pancreatic beta-cells. Insulin then binds to the insulin receptor and signals adipocytes and skeletal muscle cells to take up glucose and store it as

intracellular triglycerides and glycogen¹⁶. Additionally, insulin can signal the liver to promote gluconeogenesis and glycogenolysis (the production and release of glucose, respectively) to increase blood glucose levels¹⁶. Excess calories are stored as lipid droplets in adipocytes. When needed, adipocytes release these calories as FFAs via lipolysis to be used by tissues for energy¹⁷. Insulin regulates the concentration of circulating FFAs by slowing down lipolysis in adipocytes.

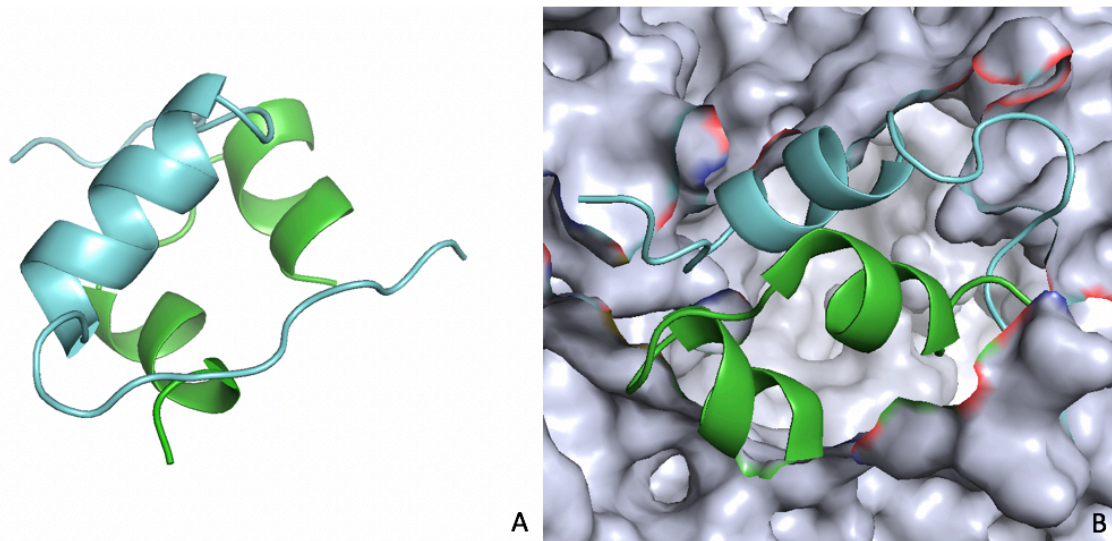


Figure 2: Insulin is a peptide hormone that consists of two heterodimers. The structure of a single insulin molecule is shown in panel A (PDB: 2HIU¹⁸). Insulin binding to the insulin receptor (grey), a receptor tyrosine kinase, is shown in panel B (PDB: 6HN5¹⁹). Individual dimers shown in blue and green. Structures were visualized using PyMol²⁰.

There can be severe health consequences when insulin fails to produce the appropriate response in a target tissue. This is known as insulin resistance (IR). IR is central to the development of metabolic disease because it is the cause of the elevated concentrations of circulating lipids (hyperlipidemia) and glucose (hyperglycemia) observed in T2DM. Understanding of IR, however, remains limited despite its importance in metabolic disease. It has been shown that IR can develop without changes in insulin signaling²¹⁻²². This lack of understanding is further highlighted by the available treatments for IR. Most treatments to

improve insulin sensitivity either lower carbohydrate absorption or affect insulin secretion in beta-cells. There have also been concerns about toxicity directed toward the two classes of drugs available to improve insulin sensitivity, biguanide metformin and thiazolidinediones (TZDs). Furthermore, many of these treatments promote weight gain and thereby worsen IR²³.

Overview of Adipocyte Biology and the Pathogenesis of Insulin Resistance

Adipose was classified as connective tissue until the 1940s¹⁷. It has become apparent that adipose is an incredibly complex organ that plays an important role in many physiological functions including energy homeostasis, reproduction, immune response, and blood pressure control (Figure 3)^{17, 24-25}. As with most organs, there are many cell types in adipose. This study will focus on the adipocytes, which can be predominantly placed in two categories: white or brown. Brown adipocytes are highly specialized, thermogenic cells that are unique to eutherian (placental) mammals²⁶. When people think of fat, they are often thinking about white adipocytes. The primary function of white adipocytes is to store lipids in a single-chambered droplet, but their physiological functions can be grouped into three categories: lipid metabolism, glucose metabolism, and endocrine function²⁷⁻²⁸. Consequently, adipocyte function, and therefore adipose function, is closely tied to the development of IR.

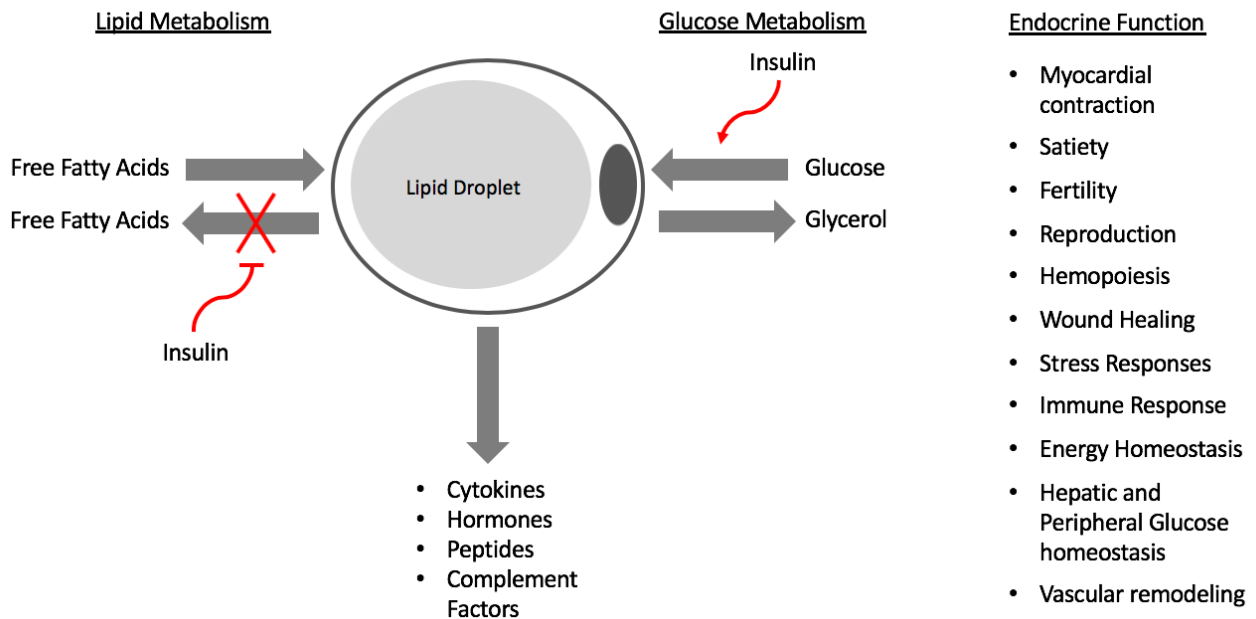


Figure 3: Adipocytes have a wide range of endocrine function. Insulin regulates glucose uptake and controls the release of free fatty acids from the lipid droplet via inhibition of lipolysis. Adapted from Morrison et. al, 2000²⁸.

As adipocytes take in more lipid, their size expands. Simultaneously, adipocytes secrete matrix proteins to maintain the broader structure of the fat depot. In turn, this matrix limits the adipocytes' capacity to expand. Overexpansion leads to hyperlipidemia (excess circulating lipid concentration) because the adipocytes can no longer take in more FFAs. FFAs are then taken up by other tissues, such as skeletal muscle, where their metabolites inhibit insulin signaling²⁹. Overexpansion of adipocytes can also lead to hypoxia and inflammation of the tissue, a common characteristic of obesity¹⁷. This chronic, low-grade inflammation leads to a release of inflammatory cytokines that inhibits insulin signaling downstream of the insulin receptor³⁰. Obesity-induced inflammation thus leads to IR and T2DM (Figure 4).

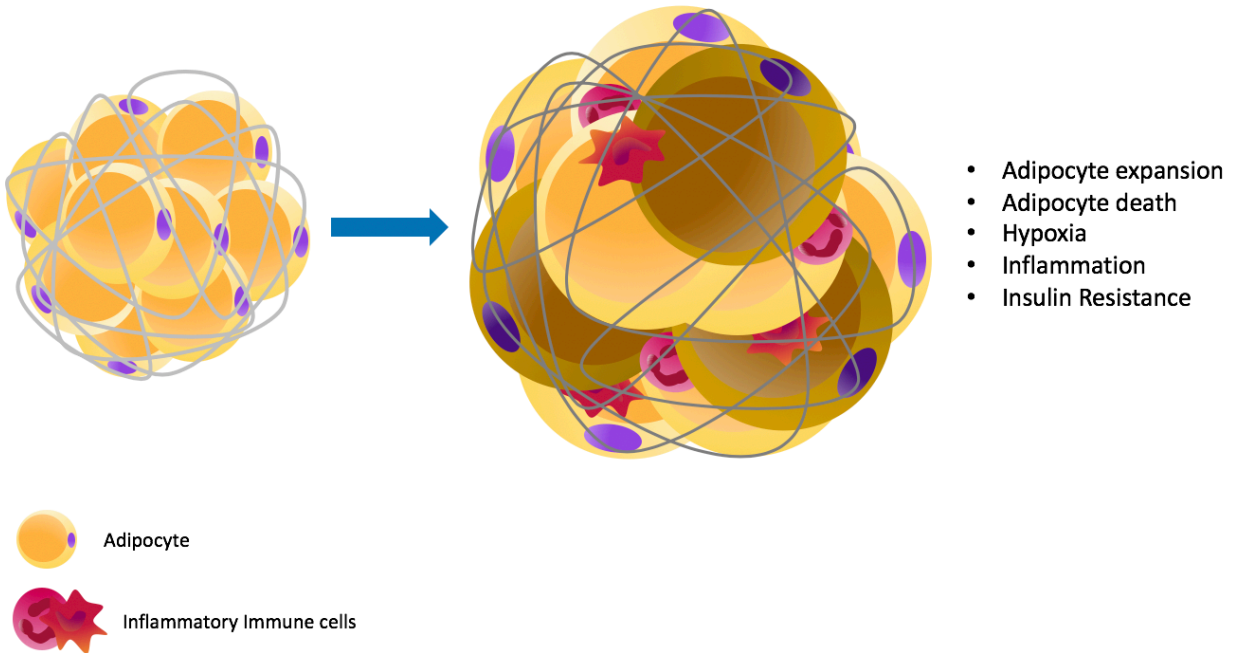


Figure 4: Obesity pathology. Adipocytes secrete matrix proteins to maintain the structure of the tissue. Expansion of the adipocyte is limited by the matrix. Over-nutrition leads to fibrotic changes in the matrix and hypoxia, inflammation, and cell death that contributes to insulin resistance. Adapted from Rosen and Spiegelman, 2014¹⁷.

Adipose Distribution and Metabolic Risk

Adipose tissue can be further classified as subcutaneous, under the skin, or visceral, surrounding organs in the abdominal cavity. Adipocyte function is depot, or location, dependent. These differences have significant implications for patient health. While there is evidence to suggest that subcutaneous fat may be inversely correlated with disease risk, visceral fat has a well-studied association with metabolic disease^{17, 31}.

Multiple methods are available to measure an individual's body fat. One of the most commonly used measures is body mass index (BMI). BMI is calculated by dividing weight (kg) by height (m) squared, making BMI a quick and non-invasive way to identify if a person is overweight. However, BMI is limited because it can only measure excess weight, not excess fat,

and it fails to account for fat distribution. Waist-to-hip ratio (WHR) provides a better measure for abdominal fat distribution. Individuals with a higher WHR have an increased predisposition to T2DM and coronary heart disease³². BMI alone is less associated with mortality than WHR, adjusted for BMI (WHRadjBMI)³³. Loci identified in WHRadjBMI genome wide association studies are enriched for adipose-specific genes and regulatory elements, and are involved in adipogenesis (differentiation of pre-adipocytes to adipocytes) and fat distribution among others³⁴. Several adipogenic signaling pathways, and important proadipogenic transcription factors such as PPAR- γ , are very well studied. Thus, the *in vitro* differentiation of preadipocytes to mature adipocytes is the most common mechanism for studying adipogenesis and adipocyte biology. As a result, our understanding of *in vivo* adipogenesis is limited¹⁷.

Genome Wide Association Studies (GWAS)

High-throughput genotyping technologies have become invaluable to understanding complex diseases like IR. One such example is Genome Wide Association Studies (GWAS) (Figure 5). GWAS genotype thousands of genomes using single-nucleotide polymorphism (SNP) microarrays. The individual genomes used in GWAS vary for a particular trait such as insulin sensitivity or height, and the allele frequencies at each SNP are calculated to see if any SNPs are over-represented in the phenotype of interest.

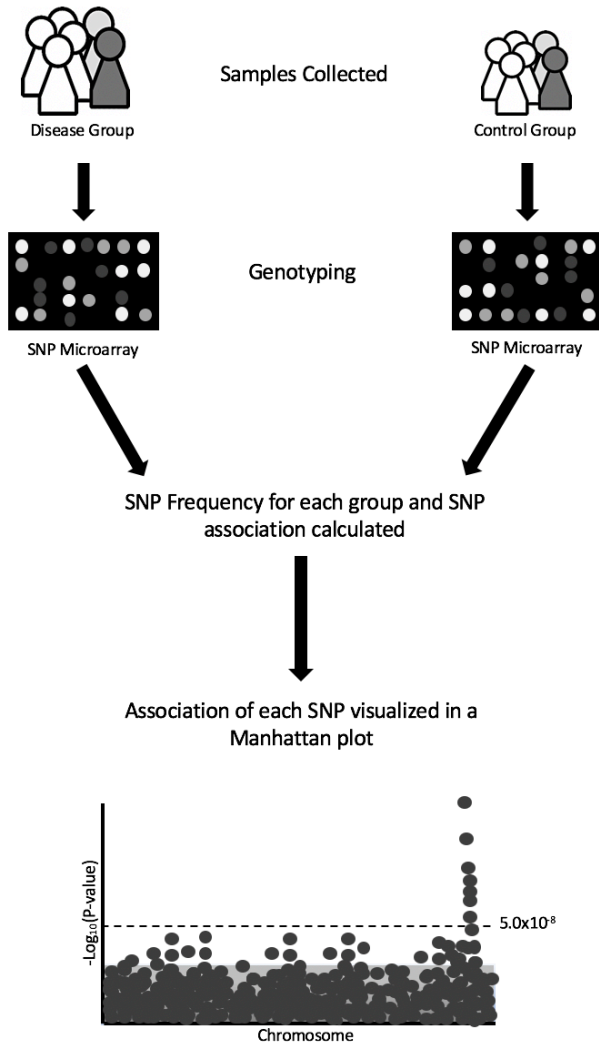


Figure 5: Genome Wide Association Studies analyze SNP frequencies between disease and control groups. Samples are genotyped using SNP microarrays. SNP frequencies and associations are calculated. Each point in a Manhattan Plot is the $-\text{Log}_{10}(\text{P-value})$ of each SNP plotted against the genomic position of the SNP. The dotted line is the threshold for significance at 5×10^{-8} .

GWAS would hopefully allow researchers to identify the genomic regions, or loci, that cause disease. Unfortunately, this is hardly ever the case since GWAS, though valuable, are limited in their scope. One limit is their inability to identify in which tissue a given loci is acting to affect the phenotype. Additionally, over 90% of the loci identified in GWAS are found in non-coding regions of the genome, which makes it difficult to know what genes are being affected and causing the phenotypic change³⁵. Enhancers, DNA regions involved in transcriptional regulation, can be right next or thousands of base pairs away from promoter regions and the

transcription start site. So a SNP in an enhancer region could be affecting genes anywhere on the chromosome. Furthermore, it is difficult to tease apart which SNPs are causal variants of a disease and which are simply neutral markers due to linkage disequilibrium (LD). LD occurs when the alleles at two or more loci in a population are associated and are more or less likely to be inherited together. These limitations to GWAS have led researchers to search for new methods to advance the interpretation and understanding of GWAS identified variants.

Identifying Genetic Regulatory Elements using Histone Modification ChIP-Seq

Chromatin immunoprecipitation-sequencing (ChIP-Seq) is a valuable tool used to study DNA-protein interactions (Figure 6). Proteins are cross-linked to the DNA, and the chromatin is sheared. Then, antibodies targeting specific transcription factors or chromatin histone modifications are incubated with the chromatin and pulled down, allowing isolation of the targeted DNA-protein complexes. The DNA fragments are reverse crosslinked and sequenced and aligned to the hg38 (human genome build 38) reference genome³⁶⁻³⁷. This study uses ChIP-Seq data sets for the chromatin histone modifications H3K27ac, H3K4me1, and H3K4me3. H3K4me1 is found on most enhancer sites, H3K27ac marks active enhancers and promoters, and H3K4me3 is a marker for active promoters³⁸⁻³⁹.

ChIP-Seq data is analyzed by “peak calling” (Figure 6). Each individual segment of DNA that is sequenced and aligned is called a read. Some regions of the genome have more reads aligned to it than others. A region is called a peak if a greater number of reads align to it than the regions surrounding it. In our ChIP-Seq data for the previously described histone marks, a region with a peak is likely an enhancer region in the case of H3K4me1, an active promoter region for H3K4me3, or an active enhancer or promoter region as in H3K27ac.

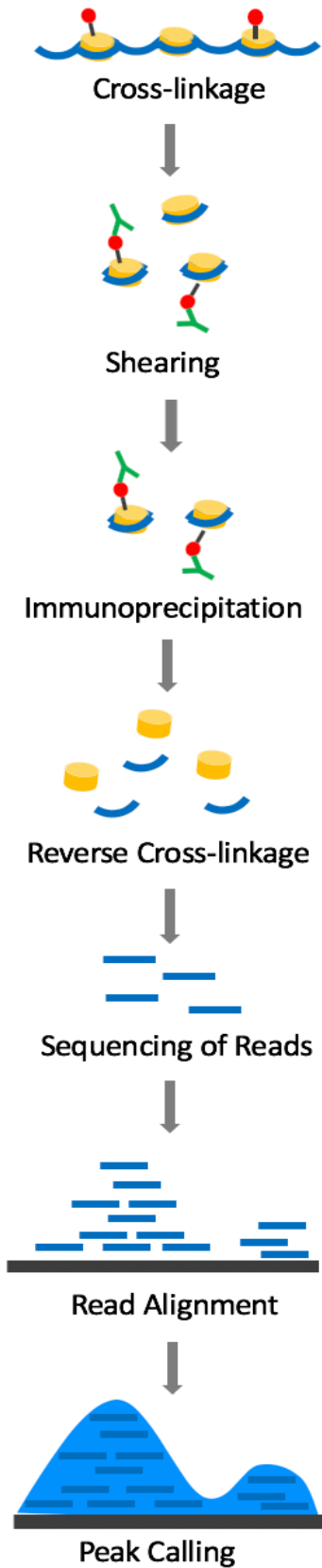


Figure 6: Using ChIP-Seq to identify regulatory elements in adipocytes. ChIP-Seq method for identifying regulatory elements. Antibodies are specific to the enhancer marks H3K27ac and H3K4me1, and promoter mark H3K4me3. ChIP-Seq data for adipocytes and ENCODE tissues were peak-called using existing peak calling pipeline.

Experimental Aims

Investigative methods such as GWAS have identified vast datasets of genomic loci that are associated with phenotypes like BMI, height, and Alzheimer's. One of the biggest challenges in interpreting GWAS is that a majority of identified loci are in non-coding regions of the DNA, suggesting an epigenetic factor to phenotype development. Epigenetics is central to cell differentiation and development. Differences in gene expression allow stem cells to go from pluripotent to say a liver or muscle cell and affects cell function throughout its life. Epigenetic variation is thus implicated in the development of different disease states, including IR. For GWAS, the question that many researchers have been working to answer is how do we find which genes are being regulated by these loci? This study uses computational techniques to pursue an answer to this question.

This analysis was conducted on ChIP-Seq data collected by the Rosen lab in 2015 and 2016 with the goal of identifying regulatory elements that differ between insulin sensitive (IS) and IR populations. *In vivo*, abdominal subcutaneous adipocyte samples were collected from individual human patients (Figure 7). In addition to the adipocyte samples from the Rosen lab, we curated ChIP-Seq data for ten other metabolically relevant tissue types— aorta, CD14+ monocytes, pancreatic islets, liver, lung, pancreas, peripheral blood mononuclear cell (PBMC), psoas muscle, and skeletal muscle— from the ENCODE genome database (Supplemental Table 1)⁴⁰⁻⁴¹. In order to identify genomic regions that regulate transcription, we target three histone marks with ChIP: H3K27ac (active enhancers and promoters), H3K4me1 (enhancers), and H3K4me3 (active promoters).

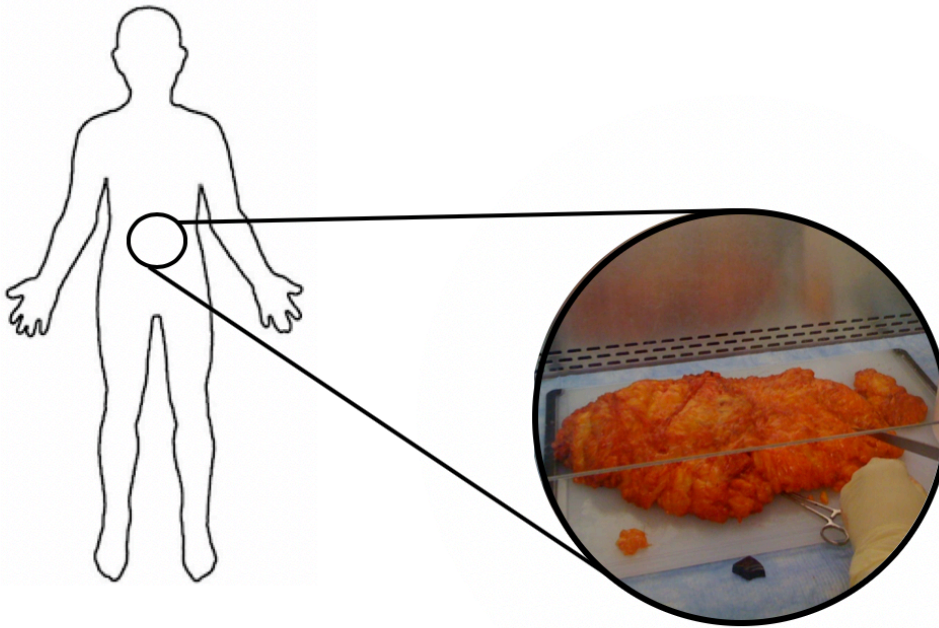


Figure 7: Adipocyte samples were isolated from human, abdominal whole-adipose samples. Adipocytes make up about half of the cells in adipose tissue and are separated from the tissue by low-speed centrifugation. Adipocytes are then lysed, and their nuclei are isolated for sequencing.

There are four major goals of this thesis. In order to begin teasing apart which genes are being regulated by a given enhancer or promoter in adipocytes, we first needed to identify a set of enhancers and promoters that were differentially enriched in adipocytes. To do this we used two methods, an Extremity analysis and differential enrichment analysis (DEA). The Extremity analysis was developed for this study to allow us to calculate relative enrichment of enhancers and promoters between individual samples, which allows for further analysis into which samples are potentially driving the enrichment of a given enhancer or promoter region. The Extremity analysis method was initially developed for, and applied to, adipocyte samples from around 30 different patients, for each mark. From this we hoped to see if epigenetics varied within adipocytes. Extremity was then adapted and applied to identify enhancers and promoters that

were enriched in adipocytes with respect to the non-adipocyte samples collected from ENCODE. Enriched enhancers and promoters were selected from the Extremity analysis for further analysis.

In order to test the validity of the Extremity analysis and further examine how the methods differ, we conducted a DEA on adipocyte vs non-adipocyte samples. DEA is well established in the field of genomics. Unlike Extremity where all samples are considered, DEA compares relative enrichment of enhancers and promoters between exactly two groups using a negative binomial distribution. The DEA should provide a similar result of adipocyte enriched peaks for the adipocyte vs non-adipocyte comparison, but the differences in how enrichment is calculated could affect which peaks are called enriched. For that reason, we compared the Extremity analysis and DEA.

A second aim of this study is to investigate if enhancers that are enriched in adipocyte samples are enriched or de-enriched in IR samples. This allowed us to begin to explore potential links between enhancer and promoter enrichment in adipocytes to enrichment and de-enrichment in IR patients. For this analysis, we ran a DEA on the adipocyte samples for every histone mark, and grouped samples by whether or not they were IR or insulin sensitive (IS) (Supplemental Table 2 and Supplemental Figure 1). This experiment allowed us to compare the peaks enriched in the adipocyte vs non-adipocyte DEA with those enriched or de-enriched in IR samples. We only compared between the two DE analyses rather than with Extremity to control for the inherent differences between the methods.

Once we established a set of adipocyte enriched enhancers and promoters from the Extremity analysis, we ran a motif enrichment analysis to see if these enhancers and promoters had enrichment of transcription factor (TF) binding sites. By seeing which TFs potentially

interact to our enriched enhancers and promoters, we began the first step to identifying the genes or signaling pathways these regions regulate.

Finally, we concluded our analysis by integrating peaks called in each tissue with existing GWAS data sets. We then partitioned the heritability the GWAS phenotypes across our adipocyte and non-adipocyte tissues. From this, we were able to compare the relative heritability of adipocytes against the non-adipocyte tissues for these phenotypes, with the aim of seeing if adipocytes have more of the explanatory power for the heritability of the GWAS traits.

Methods

Sample Collection and Selection

Subcutaneous adipose tissue samples were collected in 2015 and 2016 from healthy male and female subjects, ages 18-64, receiving abdominal surgery under IRB 2011P000079 from the plastic surgeon operating room schedule at Beth Israel Deaconess Medical Center (supplementary table 2). Samples were excluded if the subjects had a diagnosis of diabetes or were on medications that are insulin-sensitizing (such as thiazolidinediones or metformin), chromatin-modifying (valproic acid), or known to induce insulin resistance (mTOR inhibitors or systemic steroid medications). Fasting serum was collected and tested for insulin, glucose, free fatty acids, and lipid panel in a CLIA-approved lab. Body mass index (BMI) measurements were derived from electronic medical records and confirmed by self-report. Two measures of insulin resistance—the homeostasis model assessment-estimated insulin resistance index (HOMA-IR) and revised quantitative insulin sensitivity check index (QUICKI)—were calculated⁴²⁻⁴³. Female subjects in the 1st and 4th quartiles for either HOMA-IR or QUICKI and matched for age and BMI were selected for computational analysis (supplementary Figure 1).

Human Adipocyte Sample Preparation

Adipocytes were isolated from the whole tissue via enzymatic dissociation in an orbital shaker for 15 minutes, then centrifuged into a floating adipocyte supernatant. Isolated adipocytes were lysed to isolate nuclei and cross-linked in 1% formaldehyde for 3 minutes at 4°C.

Chromatin IP was performed as described in Mikkelsen et. al, 2007⁴⁴. Libraries were prepared from $1-5 \times 10^6$ nuclear equivalents and sequenced to a target depth of 20 M (million) reads using an Illumina NextSeq 500 sequencer. Output BCL files were converted to FASTQ reads using Illumina's bcl2fastq2 conversion software (VN: 2.17.1.14). Reads were aligned by Bowtie2 (VN: 2.2.9) to the hg38 human reference genome, and then filtered for duplicates by Picard^{36-37, 45-46}.

ENCODE Samples

Alignments and peak calls for histone mark ChIP-Seq data were downloaded from the ENCODE portal for whole adipose (H3K27ac only), aorta, CD14+ monocytes, pancreatic islets, liver, lung, pancreas, peripheral blood mononuclear cells (PBMC), psoas muscle, and skeletal muscle (Supplemental Table 1)⁴⁰⁻⁴¹. These tissues, with the exception of lung, were selected because they are involved in glucose homeostasis.

Adipocyte Peak Calling and Filtering

MACS2 (VN: 2.1.1) was used to call peaks in adipocyte samples⁴⁷. BEDTools2 (VN: 2.27.1) was used to merge and calculate coverage of all tissues for adipocyte peaks⁴⁸. Peaks were

finally filtered so that their length was between 100 and 10,000 base pairs and they had at least a 4:1 ratio against the whole cell extract (WCE) background. Counts for these peaks were normalized and converted to counts per million (CPM) with the edgeR R package⁴⁹⁻⁵⁰.

Differential Enrichment Analysis

Differential enrichment analyses (DEA) were performed to identify peaks with significant differences in read CPM between adipocytes and non-adipocytes and between adipocytes with IR and adipocytes with IS. To compare adipocytes and non-adipocytes, the DEA compared all adipocyte samples against all samples of the ENCODE tissues. To compare IR and IS the DEA was done on the adipocyte samples, and a batch correction was done with the edgeR `glmQLFTest` function⁴⁹⁻⁵⁰. The edgeR `exactTest` function uses a negative binomial distribution to compare peak enrichment and calculate fold-change (FC), average CPM, and a p-value⁵⁰. After adjusting the p-value into a false discovery rate (FDR), significant differentially enriched peaks in adipocytes were selected as those with an average $\log(\text{CPM}) \geq 1$, a $\log(\text{FC}) \geq 1.0$, and a $\text{FDR} \leq 0.05$. For IR, significant differentially enriched peaks were selected as those with an average $\log(\text{CPM}) \geq 2$, a $\log(\text{FC}) \geq 0.5$, and a $\text{FDR} \leq 0.25$.

To calculate the significance of the overlap between the adipocyte vs non-adipocyte DEA and the IR vs IS DEA, we used a randomized distribution. The data was randomly sampled with as many replicates as there were peaks for each histone mark. The p-value was calculated by taking the sum of the number of times the overlap of the randomized samples exceeded the experimental overlap divided by the number of randomized sampling replicates.

$$\text{p-value} = (\sum \text{randomized overlap} \geq \text{experimental overlap}) \div \text{sampling replicates}$$

Extremity Analysis

Extremity was first calculated between adipocyte samples (independent of IR status) to identify peaks enriched within adipocytes. The percent contribution of every adipocyte sample to each peak was calculated. The average percentage of each peak—the percent value that each sample would contribute to a peak if each sample had an equal contribution—was calculated. For example, the average contribution for H3K27ac peaks was

$$100\% \div 38 \approx 2.6\%$$

To calculate Extremity, the average percentage was subtracted from the highest percent contribution of each peak.

sample x, peak 1:

$$13\% - 2.6\% = 10.4\%$$

To analyze adipocyte enrichment of peaks between different tissue types, the CPM of samples for each tissue were averaged and the percent contribution of each tissue to each peak was calculated. To calculate adipocyte Extremity, the average percent contribution was subtracted from the adipocyte percent contribution.

Motif Enrichment Analysis

FIMO (Find Individual Motif Occurrences) was used to calculate motif occurrence with a maximum number of motif occurrences cutoff of 1,000,000 to ensure that all significant occurrences were kept⁵¹. AME (Analysis of Motif Enrichment) was used to calculate enrichment using a Mann-Whitney U test⁵². An E-value threshold of 1000 was used to ensure that a q-value was provided for each motif, we then performed our own filtering using a q-value ≤ 0.05 . Both programs are part of the MEME Suite⁵³.

GWAS: Linkage Disequilibrium Score Regression and Partitioning Heritability

This analysis uses the adipocyte samples, and non-adipocyte samples collected from ENCODE. ChIP-Seq peak sets used for this analysis were called in their respective tissues (I.E. adipocyte samples used adipocyte peak calls, etc.). The GWAS data sets being used are for this analysis are T2DM, WHRadjBMI, and Alzheimer's Disease^{34, 54-55}. Peak coordinates from hg38 were lifted-over to the hg19 human reference genome such that they were on the same reference genome as all SNP data³⁶⁻³⁷. To find SNPs contained in the aligned sample peaks, peaks are overlapped with known SNPs. From each tissue type, two annotations were created. The first annotation for each tissue is simply the set of SNPs overlapping any peak called in that tissue. The second annotation for each tissue is the set of SNPs overlapping all peak coordinates plus 500bp on either side. For adipocytes two additional annotations were created. These annotations represent SNPs that overlap the significantly enriched peaks from the Extremity analysis, and the significantly enriched peak coordinates plus 500bp on either side. By adding 500bp to either side of a peak, we attempt to catch SNPs that may be just beyond the bounds of a peak but are still in a GWAS locus with high heritability. We used the software program, LD Score Regression (LDSC), to calculate linkage disequilibrium (LD) scores per annotation and partition the heritability of each GWAS phenotype across these annotations⁵⁶⁻⁵⁷. LD score and partitioned heritability were calculated as described in Bulik-Sullivan et. al, 2015 and Finucane et. al, 2015, respectively. LD measures the likelihood that any given SNP appears in an annotation given the presence of all other SNPs. The LD score step calculates these LDs per annotation. These LD scores are then used to partition heritability.

Computational Analysis

Extremity and differential enrichment analyses were done using R (VN: 3.5.3) in RStudio (VN: 1.1.463)⁵⁸⁻⁵⁹. Code available in supplemental data section.

Results

In this study we set out four experimental aims to understand how epigenetic regulation in adipocytes affects human metabolic phenotype. We did this by using ChIP-Seq data for three different enhancer and promoter marks—H3K27ac, H3K4me1, and H3K4me3—to identify and analyze adipocyte enriched enhancers and promoters. The first aim was to identify a set of enhancers and promoters that were enriched in adipocytes relative to non-adipocyte tissue samples. We developed a method to identify differential enrichment, the Extremity analysis, by seeing if enhancer and promoter enrichment varied between adipocyte samples taken from individual human subjects. The Extremity analysis was then adapted to compare enhancer and promoter enrichment between adipocytes and 10 other tissues types. We then compared the Extremity analysis with a well-established method comparing adipocytes vs non-adipocytes, differential enrichment analysis, to validate the Extremity method and evaluate how the two methods differ. For the second aim, we wanted to compare adipocyte enriched enhancers and promoters from the adipocyte vs non-adipocyte DEA with IR enriched and de-enriched enhancers and promoters in DEA that compared IR vs IS adipocyte samples. This allowed us to explore a potential link between enhancer and promoter enrichment and IR/IS status in adipocytes.

Our third aim was to look for enrichment of TF binding sites in the adipocyte enriched peak sets from the Extremity analysis. We thus began the first step in identifying which genes or

pathways are being affected by the differential enrichment of enhancers and promoters in adipocytes. The Final aim of this project was to overlap the adipocyte and non-adipocyte samples with existing GWAS data sets in order to investigate the heritability of adipocytes for T2DM, WHRadjBMI, and Alzheimer's Disease, and to compare the relative heritability of the tissues for each phenotype.

Promoter and Enhancer Enrichment do not Vary Largely Between Adipocyte Samples

Human abdominal, subcutaneous adipocyte samples were isolated from individual human subjects previously by the lab, and ChIP-Seq experiments were run for the target histone marks H3K27ac, H3K4me1, and H3K4me3. To investigate if there is differential active promoter (H3K4me3 and H3K27ac) or enhancer (H3K27ac and H3K4me1) enrichment within the adipocyte samples, we designed a method to calculate how extreme the enrichment of a given peak varied from the average enrichment provided by the combined set of equally contributing samples.

In order to identify differential enrichment within the adipocyte samples, we explored three different cutoffs of percent contribution: if 2 samples contribute 90%, 3 samples contributed 75%, and if 4 samples contributed 50% or more of the total number of reads under a peak. No peaks were identified for the first cutoff of 2 samples contributing 90% or more for any of the histone marks, and only 6 peaks in H3K4me3 met the cutoff of 3 samples contributing 75% or more. The 4 samples contributing 50% or more cutoff caught 14 peaks for H3K27ac, 221 for H3K4me1, and 124 for H3K4me3 (Table 1).

Percent Contribution Cutoff	H3K27ac (%)	H3K4me1 (%)	H3K4me3 (%)
2 Samples \geq 90% contribution	0	0	0
3 Samples \geq 75% contribution	0	0	6 (7.8×10^{-3})
4 Samples \geq 50% contribution	14 (1.3×10^{-2})	221 (0.13)	124 (0.16)

Table 1: Number and percent of peaks that met enrichment percent contribution cutoffs for H3K27ac, H3K4me1, and H3K4me3 (n = 104,573; 164,608; and 77,107 respectively).

When the Extremity of peaks meeting the least stringent cutoff of 50% were plotted by average CPM, it was clear that these had low CPM and Extremity (Figure 8). It was therefore determined that enrichment of enhancers and promoters was likely not differential between adipocyte samples. In contrast, we observed that higher adipocyte Extremity was related to higher adipocyte CPM when compared to non-adipocyte samples (Figure 9).

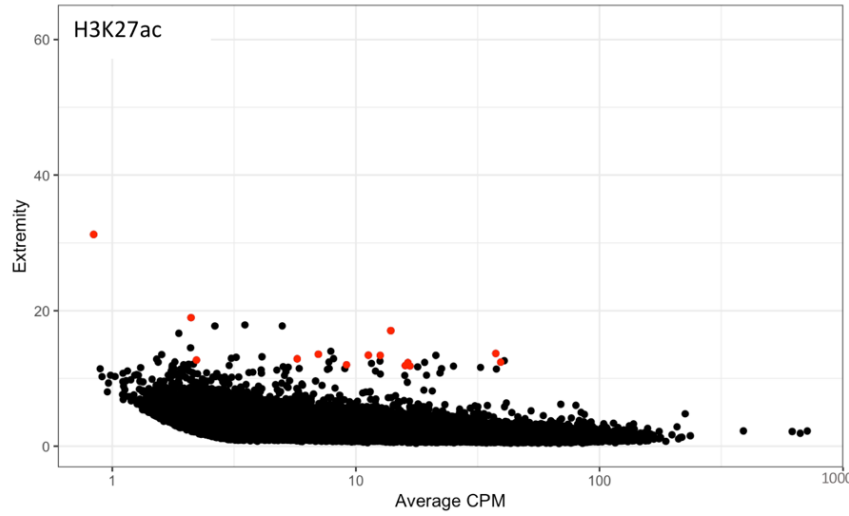
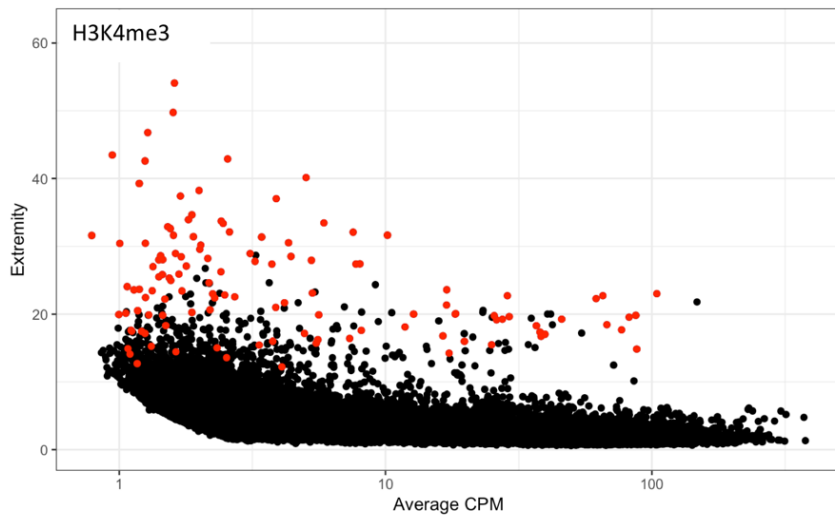
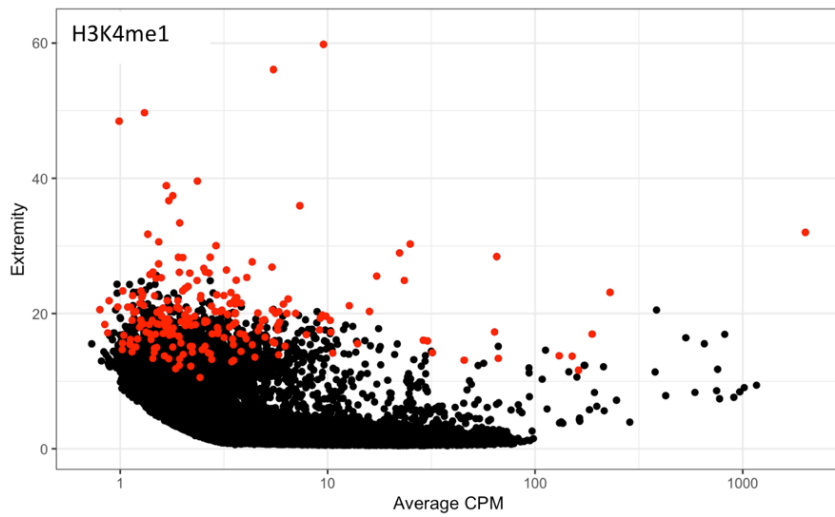


Figure 8: Enhancer (H3K27ac; H3K4me1) and promoter (H3K4me3) enrichment does not vary between individual adipocyte samples.

Distribution of Extremity against the average CPM for each peak. Percent contribution and Extremity were calculated for adipocyte samples. Red points show peaks in which 4 samples contributed 50% or more of the reads under the peak.



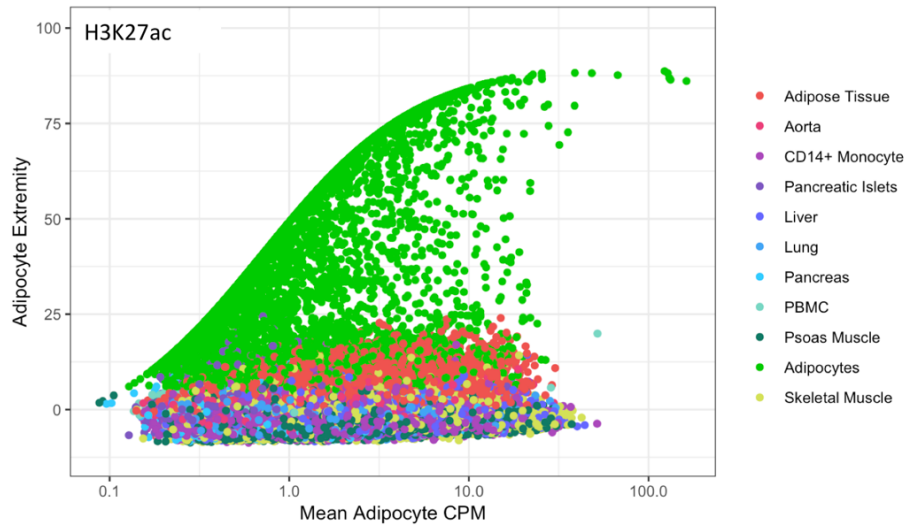
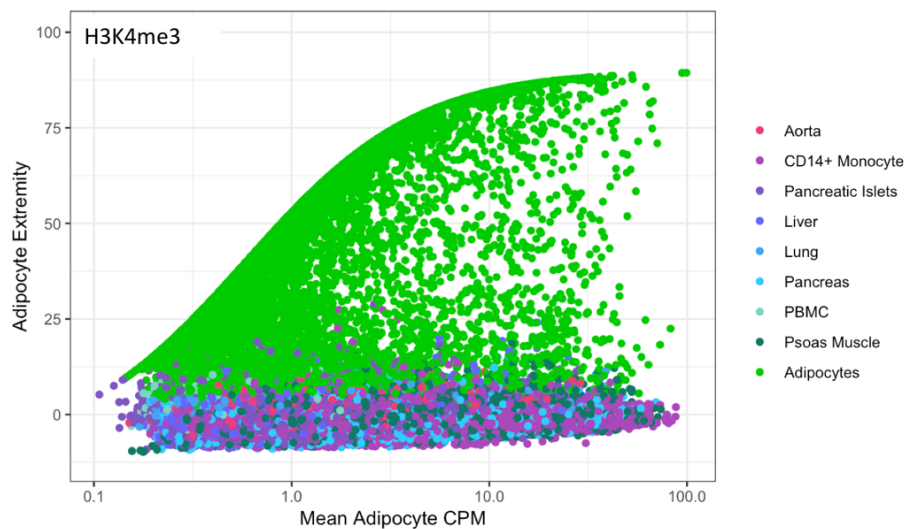
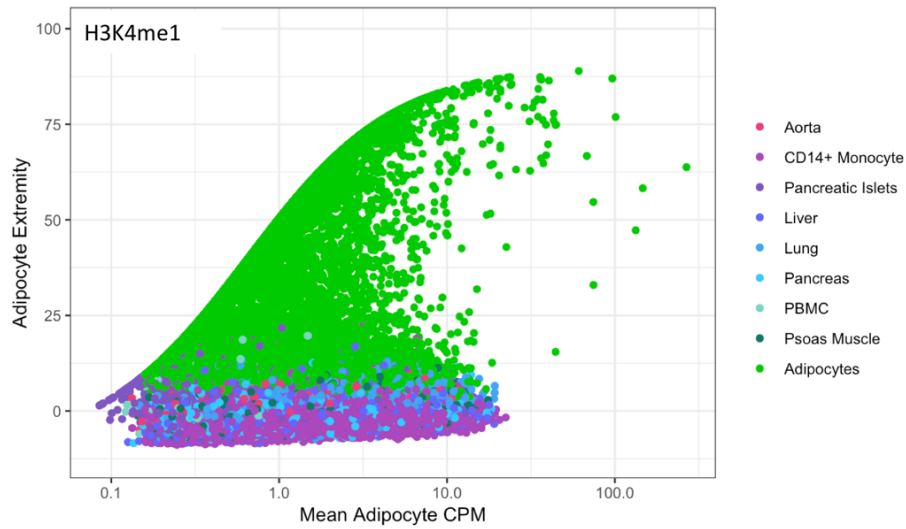


Figure 9: Adipocytes have higher representation as the top peak contributor as adipocyte Extremity increases. Distribution of average peak CPM by adipocyte Extremity. Points are colored by tissue with the highest percent contribution to the peak.



Differential Enrichment Analysis between Adipocytes and Non-adipocytes

Differential enrichment analysis (DEA) was done to compare enrichment in the adipocyte samples against the non-adipocyte samples downloaded from ENCODE. The DEA method is well established in the field of genomics. We conducted our DEA using the R package edgeR to build negative binomial models of the two test groups before using the exactTest function to calculate differential enrichment.

Peaks with a $\log(\text{FC}) \geq \pm 1$, $\text{FDR} \leq 0.05$, and $\log(\text{average CPM}) \geq 1$ were selected as differentially enriched (Up) or de-enriched (Down) in adipocytes (Figure 10).

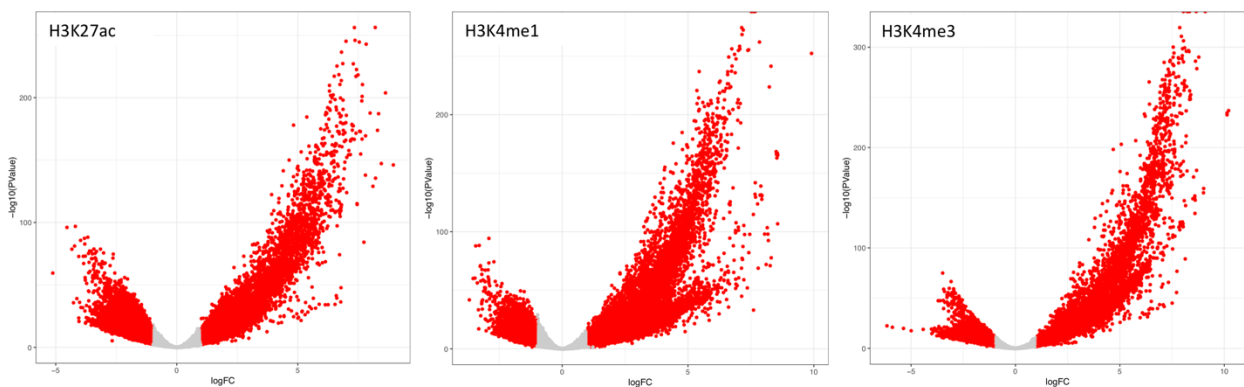


Figure 10: Volcano plots for differential enrichment analysis (DEA) of adipocytes vs nonadipocytes. DEA compared enrichment of peaks called in adipocytes between adipocyte samples and non-adipocyte samples: whole adipose (H3K27ac only), aorta, CD14+ monocyte, pancreatic islets, liver, lung, pancreas, peripheral blood mononuclear cell (PBMC), psoas muscle, and skeletal muscle. Red peaks show significant enrichment or de-enrichment ($\log(\text{FC}) \geq \pm 1$, $\text{FDR} \leq 0.05$, and $\log(\text{average CPM}) \geq 1$).

Since we wanted to identify promoters and enhancers that are enriched in adipocytes, we focused on the DEA Up peaks. In H3K27ac, 3,733 peaks were enriched in adipocytes while 1,526 and 2,106 peaks were enriched for H3K4me1 and H3K4me3, respectively (Table 2).

	H3K27ac (%)	H3K4me1 (%)	H3K4me3 (%)
Down	10678 (10.2)	9595 (5.83)	8505 (11.0)
Up	3733 (3.57)	1526 (0.93)	2106 (2.73)

Table 2: Number and percent of total peaks that are enriched (Up) and de-enriched (Down) in adipocytes for H3K27ac, H3K4me1, and H3K4me3 (n = 104,573, 164,608 and 77,107 respectively) ($\log(\text{FC}) \geq \pm 1$, $\text{FDR} \leq 0.05$, and $\log(\text{average CPM}) \geq 1$).

Adipocyte Extremity Shows Correlation with Fold-Change from DEA

In order to compare between both Extremity and DEA methods for calculating adipocyte-specific enrichment of enhancers and promoters, adipocyte Extremity was plotted against $\log(\text{FC})$. A linear regression showed strong correlation between FC and adipocyte Extremity with R^2 values of 0.82 for H3K27ac, 0.92 for H3K4me1, and 0.97 for H3K4me3 (Figure 11). This supports the validity of the Extremity analysis as a method to identify enhancer and prom

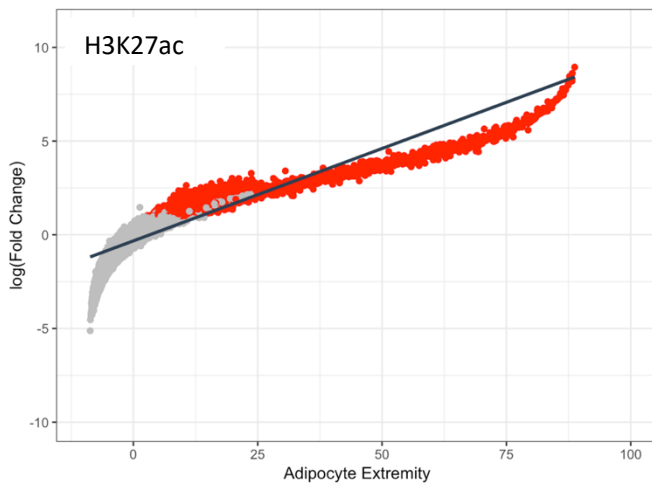
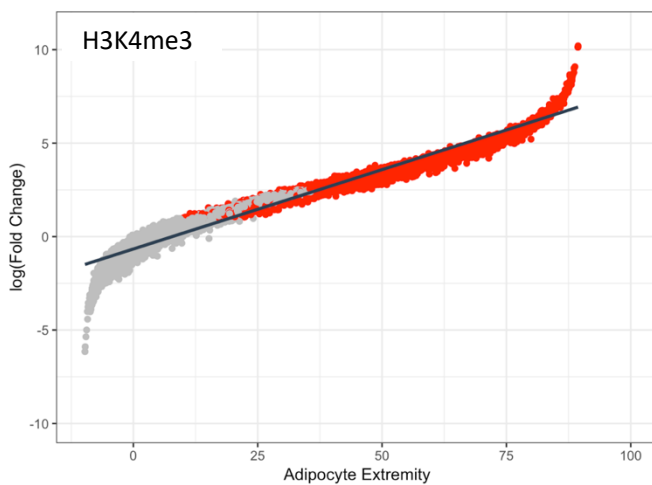
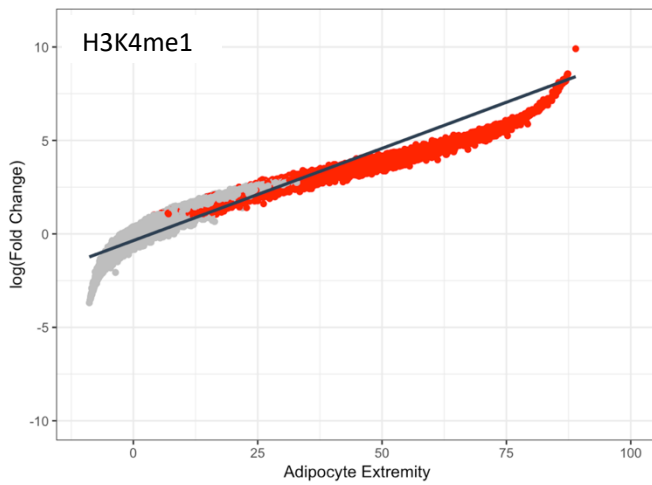


Figure 11: Comparison of adipocyte Extremity score and DEA fold change (FC). A linear regression is shown in black ($R^2 = 0.82, 0.92, \text{ and } 0.97$ respectively). Red points are significant DEA peaks ($\log(\text{FC}) \geq +1, \text{ FDR} \leq 0.05, \text{ and } \log(\text{average CPM}) \geq 1$).



Comparison of Adipocyte vs Non-adipocyte DEA and IR vs IS DEA

Obesity is known to increase risk of metabolic disorders, including IR. Thus, we investigated if there was overlap between enhancers and promoters that were enriched in adipocytes and those that are enriched or de-enriched in IR adipocyte samples. For the IR vs IS DEA, adipocyte samples were assigned IR or IS based on patient data (supplemental table 2 and supplemental figure 1). The DEA was then run on the adipocyte samples with IR and IS status as the groups to be compared. Significance for IR vs IS DEA peaks was determined by a $\log(\text{FC}) \geq \pm 0.5$, $\text{FDR} \leq 0.25$, and $\log(\text{average CPM}) \geq 2$ (Table 3).

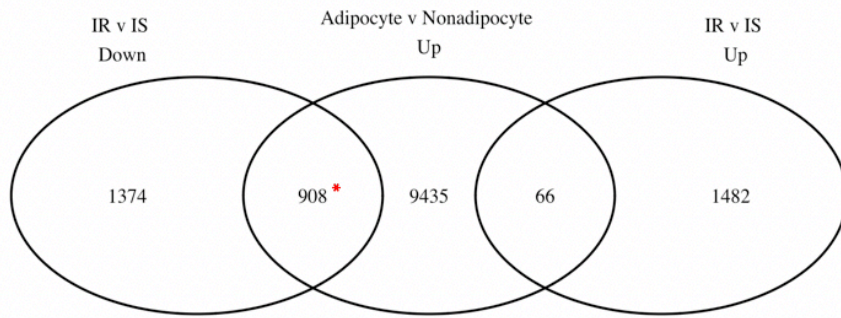
	H3K27ac (%)	H3K4me1 (%)	H3K4me3 (%)
Down	3543 (3.39)	1721 (1.04)	5086 (6.60)
Up	2567 (2.45)	5963 (3.62)	15080 (19.6)

Table 3: Number and percent of total peaks that are enriched (Up) and de-enriched (Down) in IR adipocyte samples against IS adipocyte samples for H3K27ac, H3K4me1, and H3K4me3. Significant Up and Down peaks have a $\log(\text{FC}) \geq \pm 0.5$, $\text{FDR} \leq 0.25$, and $\log(\text{average CPM}) \geq 2$ (n = 104,573; 164,608; and 77,107 respectively).

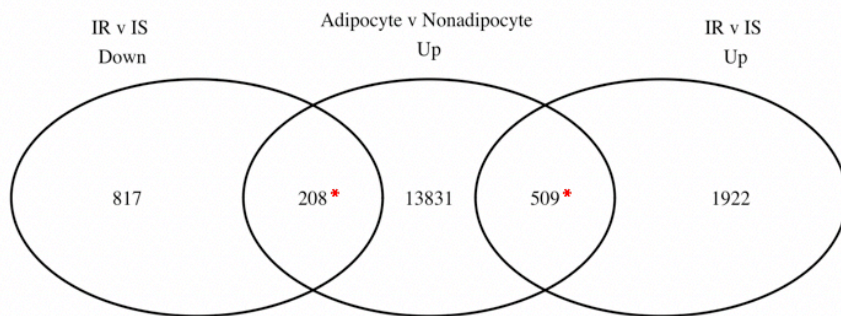
To compare overlap between the enriched sets of adipocyte DEA peaks with the enriched and de-enriched sets of IR peaks, a Venn diagram was used (Figure 12). Significance of overlap was calculated by randomized distribution and calculated a p-value using the function presented in the Methods. In each case, the number of replicates which resulted in an overlap greater than the experimental overlap was either 0 or equal to the number of replicates. Thus, each p-value was either 0 or 1. H3K4me1 showed significant overlap of adipocyte enriched peaks with both

IR enriched and de-enriched peaks (p-value = 0). H3K27ac only showed significant overlap of adipocyte enrichment with IR de-enriched peaks (p-value = 0), and H3K4me3 only showed significant overlap of adipocyte enrichment with IR enriched peaks (p-value = 0).

H3K27ac



H3K4me1



H3K4me3

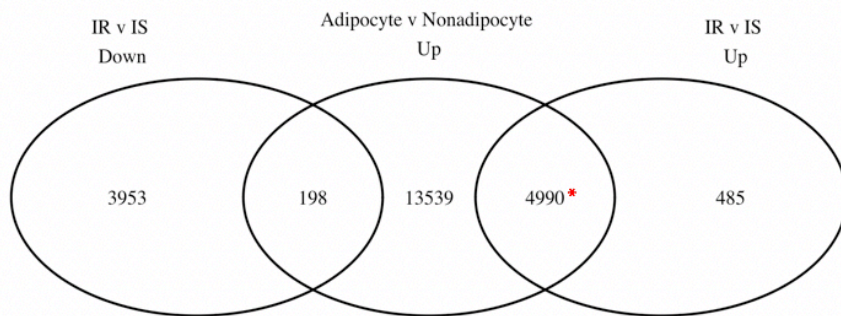


Figure 12: Comparison of significant peaks from IR vs IS DEA and adipocyte vs nonadipocyte DEA. Venn diagrams show the overlap of shared enriched peaks between adipocyte vs nonadipocyte DEA ($\log(\text{FC}) \geq 1$, $\text{FDR} \leq 0.05$, and $\log(\text{average CPM}) \geq 1$), and enriched (Up) and de-enriched (Down) peaks from the IR vs IS DEA ($\log(\text{FC}) \geq \pm 0.5$, $\text{FDR} \leq 0.25$, and $\log(\text{average CPM}) \geq 2$). Red asterisks show significant overlaps (p-value = 0 for significant overlaps).

Motif Enrichment Analysis of Peaks Selected from Adipocyte Extremity

To begin investigating the functions of the enhancers and promoters identified in the adipocyte Extremity analysis, a motif analysis was conducted using AME from the MEME Suite⁵²⁻⁵³. The primary set of adipocyte enriched peaks was determined by taking the top three deciles of adipocyte Extremity and filtering to see the numbers of peaks with adipocytes as the highest contributor to the peak (Table 4). For H3K27ac and H3K4me1, the top decile had the highest percentage of peaks with adipocytes as the top contributor, so those peaks were selected. For H3K4me3, the top two deciles had a high percentage of peaks with adipocytes as top contributor, so the two deciles were pooled. The top deciles of H3K27ac and H3K4me1, and the pooled top two deciles of H3K4me3 were then quartiled. The top quartile—2.5% of peaks for H3K27ac and H3K4me1 and 5% of peaks for H3K4me3—was selected as the primary set for the motif enrichment analysis. Peaks in the bottom 30% of adipocyte Extremity for each histone mark were chosen as the background sets. Significant motif enrichment was determined to be q-value ≤ 0.05 .

	H3K27ac	H3K4me1	H3K4me3
Decile 1 (1-0.9)	3074	4390	2234
Decile 1 # Adipocyte top contributor (%)	2015 (65.5)	3846 (87.6)	2234 (100)
Decile 2 (0.9-0.8)	3074	4392	2233
Decile 2 # Adipocyte top contributor (%)	457 (14.9)	939 (21.4)	2072 (92.8)
Decile 3 (0.8-0.7)	3074	4393	2233 (7.75)
Decile 3 # Adipocyte top contributor (%)	60 (1.95)	51 (1.16)	173

Table 4: Selection of primary peaks for motif analysis for H3K27ac, H3K4me1, and H3K4me3. H3K4me3 peaks from the first and second deciles were pooled. Peak sets used circled in red.

Motif occurrences were calculated using FIMO from the MEME Suite for each primary set of peaks. Motif enrichment was plotted against occurrences in Figure 13^{51, 53}. 15 motifs were enriched in H3K27ac, 11 were enriched in H3K4me1, and 42 were enriched in H3K4me3. Additionally, significantly enriched motifs have high numbers of occurrences.

To investigate which transcription factors (TFs) were the most enriched for each mark, the TF family and motif logo for the top 5 enriched motifs were found from the HOCOMOCO database (Table 5)⁶⁰. PPAR- γ , a regulator of adipogenesis, and PPAR- α , a regulator of lipid metabolism, are highly enriched in H3K27ac⁶¹⁻⁶². At least 2 Krüppel-like zinc finger motifs are enriched in each mark. These are involved in cell differentiation and development in mammals and are known to induce one of the two PPAR- γ receptors⁶³⁻⁶⁴.

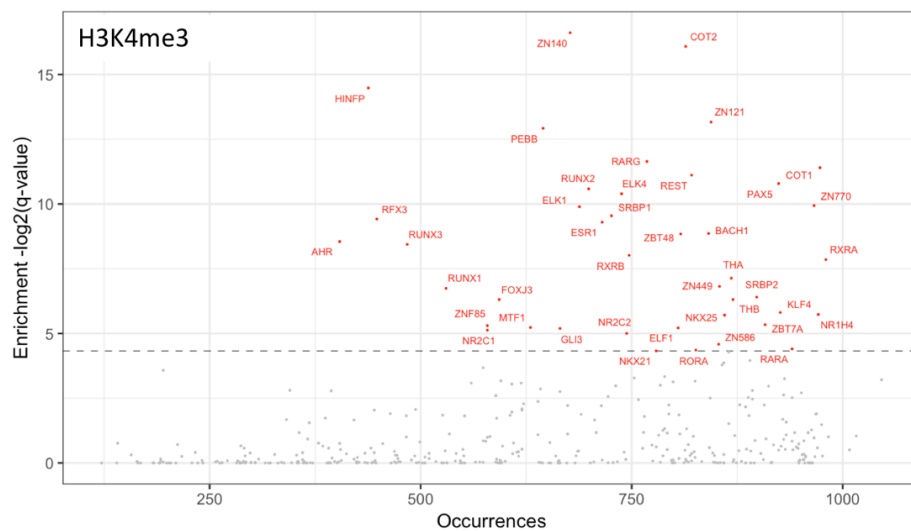
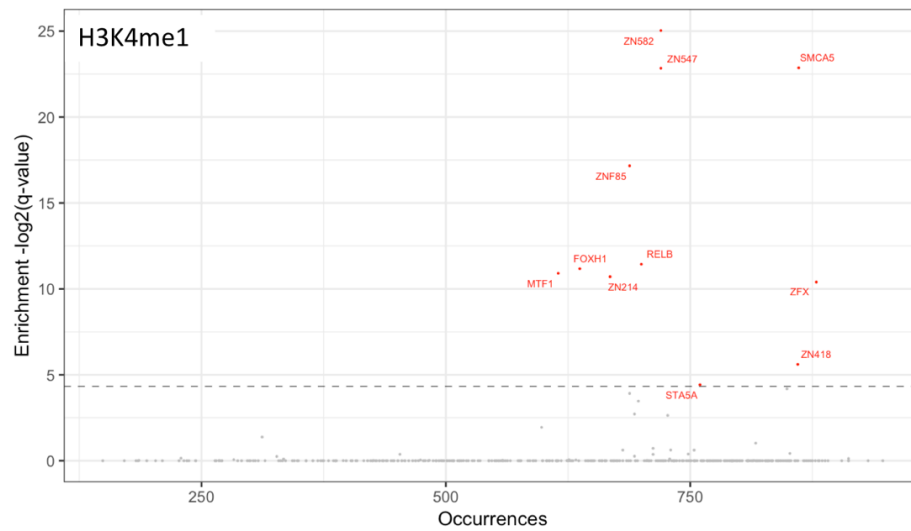
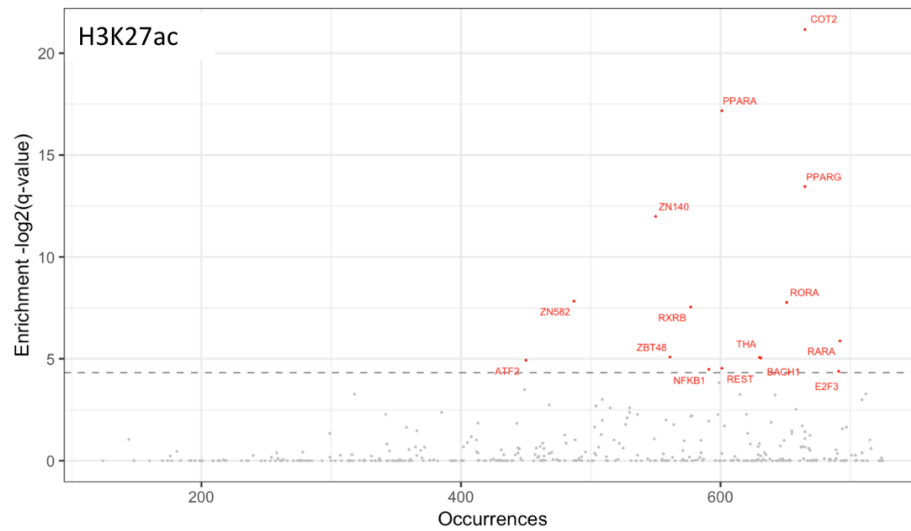


Figure 13: Motif enrichment and occurrence for H3K27ac, H3K4me1, and H3K4me3. Red points show motifs that are significantly enriched. Dotted line shows cutoff of significance ($-\log_2(q\text{-value}) \geq -\log_2(0.05)$).


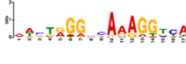


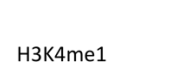


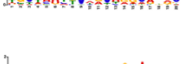

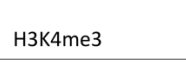


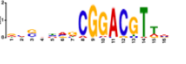
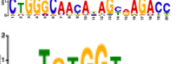
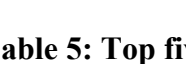
H3K27ac			
Motif	Transcription Factor (Model ID)	TF Family	q-value
	NR2F2 (COT2.0.A)	Nuclear hormone receptor family	4.27x10 ⁻⁰⁷
	PPAR-α (PPARA.0.B)	Thyroid hormone receptor-related factors	6.75x10 ⁻⁰⁶
	PPAR-γ (PPARG.0.A)	Thyroid hormone receptor-related factors	8.92x10 ⁻⁰⁵
	ZNF140 (ZNF140.0.C)	krueppel C2H2-type zinc-finger protein family	2.46x10 ⁻⁰⁴
	ZNF582 (ZNF582.0.C)	krueppel C2H2-type zinc-finger protein family	4.40x10 ⁻⁰³
H3K4me1			
Motif	Transcription Factor (Model ID)	TF Family	q-value
	ZNF582 (ZNF582.0.C)	krueppel C2H2-type zinc-finger protein family	2.91x10 ⁻⁰⁸
	SMARCA5 (SMCA5.0.C)	SNF2/RAD54 helicase family	1.31x10 ⁻⁰⁷
	ZNF547 (ZNF547.0.C)	krueppel C2H2-type zinc-finger protein family	8.92x10 ⁻⁰⁵
	ZNF140 (ZNF140.0.C)	krueppel C2H2-type zinc-finger protein family	6.82x10 ⁻⁰⁶
	RELB (RELB.0.C)	NF-kappaB-related factors	3.61x10 ⁻⁰⁴
H3K4me3			
Motif	Transcription Factor (Model ID)	TF Family	q-value
	ZNF140 (ZNF140.0.C)	Krueppel C2H2-type zinc-finger protein family	1.00x10 ⁻⁰⁵
	NR2F2 (COT2.0.A)	Nuclear hormone receptor family	1.44x10 ⁻⁰⁵
	HINFP (HINFP.0.C)	Factors with multiple dispersed zinc fingers	4.38x10 ⁻⁰⁵
	ZNF121 (ZNF121.0.C)	Krueppel C2H2-type zinc-finger protein family	1.09x10 ⁻⁰⁴
	CBFB (PEBB.0.C)	CBF-beta family	1.29x10 ⁻⁰⁴

Table 5: Top five enriched motifs identified using adipocyte Extremity method for each histone mark: H3K27ac, H3K4me1, and H3K4me3^{52-53, 60, 65}.

Heritability for GWAS Traits

We calculated LD scores and partitioned the heritability using LD score regression for tissues for existing GWAS data sets for T2DM, WHRadjBMI, and Alzheimer's Disease^{34, 54-55}. T2DM and WHRadjBMI were selected because they would likely show higher levels of heritability in adipocytes. The Alzheimer's data set was used as negative control. Four annotations were created from the peaks per tissue: SNPs overlapping all peaks within that tissue, SNPs overlapping all peak coordinates plus 500bp on either side, significantly enriched peaks from the Extremity analysis (adipocytes only), and significantly enriched peak coordinates plus 500bp on either side (adipocytes only).

In Figure 14, we plotted the z-scored coefficients of heritability for each annotation with respect to each GWAS data set in each of our three histone marks. For both T2DM and WHRadjBMI, adipocytes had high coefficient z-scores in H3K27ac and H3K4me1. Interestingly, H3K4me3 had low coefficient z-scores for adipocytes relative to the other tissue types. This means that the SNPs that overlap H3K27ac and H3K4me1 have high heritability for T2DM and WHRadjBMI, as expected. The SNPs overlapping the promoter mark H3K4me3 did not exhibit the same level of heritability. Additionally, relative heritability of significant adipocyte peaks to all adipocyte peaks was lower in T2DM and WHRadjBMI except H3K4me1 and H3K4me3 in the T2DM comparison. As expected, adipocytes had low and even negative coefficient z-scores for Alzheimer's disease.

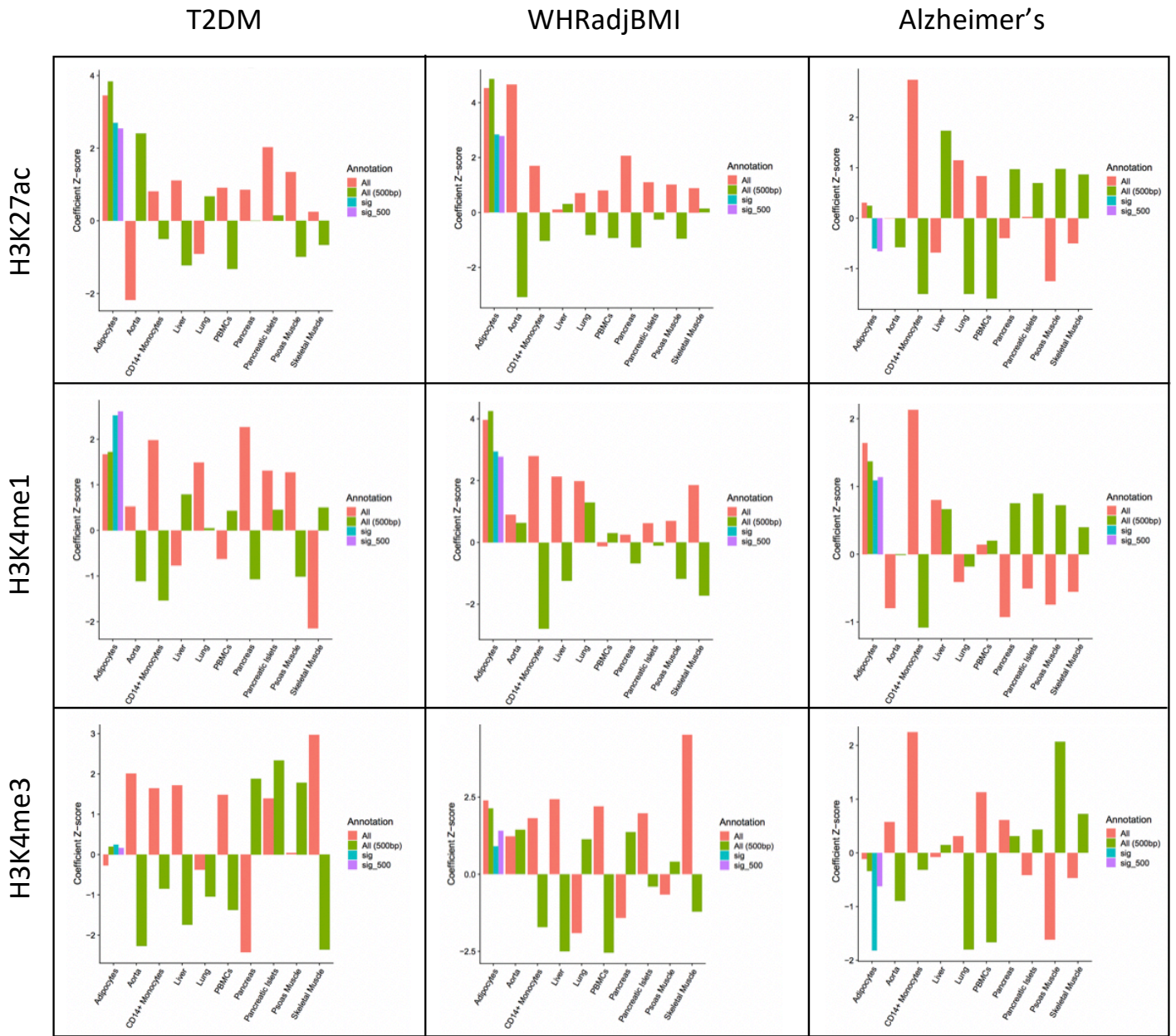


Figure 14: Coefficient z-score for heritability of T2DM, WHRadjBMI, and Alzheimer's Diseases among different tissue types. Coefficient z-score of adipocyte and non-adipocyte samples was calculated using LDSC for three GWAS analyses^{34, 54-55}. Annotations include all peaks of a tissue, all peaks+500bp on either side, significantly enriched peaks (sig) from the Extremity analysis (adipocytes only), and significantly enriched peaks+500bp on either side.

Discussion

Rates of obesity and diabetes have almost doubled since the 1980s. Type 2 diabetes mellitus (T2DM) is now the seventh leading cause of death in the United States. Though much of the blame for this rise is attributed to lifestyle changes and overnutrition, it is understood that metabolic disease risk is heritable. Advancements in genomics has allowed researchers to investigate genomic variants in obese and diabetic patients. GWAS-identified genetic variants for T2DM only explained 18% of heritable risk. Additionally, over 90% of all GWAS loci are found in non-coding regions of DNA. Strongly suggesting an epigenetic factor to inherited metabolic disease risk. This is not entirely surprising as regulation of gene expression is necessary for cell differentiation and function. Understanding how changes in gene expression has proven more difficult than expected. One of the biggest challenges faced by researchers is determining which GWAS loci are causing the disease, and which gene they are affecting to do so. This thesis is focuses on four experimental aims that are meant to begin to tackle this challenge. First, we identify sets of differentially enriched enhancer and promoter marks in adipocytes. We do this by developing an Extremity analysis, that we compare with a well-established method, DEA. Adipocyte enriched enhancers and promoters identified in the Extremity analysis are selected for a motif enrichment analysis to identify TF binding site enrichment. Identified TFs provide information on the potential pathways and genes being affected by the regulatory variants in adipocytes. Next, we compare our DE analyses to begin to compare adipocyte enrichment of enhancers and promoters to those enriched in IR patients. Additionally, we calculate the heritability of our adipocyte and non-adipocyte samples with GWAS data sets for T2DM, WHRadjBMI, and Alzheimer's Disease.

Promoter and enhancer enrichment in human abdominal, subcutaneous adipocyte samples was analyzed using ChIP-Seq. Data was collected for adipocytes from ChIP-Seq experiments in 2015 and 2016, and from the ENCODE database for 10 additional tissue types. We targeted 3 histone marks with ChIP: H3K27ac marks active promoters and enhancers, H3K4me1 marks enhancers, and H3K4me3 marks active promoters.

To begin identifying adipocyte enriched enhancers and promoters, we first applied the Extremity analysis to compare enrichment between adipocyte samples. Epigenetic markers are not expected to vary widely for a single cell type. However, most studies of adipogenesis are conducted on cultured adipocytes while our study uses *in vivo* samples. Our study therefore takes into account potential epigenetic variation that results from patient medical conditions, life experiences, and environment. In order to identify differentially enriched peaks, we set three cutoffs for percent contribution: if at least 2 samples contributed $\geq 90\%$, at least 3 samples contributed $\geq 75\%$, or at least 4 samples contributed $\geq 50\%$ of reads to a peak. Less than 0.2% of the number of peaks in each mark reached the least stringent cutoff of 4 samples contributing $\geq 50\%$ of reads. Additionally, peaks that did pass the threshold had lower counts per million (CPM) and Extremity values than the overall CPM and Extremity distribution for all peaks. Hence, we did not observe significant variation in enhancer and promoter enrichment between adipocyte samples.

The Extremity analysis was then adapted and applied to adipocyte samples and 10 metabolically relevant tissue types—whole adipose tissue (H3K27ac only), aorta, CD14+ monocytes, pancreas, pancreatic islets, liver, lung, PBMC, psoas muscle, and skeletal muscle—to identify our sets of adipocyte enriched enhancers and promoters. We also used a differential enrichment analysis (DEA) to compare against the newly developed Extremity analysis. Though

their aims were the same, to identify adipocyte enriched enhancers and promoters, the Extremity analysis and DEA approaches were quite different. While the DEA compares the relative enrichment of each peak between two groups, in this case adipocytes vs non-adipocytes, Extremity allows for the comparison of the relative enrichment of each tissue to every peak. We show that the adipocyte vs non-adipocyte DEA and the Extremity analysis are closely correlated with each other. This validates the Extremity by showing that, despite their differences, both methods provide similar results. Furthermore, the Extremity method would allow us to pursue questions that the DEA would not. One future question we would like to pursue is to try and identify if the epigenetic regulation of any tissue type is similar to adipocytes, and whether IR/IS status affects this answer. The Extremity analysis provides us with a new method to identify the relative enrichment of regulatory regions between multiple samples rather than just two groups.

That is not to say that the DEA is worth replacing. In order to examine whether enhancer or promoter enrichment in adipocytes was associated with enrichment or de-enrichment in IR samples, we overlapped the adipocyte-enriched peaks from the adipocyte vs non-adipocyte DEA with a DEA that compared IR vs IS adipocyte samples. A significant overlap was found in H3K27ac (promoters and enhancers) peaks de-enriched in IR samples, H3K4me1 (enhancers) peaks both enriched and de-enriched peaks in IR samples, and H3K4me3 (promoters) peaks enriched in IR samples. These significantly overlapping peaks would need additional research to elucidate their potential effects on obesity and IR risk. But we were able to begin analyzing the association between adipocyte epigenetics and development of IR.

A motif enrichment analysis was done on adipocyte enriched enhancers and promoters from the Extremity analysis. We hoped to identify binding sites for TFs associated with genes or pathways that are known to be adipocyte enriched or could be future targets of study. We

selected the top five enriched motif binding sites, though there were more significantly enriched binding sites that were not analyzed in this study. Motifs recognized by PPAR- γ , an important regulator of adipogenesis, and PPAR- α , a known regulator of lipid metabolism, were highly enriched in H3K27ac⁶¹⁻⁶². These are known as peroxisome proliferator-activated receptors (PPARs) and are modulated by the common diabetic drug type, thiazolidinediones. Their enriched expression in adipocytes is expected, and further supports the validity of the Extremity analysis⁶⁶. A number of motifs recognized by Krüppel-like zinc finger proteins were found in the top five for all three marks. Krüppel-like zinc fingers are involved in cell differentiation and development in mammals, and have been shown to induce one of the two PPAR- γ promoters⁶³⁻⁶⁴. Additionally, the NR2F2 gene appears in both H3K27ac and H3K4me3, and is known to be important for angiogenesis (formation of new blood vessels) and heart development⁶⁷. Expansion of any tissue requires the formation of new blood vessels to supply it with oxygen. In the case of weight gain and adipogenesis, angiogenesis has been shown to be necessary for differentiation of pre-adipocytes to adipocytes *in vivo*⁶⁸. Little is known about *in vivo* adipogenesis because most models use cultured preadipocytes and induce differentiation¹⁷. NR2F2 could be an interesting target for future study because rapidly expanding fat tissue, like rapidly dividing cancers, can outgrow its blood supply and therefore become hypoxic¹⁷. Hypoxia can then lead to metabolic disorder¹⁷. Further analysis should be conducted on the enriched TF binding sites that were not in the top five. Overall, we identified a number of potentially interesting enriched and highly occurring TF binding sites that can be further studied.

To address the final aim of our analysis we incorporate GWAS data sets to show that adipocyte biology is important in obesity related phenotypes such as T2DM and WHRadjBMI. The purpose of this experiment was more exploratory, and we therefore did not come to a clear

conclusion or result from this section. Our preliminary results met expectations. Adipocytes showed relatively high heritability in T2DM and WHRadjBMI and did not show high heritability for the Alzheimer's GWAS. The active promoter mark, H3K4me3, showed surprisingly low heritability for T2DM and WHRadjBMI. A more thorough analysis of this phenomenon is needed.

This study still bridges the gap between epigenetics and GWAS interpretation. Shungin et. al, 2015 conducted a meta-analysis of WHRadjBMI GWAS where they identified loci associated with body fat distribution⁶⁹. They determined that the identified loci were enriched in adipose tissue genes and adipocyte regulatory elements. Pathway analyses suggested these loci were involved in adipogenesis, angiogenesis, and transcriptional regulation. Shungin et. al, 2015 used literature searches and computer modeling to identify gene sets. In this study, we used motif analysis to identify TF binding sites. A chromatin study by Mikkelsen et. al, 2010 generated the chromatin state maps of cultured mouse and human pre-adipocyte samples that were differentiated into adipocytes⁷⁰. They identified distal regulatory regions in adipogenesis associated loci and used TF motif analysis to identify two regulators of adipogenesis. Our motif analysis identified binding sites for TFs associated with adipogenesis and angiogenesis like the two studies, validating our method. Furthermore, our use of *in vivo* samples is particularly important because understanding of *in vivo* adipogenesis is limited. In this study, we are thus able to present a novel and generalizable method to compare *in vivo* samples for any number of histone marks or tissue types.

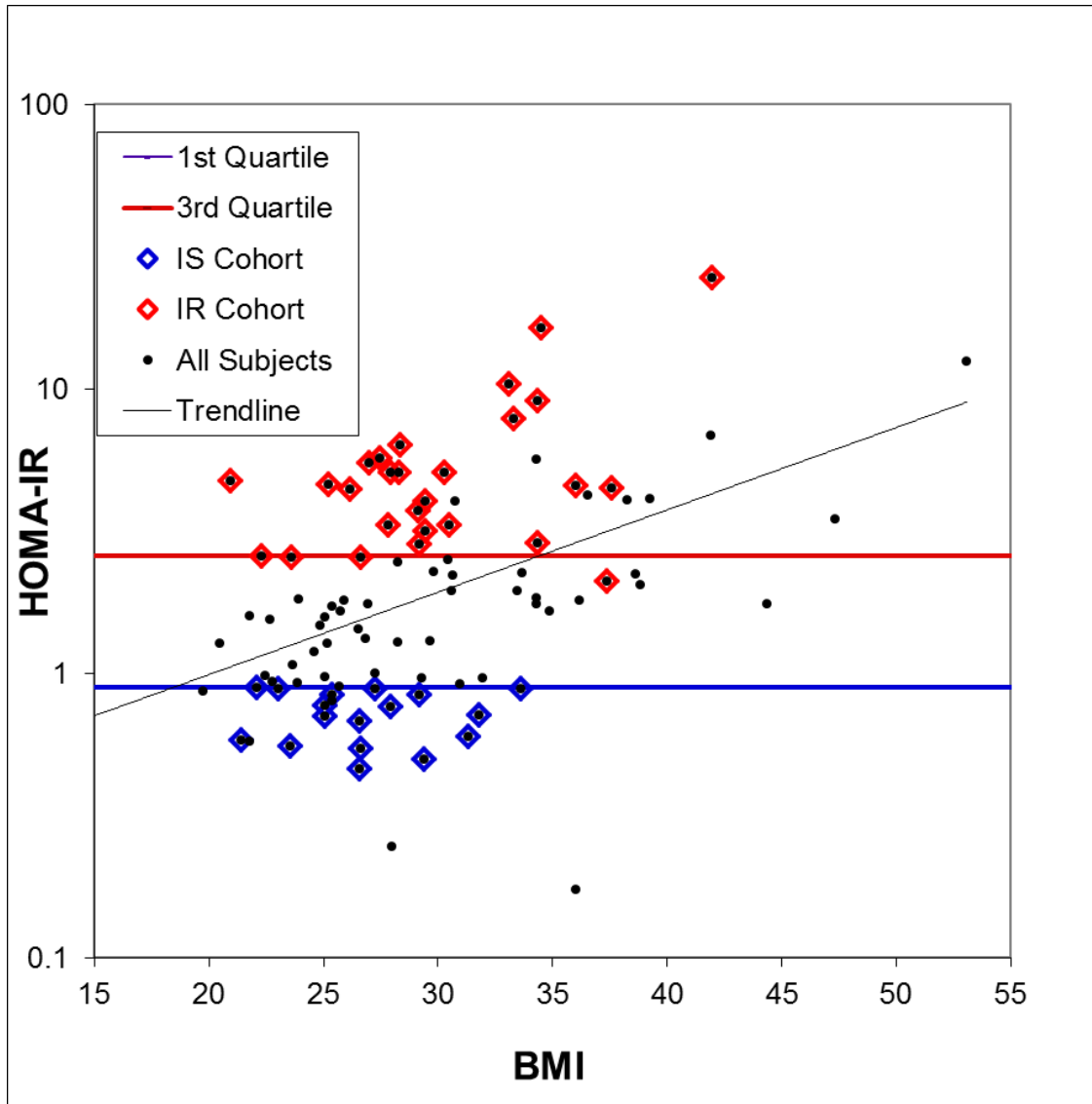
Supplemental Data

Tissue	Control	H3K27ac	H3K4me1	H3K4me3
Whole Adipose Tissue	ENCSR163VJS	ENCSR082SHT	-	-
Aorta	ENCSR116QUS, ENCSR201KGX	ENCSR322TJD, ENCSR519CFV	ENCSR325VOA, ENCSR848TLB	ENCSR957BPJ, ENCSR960EVO
CD14+ Monocyte	ENCSR000AUU, ENCSR000DWK, ENCSR444GJM	ENCSR000ASJ, ENCSR012PII	ENCSR000ASM, ENCSR400VWA	ENCSR000ASN, ENCSR000DWL, ENCSR796FCS
Endocrine Pancreas (Pancreatic Islets)	ENCSR770VVH, ENCSR928GAG	ENCSR324JDC, ENCSR492PXH	ENCSR292WQY, ENCSR817QHJ	ENCSR438NCW, ENCSR884EVT, ENCSR957UQS
Liver	ENCSR059QYS, ENCSR236NLU, ENCSR687HYO, ENCSR942ZRO	ENCSR230IMS, ENCSR678LND	ENCSR111OHT, ENCSR203RKZ, ENCSR218ZMU, ENCSR642HII	ENCSR458WIH, ENCSR520BUX, ENCSR795VEN, ENCSR803JYI
Lung	ENCSR061VJM, ENCSR494FGC, ENCSR577BCL, ENCSR724XIL	ENCSR067BMB, ENCSR540ADS, ENCSR550WUX	ENCSR356ANC, ENCSR575SWA, ENCSR953XVZ	ENCSR466DZW, ENCSR500GXT
Pancreas	ENCSR480NNC, ENCSR503BIB	ENCSR402HFW, ENCSR612BWE	ENCSR449PVI, ENCSR984UHU	ENCSR315LPR, ENCSR747VED
Peripheral Blood Mononuclear Cell (PBMC)	ENCSR585UEE, ENCSR837ART, ENCSR913DKN	ENCSR105EMQ, ENCSR156XNC, ENCSR615HXA, ENCSR625BDY	ENCSR336ZSZ, ENCSR420EWO, ENCSR482QXO	ENCSR206JRX, ENCSR275EAG, ENCSR368YPC, ENCSR443SLY
Psoas Muscle	ENCSR139WMA, ENCSR163UEW, ENCSR688CIB	ENCSR250NHD, ENCSR367WYJ, ENCSR791ISZ	ENCSR410UUH, ENCSR700NGJ	ENCSR245BEV, ENCSR949OYZ
Skeletal Muscle	ENCSR211LEQ, ENCSR268LIX, ENCSR835ARG	ENCSR329FXI	ENCSR146JFX, ENCSR668NXG, ENCSR823QYQ	ENCSR238LEG, ENCSR346KKE, ENCSR767NIF

Supplemental Table 1: Experiment identifiers for tissue samples downloaded from the ENCODE portal⁴⁰⁻⁴¹.

	Total Cohort	IS Cohort (18)	IR Cohort (27)
Age	48	47.8	51
% Female	95%	100%	100%
BMI (Body Mass Index)	28.1	26.8	30.1*
HOMA-IR	1.71	0.71	5.78*
Total Cholesterol	181	184.5	175
TG (triglycerides)	86	81	135*
HDL (high-density lipoprotein)	61	73.7	54.1*
LDL (low-density lipoprotein)	101	94.3	97.6

Supplemental Table 2: Summary of cohort metadata. Red asterisks show values that differ significantly from the IS cohort values. (**p values?**)



Supplemental Figure 1: Insulin resistance (IR) and insulin sensitivity (IS) cohort selection. Individuals in the top and bottom quartiles of HOMA-IR were selected for the IR and IS cohort respectively.

Individual peak contribution

```
library(tibble)
library(plyr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(readr)
library(ggplot2)
library(VennDiagram)

## Loading required package: grid

## Loading required package: futile.logger

knitr::opts_chunk$set(root.dir = "/Volumes/broad_rosemlab_archive/Projects/Linus-Human-Ad/Analysis/Individual-Peak-Characterization-no-137/peaks_by_tissue/for_thesis/")

#Load Files
#Create/format dataframes
countfile <- read_tsv(".././cpm.tsv")
count_table <- as.data.frame(countfile[,-c(1)])
rownames(count_table) <- countfile$Name

peakfile <- read_tsv(".././new_peaks_merged_filt_small.bed", col_types = 'ciicidciciic', col_names = FALSE)
peak_table <- as.data.frame(peakfile[,-c(1:3,5:11)])
rownames(peak_table) <- peak_table[,1]

metadata <- read.csv(".././Copy of ChIP-Seq Library Set - Sheet1.csv")
metadata$Sample <- gsub("EPI", "", metadata$Sample)
rownames(metadata) <- metadata$Sample

unlog_counts <- 2^count_table
print(head(unlog_counts))

##
##           X120_27ac X121_27ac X123_27ac X124_27ac X129_27ac
## 1|chr1:9857-10598   7.7812396 10.196485 13.177456  8.9382971  8.111676
## 6|chr1:28574-29849   9.9176616  9.317869 10.056107 12.9062681  8.514961
## 8|chr1:136444-137017 0.6643429  4.141060  1.474269  0.9794203  1.101905
## 11|chr1:171222-173171 3.7580910 10.556063  7.727491  2.1584565  3.010493
## 12|chr1:173263-174714 2.5140267  5.735821  6.588728  3.5553707  2.158456
## 13|chr1:180603-181120 4.9933222  7.160201  7.727491  6.5887281  7.464264
##           X141_27ac X142_27ac X144_27ac X157_27ac X158_27ac
## 1|chr1:9857-10598     9.382680  9.579830  6.147501  7.781240  9.126110
## 6|chr1:28574-29849    14.025692 10.338823  9.513657 13.177456 12.466633
```



```

## 8|chr1:136444-137017 1.658639 1.802501 1.164734 1.802501 1.356604
## 11|chr1:171222-173171 7.568461 6.680703 7.260153 7.210004 9.382680
## 12|chr1:173263-174714 5.979397 3.271608 5.028053 2.969047 3.410540
## 13|chr1:180603-181120 7.210004 7.412704 4.531536 4.287094 5.028053
##
## X160_27ac X163_27ac X168_27ac X170_27ac X172_27ac_1
## 1|chr1:9857-10598 7.889862 5.169411 8.8152409 12.041974 9.317869
## 6|chr1:28574-29849 7.361501 7.061624 10.7778686 18.252219 9.126110
## 8|chr1:136444-137017 1.613284 1.945310 0.8888427 3.073750 1.526259
## 11|chr1:171222-173171 6.821079 5.314743 16.4498212 16.564239 4.141060
## 12|chr1:173263-174714 3.863745 4.141060 12.7285837 6.680703 2.281527
## 13|chr1:180603-181120 5.098243 4.141060 4.5630549 6.498019 5.617780
##
## X178_27ac X180_27ac X181_27ac X188_27ac X190_27ac
## 1|chr1:9857-10598 17.267652 13.737047 13.454343 11.794154 12.640661
## 6|chr1:28574-29849 30.484416 18.507011 10.196485 21.258973 16.449821
## 8|chr1:136444-137017 4.823231 2.329467 2.657372 4.856780 2.770219
## 11|chr1:171222-173171 21.406841 7.568461 13.177456 9.447941 5.856343
## 12|chr1:173263-174714 8.693879 3.555371 7.310652 5.241574 3.138336
## 13|chr1:180603-181120 8.456144 6.020987 8.456144 5.063026 5.979397
##
## X193_27ac X195_27ac X199_27ac X202_27ac X203_27ac
## 1|chr1:9857-10598 10.338823 11.631780 11.471642 13.086433 16.795467
## 6|chr1:28574-29849 27.095850 23.102867 18.379174 26.908685 27.474094
## 8|chr1:136444-137017 1.986185 1.705270 1.815038 2.694467 2.531513
## 11|chr1:171222-173171 10.777869 10.126053 8.574188 11.876189 7.061624
## 12|chr1:173263-174714 6.773962 5.028053 4.084049 4.723971 2.989698
## 13|chr1:180603-181120 4.823231 4.890561 6.634556 7.621104 8.282119
##
## X205_27ac X206_27ac X207_27ac X209_27ac X210_27ac
## 1|chr1:9857-10598 9.849155 5.979397 11.392402 8.456144 10.056107
## 6|chr1:28574-29849 23.752377 18.252219 26.354913 18.000936 23.917588
## 8|chr1:136444-137017 1.248331 1.283426 1.729074 3.226567 1.347234
## 11|chr1:171222-173171 6.680703 10.126053 8.693879 13.547925 7.674113
## 12|chr1:173263-174714 3.863745 5.133704 3.630077 4.346939 3.944931
## 13|chr1:180603-181120 5.278032 3.630077 6.364292 3.555371 6.147501
##
## X211_27ac X212_27ac X213_27ac X214_27ac X216_27ac
## 1|chr1:9857-10598 6.147501 8.876556 17.630482 7.061624 28.640802
## 6|chr1:28574-29849 14.928528 20.677645 17.267652 13.928809 22.943284
## 8|chr1:136444-137017 1.375542 2.584706 4.438278 1.717131 1.765406
## 11|chr1:171222-173171 7.260153 18.000936 14.221483 16.111289 9.917662
## 12|chr1:173263-174714 4.594793 11.235559 6.020987 8.168097 3.052518
## 13|chr1:180603-181120 4.027822 6.020987 8.815241 3.837056 17.029923
##
## X220_27ac X221_27ac X222_27ac
## 1|chr1:9857-10598 8.339726 10.196485 13.454343
## 6|chr1:28574-29849 26.354913 15.889480 18.000936
## 8|chr1:136444-137017 1.006956 3.732132 1.569168
## 11|chr1:171222-173171 12.041974 17.148375 12.817118
## 12|chr1:173263-174714 5.314743 8.876556 5.502167
## 13|chr1:180603-181120 3.732132 7.260153 7.061624

```

Calculate the percent contribution of each sample to each peak

```

col_sumd<- colSums(unlog_counts)

rows_sumd<- rowSums(unlog_counts)

perc_cont <- ( unlog_counts/ rows_sumd) * 100;
colnames(perc_cont)<-colnames(count_table)
rownames(perc_cont)<-rownames(count_table)
print(perc_cont[1,c(1,2)])

## X120_27ac X121_27ac
## 1|chr1:9857-10598 1.889771 2.476344

```

Subsetting peaks by percent contribution

Create a list of the peaks that whose top 2 contributors contribute over 90% of the counts for the peak, whose top 3 contributors contribute over 75% of the counts, and whose top 4 contributors contribute over 50% of the counts.

```
ninety_two<- c()
seventyfive_three<- c()
fifty_four<- c()

for(p in rownames(perc_cont))
{
  sor_row <- sort(perc_cont[p,], decreasing=TRUE)

  if(sum(sor_row[1:2]) >= 90)
  {
    ninety_two <- c(ninety_two, rownames(sor_row))
    next
  }

  if(sum(sor_row[1:3]) >= 75)
  {
    seventyfive_three <- c(seventyfive_three, rownames(sor_row))
    next
  }

  if(sum(sor_row[1:4]) >= 50)
  {
    fifty_four<- c(fifty_four, rownames(sor_row))
  }
}

seventyfive_three<- c(seventyfive_three,ninety_two)
fifty_four<- c(fifty_four,seventyfive_three)

print(length(ninety_two))

## [1] 0

print(length(seventyfive_three))

## [1] 0

print(length(fifty_four))

## [1] 14
```

Extremity

Average percent contribution

```
avg_perc<- 100/ncol(perc_cont)
print(avg_perc)

## [1] 2.631579
```

Make a dataframe with the maximum percent contribution for each peak

```
max_perc <- as.data.frame(apply(perc_cont,1,max))
colnames(max_perc) <- c("Max")
rownames(max_perc) <- rownames(count_table)
```

```
print(head(max_perc))
```

```
##                               Max
## 1|chr1:9857-10598      6.955777
## 6|chr1:28574-29849    4.741401
## 8|chr1:136444-137017  6.041754
## 11|chr1:171222-173171 5.712784
## 12|chr1:173263-174714 6.489837
## 13|chr1:180603-181120 7.176264
```

calculate the extremity of each peak maximum percent contribution - average percent contribution

```
max_avg <- as.data.frame(apply(max_perc, 1, function(x){x - avg_perc}))
colnames(max_avg) <- c('Extremity')
print(head(max_avg))
```

```
##                               Extremity
## 1|chr1:9857-10598      4.324198
## 6|chr1:28574-29849    2.109822
## 8|chr1:136444-137017  3.410175
## 11|chr1:171222-173171 3.081205
## 12|chr1:173263-174714 3.858258
## 13|chr1:180603-181120 4.544685
```

Calculate the average number of counts for each peak and add the column to the dataframe with extremity

```
average_counts <- rowMeans(unlog_counts)
```

```
max_avg <- cbind(average_counts,max_avg)
print(head(max_avg))
```

```
##                               average_counts Extremity
## 1|chr1:9857-10598      10.835674  4.324198
## 6|chr1:28574-29849    16.919503  2.109822
## 8|chr1:136444-137017  2.115445  3.410175
## 11|chr1:171222-173171  9.861004  3.081205
## 12|chr1:173263-174714  5.161343  3.858258
## 13|chr1:180603-181120  6.244975  4.544685
```

Make a histogram of of extremity Peaks in the fifty_four list are highlighted in red

```
hist<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[fifty_four,], color="Red", binwidth = 0.5)
```

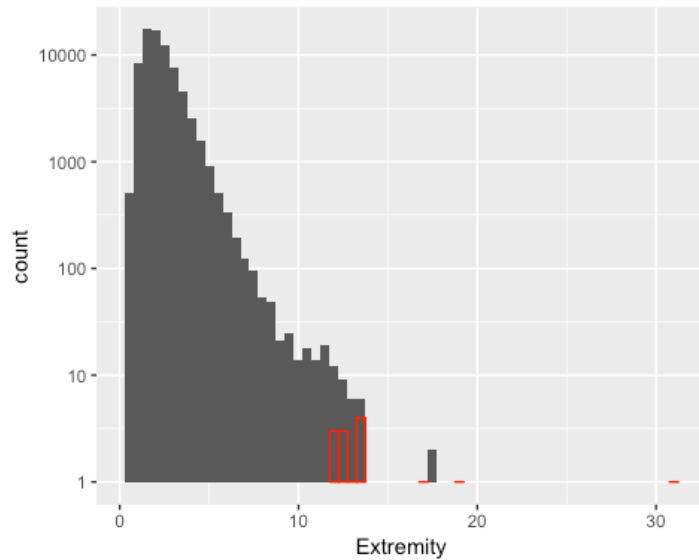
```
show(hist)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

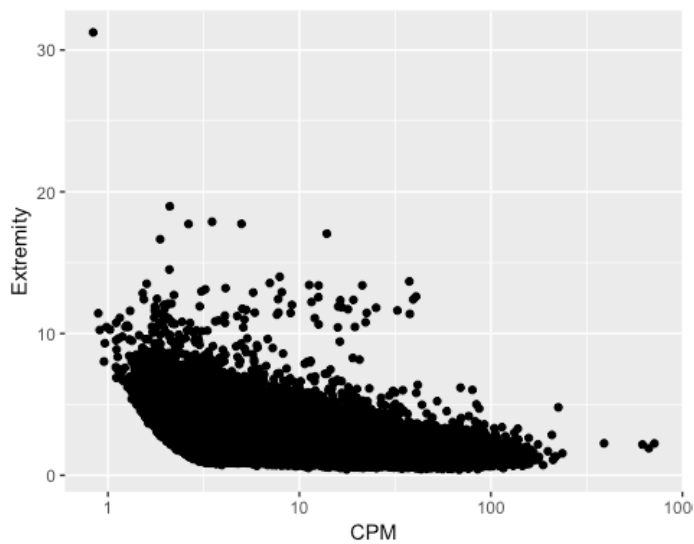
```
## Warning: Removed 27 rows containing missing values (geom_bar).
```

```
## Warning: Removed 55 rows containing missing values (geom_bar).
```

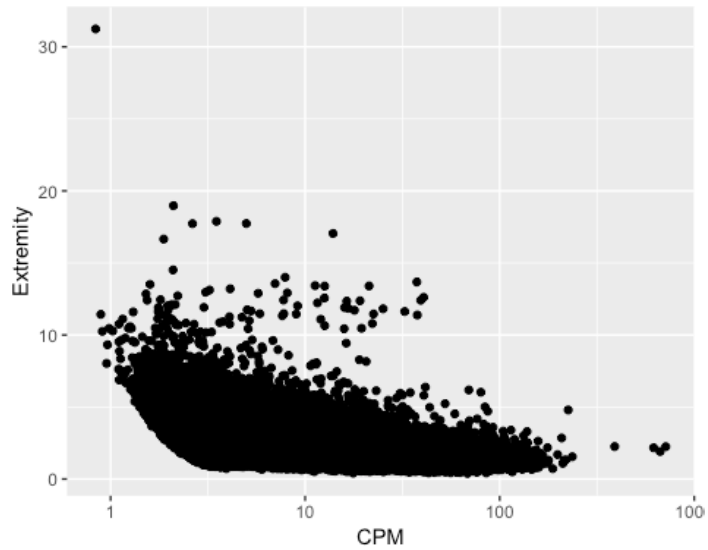


Make a scatter plot of extremity by average counts

```
sctrplot_nt<-ggplot(max_avg,aes(x=average_counts, y=Extremity)) +
  geom_point() +
  scale_x_continuous(trans="log10") +
  geom_point(data=max_avg[ninety_two,], color="Red")+
  labs(x = "CPM")
show(sctrplot_nt)
```

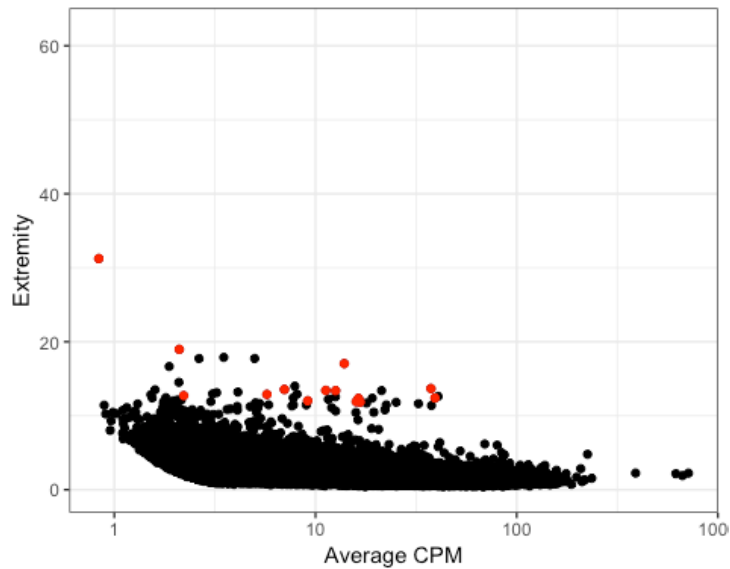


```
sctrplot_st<-ggplot(max_avg,aes(x=average_counts, y=Extremity)) +
  geom_point() +
  scale_x_continuous(trans="log10") +
  geom_point(data=max_avg[seventyfive_three,], color="Red")+
  labs(x = "CPM")
show(sctrplot_st)
```



```
sctrplot_ff<-ggplot(max_avg,aes(x=average_counts, y=Extremity)) +
  geom_point() +
  theme_bw() +
  scale_x_continuous(trans="log10") +
  geom_point(data=max_avg[fifty_four,], color="Red")+
  labs(x = "Average CPM")+
  ylim(0,62)
```

```
show(sctrplot_ff)
```



absolute cutoffs

Comparing extremity cutoff methods

Set extremity and average count cutoffs print the number of peaks in each cutoff

```
ten_eight_peaks<-subset(max_avg, Extremity>= 10 & average_counts>=8, select=c("average_counts"
, "Extremity"))
print(nrow(ten_eight_peaks))

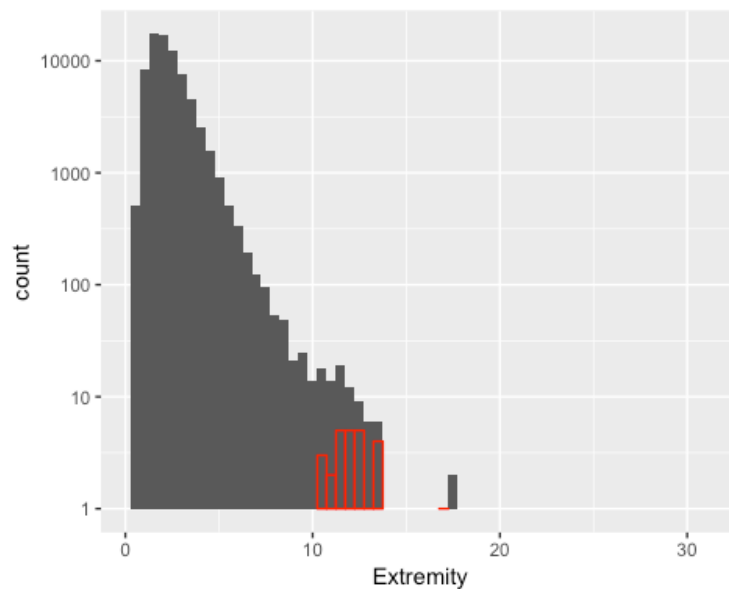
## [1] 26
```

Create a histogram of the extremity ten_eight_peaks highlighted in red

```
hist_ten_eight_peaks<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[rownames(ten_eight_peaks),], color="Red", binwidth = 0.5)

show(hist_ten_eight_peaks)

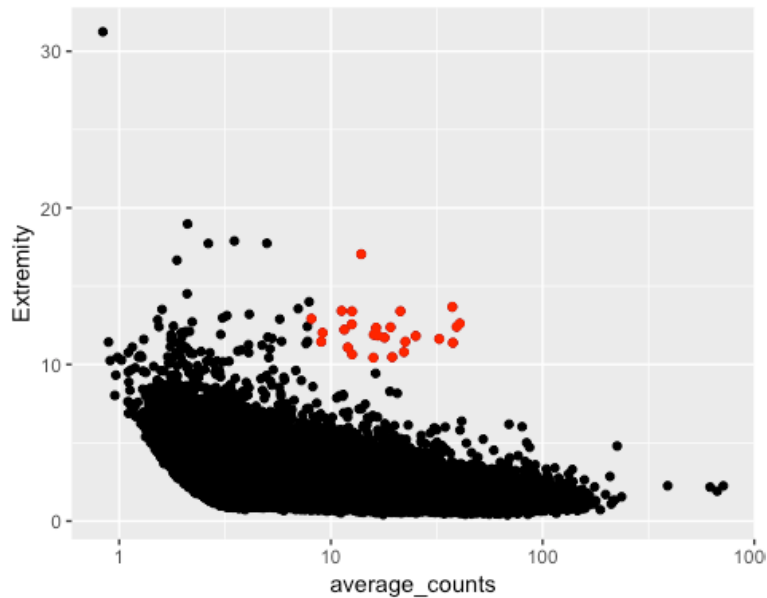
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 54 rows containing missing values (geom_bar).
```



Create a scatter plot of extremity by average counts ten_eight_peaks highlighted in red

```
sctrplot_ten_eight_peaks<-ggplot(max_avg, aes(x=average_counts, y=Extremity)) +
  geom_point() +
  scale_x_continuous(trans="log10") +
  geom_point(data=max_avg[rownames(ten_eight_peaks),], color="Red")

show(sctrplot_ten_eight_peaks)
```



sort max_avg to subset for cutoffs

```
maxavg_sorted<- max_avg[order(-max_avg$Extremity),]
print(head(maxavg_sorted))
```

##		average_counts	Extremity
##	215934 chrUn_KI270752v1:0-384	0.8372687	31.24382
##	196483 chr7:155573467-155574124	2.1031557	18.97461
##	55692 chr12:132155524-132157373	3.4994279	17.88630
##	97850 chr18:78652742-78655392	4.9849780	17.73796
##	184135 chr6:165618912-165619863	2.6357871	17.72883
##	160791 chr5:414799-415907	13.9046234	17.04539

Top n cutoffs

Comparing extremity cutoff methods

Set extremity cutoffs print the number of peaks in each cutoff

Take the top n number of peaks of the max_avg dataframe sorted by extremity (dataframe = "maxavg_sorted")

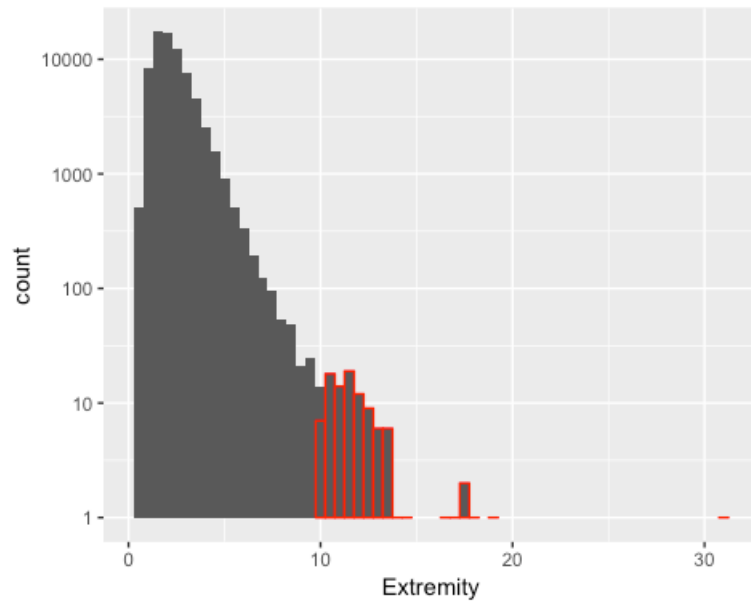
```
top_100<-rownames(maxavg_sorted[c(1:100),])
top_500<-rownames(maxavg_sorted[c(1:500),])
top_1000<-rownames(maxavg_sorted[c(1:1000),])
```

Create histograms of extremity with top n cutoff peaks highlighted in red

```
hist_top_100<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[top_100,], color="Red", binwidth = 0.5)

show(hist_top_100)
```

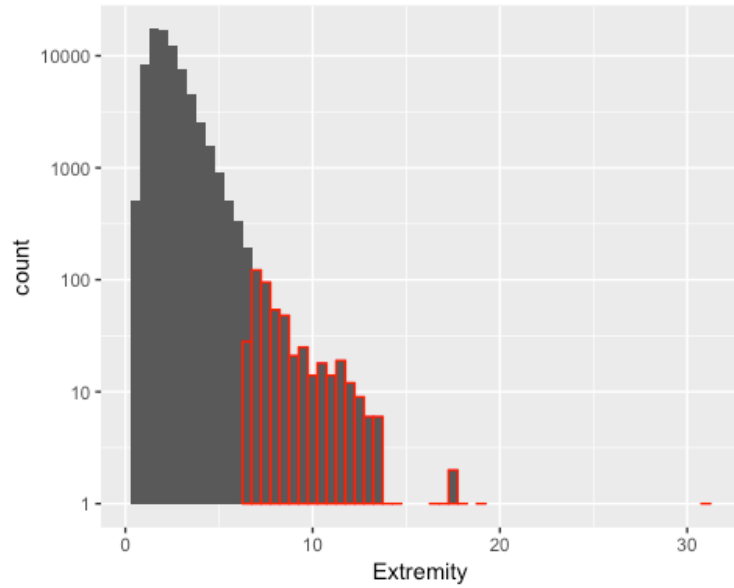
```
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 46 rows containing missing values (geom_bar).
```



```
hist_top_500<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[top_500,], color="Red", binwidth = 0.5)
```

```
show(hist_top_500)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 39 rows containing missing values (geom_bar).
```

```
hist_top_1000<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[top_1000,], color="Red", binwidth = 0.5)
```

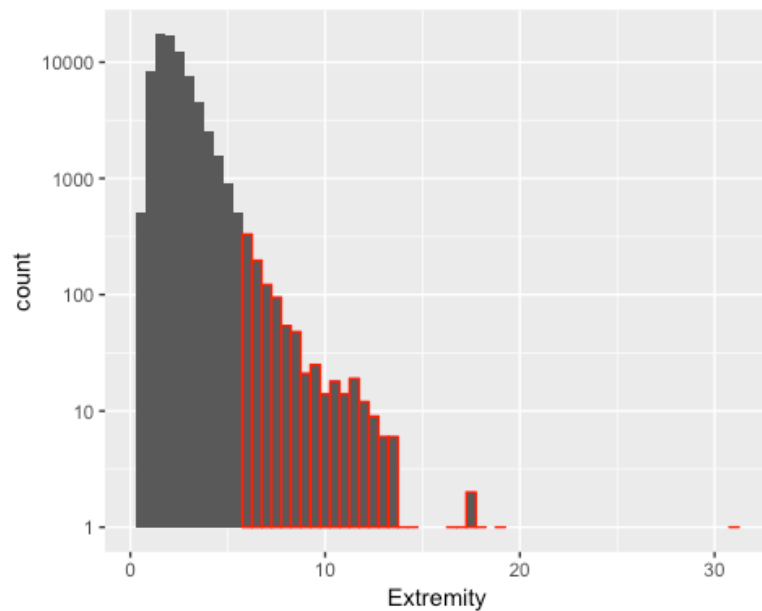
```
show(hist_top_1000)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

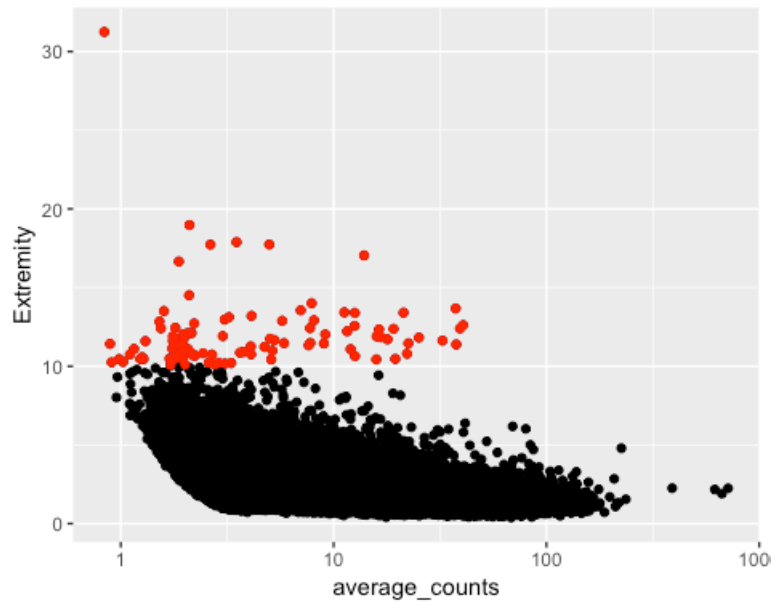
```
## Warning: Removed 27 rows containing missing values (geom_bar).
```

```
## Warning: Removed 38 rows containing missing values (geom_bar).
```

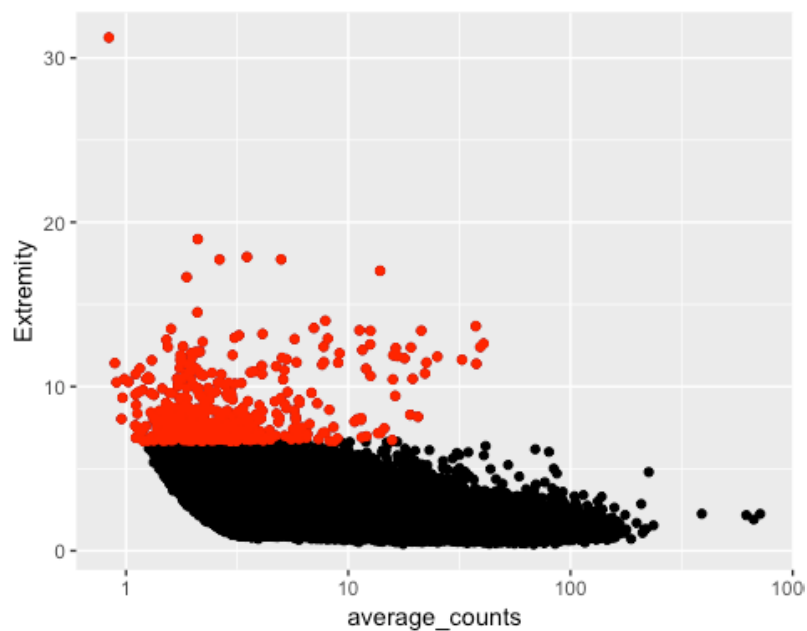


Create scatterplots of extremity by average counts

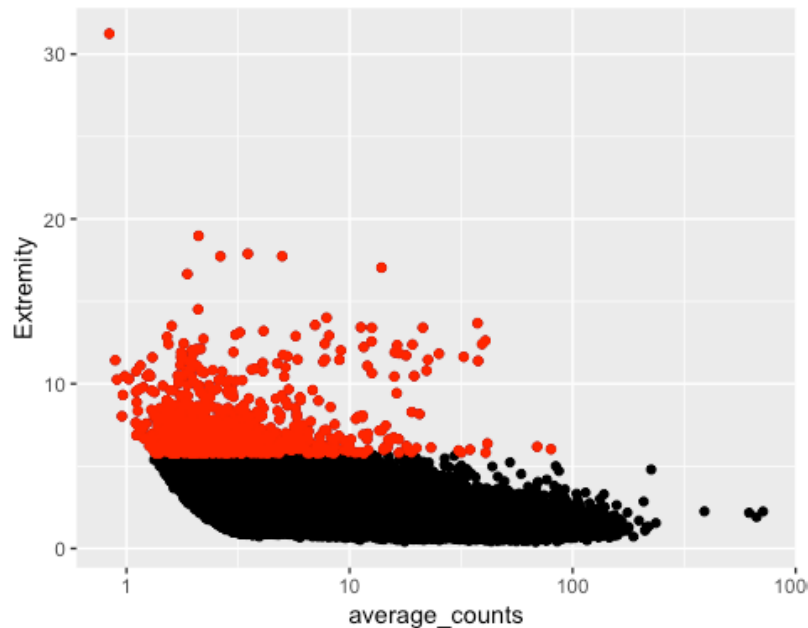
```
sctrplot_top_100<-ggplot(max_avg,aes(x=average_counts, y=Extremity)) +  
  geom_point() +  
  scale_x_continuous(trans="log10") +  
  geom_point(data=max_avg[top_100,], color="Red")  
  
show(sctrplot_top_100)
```



```
sctrplot_top_500<-ggplot(max_avg,aes(x=average_counts, y=Extremity)) +  
  geom_point() +  
  scale_x_continuous(trans="log10") +  
  geom_point(data=max_avg[top_500,], color="Red")  
  
show(sctrplot_top_500)
```



```
sctrplot_top_1000<-ggplot(max_avg,aes(x=average_counts, y=Extremity)) +
  geom_point() +
  scale_x_continuous(trans="log10") +
  geom_point(data=max_avg[top_1000,], color="Red")
show(sctrplot_top_1000)
```



percentile cutoffs

Comparing extremity cutoff methods

Set extremity cutoffs print the number of peaks in each cutoff

Take the top x percent of peaks of the max_avg dataframe sorted by extremity (dataframe = "maxavg_sorted")

```
quarter_perc<-nrow(maxavg_sorted)*0.0025
quarter_perc_extr<-rownames(maxavg_sorted[c(1:quarter_perc),])
print(length(quarter_perc_extr))

## [1] 186

half_perc<-nrow(maxavg_sorted)*0.005
half_perc_extr<-rownames(maxavg_sorted[c(1:half_perc),])
print(length(half_perc_extr))

## [1] 372

one_perc<-nrow(maxavg_sorted)*0.01
one_perc_extr<-rownames(maxavg_sorted[c(1:one_perc),])
print(length(one_perc_extr))

## [1] 744

two_perc<-nrow(maxavg_sorted)*0.02
two_perc_extr<-rownames(maxavg_sorted[c(1:two_perc),])
print(length(two_perc_extr))
```

```
## [1] 1488

three_perc<-nrow(maxavg_sorted)*0.03
three_perc_extr<-rownames(maxavg_sorted[c(1:three_perc),])
print(length(three_perc_extr))

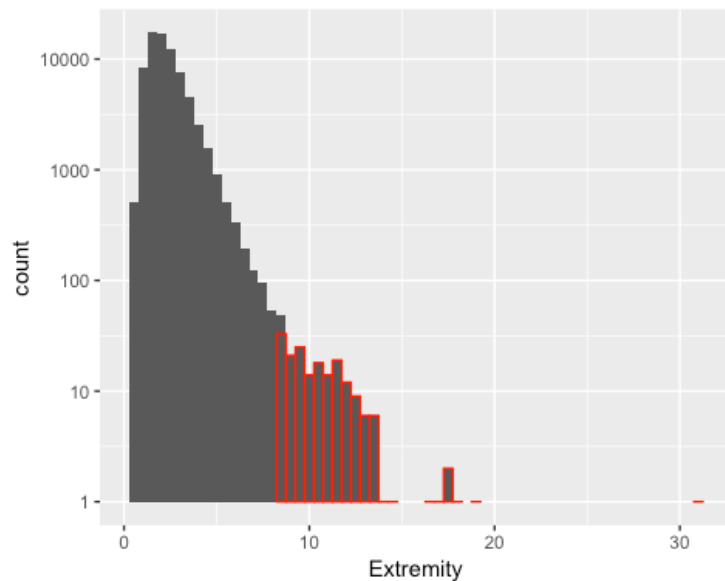
## [1] 2232
```

Create histograms of extremity with percent cutoff peaks highlighted in red

```
hist_quarter_perc<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[quarter_perc_extr,], color="Red", binwidth = 0.5)

show(hist_quarter_perc)

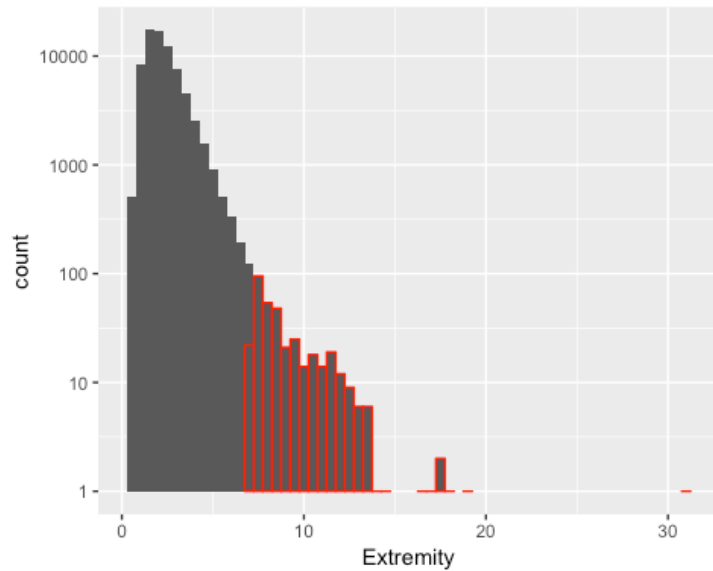
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 43 rows containing missing values (geom_bar).
```



```
hist_half_perc<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[half_perc_extr,], color="Red", binwidth = 0.5)

show(hist_half_perc)

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 40 rows containing missing values (geom_bar).
```



```
hist_one_perc<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[one_perc_extr,], color="Red", binwidth = 0.5)
```

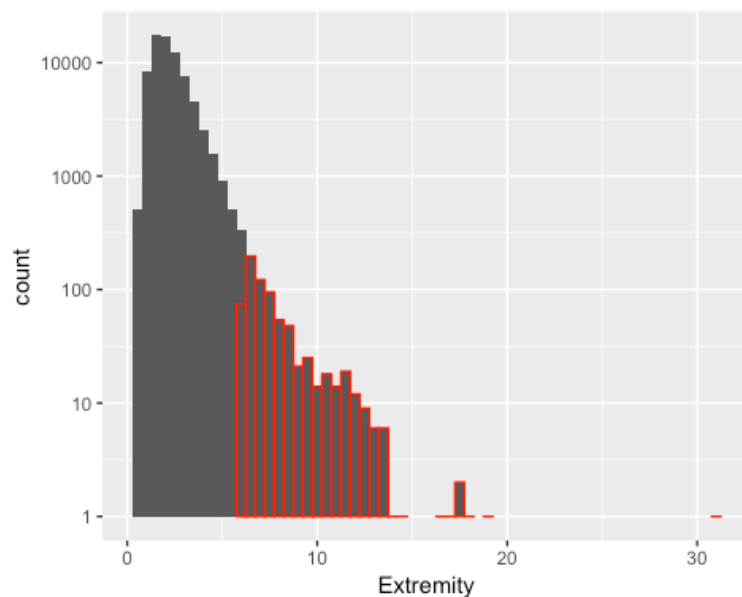
```
show(hist_one_perc)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 27 rows containing missing values (geom_bar).
```

```
## Warning: Removed 38 rows containing missing values (geom_bar).
```



```
hist_two_perc<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
```

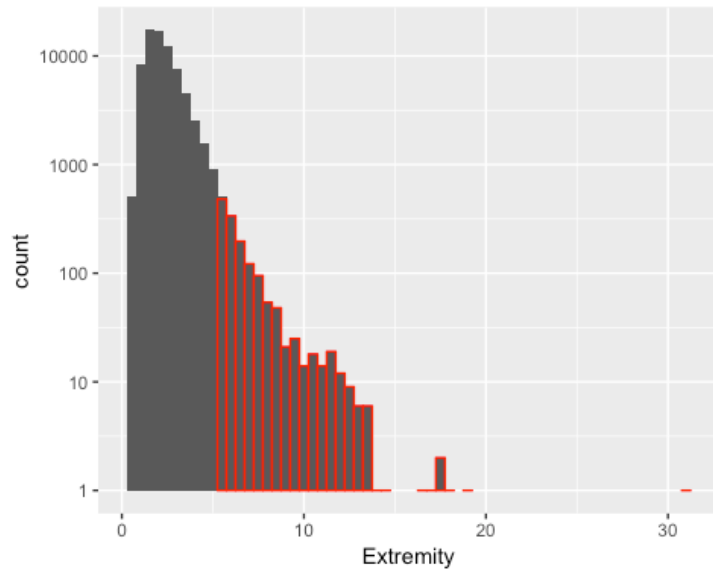
```

scale_y_continuous(trans="log10") +
geom_histogram(data=max_avg[two_perc_extr,], color="Red", binwidth = 0.5)

show(hist_two_perc)

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 37 rows containing missing values (geom_bar).

```



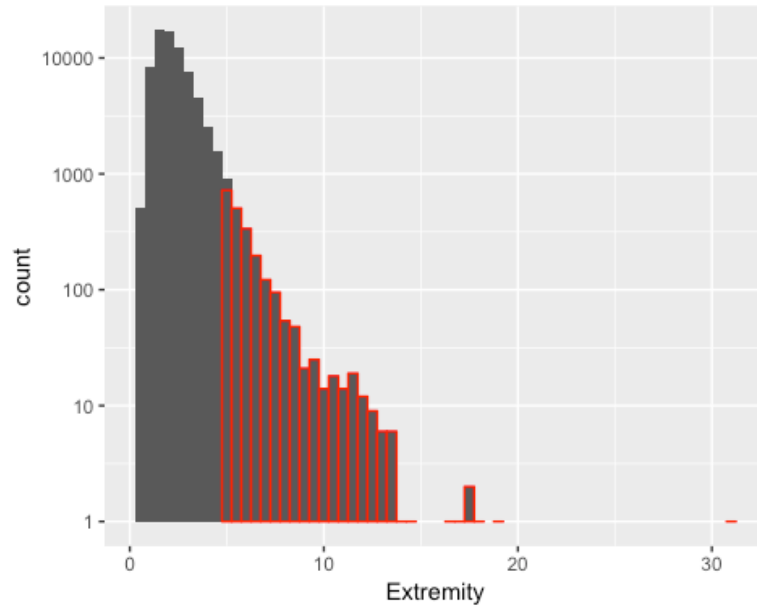
```

hist_three_perc<-ggplot(max_avg, aes(Extremity)) +
  geom_histogram(binwidth = 0.5) +
  scale_y_continuous(trans="log10") +
  geom_histogram(data=max_avg[three_perc_extr,], color="Red", binwidth = 0.5)

show(hist_three_perc)

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 27 rows containing missing values (geom_bar).
## Warning: Removed 36 rows containing missing values (geom_bar).

```



Venn Diagram: Absolute cutoffs v Top n Cutoffs

Calculate the areas of each cutoff and the overlap between cutoffs

```
print(nrow(ten_eight_peaks))
```

```
## [1] 26
```

```
print(length(top_100))
```

```
## [1] 100
```

```
print(length(top_500))
```

```
## [1] 500
```

```
print(length(top_1000))
```

```
## [1] 1000
```

```
print(length(intersect(rownames(ten_eight_peaks), top_100)))
```

```
## [1] 26
```

```
print(length(intersect(rownames(ten_eight_peaks), top_500)))
```

```
## [1] 26
```

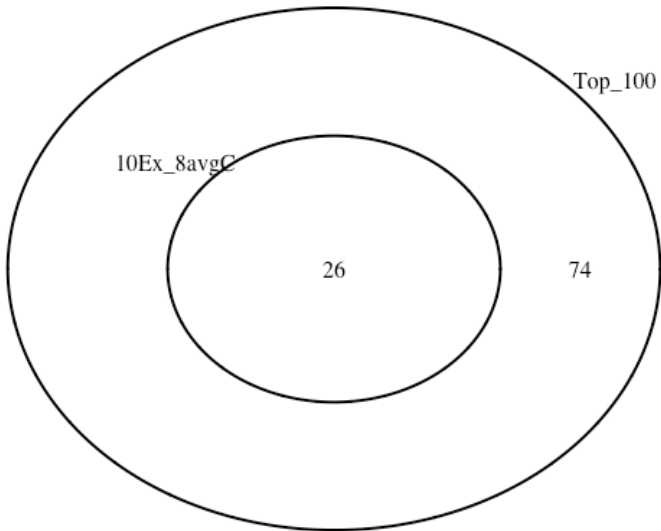
```
print(length(intersect(rownames(ten_eight_peaks), top_1000)))
```

```
## [1] 26
```

Create Venn diagram of Absolute cutoffs to top n cutoffs

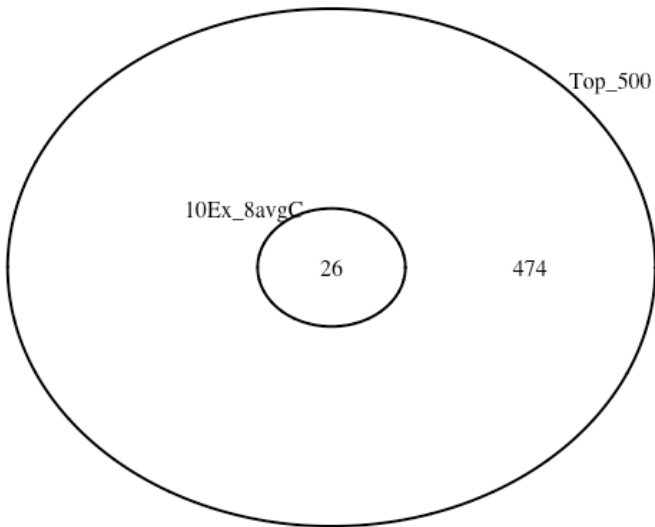
```
grid.newpage()
```

```
draw.pairwise.venn(area1 = 26, area2 = 100, cross.area = 26, category = c("10Ex_8avgC", "Top_100"))
```



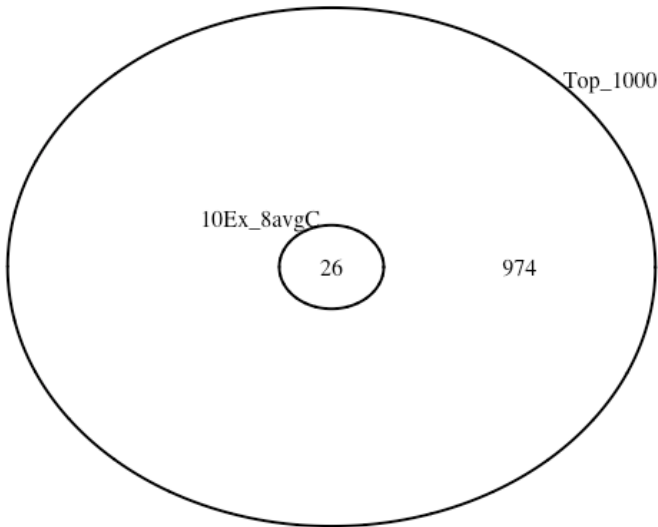
```
## (polygon[GRID.polygon.763], polygon[GRID.polygon.764], polygon[GRID.polygon.765], polygon[GRID.polygon.766], text[GRID.text.767], text[GRID.text.768], text[GRID.text.769], text[GRID.text.770])
```

```
grid.newpage()
draw.pairwise.venn(area1 = 26, area2 = 500, cross.area = 26, category = c("10Ex_8avgC", "Top_500"))
```



```
## (polygon[GRID.polygon.771], polygon[GRID.polygon.772], polygon[GRID.polygon.773], polygon[GRID.polygon.774], text[GRID.text.775], text[GRID.text.776], text[GRID.text.777], text[GRID.text.778])
```

```
grid.newpage()
draw.pairwise.venn(area1 = 26, area2 = 1000, cross.area = 26, category = c("10Ex_8avgC", "Top_1000"))
```

```
## (polygon[GRID.polygon.779], polygon[GRID.polygon.780], polygon[GRID.polygon.781], polygon[GRID.polygon.782], text[GRID.text.783], text[GRID.text.784], text[GRID.text.785], text[GRID.text.786])
```

Venn diagram: Absolute cutoffs v percent cutoffs

calculate areas of the cutoffs and their intersects

```
print(nrow(ten_eight_peaks))
## [1] 26
print(length(quarter_perc_extr))
## [1] 186
print(length(half_perc_extr))
## [1] 372
print(length(one_perc_extr))
## [1] 744
print(length(two_perc_extr))
## [1] 1488
print(length(three_perc_extr))
## [1] 2232
print(length(intersect(rownames(ten_eight_peaks), quarter_perc_extr)))
## [1] 26
print(length(intersect(rownames(ten_eight_peaks), half_perc_extr)))
## [1] 26
print(length(intersect(rownames(ten_eight_peaks), one_perc_extr)))
```

```
## [1] 26
print(length(intersect(rownames(ten_eight_peaks), two_perc_extr)))
## [1] 26
print(length(intersect(rownames(ten_eight_peaks), three_perc_extr)))
## [1] 26
```

NOTE: I think I should choose a avg count cutoff then do the overlap bc taking the top n of the sorted list gives a lot of overlap anyway.

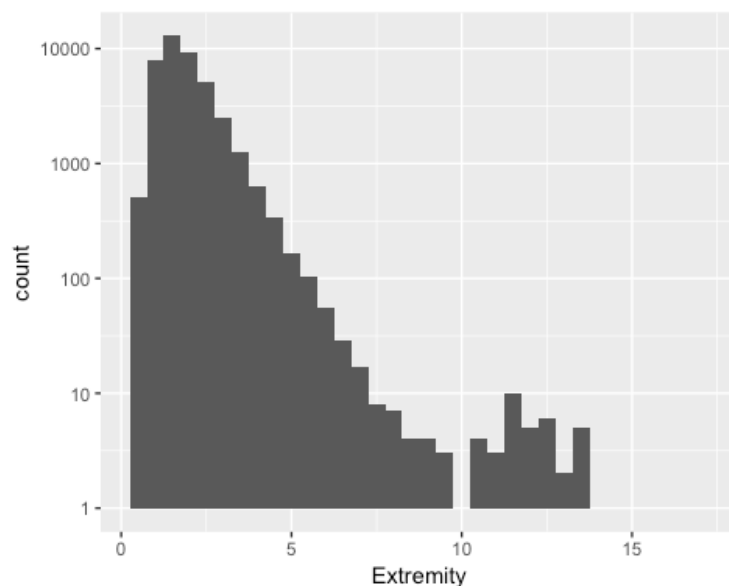
Cutoff by number of counts

Decided I should cutoff by average counts before I decide which method to cutoff extremity by

```
avgC_5<-subset(maxavg_sorted, maxavg_sorted$average_counts >= 5)
print(nrow(avgC_5))
## [1] 40822
```

Create a histogram of the peaks within the average count cutoffs

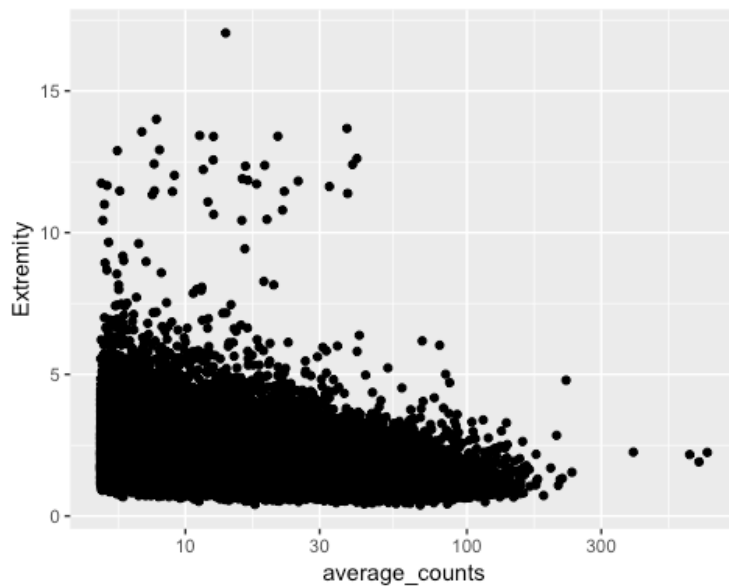
```
hist_avg5<-ggplot(data = avgC_5, aes(Extremity))+
  geom_histogram(binwidth = 0.5)+
  scale_y_continuous(trans = "log10")
show(hist_avg5)
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 6 rows containing missing values (geom_bar).
```



Create a scatter plot of the extremity by the average counts of the average count cutoff

```
sctrplot_avg5<-ggplot(data = avgC_5, aes(y=Extremity, x=average_counts))+
  geom_point()+
```

```
scale_x_continuous(trans = "log10")
show(sctrplot_avg5)
```



Finding an inflection point

Want to plot the average extremity value for different intervals (0-1000, by 50) do this for all average counts, average counts ≥ 8 , and average counts ≥ 5

Subset data for the average counts cutoffs

```
print(nrow(max_perc))
## [1] 74412

print(nrow(avgC_5))
## [1] 40822

avgC_8<-subset(maxavg_sorted, maxavg_sorted$average_counts >= 8)
print(nrow(avgC_8))
## [1] 28466
```

Create dataframes of the extremity interval and the average extremity

```
intervals<-seq(50, 1000, by = 50)

avgC_all_ex<-data.frame(matrix(data = NA, nrow = length(intervals), ncol = 2 ))
colnames(avgC_all_ex)<-c("Interval", "Average_Extremity")
avgC_all_ex$Interval<-intervals

avgC_8_ex<-data.frame(matrix(data = NA, nrow = length(intervals), ncol = 2 ))
colnames(avgC_8_ex)<-c("Interval", "Average_Extremity")
avgC_8_ex$Interval<-intervals

avgC_5_ex<-data.frame(matrix(data = NA, nrow = length(intervals), ncol = 2 ))
colnames(avgC_5_ex)<-c("Interval", "Average_Extremity")
avgC_5_ex$Interval<-intervals
```

```
print(avgC_all_ex[c(1:3),])
```

```
##   Interval Average_Extremity
## 1      50                NA
## 2     100                NA
## 3     150                NA
```

Calculate the average extremity for each interval and populate the dataframes

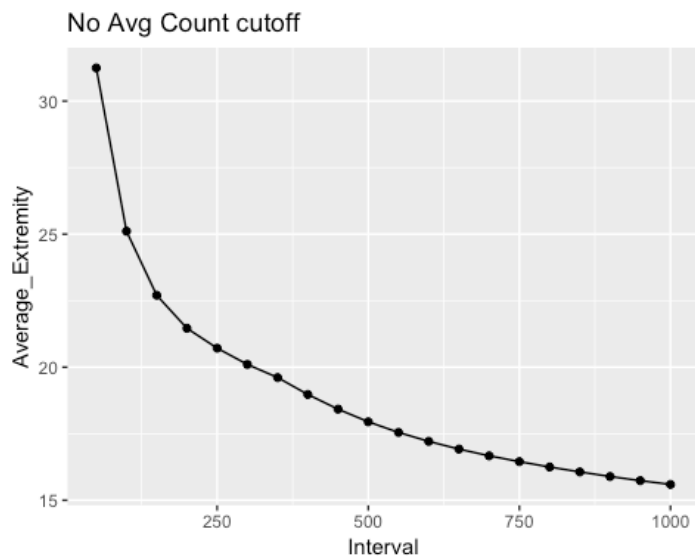
```
for(n in rownames(avgC_all_ex))
{
  int<-avgC_all_ex[n,]$Interval
  avgC_all_ex[n,]$Average_Extremity<-mean(maxavg_sorted[c(1:n),]$Extremity)

  avgC_8_ex[n,]$Average_Extremity<-mean(avgC_8[c(1:n),]$Extremity)

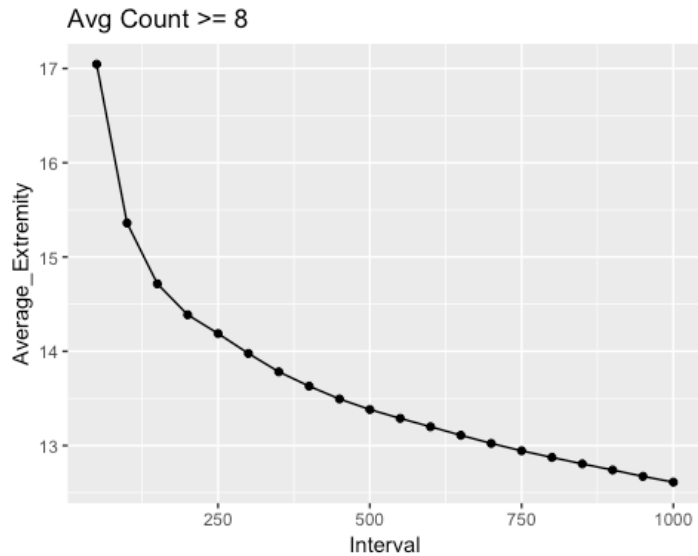
  avgC_5_ex[n,]$Average_Extremity<-mean(avgC_5[c(1:n),]$Extremity)
}
```

Create plots of the average extremity intervals for each average count cutoff

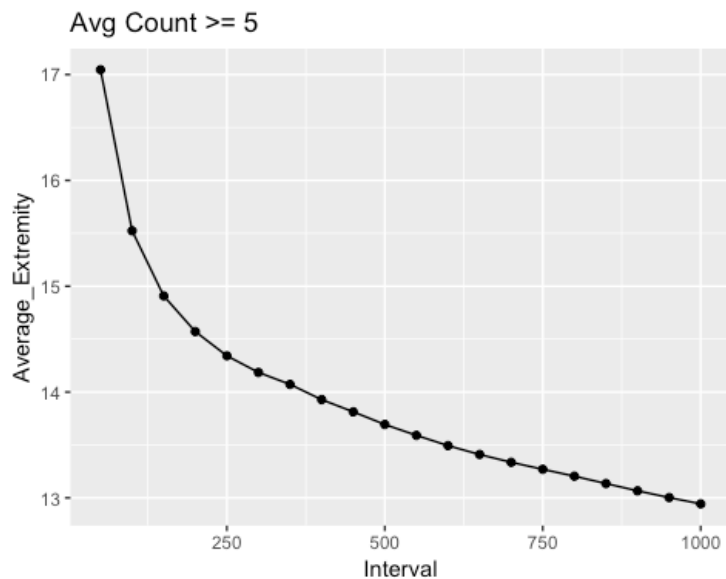
```
sctrplot_avgC_all<-ggplot(avgC_all_ex,aes(x= Interval, y=Average_Extremity)) +
  geom_point() +
  geom_line() +
  labs(title = "No Avg Count cutoff")
show(sctrplot_avgC_all)
```



```
sctrplot_8_all<-ggplot(avgC_8_ex,aes(x= Interval, y=Average_Extremity)) +
  geom_point() +
  geom_line() +
  labs(title = "Avg Count >= 8")
show(sctrplot_8_all)
```



```
sctrplot_5_all<-ggplot(avgC_5_ex,aes(x= Interval, y=Average_Extremity)) +
  geom_point() +
  geom_line() +
  labs(title = "Avg Count >= 5")
show(sctrplot_5_all)
```



It looks like the plot of average extremity intervals for peaks with an avg count of ≥ 8 may have an inflection point at interval 200

Re-do interval analysis for the same set of peaks with avg counts ≥ 8 , but with intervals of (0-300, by 1) Create dataframes of the extremity interval and the average extremity

```
intervals_2<-seq(1,300, by = 1)
avgC_8_300<-data.frame(matrix(data = NA, nrow = length(intervals_2), ncol = 2 ))
colnames(avgC_8_300)<-c("Interval", "Average_Extremity")
avgC_8_300$Interval<-intervals_2
```

```
print(avgC_8_300[c(1:3),])
```

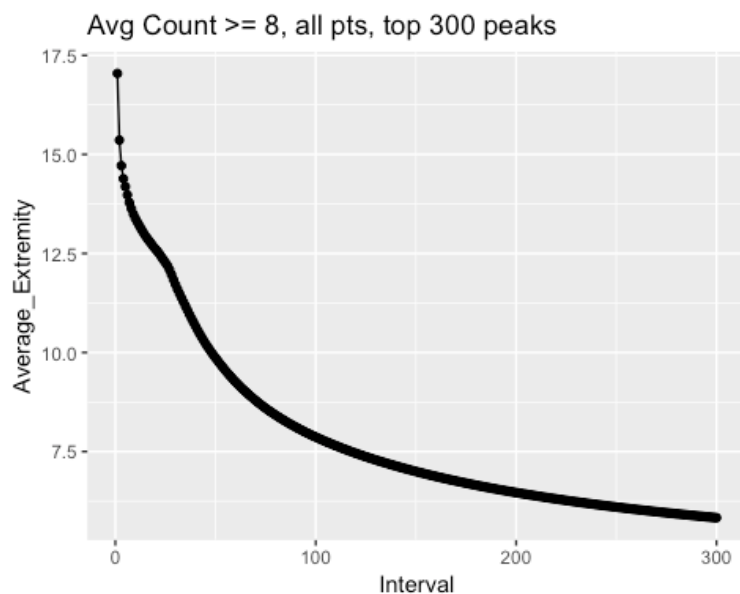
```
##   Interval Average_Extremity
## 1         1                 NA
## 2         2                 NA
## 3         3                 NA
```

Calculate the average extremity for each interval and populate the dataframes

```
for(m in rownames(avgC_8_300))
{
  avgC_8_300[m,]$Average_Extremity<-mean(avgC_8[c(1:m),]$Extremity)
}
```

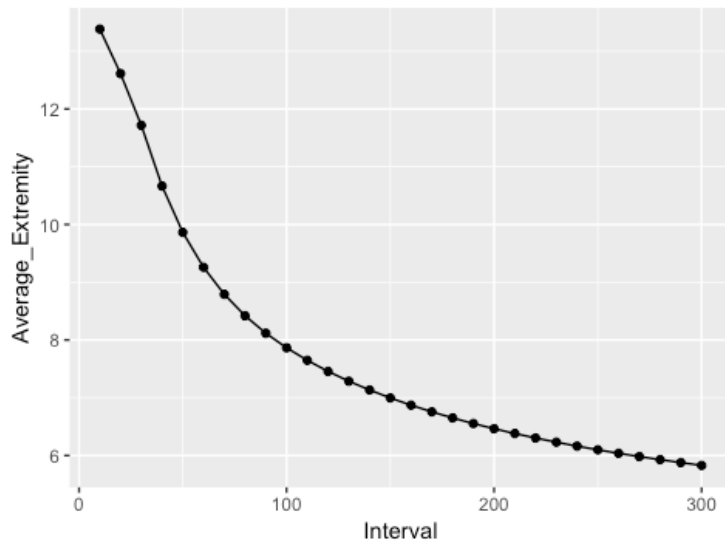
Create plot of the average extremity intervals of peaks with avg counts ≥ 8 plot all points plot only every 10th point

```
sctrplot8_300_all<-ggplot(avgC_8_300,aes(x= Interval, y=Average_Extremity)) +
  geom_point() +
  geom_line() +
  labs(title = "Avg Count  $\geq 8$ , all pts, top 300 peaks")
show(sctrplot8_300_all)
```



```
sctrplot8_300_10<-ggplot(avgC_8_300[seq(0, 300, by = 10),],aes(x= Interval, y=Average_Extremity)) +
  geom_point() +
  geom_line() +
  labs(title = "Avg Count  $\geq 8$ , every 10th pt, top 300 peaks")
show(sctrplot8_300_10)
```

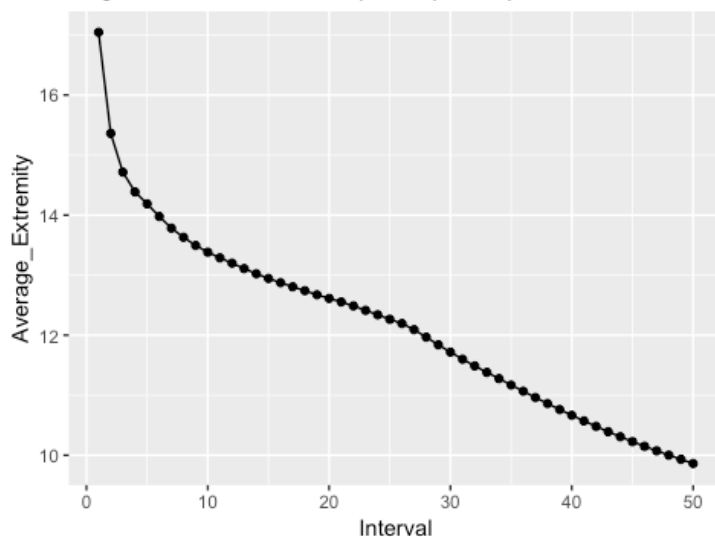
Avg Count >= 8, every 10th pt, top 300 peaks



Try to find the exact point where the curve above kinks to see which peak should be the cut off plot the average extremity for each interval for the top 50 peaks

```
sctrplot8_300_50<-ggplot(avgC_8_300[c(1:50)],,aes(x= Interval, y=Average_Extremity)) +  
  geom_point() +  
  geom_line() +  
  labs(title = "Avg Count >= 8, first 50 pts, top 300 peaks")  
  
show(sctrplot8_300_50)
```

Avg Count >= 8, first 50 pts, top 300 peaks



Try to find exact point where the curve kinks appears to be between 25 and 27 so will calculate the difference in y

```
print(avgC_8_300[23,]$Average_Extremity - avgC_8_300[24,]$Average_Extremity)  
  
## [1] 0.07374189  
  
print(avgC_8_300[24,]$Average_Extremity - avgC_8_300[25,]$Average_Extremity)
```

```
## [1] 0.07491177
print(avgC_8_300[25,]$Average_Extremity - avgC_8_300[26,]$Average_Extremity)
## [1] 0.07041103
print(avgC_8_300[26,]$Average_Extremity - avgC_8_300[27,]$Average_Extremity)
## [1] 0.102377
print(avgC_8_300[27,]$Average_Extremity - avgC_8_300[28,]$Average_Extremity)
## [1] 0.1251902
print(avgC_8_300[29,]$Average_Extremity - avgC_8_300[30,]$Average_Extremity)
## [1] 0.1227083
```


Adipocyte extremity

```
knitr::opts_chunk$set(root.dir = "/Volumes/broad_rosenlab_archive/Projects/Linus-Human-Ad/Analysis/Individual-Peak-Characterization-no-137/peaks_by_tissue/for_thesis/")
library(tibble)
library(plyr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(readr)
library(tidyr)
library(ggplot2)
library(VennDiagram)

## Loading required package: grid

## Loading required package: futile.logger

Library(pheatmap)
Library(RColorBrewer)
Library(reshape2)
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##   smiths
```

adipocytes

```
#Load cmp files
ac_countfile <- read_tsv("../H3K27ac/cpm-Linus.tsv")
ac_count_table <- as.data.frame(ac_countfile[,-c(61)])
rownames(ac_count_table) <- ac_countfile$name

#some of the tissues have multiple samples. The samples will be averaged for each tissue
#only have adipose tissue sample for H3K27ac
ac_adipose_samples<-ac_count_table[,grepl("adipose", colnames(ac_count_table))]
ac_scadip_rosen_samples<-ac_count_table[,grepl("Linus", colnames(ac_count_table))]
ac_CD14_samples<-ac_count_table[,grepl("CD14", colnames(ac_count_table))]
ac_aorta_samples<-ac_count_table[,grepl("aorta", colnames(ac_count_table))]
ac_endo_samples<-ac_count_table[,grepl("endocrine", colnames(ac_count_table))]
ac_liver_samples<-ac_count_table[,grepl("liver", colnames(ac_count_table))]
ac_lung_samples<-ac_count_table[,grepl("lung", colnames(ac_count_table))]
ac_panc_samples<-ac_count_table[,grepl("^pancreas_", colnames(ac_count_table))]
ac_pbmc_samples<-ac_count_table[,grepl("peripheral", colnames(ac_count_table))]
```

```

ac_psoas_samples<-ac_count_table[,grep1("psoas", colnames(ac_count_table))]
ac_skel_samples<-ac_count_table[,grep1("skeletal", colnames(ac_count_table))]

#make dataframes for the sample averages
ac_avg<-data.frame(matrix(data = NA, nrow = nrow(ac_count_table), ncol = 11 ))
  rownames(ac_avg)<-rownames(ac_count_table)
  colnames(ac_avg)<-c("ac_adipose_samples","ac_aorta_samples","ac_CD14_samples","ac_endo_samp
les","ac_scadip_rosen_samples","ac_liver_samples","ac_lung_samples","ac_panc_samples","ac_pbmc
samples","ac_psoas_samples" ,"ac_skel_samples" )

#Set colors for each tissue
ac_tiss_colors<-c("green3","orange1","firebrick3", "coral","yellow1", "maroon2","slateblue1","
hotpink1", "purple","pink1","blue")
names(ac_tiss_colors)<-c("ac_adipose_samples","ac_scadip_rosen_samples","ac_CD14_samples","ac
aorta_samples","ac_endo_samples","ac_liver_samples","ac_lung_samples","ac_panc_samples","ac_pb
mc_samples","ac_psoas_samples" ,"ac_skel_samples" )

#take the average cpm for each peak of each tissue and put into dataframe
ac_avg[,"ac_adipose_samples"]<-apply(ac_adipose_samples, 1, mean)
ac_avg[,"ac_aorta_samples"]<-apply(ac_aorta_samples, 1, mean)
ac_avg[,"ac_CD14_samples"]<-apply(ac_CD14_samples, 1, mean)
ac_avg[,"ac_endo_samples"]<-apply(ac_endo_samples, 1, mean)
ac_avg[,"ac_scadip_rosen_samples"]<-apply(ac_scadip_rosen_samples, 1, mean)
ac_avg[,"ac_liver_samples"]<-apply(ac_liver_samples, 1, mean)
ac_avg[,"ac_lung_samples"]<-apply(ac_lung_samples, 1, mean)
ac_avg[,"ac_panc_samples"]<-apply(ac_panc_samples, 1, mean)
ac_avg[,"ac_pbmc_samples"]<-apply(ac_pbmc_samples, 1, mean)
ac_avg[,"ac_psoas_samples"]<-apply(ac_psoas_samples, 1, mean)
ac_avg[,"ac_skel_samples"]<-ac_skel_samples

#calculate percent contribution of each tissue to the peak
ac_rows_sumd<- rowSums(ac_avg)
ac_perc_cont <- ( ac_avg/ ac_rows_sumd) * 100;
  colnames(ac_perc_cont)<-colnames(ac_avg)
  rownames(ac_perc_cont)<-rownames(ac_avg)
  head(ac_perc_cont)

##          ac_adipose_samples ac_aorta_samples ac_CD14_samples
## 1|chr1:9607-10848          1.436791          1.201260          1.250284
## 2|chr1:11688-12478          5.298489          4.429913          4.610702
## 7|chr1:20370-20883          5.954186          4.978123          5.181285
## 8|chr1:21105-22873          3.009118          2.515837          2.618510
## 10|chr1:28463-30061         1.322220          1.105470          1.499070
## 11|chr1:34505-34972         6.140357          5.133775          5.343289
##          ac_endo_samples ac_scadip_rosen_samples
## 1|chr1:9607-10848          7.192398          54.51353
## 2|chr1:11688-12478          8.136700          42.34462
## 7|chr1:20370-20883          9.143631          35.20966
## 8|chr1:21105-22873          4.620995          67.25635
## 10|chr1:28463-30061         2.030487          85.26380
## 11|chr1:34505-34972         9.429527          33.18384
##          ac_liver_samples ac_lung_samples ac_panc_samples
## 1|chr1:9607-10848          1.250211          1.438300          2.118103
## 2|chr1:11688-12478          4.610434          5.304052          7.810977
## 7|chr1:20370-20883          5.180983          5.960438          8.777599
## 8|chr1:21105-22873          2.618358          3.012278          4.436011
## 10|chr1:28463-30061         1.150519          1.323609          1.949204
## 11|chr1:34505-34972         5.342978          6.146805          9.052051
##          ac_pbmc_samples ac_psoas_samples ac_skel_samples
## 1|chr1:9607-10848          26.072273          2.305883          1.220966
## 2|chr1:11688-12478          4.448074          8.503459          4.502583

```

```

## 7|chr1:20370-20883      4.998532      9.555778      5.059786
## 8|chr1:21105-22873      2.526151      4.829286      2.557107
## 10|chr1:28463-30061     1.110002      2.122010      1.123605
## 11|chr1:34505-34972     5.154822      9.854561      5.217991

#calculate the average percent contribution
ac_avg_perc<- 100/ncol(ac_perc_cont)
print(ac_avg_perc)

## [1] 9.090909

#Make a dataframe with the adipocyte percent contribution for each peak
ac_adip_perc <- as.data.frame(ac_perc_cont$ac_scadip_rosen_samples)
  colnames(ac_adip_perc) <- c("adipocytes")
  rownames(ac_adip_perc) <- rownames(ac_count_table)
head(ac_adip_perc)

##                adipocytes
## 1|chr1:9607-10848      54.51353
## 2|chr1:11688-12478     42.34462
## 7|chr1:20370-20883     35.20966
## 8|chr1:21105-22873     67.25635
## 10|chr1:28463-30061    85.26380
## 11|chr1:34505-34972    33.18384

#calculate the adipocyte extremity of each peak
#adipocyte percent contribution - average percent contribution
ac_adip_avg <- as.data.frame(apply(ac_adip_perc, 1, function(x){x - ac_avg_perc}))
  colnames(ac_adip_avg) <- c('Extremity')
  head(ac_adip_avg)

##                Extremity
## 1|chr1:9607-10848      45.42262
## 2|chr1:11688-12478     33.25371
## 7|chr1:20370-20883     26.11875
## 8|chr1:21105-22873     58.16544
## 10|chr1:28463-30061    76.17290
## 11|chr1:34505-34972    24.09293

#Calculate the average number of counts for each peak and add the column to the dataframe with
#extremity
ac_average_counts <- rowMeans(ac_count_table)
ac_adip_avg <- cbind(ac_average_counts,ac_adip_avg)
  head(ac_adip_avg)

##                ac_average_counts Extremity
## 1|chr1:9607-10848      1.6769201  45.42262
## 2|chr1:11688-12478     0.3606192  33.25371
## 7|chr1:20370-20883     0.2737637  26.11875
## 8|chr1:21105-22873     0.9606769  58.16544
## 10|chr1:28463-30061    2.7223289  76.17290
## 11|chr1:34505-34972     0.2524841  24.09293

#Make a table to see which tissues are the top three contributors to each peak
ac_conts<-data.frame(matrix(data = NA, nrow = length(rownames(ac_perc_cont)), ncol = 3))
  colnames(ac_conts)<- c("cont 1","cont 2","cont 3")
  rownames(ac_conts)<-rownames(ac_perc_cont)

for(o in rownames(ac_perc_cont))
{
  sor_row <- sort(ac_perc_cont[o,], decreasing=TRUE)

```

```

ac_conts[o,1]<-colnames(sor_row[1])
ac_conts[o,2]<-colnames(sor_row[2])
ac_conts[o,3]<-colnames(sor_row[3])
}

head(ac_conts)

##                               cont 1          cont 2
## 1|chr1:9607-10848  ac_scadip_rosen_samples  ac_pbmc_samples
## 2|chr1:11688-12478  ac_scadip_rosen_samples  ac_psoas_samples
## 7|chr1:20370-20883  ac_scadip_rosen_samples  ac_psoas_samples
## 8|chr1:21105-22873  ac_scadip_rosen_samples  ac_psoas_samples
## 10|chr1:28463-30061 ac_scadip_rosen_samples  ac_psoas_samples
## 11|chr1:34505-34972 ac_scadip_rosen_samples  ac_psoas_samples
##                               cont 3
## 1|chr1:9607-10848  ac_endo_samples
## 2|chr1:11688-12478  ac_endo_samples
## 7|chr1:20370-20883  ac_endo_samples
## 8|chr1:21105-22873  ac_endo_samples
## 10|chr1:28463-30061 ac_endo_samples
## 11|chr1:34505-34972 ac_endo_samples

#Create name and group column for each peak in adip_avg table to top contributing tissue
ac_adip_avg$name<-rownames(ac_adip_avg)

ac_adip_avg$group<-ac_conts$`cont 1`

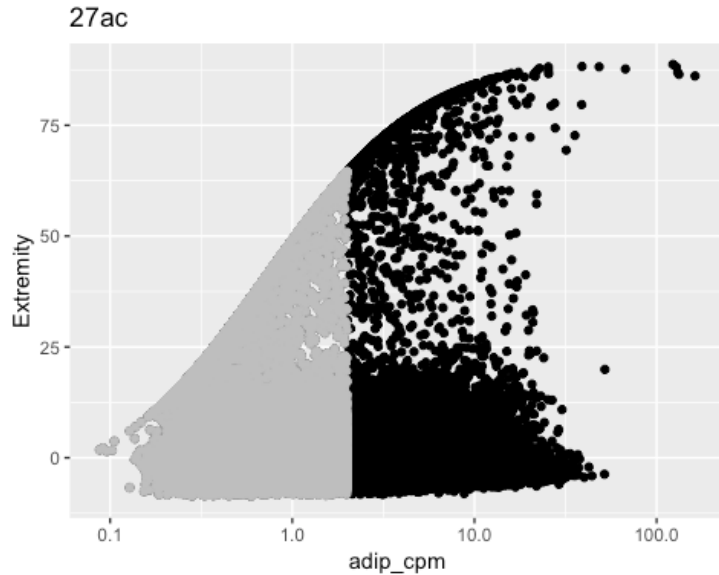
#Add adipocyte counts to adip_avg data frames
ac_adip_avg$adip_cpm<-ac_avg$ac_scadip_rosen_samples

#Calculate absolute extremity
##this is the extremity where the average percent contribution is subtracted from the highest
percent contribution, regardless of which tissue is the top contributor
ac_max_perc <- as.data.frame(apply(ac_perc_cont,1,max))
  colnames(ac_max_perc) <- c("Max")
  rownames(ac_max_perc) <- rownames(ac_count_table)
  head(ac_max_perc)

##                               Max
## 1|chr1:9607-10848  54.51353
## 2|chr1:11688-12478  42.34462
## 7|chr1:20370-20883  35.20966
## 8|chr1:21105-22873  67.25635
## 10|chr1:28463-30061 85.26380
## 11|chr1:34505-34972 33.18384

#setting cpm cutoff to the cpm cutoff used in DEA, 2
ac_cpm_sctr<-ggplot(ac_adip_avg,aes(x=adip_cpm, y=Extremity)) +
  geom_point() +
  scale_x_continuous(trans="log10") +
  geom_point(data= filter(ac_adip_avg, adip_cpm < 2), color="Grey")+
  labs(title = "27ac")
show(ac_cpm_sctr)

```

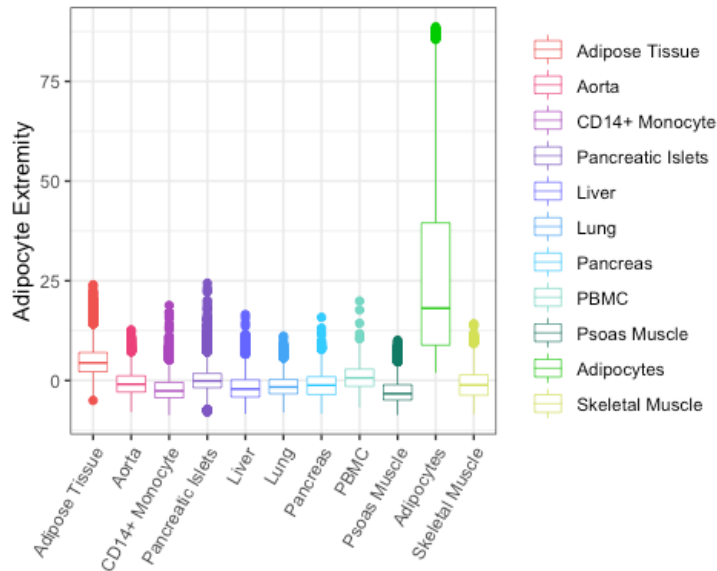


```
#calculating quartiles after removing peaks with cpm Less than 2
ac_forqrts<-filter(ac_adip_avg,adip_cpm > 2)
rownames(ac_forqrts)<-ac_forqrts$name

ac_quarts<-quantile(ac_forqrts[, "Extremity"], probs = c(0.25, 0.35, 0.5, 0.65, 0.75, 0.7, 0.8,
0.9))
print(ac_quarts)

##      25%      35%      50%      65%      75%      70%      80%
## -3.759073 -2.840554 -1.046131  1.438466  3.699154  2.462913  5.189570
##      90%
##  9.920687

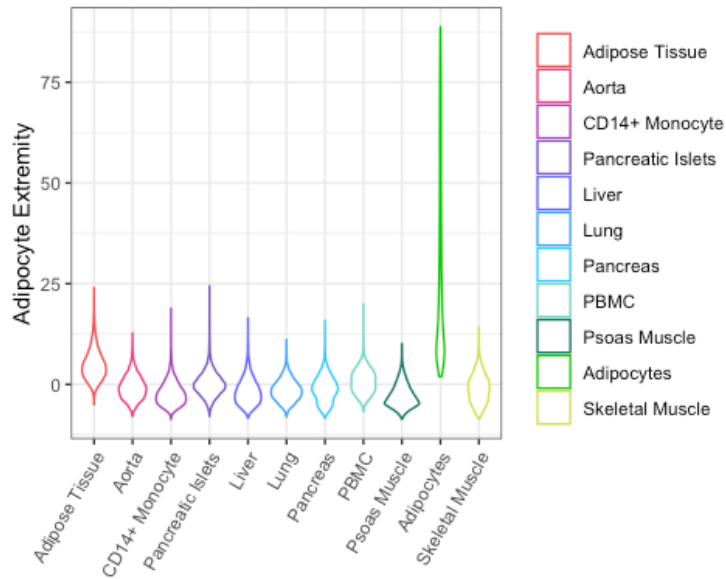
#make boxplots for adipocyte extremity, color/tissue from tissue with the highest contribution
to the peak
ac_box<-ggplot(ac_adip_avg, aes(x= group, y = Extremity, color = group))+
  geom_boxplot(size = 0.25)+
  theme_bw()+
  labs(y = "Adipocyte Extremity", x = NULL , color=NULL)+
  scale_x_discrete(labels=c("ac_adipose_samples" = "Adipose Tissue", "ac_aorta_samples" = "Aor
ta", "ac_CD14_samples" = "CD14+ Monocyte", "ac_endo_samples" = "Pancreatic Islets", "ac_liver
_samples" = "Liver", "ac_lung_samples" = "Lung", "ac_panc_samples" = "Pancreas", "ac_pbmc_samp
les" = "PBMC", "ac_psoas_samples" = "Psoas Muscle", "ac_scadip_rosen_samples" = "Adipocytes", "
ac_skel_samples" = "Skeletal Muscle"))+
  scale_color_manual(values = c("ac_adipose_samples" = "#EF5350", "ac_aorta_samples" = "#EC407
A", "ac_CD14_samples" = "#AB47BC", "ac_endo_samples" = "#7E57C2", "ac_liver_samples" = "#6666
FF", "ac_lung_samples" = "#42A5F5", "ac_panc_samples" = "#33CCFF", "ac_pbmc_samples" = "#76D7C
4", "ac_psoas_samples" = "#117A65", "ac_scadip_rosen_samples" = "#00CC00", "ac_skel_samples" =
"#D4E157"),
  labels = c( "Adipose Tissue", "Aorta", "CD14+ Monocyte", "Pancreatic Isle
ts", "Liver", "Lung", "Pancreas", "PBMC", "Psoas Muscle", "Adipocytes", "Skeletal Muscle"))+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
show(ac_box)
```



```

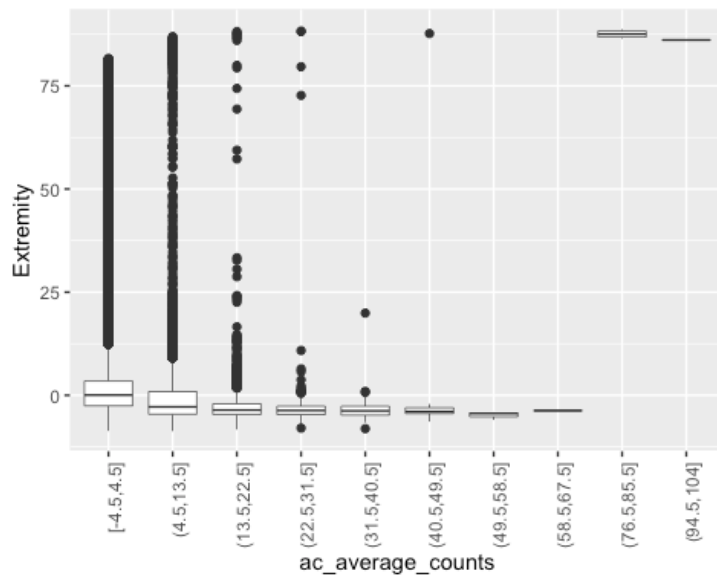
#violin plot adipocyte extremity, by tissue
ac_viol<-ggplot(ac_adip_avg,aes(x= group, y = Extremity, color = group))+
  geom_violin()+
  theme_bw()+
  labs(y = "Adipocyte Extremity", x = NULL , color=NULL)+
  scale_x_discrete(labels=c("ac_adipose_samples" = "Adipose Tissue", "ac_aorta_samples" = "Aor
ta", "ac_CD14_samples" = "CD14+ Monocyte", "ac_endo_samples" = "Pancreatic Islets", "ac_liver
_samples" = "Liver", "ac_lung_samples" = "Lung", "ac_panc_samples" = "Pancreas", "ac_pbmc_samp
les" = "PBMC", "ac_psoas_samples" = "Psoas Muscle", "ac_scadip_rosen_samples" = "Adipocytes", "
ac_skel_samples" = "Skeletal Muscle"))+
  scale_color_manual(values = c("ac_adipose_samples" = "#EF5350", "ac_aorta_samples" = "#EC407
A", "ac_CD14_samples" = "#AB47BC", "ac_endo_samples" = "#7E57C2", "ac_liver_samples" = "#6666
FF", "ac_lung_samples" = "#42A5F5", "ac_panc_samples" = "#33CCFF", "ac_pbmc_samples" = "#76D7C
4", "ac_psoas_samples" = "#117A65", "ac_scadip_rosen_samples" = "#00CC00", "ac_skel_samples" =
"#D4E157"),
                    labels = c("Adipose Tissue", "Aorta", "CD14+ Monocyte", "Pancreatic Isle
ts", "Liver", "Lung", "Pancreas", "PBMC", "Psoas Muscle", "Adipocytes", "Skeletal Muscle"))+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
show(ac_viol)

```



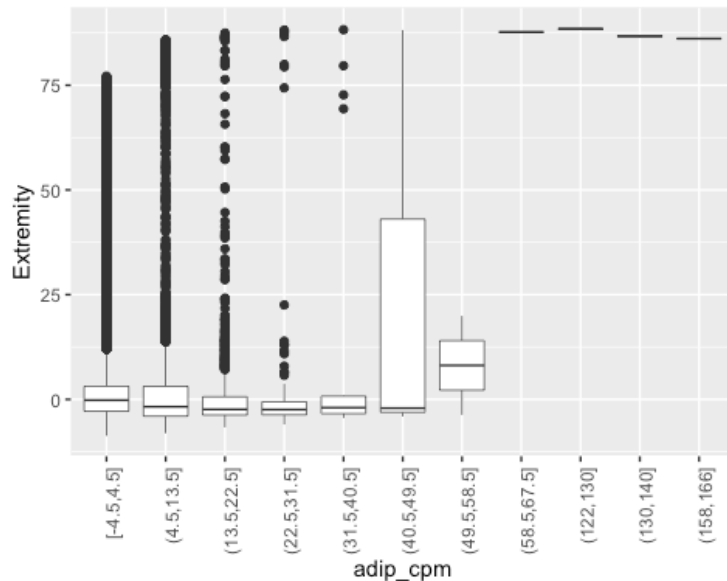
#boxplot adipocyte extremity, bin by count

```
ac_c_box<-ggplot(ac_adip_avg, aes(x= ac_average_counts, y = Extremity))+
  geom_boxplot(size = 0.25,aes(cut_width(ac_average_counts, 9)))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
show(ac_c_box)
```



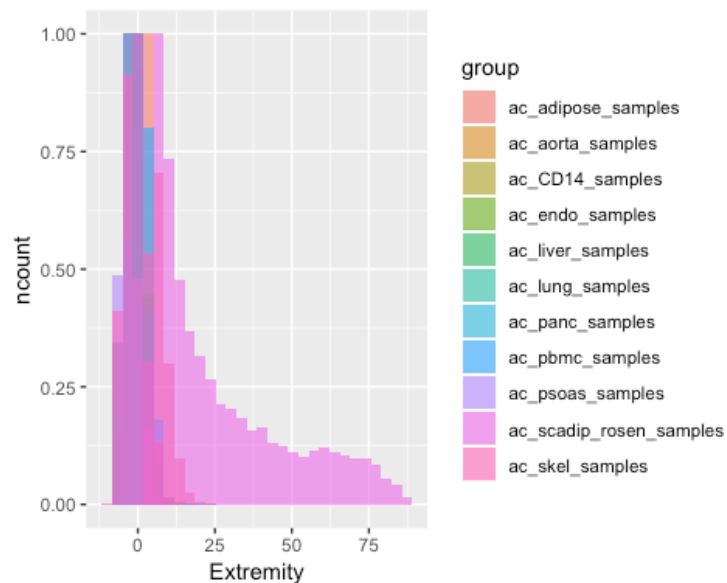
#boxplot adipocyte extremity, bin by count, adipocyte cpm

```
ac_c_box<-ggplot(ac_adip_avg, aes(x= adip_cpm, y = Extremity))+
  geom_boxplot(size = 0.25,aes(cut_width(adip_cpm, 9)))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
show(ac_c_box)
```



```
#make histogram of adipocyte extremity, colored by tissue with the top contribution to a peak
ac_adip_hist<-ggplot(ac_adip_avg, aes(Extremity, fill = group))+
  geom_histogram(aes(y = ..ncount..),position = "identity", alpha = 0.6)
show(ac_adip_hist)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

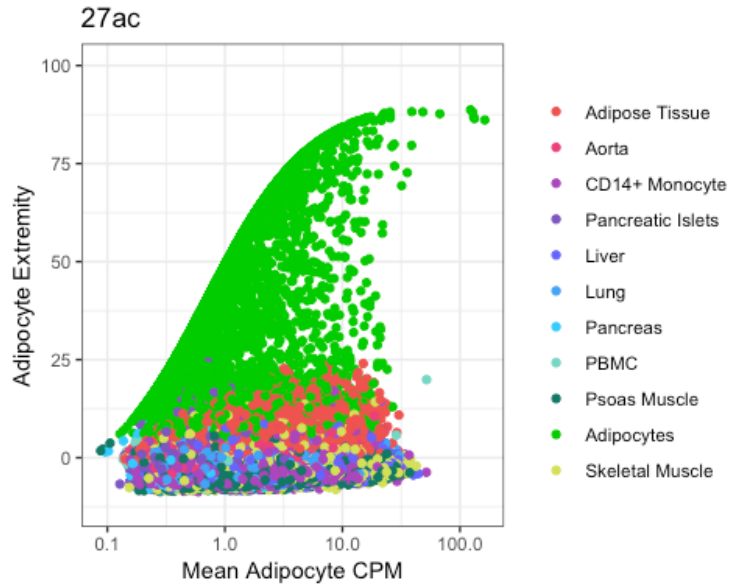


```
#Make a scatter plot of extremity by average counts
```

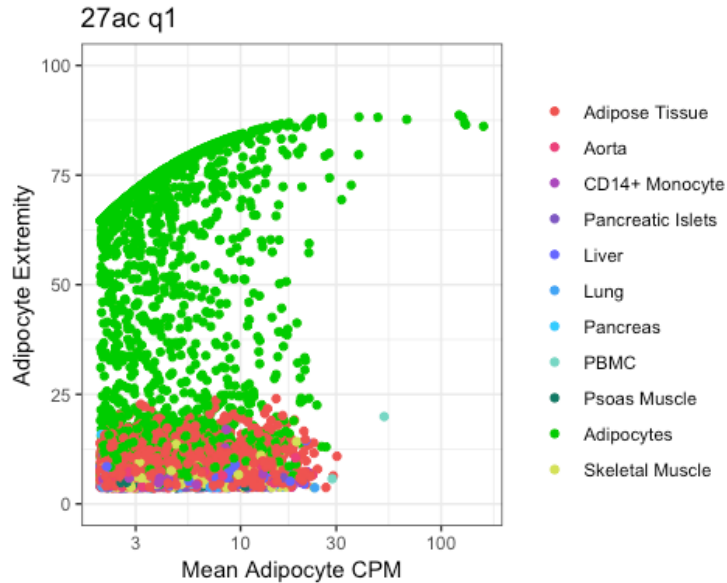
```
ac_adip_sctr<-ggplot(ac_adip_avg, aes(x=adip_cpm, y=Extremity, color = group)) +
  geom_point()+
  theme_bw()+
  scale_x_continuous(trans="log10") +
  labs(title = "27ac", y = "Adipocyte Extremity", x = "Mean Adipocyte CPM", color = NULL)+
  scale_color_manual(values = c("ac_adipose_samples" = "#EF5350", "ac_aorta_samples" = "#EC407A", "ac_CD14_samples" = "#AB47BC", "ac_endo_samples" = "#7E57C2", "ac_liver_samples" = "#6666FF", "ac_lung_samples" = "#42A5F5", "ac_panc_samples" = "#33CCFF", "ac_pbmc_samples" = "#76D7C"))
```



```
4", "ac_psoas_samples" = "#117A65", "ac_scadip_rosen_samples" = "#00CC00", "ac_skel_samples" =
"#D4E157"),
      labels = c("Adipose Tissue", "Aorta", "CD14+ Monocyte", "Pancreatic Islets", "Liver", "Lung",
"Pancreas", "PBMC", "Psoas Muscle", "Adipocytes", "Skeletal Muscle"))+
      ylim(-12, 100)
show(ac_adip_sctr)
```



```
#plot top adipocyte extremity
#1-0.9
ac_q1<-ggplot(filter(ac_forqrts, Extremity > ac_quarts[[5]]), aes(x=adip_cpm, y=Extremity, color = group)) +
  geom_point() +
  theme_bw()+
  scale_x_continuous(trans="log10") +
  labs(title = "27ac q1", y = "Adipocyte Extremity", x = "Mean Adipocyte CPM", color = NULL)+
  scale_color_manual(values = c("ac_adipose_samples" = "#EF5350", "ac_aorta_samples" = "#EC407A", "ac_CD14_samples" = "#AB47BC", "ac_endo_samples" = "#7E57C2", "ac_liver_samples" = "#6666FF", "ac_lung_samples" = "#42A5F5", "ac_panc_samples" = "#33CCFF", "ac_pbmc_samples" = "#76D7C4", "ac_psoas_samples" = "#117A65", "ac_scadip_rosen_samples" = "#00CC00", "ac_skel_samples" = "#D4E157"),
      labels = c("Adipose Tissue", "Aorta", "CD14+ Monocyte", "Pancreatic Islets", "Liver", "Lung",
"Pancreas", "PBMC", "Psoas Muscle", "Adipocytes", "Skeletal Muscle"))+
  ylim(0, 100)
show(ac_q1)
```



```
#number of peaks with adipocytes as contributor 1 at each quartile
print(c("27ac",
      length(rownames(filter(ac_forqrts, Extremity > ac_quarts[[3]] & (group == "ac_scadip_rosen_samples" | group == "ac_adipose_samples")))),
      length(rownames(filter(ac_forqrts, Extremity < ac_quarts[[3]] & Extremity > ac_quarts[[2]] & group == "ac_scadip_rosen_samples"))),
      length(rownames(filter(ac_forqrts, Extremity < ac_quarts[[2]] & Extremity > ac_quarts[[1]] & group == "ac_scadip_rosen_samples"))),
      length(rownames(filter(ac_forqrts, Extremity < ac_quarts[[1]] & group == "ac_scadip_rosen_samples"))),
      length(rownames(filter(ac_forqrts, Extremity > ac_quarts[[4]] & (group == "ac_scadip_rosen_samples" | group == "ac_adipose_samples"))))))

## [1] "27ac" "7174" "0" "0" "0" "6865"

#total number of peaks in each top quartile
show(length(rownames(filter(ac_forqrts, Extremity > ac_quarts[[4]] )))

## [1] 10759

show(length(rownames(filter(ac_forqrts, Extremity > ac_quarts[[3]] )))

## [1] 15369

show(length(rownames(filter(ac_forqrts, Extremity < ac_quarts[[1]] )))

## [1] 7685
```

trying different ways to subset for primary and background peaks for FIMO and AME

```
#take control peaks from middle 30% adipocyte extremity
ac_30_ctr<-filter(ac_forqrts, Extremity > ac_quarts[[2]] & Extremity < ac_quarts[[4]])
ac_30_ctr$adip_cpm<-round(ac_30_ctr$adip_cpm, 0)

print(c("control, no cpm filter",length(rownames(ac_30_ctr))))

## [1] "control, no cpm filter" "9221"
```

#get peaks in the top 10%, the next top 10%, and the following top 10% for the primary sets. With no tissue filtering, with just adipocyte samples, or with adipocyte and adipose peaks

```

ac_91_prim<-filter(ac_forqrts, Extremity > ac_quarts[[8]])
ac_91_prim$adip_cpm<-round(ac_91_prim$adip_cpm, 0)
ac_91_a_prim<-filter(ac_forqrts, Extremity > ac_quarts[[8]] & group == "ac_scadip_rosen_sampl
es")
ac_91_a_prim$adip_cpm<-round(ac_91_a_prim$adip_cpm, 0)
ac_91_aa_prim<-filter(ac_forqrts, Extremity > ac_quarts[[8]] & (group == "ac_scadip_rosen_samp
les" | group == "ac_adipose_samples"))
ac_91_aa_prim$adip_cpm<-round(ac_91_aa_prim$adip_cpm, 0)
ac_89_prim<-filter(ac_forqrts, Extremity > ac_quarts[[7]] & Extremity < ac_quarts[[8]])
ac_89_prim$adip_cpm<-round(ac_89_prim$adip_cpm, 0)
ac_89_a_prim<-filter(ac_forqrts, Extremity > ac_quarts[[7]] & Extremity < ac_quarts[[8]] & gro
up == "ac_scadip_rosen_samples")
ac_89_a_prim$adip_cpm<-round(ac_89_a_prim$adip_cpm, 0)
ac_89_aa_prim<-filter(ac_forqrts, Extremity > ac_quarts[[7]] & Extremity < ac_quarts[[8]] & (g
roup == "ac_scadip_rosen_samples" | group == "ac_adipose_samples"))
ac_89_aa_prim$adip_cpm<-round(ac_89_aa_prim$adip_cpm, 0)
ac_78_prim<-filter(ac_forqrts, Extremity > ac_quarts[[6]] & Extremity < ac_quarts[[7]])
ac_78_prim$adip_cpm<-round(ac_78_prim$adip_cpm, 0)
ac_78_a_prim<-filter(ac_forqrts, Extremity > ac_quarts[[6]] & Extremity < ac_quarts[[7]] & gro
up == "ac_scadip_rosen_samples")
ac_78_a_prim$adip_cpm<-round(ac_78_a_prim$adip_cpm, 0)
ac_78_aa_prim<-filter(ac_forqrts, Extremity > ac_quarts[[6]] & Extremity < ac_quarts[[7]] & (g
roup == "ac_scadip_rosen_samples" | group == "ac_adipose_samples"))
ac_78_aa_prim$adip_cpm<-round(ac_78_aa_prim$adip_cpm, 0)

print(c(length(rownames(ac_91_prim)), length(rownames(ac_91_a_prim)), length(rownames(ac_91_aa
_prim)), length(rownames(ac_89_prim)), length(rownames(ac_89_a_prim)), length(rownames(ac_89_a
a_prim)), length(rownames(ac_78_prim)), length(rownames(ac_78_a_prim)), length(rownames(ac_78
_aa_prim))))

## [1] 3074 2015 3012 3074 457 2335 3074 60 1194

#Decided that the control (background) peaks would not be filtered by CPM or top contributor a
nd will be in the middle 30% of adipocyte extremity
#create control peak .BED files for motif analysis

ac_30ctr_bed<-data.frame(matrix(data = NA, nrow = length(rownames(ac_30_ctr)), ncol = 4))
rownames(ac_30ctr_bed)<-ac_30_ctr$name
colnames(ac_30ctr_bed)<-c("chr", "start", "stop", "name")

ac_30ctr_bed$name<-paste0("27ac_", rownames(ac_30ctr_bed))
ac_30ctr_bed$chr<-gsub("^.*\\|(chr.*):.*$", "\\1", rownames(ac_30ctr_bed))
ac_30ctr_bed$start<-gsub("^.*:(\\d+)-\\d+$", "\\1", rownames(ac_30ctr_bed))
ac_30ctr_bed$stop<-gsub("^.*:\\d+-\\d+$", "\\1", rownames(ac_30ctr_bed))

write_tsv(ac_30ctr_bed, "27ac_30ctr_adip_peaks.bed", col_names = F)

#calculate new quantiles to decile peaks by adipocyte extremity
ac_dec<-quantile(ac_forqrts[, "Extremity"], probs = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0
.9))
print(ac_dec)

##          10%          20%          30%          40%          50%          60%
## -5.1058859 -4.2004137 -3.3145153 -2.2956129 -1.0461312  0.5170884
##          70%          80%          90%
##  2.4629130  5.1895705  9.9206867

#Filter the peaks by decile
ac_dec1_conts<-filter(ac_forqrts, Extremity > ac_dec[[9]])$name %>% ac_perc_cont[.,]
ac_dec2_conts<-filter(ac_forqrts, Extremity < ac_dec[[9]] & Extremity > ac_dec[[8]])$name %>%
ac_perc_cont[.,]

```

```

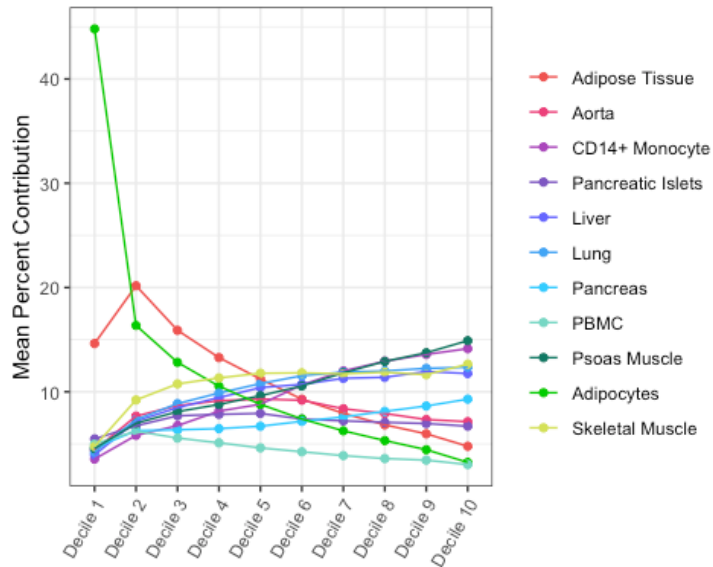
ac_dec3_conts<-filter(ac_forqrts, Extremity < ac_dec[[8]] & Extremity > ac_dec[[7]])$name %>%
ac_perc_cont[.,]
ac_dec4_conts<-filter(ac_forqrts, Extremity < ac_dec[[7]] & Extremity > ac_dec[[6]])$name %>%
ac_perc_cont[.,]
ac_dec5_conts<-filter(ac_forqrts, Extremity < ac_dec[[6]] & Extremity > ac_dec[[5]])$name %>%
ac_perc_cont[.,]
ac_dec6_conts<-filter(ac_forqrts, Extremity < ac_dec[[5]] & Extremity > ac_dec[[4]])$name %>%
ac_perc_cont[.,]
ac_dec7_conts<-filter(ac_forqrts, Extremity < ac_dec[[4]] & Extremity > ac_dec[[3]])$name %>%
ac_perc_cont[.,]
ac_dec8_conts<-filter(ac_forqrts, Extremity < ac_dec[[3]] & Extremity > ac_dec[[2]])$name %>%
ac_perc_cont[.,]
ac_dec9_conts<-filter(ac_forqrts, Extremity < ac_dec[[2]] & Extremity > ac_dec[[1]])$name %>%
ac_perc_cont[.,]
ac_dec10_conts<-filter(ac_forqrts, Extremity < ac_dec[[1]] )$name %>% ac_perc_cont[.,]

#Calculate the mean percent contribution of each tissue
ac_dec01_m<-apply(ac_dec1_conts, 2, mean)
ac_dec02_m<-apply(ac_dec2_conts, 2, mean)
ac_dec03_m<-apply(ac_dec3_conts, 2, mean)
ac_dec04_m<-apply(ac_dec4_conts, 2, mean)
ac_dec05_m<-apply(ac_dec5_conts, 2, mean)
ac_dec06_m<-apply(ac_dec6_conts, 2, mean)
ac_dec07_m<-apply(ac_dec7_conts, 2, mean)
ac_dec08_m<-apply(ac_dec8_conts, 2, mean)
ac_dec09_m<-apply(ac_dec9_conts, 2, mean)
ac_dec10_m<-apply(ac_dec10_conts, 2, mean)

ac_dec_m<-as.data.frame(rbind(ac_dec01_m, ac_dec02_m, ac_dec03_m, ac_dec04_m, ac_dec05_m, ac_d
ec06_m, ac_dec07_m, ac_dec08_m, ac_dec09_m, ac_dec10_m))
ac_dec_m$decile<-rownames(ac_dec_m)
ac_dec_m<-gather(ac_dec_m, -decile, key = "variable", value = "value")

#plot a line plot of the mean percent contribution of each tissue
ac_dec_m_line<-ggplot(ac_dec_m, aes(x = decile,y = value, color = variable))+
  geom_point()+
  geom_line(aes(group = variable))+
  theme_bw()+
  labs(y = "Mean Percent Contribution", x = NULL , color=NULL)+
  scale_x_discrete(labels=c("ac_dec01_m" = "Decile 1", "ac_dec02_m" = "Decile 2", "ac_dec03_m
" = "Decile 3", "ac_dec04_m" = "Decile 4", "ac_dec05_m" = "Decile 5", "ac_dec06_m" = "Decile
6", "ac_dec07_m" = "Decile 7", "ac_dec08_m" = "Decile 8", "ac_dec09_m" = "Decile 9", "ac_dec10_m
" = "Decile 10"))+
  scale_color_manual(values = c("ac_adipose_samples" = "#EF5350", "ac_aorta_samples" = "#E407
A", "ac_CD14_samples" = "#AB47BC", "ac_endo_samples" = "#7E57C2", "ac_liver_samples" = "#6666
FF", "ac_lung_samples" = "#42A5F5", "ac_panc_samples" = "#33CCFF", "ac_pbmc_samples" = "#76D7C
4", "ac_psoas_samples" = "#117A65", "ac_scadip_rosen_samples" = "#00CC00", "ac_skel_samples" =
"#D4E157"),
                    labels = c( "Adipose Tissue", "Aorta", "CD14+ Monocyte", "Pancreatic Isle
ts", "Liver", "Lung", "Pancreas", "PBMC", "Psoas Muscle", "Adipocytes", "Skeletal Muscle"))+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
show(ac_dec_m_line)

```



```
#combined decile contributions into one data frame
```

```
ac_dec1_conts$name<-rownames(ac_dec1_conts)
ac_dec2_conts$name<-rownames(ac_dec2_conts)
ac_dec3_conts$name<-rownames(ac_dec3_conts)
ac_dec4_conts$name<-rownames(ac_dec4_conts)
ac_dec5_conts$name<-rownames(ac_dec5_conts)
ac_dec6_conts$name<-rownames(ac_dec6_conts)
ac_dec7_conts$name<-rownames(ac_dec7_conts)
ac_dec8_conts$name<-rownames(ac_dec8_conts)
ac_dec9_conts$name<-rownames(ac_dec9_conts)
ac_dec10_conts$name<-rownames(ac_dec10_conts)
```

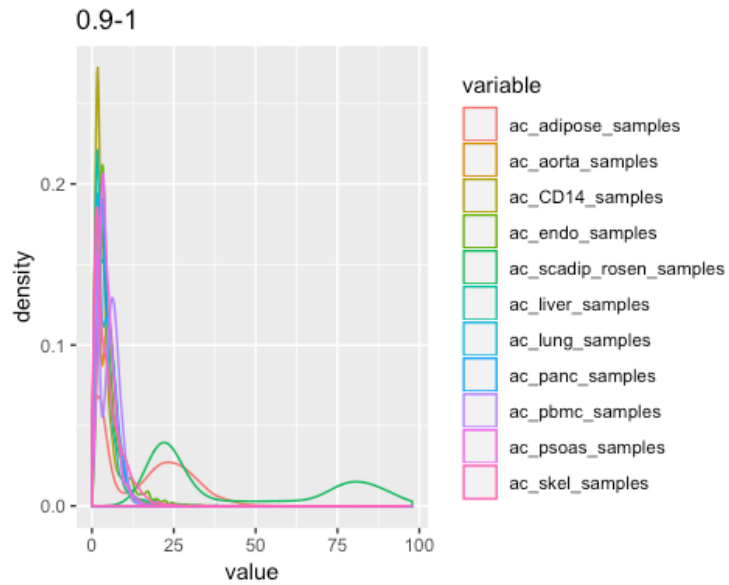
```
ac_b30_ctr<-rbind(ac_dec8_conts,ac_dec9_conts,ac_dec10_conts)
rownames(ac_b30_ctr)<-ac_b30_ctr$name
```

```
#melt data frame
```

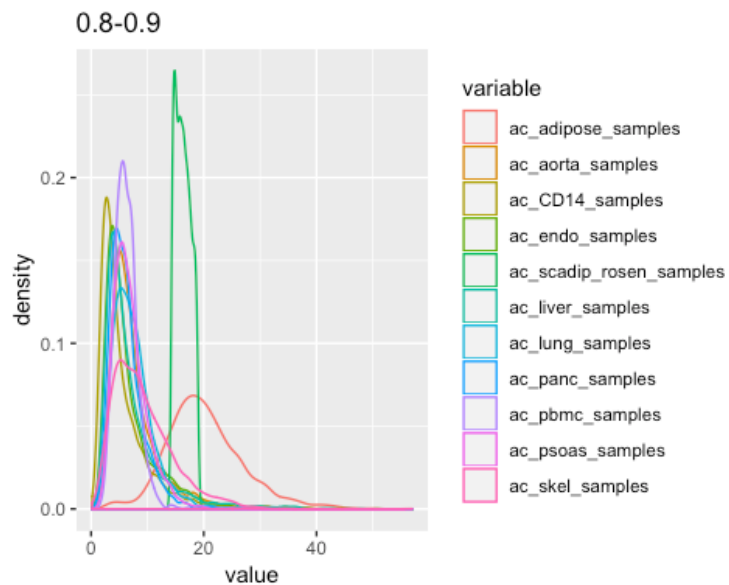
```
ac_dec1_conts<-melt(ac_dec1_conts, id = "name")
ac_dec2_conts<-melt(ac_dec2_conts, id = "name")
ac_dec3_conts<-melt(ac_dec3_conts, id = "name")
ac_dec4_conts<-melt(ac_dec4_conts, id = "name")
ac_dec5_conts<-melt(ac_dec5_conts, id = "name")
ac_dec6_conts<-melt(ac_dec6_conts, id = "name")
ac_dec7_conts<-melt(ac_dec7_conts, id = "name")
ac_dec8_conts<-melt(ac_dec8_conts, id = "name")
ac_dec9_conts<-melt(ac_dec9_conts, id = "name")
ac_dec10_conts<-melt(ac_dec10_conts, id = "name")
```

```
#Plot distribution of mean average percent contribution on each tissue in each decile
```

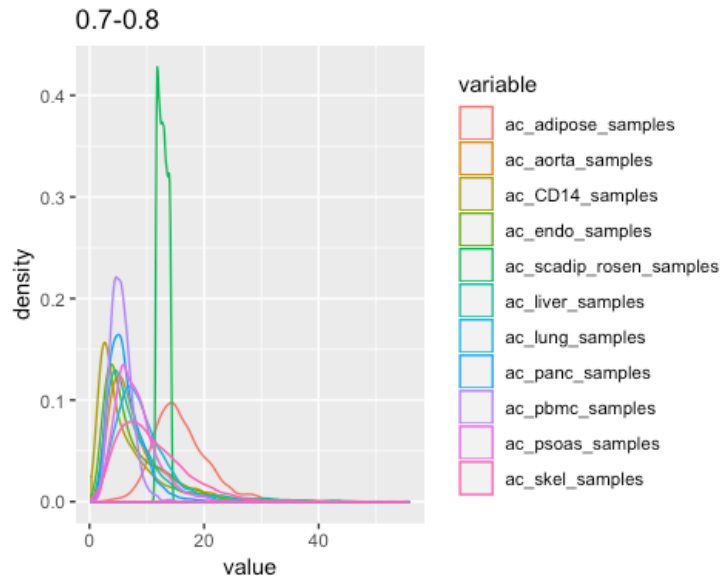
```
ac_dec1_hist<-ggplot(ac_dec1_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.9-1")
show(ac_dec1_hist)
```



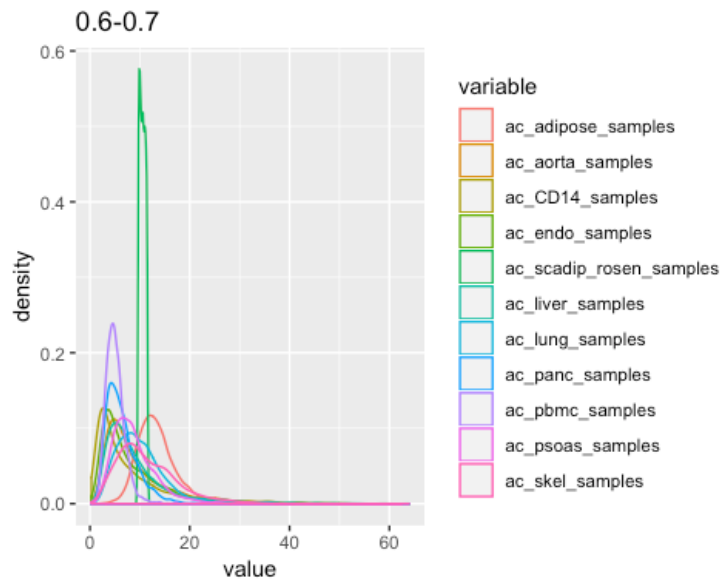
```
ac_dec2_hist<-ggplot(ac_dec2_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.8-0.9")
show(ac_dec2_hist)
```



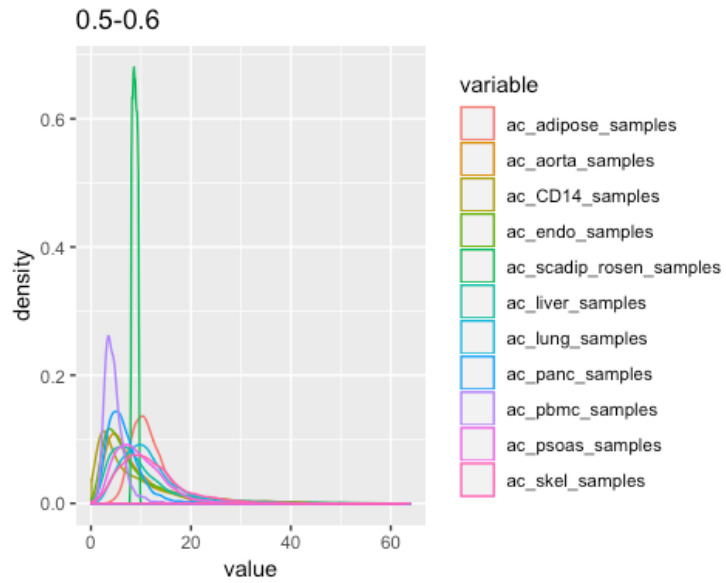
```
ac_dec3_hist<-ggplot(ac_dec3_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.7-0.8")
show(ac_dec3_hist)
```



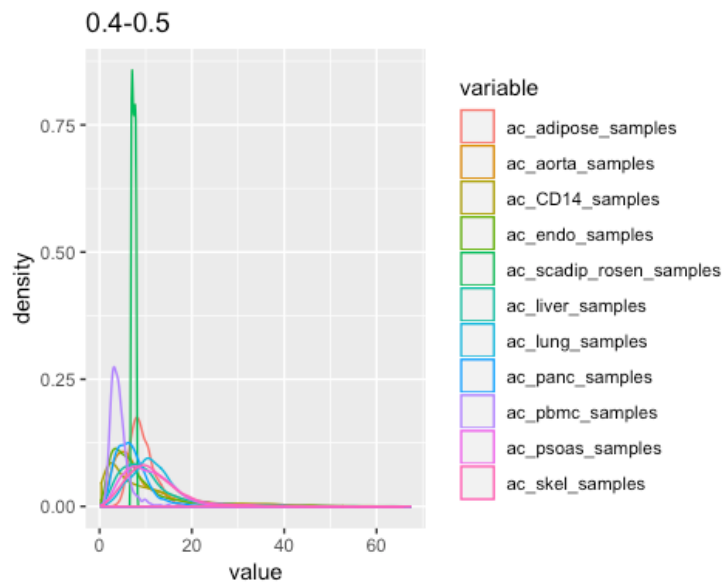
```
ac_dec4_hist<-ggplot(ac_dec4_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.6-0.7")
show(ac_dec4_hist)
```



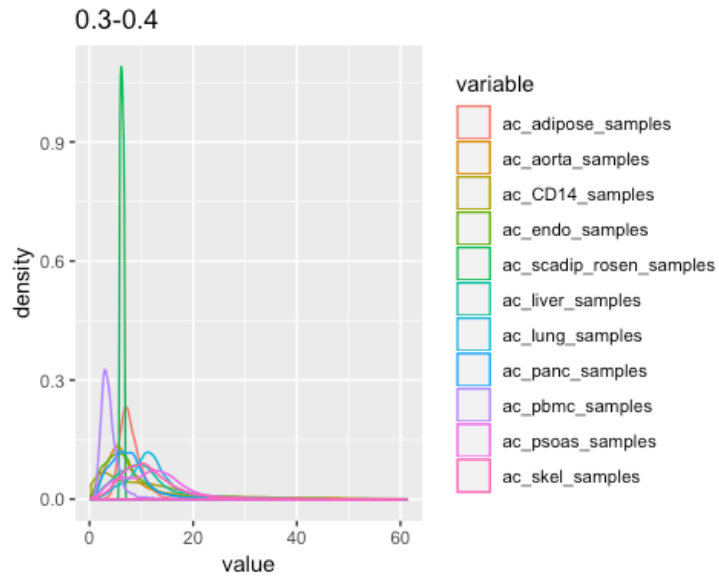
```
ac_dec5_hist<-ggplot(ac_dec5_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.5-0.6")
show(ac_dec5_hist)
```



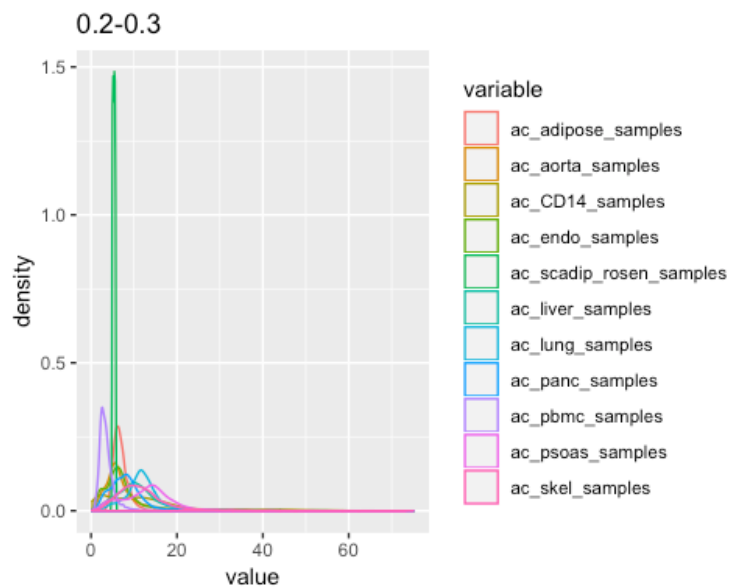
```
ac_dec6_hist<-ggplot(ac_dec6_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.4-0.5")
show(ac_dec6_hist)
```



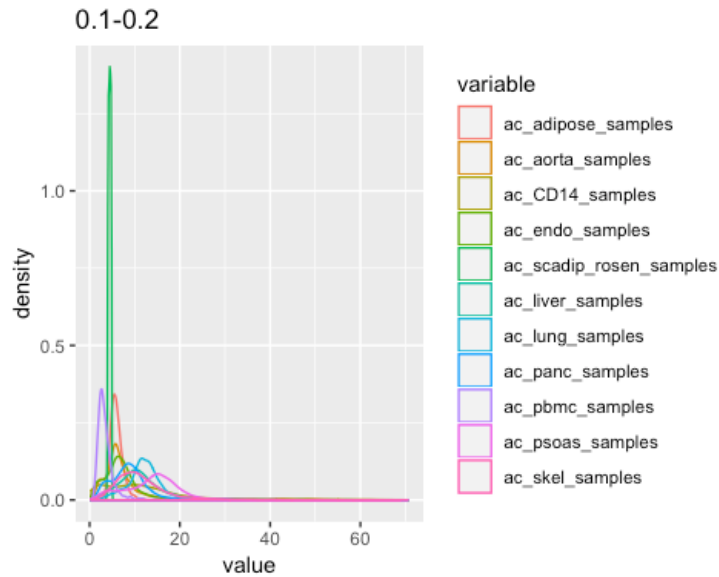
```
ac_dec7_hist<-ggplot(ac_dec7_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.3-0.4")
show(ac_dec7_hist)
```

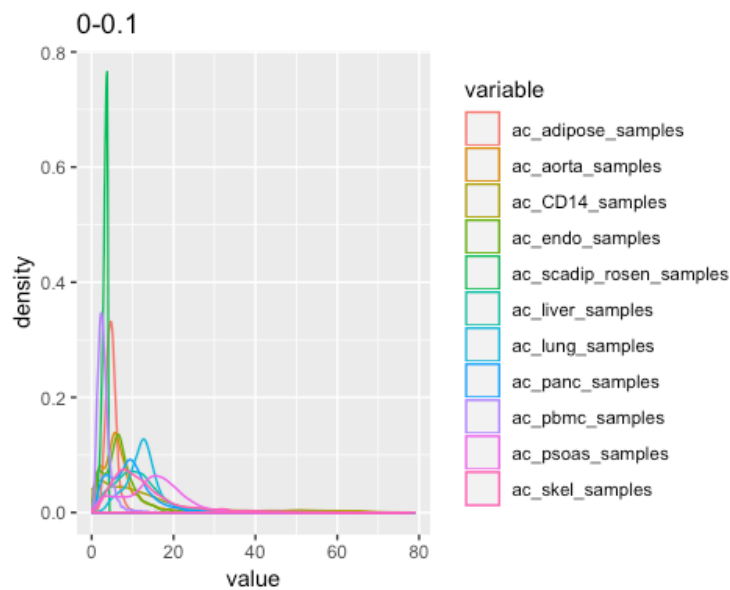
```
ac_dec8_hist<-ggplot(ac_dec8_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.2-0.3")
show(ac_dec8_hist)
```



```
ac_dec9_hist<-ggplot(ac_dec9_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0.1-0.2")
show(ac_dec9_hist)
```



```
ac_dec10_hist<-ggplot(ac_dec10_conts, aes(x = value, color = variable))+
  geom_density(alpha = 0.5)+
  labs(title = "0-0.1")
show(ac_dec10_hist)
```



```
#filter peaks to get the top decile, 1-0.9
ac_dec1_table<-filter(ac_forqrts, Extremity > ac_dec[[9]])
write_tsv(ac_dec1_table, "27ac_decile1.tsv")

print(length(rownames(ac_dec1_table)))

## [1] 3074

#calculate quartiles of adipocyte extremity for the top decile, 1-0.9
ac_dec1_quart<-quantile(ac_dec1_table[, "Extremity"], probs = c(0.25, 0.5, 0.75))

print(ac_dec1_quart)
```

```

##      25%      50%      75%
## 12.19604 17.21423 67.57033

#filter the top decile to make tables for each quartile
ac_dec1_table1<-filter(ac_dec1_table, Extremity > ac_dec1_quart[[3]])
ac_dec1_table2<-filter(ac_dec1_table, Extremity > ac_dec1_quart[[2]] & Extremity < ac_dec1_qua
rt[[3]])
ac_dec1_table3<-filter(ac_dec1_table, Extremity > ac_dec1_quart[[1]] & Extremity < ac_dec1_qua
rt[[2]])
ac_dec1_table4<-filter(ac_dec1_table, Extremity < ac_dec1_quart[[1]])

print(c("27ac",
      length(rownames(ac_dec1_table1)),
      length(rownames(ac_dec1_table2)),
      length(rownames(ac_dec1_table3)),
      length(rownames(ac_dec1_table4))
    ))

## [1] "27ac" "769" "768" "768" "769"

#Make .BED files for each quartile of the top decile, and the full top decile
ac_dec1_t1_bed<-data.frame(matrix(data = NA, nrow = length(rownames(ac_dec1_table1)), ncol = 4
))
  rownames(ac_dec1_t1_bed)<-ac_dec1_table1$name
  colnames(ac_dec1_t1_bed)<-c("chr", "start", "stop", "name")

  ac_dec1_t1_bed$name<-paste0("27ac_", rownames(ac_dec1_t1_bed))
  ac_dec1_t1_bed$chr<-gsub("^.*\\|(chr.*):.*$", "\\1", rownames(ac_dec1_t1_bed))
  ac_dec1_t1_bed$start<-gsub("^.*:(\\d+)-\\d+$", "\\1", rownames(ac_dec1_t1_bed))
  ac_dec1_t1_bed$stop<-gsub("^.*:\\d+-(\\d+)$", "\\1", rownames(ac_dec1_t1_bed))

ac_dec1_t2_bed<-data.frame(matrix(data = NA, nrow = length(rownames(ac_dec1_table2)), ncol = 4
))
  rownames(ac_dec1_t2_bed)<-ac_dec1_table2$name
  colnames(ac_dec1_t2_bed)<-c("chr", "start", "stop", "name")

  ac_dec1_t2_bed$name<-paste0("27ac_", rownames(ac_dec1_t2_bed))
  ac_dec1_t2_bed$chr<-gsub("^.*\\|(chr.*):.*$", "\\1", rownames(ac_dec1_t2_bed))
  ac_dec1_t2_bed$start<-gsub("^.*:(\\d+)-\\d+$", "\\1", rownames(ac_dec1_t2_bed))
  ac_dec1_t2_bed$stop<-gsub("^.*:\\d+-(\\d+)$", "\\1", rownames(ac_dec1_t2_bed))

ac_dec1_t3_bed<-data.frame(matrix(data = NA, nrow = length(rownames(ac_dec1_table3)), ncol = 4
))
  rownames(ac_dec1_t3_bed)<-ac_dec1_table3$name
  colnames(ac_dec1_t3_bed)<-c("chr", "start", "stop", "name")

  ac_dec1_t3_bed$name<-paste0("27ac_", rownames(ac_dec1_t3_bed))
  ac_dec1_t3_bed$chr<-gsub("^.*\\|(chr.*):.*$", "\\1", rownames(ac_dec1_t3_bed))
  ac_dec1_t3_bed$start<-gsub("^.*:(\\d+)-\\d+$", "\\1", rownames(ac_dec1_t3_bed))
  ac_dec1_t3_bed$stop<-gsub("^.*:\\d+-(\\d+)$", "\\1", rownames(ac_dec1_t3_bed))

ac_dec1_t4_bed<-data.frame(matrix(data = NA, nrow = length(rownames(ac_dec1_table4)), ncol = 4
))
  rownames(ac_dec1_t4_bed)<-ac_dec1_table4$name
  colnames(ac_dec1_t4_bed)<-c("chr", "start", "stop", "name")

  ac_dec1_t4_bed$name<-paste0("27ac_", rownames(ac_dec1_t4_bed))
  ac_dec1_t4_bed$chr<-gsub("^.*\\|(chr.*):.*$", "\\1", rownames(ac_dec1_t4_bed))
  ac_dec1_t4_bed$start<-gsub("^.*:(\\d+)-\\d+$", "\\1", rownames(ac_dec1_t4_bed))
  ac_dec1_t4_bed$stop<-gsub("^.*:\\d+-(\\d+)$", "\\1", rownames(ac_dec1_t4_bed))

```

```

write_tsv(ac_dec1_t1_bed, "27ac_dec1_t1_adip_peaks.bed", col_names = F)
write_tsv(ac_dec1_t2_bed, "27ac_dec1_t2_adip_peaks.bed", col_names = F)
write_tsv(ac_dec1_t3_bed, "27ac_dec1_t3_adip_peaks.bed", col_names = F)
write_tsv(ac_dec1_t4_bed, "27ac_dec1_t4_adip_peaks.bed", col_names = F)
write_tsv(rbind(ac_dec1_t1_bed, ac_dec1_t2_bed), "27ac_dec1_t12_adip_peaks.bed", col_names = F)
)
write_tsv(rbind(ac_dec1_t1_bed, ac_dec1_t2_bed, ac_dec1_t3_bed), "27ac_dec1_t123_adip_peaks.bed", col_names = F)
write_tsv(rbind(ac_dec1_t1_bed, ac_dec1_t2_bed, ac_dec1_t3_bed, ac_dec1_t4_bed), "27ac_dec1_t1234_adip_peaks.bed", col_names = F)

#DControl (background) peaks would not be filtered by CPM or top contributor and will be in the bottom 30% of adipocyte extremity
#create control peak .BED files for motif analysis

ac_b30ctr_bed<-data.frame(matrix(data = NA, nrow = length(rownames(ac_b30_ctr)), ncol = 4))
rownames(ac_b30ctr_bed)<-ac_b30_ctr$name
colnames(ac_b30ctr_bed)<-c("chr", "start", "stop", "name")

ac_b30ctr_bed$name<-paste0("27ac_", rownames(ac_b30ctr_bed))
ac_b30ctr_bed$chr<-gsub("^.*\\|(chr.*)\\. *$", "\\1", rownames(ac_b30ctr_bed))
ac_b30ctr_bed$start<-gsub("^.*:(\\d+)-\\d+$$", "\\1", rownames(ac_b30ctr_bed))
ac_b30ctr_bed$stop<-gsub("^.*:\\d+-(\\d+)$", "\\1", rownames(ac_b30ctr_bed))

write_tsv(ac_b30ctr_bed, "27ac_b30ctr_adip_peaks.bed", col_names = F)

#write files with all peaks
write_tsv(ac_adip_avg, "27ac_full_extrem.tsv")

```

Comparisons between DEA and Extremity

Comparison between DEAs and Adipocyte Extremity

```
knitr::opts_chunk$set(root.dir = "/Volumes/broad_rosenlab_archive/Projects/Linus-Human-Ad/Analysis/Individual-Peak-Characterization-no-137/peaks_by_tissue/for_thesis")

dbl_min <- .Machine$double.xmin

library(tibble)
library(plyr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(readr)
library(tidyr)
library(ggplot2)
library(VennDiagram)

## Loading required package: grid

## Loading required package: futile.logger

library(pheatmap)
library(RColorBrewer)
library(reshape2)

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##   smiths

library(eulerr)

#Load files from extremity analysis, IR/IS DEA, and Adipocyte/Nonadipocyte DEA
ac_all_extrem<-read_tsv("../27ac_full_extrem.tsv")

ac_iris_dea<-read.table("../H3K27ac/dea_iris_rt0/1_Init/IS_IR.txt", header = T) %>%
  mutate(cpm1 = logCPM >= 1, cpm2 = logCPM >= 2)

ac_rt0_dea<-read.table("../H3K27ac/dea_27ac/1_Init/nonadipocyte_adipocyte.txt", header = T) %>%
  inner_join(y = select(ac_iris_dea, genes, cpm1), by = "genes")
```

```

#get extremities for DEA peaks
ac_extrem_iris<-inner_join(x = ac_all_extrem, y = ac_iris_dea, by = c("name" = "genes")) %>%
  mutate(fc_1_05_color = logFC >= 1 & FDR <= 0.05, fc_5_25_color = logFC >= 0.5 & FDR <= 0.25)

ac_extrem_rt0<-inner_join(x = ac_all_extrem, y = ac_rt0_dea, by = c("name" = "genes")) %>%
  mutate(fc_1_05_color = logFC >= 1 & FDR <= 0.05 & cpm1, fc_5_25_color = logFC >= 0.5 & FDR <
= 0.25)

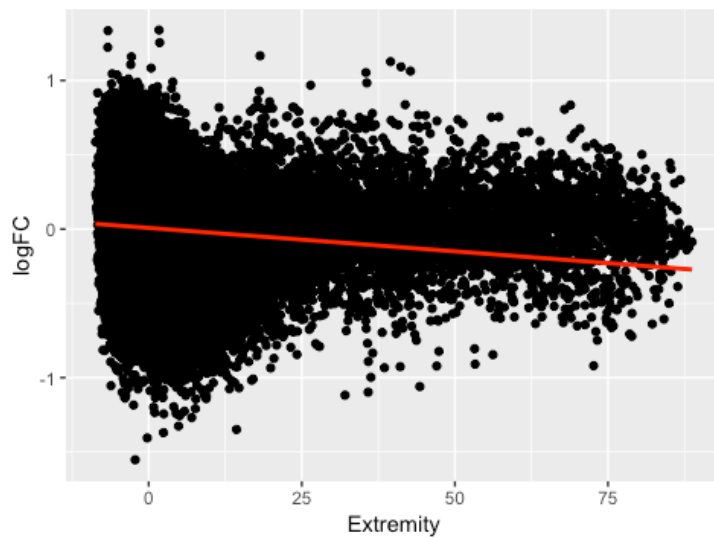
#calculate linear regression for plots
ac_extrem_iris_lm<-lm( logFC ~ Extremity, data = ac_extrem_iris)

ac_extrem_rt0_lm<-lm( logFC ~ Extremity, data = ac_extrem_rt0)

#plotting extremity against FC with linear regression
ggplot(ac_extrem_iris_lm$model, aes_string(x = names(ac_extrem_iris_lm$model)[2], y = names(ac_
_extrem_iris_lm$model)[1])) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  labs(title = paste("27ac_irisDEA",
                    "Adj R2 = ", signif(summary(ac_extrem_iris_lm)$adj.r.squared, 5),
                    "Intercept =", signif(ac_extrem_iris_lm$coef[[1]], 5 ),
                    " Slope =", signif(ac_extrem_iris_lm$coef[[2]], 5),
                    " P =", signif(summary(ac_extrem_iris_lm)$coef[2,4], 5)))

```

27ac_irisDEA Adj R2 = 0.016098 Intercept = 0.0078656

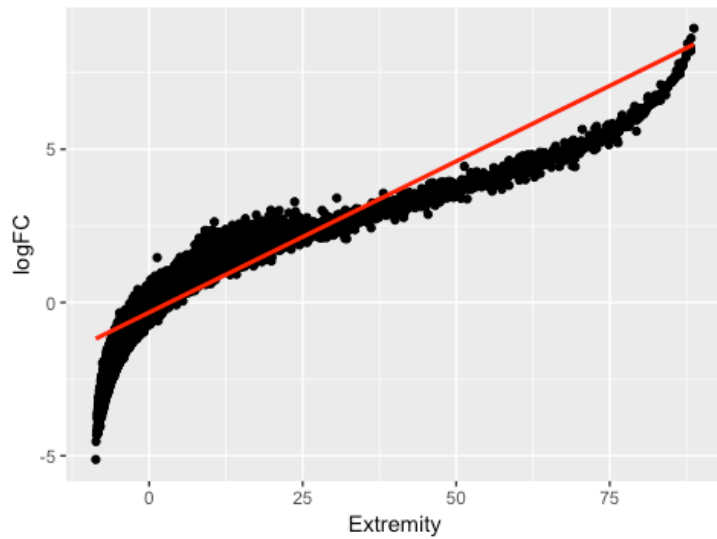


```

ggplot(ac_extrem_rt0_lm$model, aes_string(x = names(ac_extrem_rt0_lm$model)[2], y = names(ac_e
xtrem_rt0_lm$model)[1])) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  labs(title = paste("27ac_adipDEA",
                    "Adj R2 = ", signif(summary(ac_extrem_rt0_lm)$adj.r.squared, 5),
                    "Intercept =", signif(ac_extrem_rt0_lm$coef[[1]], 5 ),
                    " Slope =", signif(ac_extrem_rt0_lm$coef[[2]], 5),
                    " P =", signif(summary(ac_extrem_rt0_lm)$coef[2,4], 5)))

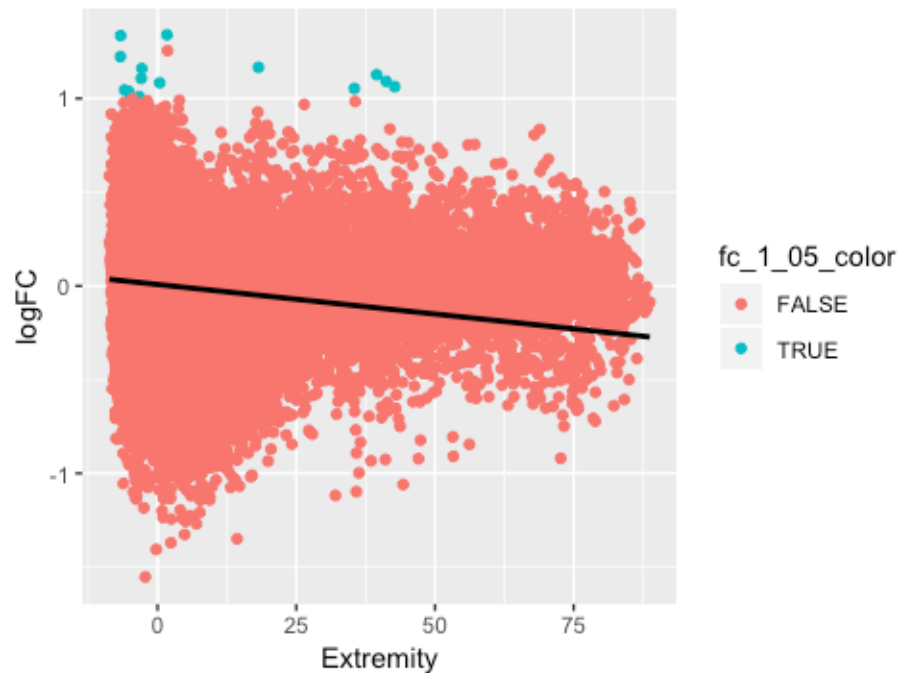
```

27ac_adipDEA Adj R2 = 0.81949 Intercept = -0.32501



```
#plot adipocyte extremity by FDR (< 0.05 or 0.25) fold change (FC) (> 1 or 0.5)
ac_105_iris_sctr<-ggplot(ac_extrem_iris, aes(y = logFC, x = Extremity, color = fc_1_05_color))
+
  geom_point()+
  labs(title = "27ac Adip Extrem & IR/IS DEA, FC >= 1, FDR <= 0.05")+
  geom_smooth(method = "lm", col = "black")
show(ac_105_iris_sctr)
```

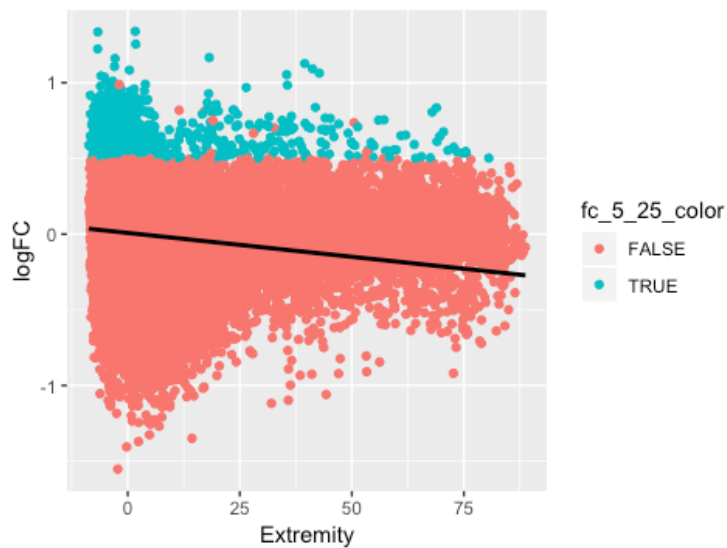
27ac Adip Extrem & IR/IS DEA, FC >= 1, FDR <= 0.05



```
ac_525_iris_sctr<-ggplot(ac_extrem_iris, aes(y = logFC, x = Extremity, color = fc_5_25_color))
+
  geom_point()+
  labs(title = "27ac Adip Extrem & IR/IS DEA, FC >= 0.5, FDR <= 0.25")+
```

```
geom_smooth(method = "lm", col = "black")
show(ac_525_iris_sctr)
```

27ac Adip Extrem & IR/IS DEA, FC >= 0.5, FDR <= 0.25



```
#plot adipocyte extremity by IR v IS DEA FDR, FDR (< 0.05 or 0.25) fold change (FC) (> 1 or 0.5)
```

```
ac_105_iris_sctr<-ggplot(ac_extrem_iris, aes(y = FDR+dbl_min, x = Extremity, color = fc_1_05_color))+
```

```
  geom_point()+
```

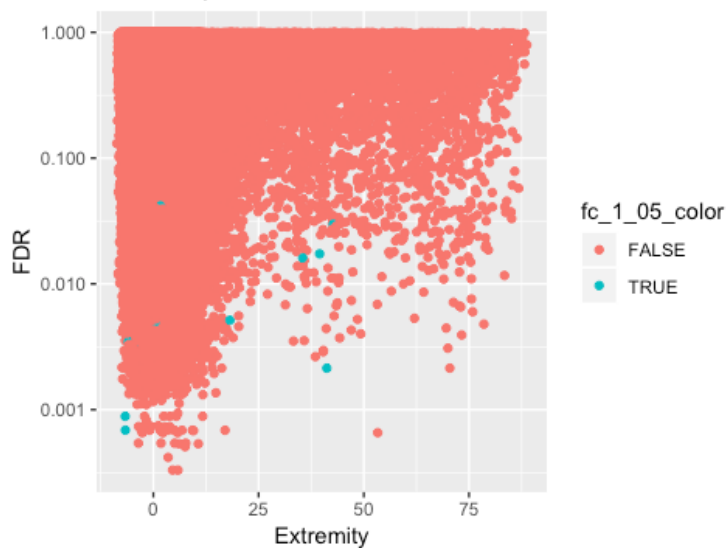
```
  labs(title = "27ac Adip Extrem & IR/IS DEA, FC >= 1, FDR <= 0.05 ", y = "FDR")+
```

```
  scale_y_continuous(trans = "log10")
```

```
  #geom_smooth(method = "Lm", col = "black")
```

```
show(ac_105_iris_sctr)
```

27ac Adip Extrem & IR/IS DEA, FC >= 1, FDR <= 0.05



```
ac_525_iris_sctr<-ggplot(ac_extrem_iris, aes(y = FDR+dbl_min, x = Extremity, color = fc_5_25_color))+
```

```
  geom_point()+
```

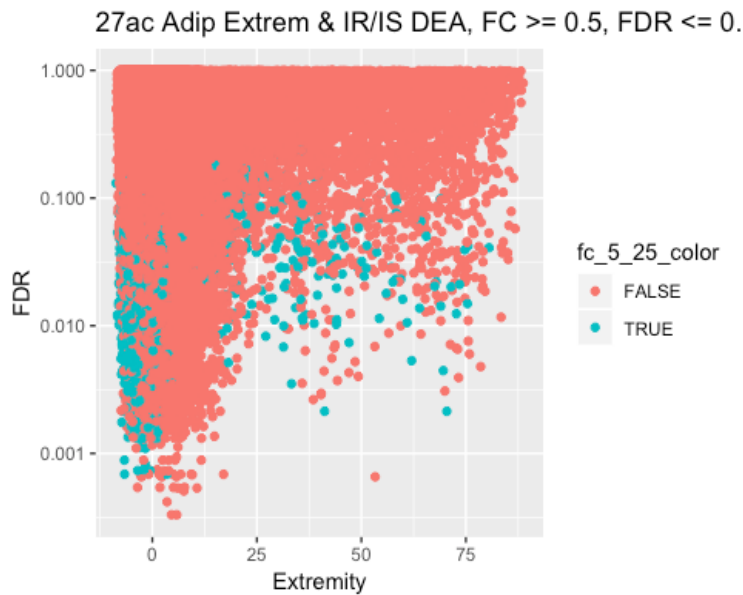
```
  labs(title = "27ac Adip Extrem & IR/IS DEA, FC >= 0.5, FDR <= 0.25 ", y = "FDR")+
```



```

scale_y_continuous(trans = "log10")
#geom_smooth(method = "lm", col = "black")
show(ac_525_iris_sctr)

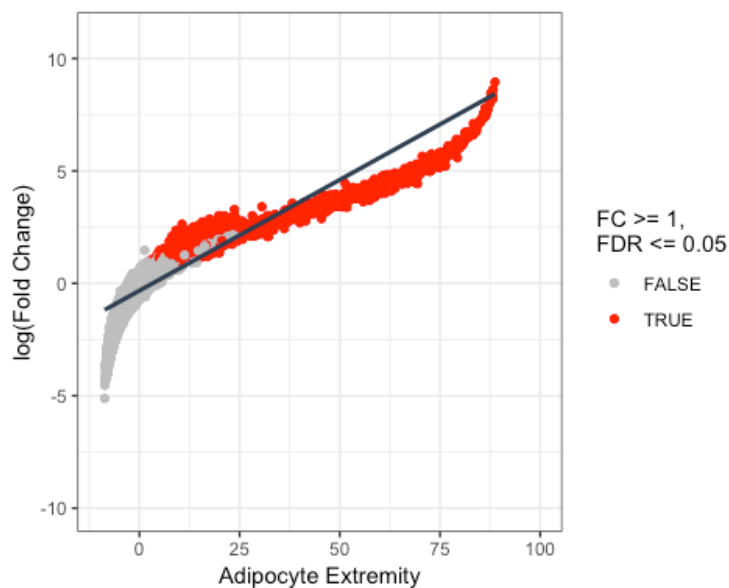
```



```

#plot adipocyte extremity by adipocyte v non-adipocyte DEA FDR, FDR (< 0.05) fold change (FC) (> 1)
ac_105_rt0_sctr<-ggplot(ac_extrem_rt0, aes(y = logFC, x = Extremity, color = fc_1_05_color))+
  geom_point()+
  theme_bw()+
  labs(x = "Adipocyte Extremity", y = "log(Fold Change)", color = NULL)+
  geom_smooth(method = "lm", col = "#2E4053")+
  scale_color_manual(values = c("FALSE" = "grey", "TRUE" = "red"), name = "FC >= 1,\nFDR <= 0.05")+
  ylim(-10, 11)+
  xlim(-10, 100)
show(ac_105_rt0_sctr)

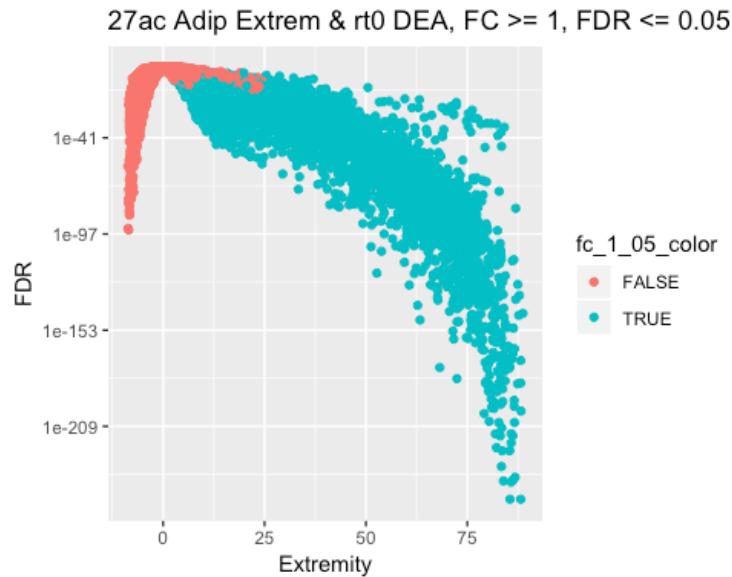
```



```

#plot adipocyte extremity by FDR
ac_105_rt0_sctr<-ggplot(ac_extrem_rt0, aes(y = FDR+dbl_min, x = Extremity, color = fc_1_05_color))
  geom_point()+
  labs(title = "27ac Adip Extrem & rt0 DEA, FC >= 1, FDR <= 0.05", y = "FDR")+
  scale_y_continuous(trans = "log10")
  # geom_smooth(method = "Lm", col = "black")
show(ac_105_rt0_sctr)

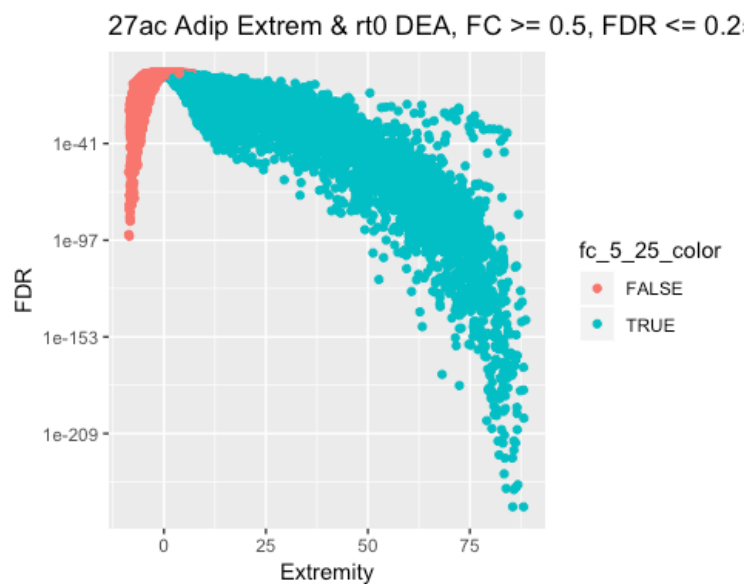
```



```

ac_525_rt0_sctr<-ggplot(ac_extrem_rt0, aes(y = FDR+dbl_min, x = Extremity, color = fc_5_25_color))
  geom_point()+
  labs(title = "27ac Adip Extrem & rt0 DEA, FC >= 0.5, FDR <= 0.25", y = "FDR")+
  scale_y_continuous(trans = "log10")
  # geom_smooth(method = "Lm", col = "black")
show(ac_525_rt0_sctr)

```

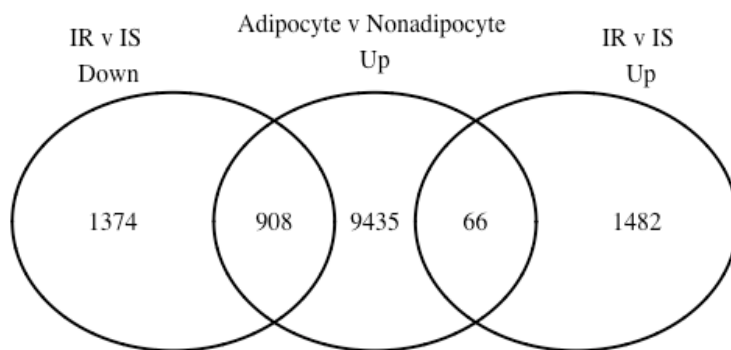


```

#plot venn diagram of overlap between the adipocyte v non-adipocyte DEA and the IR v IS DEA
ac_venn_105_list <- list(
  "IR v IS \nUp" = filter(ac_iris_dea, logFC >= 0.5 & FDR <= 0.25 & cpm2)$genes,
  "IR v IS \nDown" = filter(ac_iris_dea, logFC <= -0.5 & FDR <= 0.25 & cpm2)$genes,
  "Adipocyte v Nonadipocyte \nUp" = filter(ac_rt0_dea, logFC >= 1 & FDR <= 0.05 & cpm1)$genes
)

ac_venn_105 <- venn.diagram(ac_venn_105_list, filename = NULL, cat.pos = c(-15, 0, 15), cat.dist = c(0.11,0.11, 0.04))
grid.newpage()
grid.draw(ac_venn_105)

```



Adipocyte Specific Motif Analysis

Motif Enrichment Analysis #Ran FIMO on adipocyte peaks (peaks from extremity analysis where adipocytes were the top contributor, had a cpm greater than the mean of the average counts, and an adipocyte extremity in the top decile). Ran AME on the top decile for H3K27ac and H3K4me1, and the top two deciles for H3K4me3. Control peaks were selected as the peaks in the bottom 30% of adipocyte extremity. Reference: "tiss_spec_extremity.Rmd"

```
knitr::opts_chunk$set(root.dir = "/Volumes/broad_rosenlab_archive/Projects/Linus-Human-Ad/Analysis/Individual-Peak-Characterization-no-137/peaks_by_tissue/for_thesis")
library(tibble)
library(plyr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(readr)
library(tidyr)
library(ggplot2)
library(VennDiagram)

## Loading required package: grid

## Loading required package: futile.logger

library(pheatmap)
library(RColorBrewer)
library(reshape2)

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##   smiths

library(venn)
library(ggrepel)

#Load bed files
ac_d1_t1_bed <- read_tsv( "../27ac_dec1_t1_adip_peaks.bed", col_names = c("chr","start","stop",
,"peak_name")) %>% separate(peak_name, c(NA,"name"), sep = "\\|")

## Parsed with column specification:
## cols(
##   chr = col_character(),
##   start = col_double(),
```

```

## stop = col_double(),
## peak_name = col_character()
## )

ac_d1_t2_bed <- read_tsv("../27ac_dec1_t2_adip_peaks.bed", col_names = c("chr", "start", "stop",
"peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

## Parsed with column specification:
## cols(
## chr = col_character(),
## start = col_double(),
## stop = col_double(),
## peak_name = col_character()
## )

ac_d1_t3_bed <- read_tsv("../27ac_dec1_t3_adip_peaks.bed", col_names = c("chr", "start", "stop",
"peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

## Parsed with column specification:
## cols(
## chr = col_character(),
## start = col_double(),
## stop = col_double(),
## peak_name = col_character()
## )

ac_d1_t4_bed <- read_tsv("../27ac_dec1_t4_adip_peaks.bed", col_names = c("chr", "start", "stop",
"peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

## Parsed with column specification:
## cols(
## chr = col_character(),
## start = col_double(),
## stop = col_double(),
## peak_name = col_character()
## )

ac_d1_t1_t1_bed <- read_tsv("../27ac_dec1_t1_t1_adip_peaks.bed", col_names = c("chr", "start",
"stop", "peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

## Parsed with column specification:
## cols(
## chr = col_character(),
## start = col_double(),
## stop = col_double(),
## peak_name = col_character()
## )

ac_d1_t1_t2_bed <- read_tsv("../27ac_dec1_t1_t2_adip_peaks.bed", col_names = c("chr", "start",
"stop", "peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

## Parsed with column specification:
## cols(
## chr = col_character(),
## start = col_double(),
## stop = col_double(),
## peak_name = col_character()
## )

ac_d1_t1_t3_bed <- read_tsv("../27ac_dec1_t1_t3_adip_peaks.bed", col_names = c("chr", "start",
"stop", "peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

```

```

## Parsed with column specification:
## cols(
##   chr = col_character(),
##   start = col_double(),
##   stop = col_double(),
##   peak_name = col_character()
## )

ac_d1_t1_t4_bed <- read_tsv("../27ac_dec1_t1_t4_adip_peaks.bed", col_names = c("chr", "start",
"stop", "peak_name")) %>% separate(peak_name, c(NA, "name"), sep = "\\|")

## Parsed with column specification:
## cols(
##   chr = col_character(),
##   start = col_double(),
##   stop = col_double(),
##   peak_name = col_character()
## )

ac_peaks <- unique(c(ac_d1_t1_t1_bed$name, ac_d1_t1_bed$name, ac_d1_t2_bed$name, ac_d1_t3_bed$
name, ac_d1_t4_bed$name))
print(head(ac_d1_t1_bed))

## # A tibble: 6 x 4
##   chr      start      stop name
##   <chr>   <dbl>   <dbl> <chr>
## 1 chr1    28463    30061 chr1:28463-30061
## 2 chr1    171022   174906 chr1:171022-174906
## 3 chr1    198960   200610 chr1:198960-200610
## 4 chr1     529027   533437 chr1:529027-533437
## 5 chr1  16664042 16667941 chr1:16664042-16667941
## 6 chr1  16739294 16744474 chr1:16739294-16744474

#Load FIMO tsv files
#Load FIMO table for each mark one at a time and filter one at a time because full fimo files
too large to run locally, after filtering remove the full fimo table below
ac_fimo <- read_tsv("../fimo-27ac/fimo.tsv", col_types = "c_c_____", comment = "#") %>%
  separate(sequence_name, c(NA, "name"), sep = "\\|") %>%
  filter(name %in% ac_peaks)

#create list of peaks and their sizes for GWAS
ac_peaks<-as_tibble( x = ac_peaks)

## Warning: Calling `as_tibble()` on a vector is discouraged, because the behavior is likely t
o change in the future. Use `tibble::enframe(name = NULL)` instead.
## This warning is displayed once per session.

ac_peaks$name<-ac_peaks$value
ac_peaks<-separate(ac_peaks, value, c("chr", "coords"), sep = ":") %>%
  separate(coords, c("start", "stop"), sep = "-") %>%
  mutate_each( as.numeric, start, stop) %>%
  mutate(length = stop - start)
ac_peaks<-ac_peaks[,4:5]

write_tsv(ac_peaks, "27ac_decile_peaks.tsv")

#Load AME tsv files
ac_d1_t1_ame <- read_tsv("../ame-27ac_dec1_t1/ame.tsv", comment = "#")
ac_d1_t2_ame <- read_tsv("../ame-27ac_dec1_t2/ame.tsv", comment = "#")
ac_d1_t3_ame <- read_tsv("../ame-27ac_dec1_t3/ame.tsv", comment = "#")
ac_d1_t4_ame <- read_tsv("../ame-27ac_dec1_t4/ame.tsv", comment = "#")

```

```

ac_d1_t12_ame <- read_tsv("../ame-27ac_dec1_t12/ame.tsv", comment = "#")
ac_d1_t123_ame <- read_tsv("../ame-27ac_dec1_t123/ame.tsv", comment = "#")
ac_d1_t1234_ame <- read_tsv("../ame-27ac_dec1_t1234/ame.tsv", comment = "#")
ac_d1_t1_t1_ame <- read_tsv("../ame-27ac_dec1_t1_t1/ame.tsv", comment = "#")
ac_d1_t1_t2_ame <- read_tsv("../ame-27ac_dec1_t1_t2/ame.tsv", comment = "#")
ac_d1_t1_t3_ame <- read_tsv("../ame-27ac_dec1_t1_t3/ame.tsv", comment = "#")
ac_d1_t1_t4_ame <- read_tsv("../ame-27ac_dec1_t1_t4/ame.tsv", comment = "#")

#filter FIMO for peaks, then remove full fimo data frame
ac_d1_t1_fimo <- filter(ac_fimo, name %in% ac_d1_t1_bed$name)
ac_d1_t2_fimo <- filter(ac_fimo, name %in% ac_d1_t2_bed$name)
ac_d1_t3_fimo <- filter(ac_fimo, name %in% ac_d1_t3_bed$name)
ac_d1_t4_fimo <- filter(ac_fimo, name %in% ac_d1_t4_bed$name)
ac_d1_t1_t1_fimo <- filter(ac_fimo, name %in% ac_d1_t1_t1_bed$name)
ac_d1_t12_fimo <- rbind(ac_d1_t1_fimo, ac_d1_t2_fimo)
ac_d1_t123_fimo <- rbind(ac_d1_t1_fimo, ac_d1_t2_fimo, ac_d1_t3_fimo)
ac_d1_t1234_fimo <- rbind(ac_d1_t1_fimo, ac_d1_t2_fimo, ac_d1_t3_fimo, ac_d1_t4_fimo)
ac_d1_t1_t2_fimo <- filter(ac_d1_t1_fimo, name %in% ac_d1_t1_t2_bed$name)
ac_d1_t1_t3_fimo <- filter(ac_d1_t1_fimo, name %in% ac_d1_t1_t3_bed$name)
ac_d1_t1_t4_fimo <- filter(ac_d1_t1_fimo, name %in% ac_d1_t1_t4_bed$name)
rm(ac_fimo)

#calculate the number of unique peaks a motif occurs in
ac_d1_t1_occur <- ac_d1_t1_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t2_occur <- ac_d1_t2_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t3_occur <- ac_d1_t3_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t4_occur <- ac_d1_t4_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t12_occur <- ac_d1_t12_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t123_occur <- ac_d1_t123_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t1234_occur <- ac_d1_t1234_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t1_t1_occur <- ac_d1_t1_t1_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t1_t2_occur <- ac_d1_t1_t2_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t1_t3_occur <- ac_d1_t1_t3_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())
ac_d1_t1_t4_occur <- ac_d1_t1_t4_fimo %>% select(motif_id, name) %>% unique() %>% group_by(motif_id)
%>% summarize(occurrences = n())

#get q-values
ac_d1_t1_evo <- ac_d1_t1_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join(ac_d1_t1_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t2_evo <- ac_d1_t2_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join(ac_d1_t2_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t3_evo <- ac_d1_t3_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join(ac_d1_t3_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t4_evo <- ac_d1_t4_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join(ac_d1_t4_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t12_evo <- ac_d1_t12_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join(ac_d1_t12_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t123_evo <- ac_d1_t123_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join(ac_d1_t123_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t1234_evo <- ac_d1_t1234_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join

```

```

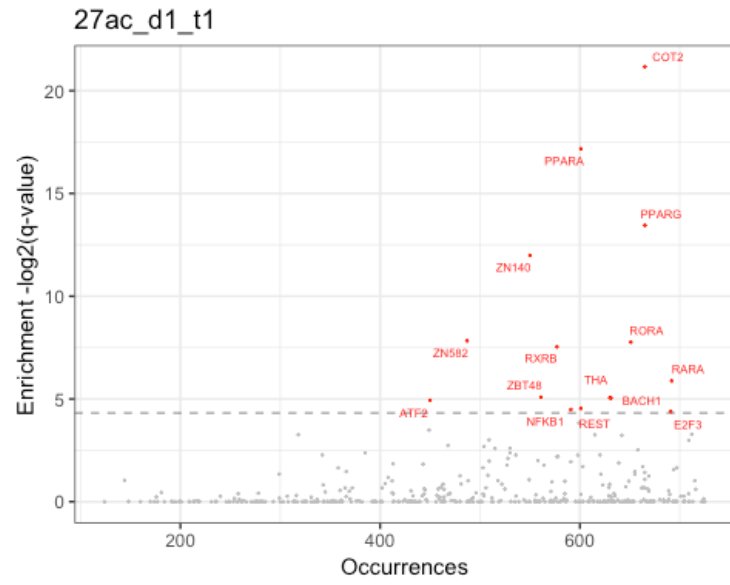
(ac_d1_t1234_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t1_t1_evo <- ac_d1_t1_t1_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join
(ac_d1_t1_t1_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t1_t2_evo <- ac_d1_t1_t2_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join
(ac_d1_t1_t2_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t1_t3_evo <- ac_d1_t1_t3_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join
(ac_d1_t1_t3_occur, by = c("motif_ID" = "motif_id"))
ac_d1_t1_t4_evo <- ac_d1_t1_t4_ame %>% select(motif_ID, qValue = `adj_p-value`) %>% inner_join
(ac_d1_t1_t4_occur, by = c("motif_ID" = "motif_id"))

#take the -Log2 of the q-values
ac_d1_t1_evo <- mutate(ac_d1_t1_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("motif_
ID"), sep = "\\.", extra = "drop")
ac_d1_t2_evo <- mutate(ac_d1_t2_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("motif_
ID"), sep = "\\.", extra = "drop")
ac_d1_t3_evo <- mutate(ac_d1_t3_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("motif_
ID"), sep = "\\.", extra = "drop")
ac_d1_t4_evo <- mutate(ac_d1_t4_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("motif_
ID"), sep = "\\.", extra = "drop")
ac_d1_t12_evo <- mutate(ac_d1_t12_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("moti
f_ID"), sep = "\\.", extra = "drop")
ac_d1_t123_evo <- mutate(ac_d1_t123_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("mo
tif_ID"), sep = "\\.", extra = "drop")
ac_d1_t1234_evo <- mutate(ac_d1_t1234_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("
motif_ID"), sep = "\\.", extra = "drop")
ac_d1_t1_t1_evo <- mutate(ac_d1_t1_t1_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("
motif_ID"), sep = "\\.", extra = "drop")
ac_d1_t1_t2_evo <- mutate(ac_d1_t1_t2_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("
motif_ID"), sep = "\\.", extra = "drop")
ac_d1_t1_t3_evo <- mutate(ac_d1_t1_t3_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("
motif_ID"), sep = "\\.", extra = "drop")
ac_d1_t1_t4_evo <- mutate(ac_d1_t1_t4_evo, loggedQ = -log2(qValue)) %>% separate(motif_ID, c("
motif_ID"), sep = "\\.", extra = "drop")

#create color column by sig threshold -Log2(0.05)
ac_d1_t1_evo <- mutate(ac_d1_t1_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t2_evo <- mutate(ac_d1_t2_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t3_evo <- mutate(ac_d1_t3_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t4_evo <- mutate(ac_d1_t4_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t12_evo <- mutate(ac_d1_t12_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t123_evo <- mutate(ac_d1_t123_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t1234_evo <- mutate(ac_d1_t1234_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t1_t1_evo <- mutate(ac_d1_t1_t1_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t1_t2_evo <- mutate(ac_d1_t1_t2_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t1_t3_evo <- mutate(ac_d1_t1_t3_evo, colors = loggedQ >= -log2(0.05))
ac_d1_t1_t4_evo <- mutate(ac_d1_t1_t4_evo, colors = loggedQ >= -log2(0.05))

#plot enrichment (q-value) vs occurrence
ac_d1_t1_evo_sctr <- ggplot(data = ac_d1_t1_evo, aes( occurrences,loggedQ, color = colors)) +
  theme_bw()+
  theme(legend.position = "none")+
  geom_hline(yintercept = -log2(0.05), color = "black", linetype = "dashed", size = 0.25, alph
a = 0.8) +
  geom_point(size = 0.1) +
  geom_text_repel(data = filter(ac_d1_t1_evo, colors == TRUE), aes(label = motif_ID), size = 2
) +
  labs(title = "27ac_d1_t1", y = "Enrichment -log2(q-value)", x = "Occurrences")+
  scale_color_manual(values = c("FALSE" = "grey", "TRUE" = "red"))
show(ac_d1_t1_evo_sctr)

```

References

1. Hales, C. M.; Carroll, M. D.; Fryar, C. D.; Ogden, C. L., Prevalence of obesity among adults and youth: United States, 2015-2016. *NCHS Data Brief* **2017**, 288.
2. Mann, T.; Tomiyama, A. J.; Westling, E.; Lew, A.-M.; Samuels, B.; Chatman, J., Medicare's search for effective obesity treatments: diets are not the answer. *Am. Psychol.* **2007**, 62 (3), 220-233.
3. Kopelman, P., Health risks associated with overweight and obesity. *Obes. Rev.* **2007**, 8 (1), 13-17.
4. Kissebah, A. H.; Freedman, D. S.; Peiris, A. N., Health risks of obesity. *Med. Clin. North Am.* **1989**, 73 (1), 111-138.
5. Centers for Disease Control, National Diabetes Statistics Report, 2017; Estimates of Diabetes and Its Burden in the United States. **2017**.
6. Centers for Disease Control; Division of Diabetes Translation, Maps of trends in diagnosed diabetes and obesity. **2017**.
7. Wooley, S. C.; Garner, D. M., Obesity treatment: the high cost of false hope. *J. Am. Diet. Assoc.* **1991**, 91 (10), 1248-1251.
8. Duarte, C.; Pinto-Gouveia, J.; Ferreira, C., Ashamed and fused with body image and eating: Binge eating as an avoidance strategy. *Clin. Psychol. Psychother.* **2017**, 24 (1), 195-202.
9. Nutter, S.; Russell-Mayhew, S.; Arthur, N.; Ellard, J. H., Weight bias as a social justice issue: A call for dialogue. *Can. Psychol.* **2018**, 59 (1), 89-99.
10. Poulsen, P.; Kyvik, K. O.; Vaag, A.; Beck-Nielsen, H., Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. *Diabetologia* **1999**, 42 (2), 139-145.
11. Jimenez-Chillaron, J. C.; Isganaitis, E.; Charalambous, M.; Gesta, S.; Pentinat-Pelegrin, T.; Faucette, R. R.; Otis, J. P.; Chow, A.; Diaz, R.; Ferguson-Smith, A., Intergenerational transmission of glucose intolerance and obesity by in utero undernutrition in mice. *Diabetes* **2009**, 58 (2), 460-468.
12. Barnett, A.; Eff, C.; Leslie, R. D.; Pyke, D., Diabetes in identical twins. *Diabetologia* **1981**, 20 (2), 87-93.
13. Mahajan, A.; Taliun, D.; Thurner, M.; Robertson, N. R.; Torres, J. M.; Rayner, N. W.; Payne, A. J.; Steinthorsdottir, V.; Scott, R. A.; Grarup, N., Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **2018**, 50 (11), 1505-1513.
14. Gingerich, P. D., Rates of evolution on the time scale of the evolutionary process. *Genetica* **2001**, 112-113 (1), 127-144.
15. Wilcox, G., Insulin and insulin resistance. *Clin. Biochem. Rev.* **2005**, 26 (2), 19-39.
16. Khan, A.; Pessin, J., Insulin regulation of glucose uptake: a complex interplay of intracellular signalling pathways. *Diabetologia* **2002**, 45 (11), 1475-1483.
17. Rosen, E. D.; Spiegelman, B. M., What we talk about when we talk about fat. *Cell* **2014**, 156 (1-2), 20-44.
18. Hua, Q.-X.; Gozani, S. N.; Chance, R. E.; Hoffmann, J. A.; Frank, B. H.; Weiss, M. A., Structure of a protein in a kinetic trap. *Nat. Struct. Biol.* **1995**, 2 (2), 129-138.

19. Weis, F.; Menting, J. G.; Margetts, M. B.; Chan, S. J.; Xu, Y.; Tennagels, N.; Wohlfart, P.; Langer, T.; Muller, C. W.; Dreyer, M. K., et al., The signalling conformation of the insulin receptor ectodomain. *Nat. Commun.* **2018**, *9* (1), 4420.
20. *PyMol*, version 2.2.3; Schrodinger: Cambridge, MA, 2018.
21. JeBailey, L.; Wanono, O.; Niu, W.; Roessler, J.; Rudich, A.; Klip, A., Ceramide-and oxidant-induced insulin resistance involve loss of insulin-dependent Rac-activation and actin remodeling in muscle cells. *Diabetes* **2007**, *56* (2), 394-403.
22. Hoehn, K. L.; Hohnen-Behrens, C.; Cederberg, A.; Wu, L. E.; Turner, N.; Yuasa, T.; Ebina, Y.; James, D. E., IRS1-independent defects define major nodes of insulin resistance. *Cell Metab.* **2008**, *7* (5), 421-433.
23. Ness-Abramof, R.; Apovian, C. M., Drug-induced weight gain. *Drugs of today* **2005**, *41* (8), 547.
24. Trayhurn, P., Endocrine and signalling role of adipose tissue: new perspectives on fat. *Acta Physiol. Scand.* **2005**, *184* (4), 285-293.
25. Rosen, E. D.; Spiegelman, B. M., Adipocytes as regulators of energy balance and glucose homeostasis. *Nature* **2006**, *444* (7121), 847-853.
26. Hayward, J. S.; Lisson, P. A., Evolution of brown fat: its absence in marsupials and monotremes. *Can. J. Zool.* **1992**, *70* (1), 171-179.
27. Avram, A. S.; Avram, M. M.; James, W. D., Subcutaneous fat in normal and diseased states: 2. Anatomy and physiology of white and brown adipose tissue. *J. Am. Acad. Dermatol.* **2005**, *53* (4), 671-683.
28. Morrison, R. F.; Farmer, S. R., Hormonal signaling and transcriptional control of adipocyte differentiation. *J. Nutr.* **2000**, *130* (12), 3116S-3121S.
29. Yu, C.; Chen, Y.; Cline, G. W.; Zhang, D.; Zong, H.; Wang, Y.; Bergeron, R.; Kim, J. K.; Cushman, S. W.; Cooney, G. J., et al., Mechanism by which fatty acids inhibit insulin activation of IRS-1 associated phosphatidylinositol 3-kinase activity in muscle. *J. Biol. Chem.* **2002**, *277* (52), 50230-50236.
30. Wellen, K. E.; Hotamisligil, G. S., Inflammation, stress, and diabetes. *The Journal of clinical investigation* **2005**, *115* (5), 1111-1119.
31. Lee, M.-J.; Wu, Y.; Fried, S. K., Adipose tissue heterogeneity: implication of depot differences in adipose tissue for obesity complications. *Mol. Aspects Med.* **2013**, *34* (1), 1-11.
32. Emdin, C. A.; Khera, A. V.; Natarajan, P.; Klarin, D.; Zekavat, S. M.; Hsiao, A. J.; Kathiresan, S., Genetic association of waist-to-hip ratio with cardiometabolic traits, type 2 diabetes, and coronary heart disease. *JAMA* **2017**, *317* (6), 626-634.
33. Pischon, T.; Boeing, H.; Hoffmann, K.; Bergmann, M.; Schulze, M. B.; Overvad, K.; Van Der Schouw, Y.; Spencer, E.; Moons, K.; Tjønneland, A., et al., General and abdominal adiposity and risk of death in Europe. *N. Engl. J. Med.* **2008**, *359* (20), 2105-2120.
34. Shungin, D.; Winkler, T. W.; Croteau-Chonka, D. C.; Ferreira, T.; Locke, A. E.; Magi, R.; Strawbridge, R. J.; Pers, T. H.; Fischer, K.; Justice, A. E., et al., New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **2015**, *518* (7538), 187-196.
35. Tak, Y. G.; Farnham, P. J., Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenet. Chromatin* **2015**, *8* (1), 57.

36. Haeussler, M.; Zweig, A. S.; Tyner, C.; Speir, M. L.; Rosenbloom, K. R.; Raney, B. J.; Lee, C. M.; Lee, B. T.; Hinrichs, A. S.; Gonzalez, J. N., et al., The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **2019**, *47* (D1), D853-D858.
37. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W., et al., Initial sequencing and analysis of the human genome. *Nature* **2001**, *409* (6822), 860-921.
38. Calo, E.; Wysocka, J., Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **2013**, *49* (5), 825-837.
39. Creighton, M. P.; Cheng, A. W.; Welstead, G. G.; Kooistra, T.; Carey, B. W.; Steine, E. J.; Hanna, J.; Lodato, M. A.; Frampton, G. M.; Sharp, P. A., et al., Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (50), 21931-21936.
40. Consortium, E. P., An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489* (7414), 57-74.
41. Davis, C. A.; Hitz, B. C.; Sloan, C. A.; Chan, E. T.; Davidson, J. M.; Gabdank, I.; Hilton, J. A.; Jain, K.; Baymuradov, U. K.; Narayanan, A. K., et al., The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **2018**, *46* (D1), D794-D801.
42. Matthews, D.; Hosker, J.; Rudenski, A.; Naylor, B.; Treacher, D.; Turner, R., Homeostasis model assessment: insulin resistance and β -cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **1985**, *28* (7), 412-419.
43. Katz, A.; Nambi, S. S.; Mather, K.; Baron, A. D.; Follmann, D. A.; Sullivan, G.; Quon, M. J., Quantitative insulin sensitivity check index: a simple, accurate method for assessing insulin sensitivity in humans. *J. Clin. Endocrinol. Metab.* **2000**, *85* (7), 2402-2410.
44. Mikkelsen, T. S.; Ku, M.; Jaffe, D. B.; Issac, B.; Lieberman, E.; Giannoukos, G.; Alvarez, P.; Brockman, W.; Kim, T. K.; Koche, R. P., et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **2007**, *448* (7153), 553-560.
45. *Picard Toolkit*, version 2.0.1; Broad Institute: Cambridge, MA, 2019.
46. Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9* (4), 357-359.
47. Zhang, Y.; Liu, T.; Meyer, C. A.; Eeckhoute, J.; Johnson, D. S.; Bernstein, B. E.; Nusbaum, C.; Myers, R. M.; Brown, M.; Li, W., et al., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **2008**, *9* (9), R137.
48. Quinlan, A. R.; Hall, I. M., BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26* (6), 841-842.
49. McCarthy, D. J.; Chen, Y.; Smyth, G. K., Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **2012**, *40* (10), 4288-4297.
50. Robinson, M. D.; McCarthy, D. J.; Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26* (1), 139-140.
51. Grant, C. E.; Bailey, T. L.; Noble, W. S., FIMO: scanning for occurrences of a given motif. *Bioinformatics* **2011**, *27* (7), 1017-1018.
52. McLeay, R. C.; Bailey, T. L., Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinf.* **2010**, *11* (1), 165.

53. Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S., MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202-208.
54. Morris, A. P.; Voight, B. F.; Teslovich, T. M.; Ferreira, T.; Segre, A. V.; Steinthorsdottir, V.; Strawbridge, R. J.; Khan, H.; Grallert, H.; Mahajan, A., Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **2012**, *44* (9), 981-990.
55. Lambert, J.-C.; Ibrahim-Verbaas, C. A.; Harold, D.; Naj, A. C.; Sims, R.; Bellenguez, C.; Jun, G.; DeStefano, A. L.; Bis, J. C.; Beecham, G. W., Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **2013**, *45* (12), 1452-1458.
56. Bulik-Sullivan, B. K.; Loh, P.-R.; Finucane, H. K.; Ripke, S.; Yang, J.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson, N.; Daly, M. J.; Price, A. L.; Neale, B. M., LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **2015**, *47* (3), 291-295.
57. Finucane, H. K.; Bulik-Sullivan, B.; Gusev, A.; Trynka, G.; Reshef, Y.; Loh, P. R.; Anttila, V.; Xu, H.; Zang, C.; Farh, K., et al., Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **2015**, *47* (11), 1228-1235.
58. *R: A Language and Environment for Statistical Computing*, version 3.5.3; R Foundation for Statistical Computing: Vienna, Austria, 2019.
59. *RStudio: Integrated Development for R*, version 1.1.463; RStudio: Boston, MA, 2015.
60. Kulakovskiy, I. V.; Vorontsov, I. E.; Yevshin, I. S.; Sharipov, R. N.; Fedorova, A. D.; Rumynskiy, E. I.; Medvedeva, Y. A.; Magana-Mora, A.; Bajic, V. B.; Papatsenko, D. A., et al., HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **2018**, *46* (D1), D252-D259.
61. Haluzik, M.; Haluzik, M., PPAR-alpha and insulin sensitivity. *Physiol. Res.* **2006**, *55* (2), 115-122.
62. Spiegelman, B. M., PPAR-gamma: adipogenic regulator and thiazolidinedione receptor. *Diabetes* **1998**, *47* (4), 507-514.
63. Mori, T.; Sakaue, H.; Iguchi, H.; Gomi, H.; Okada, Y.; Takashima, Y.; Nakamura, K.; Nakamura, T.; Yamauchi, T.; Kubota, N., et al., Role of Kruppel-like factor 15 (KLF15) in transcriptional regulation of adipogenesis. *J. Biol. Chem.* **2005**, *280* (13), 12867-12875.
64. Rosen, E. D.; MacDougald, O. A., Adipocyte differentiation from the inside out. *Nat. Rev. Mol. Cell Biol.* **2006**, *7* (12), 885-896.
65. Consortium, U., UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506-D515.
66. Kersten, S.; Desvergne, B.; Wahli, W., Roles of PPARs in health and disease. *Nature* **2000**, *405* (6785), 421-424.
67. Pereira, F. A.; Qiu, Y.; Zhou, G.; Tsai, M.-J.; Tsai, S. Y., The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. *Genes Dev.* **1999**, *13* (8), 1037-1049.
68. Fukumura, D.; Ushiyama, A.; Duda, D. G.; Xu, L.; Tam, J.; Krishna, V.; Chatterjee, K.; Garkavtsev, I.; Jain, R. K., Paracrine regulation of angiogenesis and adipocyte differentiation during in vivo adipogenesis. *Circ. Res.* **2003**, *93* (9), e88-e97.

69. Shungin, D.; Winkler, T. W.; Croteau-Chonka, D. C.; Ferreira, T.; Locke, A. E.; Mägi, R.; Strawbridge, R. J.; Pers, T. H.; Fischer, K.; Justice, A. E., New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **2015**, *518* (7538), 187.
70. Mikkelsen, T. S.; Xu, Z.; Zhang, X.; Wang, L.; Gimble, J. M.; Lander, E. S.; Rosen, E. D., Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **2010**, *143* (1), 156-169.