Preliminary Exam Proposal   **Name:** Chinmay Raut **Dissertation Advisor:** Elizabeth Speliotes
Committee: Margit Burmeister, Ryan Mills, Stephen CJ Parker, Maureen Sartor

Studying the Effects of Rare Variants in Metabolic Syndrome

Specific Aims

Metabolic syndrome (MetS) is a combination of 5 health risk factors, including abdominal obesity, high blood pressure, high triglycerides, low high-density lipoprotein (HDL) cholesterol, and high fasting blood sugar, that together increase the risk for cardiovascular disease, type 2 diabetes, stroke, and other chronic conditions [1]. MetS diagnosis is based on thresholds for these 5 quantitative traits, and individuals must surpass the thresholds on 3/5 of these traits for diagnosis making this a complex disease to study. Up to 1 in 3 Americans currently satisfy the MetS conditions making it a significant public health challenge [1]. The underlying mechanisms of metabolic syndrome are not fully understood but are thought to involve a genetic component with MetS having an estimated heritability of about 20-30% [2]. There is a need for further research to determine the underlying causes of MetS and to develop effective strategies for prevention and treatment.

Previous studies have identified hundreds of associations between common genetic variants and MetS using Genome-Wide Association Studies (GWAS) on databases of up to 300,000 individuals [3]. However, little is known about the effects of rare variants on MetS, and no large-scale ($>$100,000 individuals) rare-variant associations have been conducted on MetS [4]. Rare genetic variants (minor allele frequency $<$0.05) are worth studying for three reasons: they make up most of the sites of variation across the genome [5], they are predicted to result in significant phenotypic changes if they reside in protein-coding regions [5], and studying effects in protein-coding regions also provides a direct route for future functional studies and therapeutic targets [6]. With the integration of Next-Generation Sequencing and longer read technology, genetic repositories like the UK Biobank (UKBB) recently released their Whole Exome Sequencing (WES) data (26 million variants across 450,000+ individuals) resulting in high-quality data on low-frequency genetic variants [5]. Currently, studies on rare variants in GWAS are limited by power due to low sample counts of individuals within a selected rare variant. A series of approaches have been developed to combat these limitations including the construction of the optimal Sequence Kernel Association Test (SKAT-O) [7] which combines a set of rare variants together to increase the power of the association test. When the SKAT-O is paired with a variant annotator like the WGS Annotator (WGSA) [8] this results in the direct implication of genes to an outcome.

To understand the role genetic variants have on MetS, we want to identify and study its associations with rare, protein-coding variations using data from Biobanks. Exonic variants have the potential to alter proteins which then directly affect MetS phenotypes. Therefore, studying protein-altering variants offers direct insight into the biological mechanisms of MetS. We hypothesize that using WGS and WES data from large Biobanks such as the UKBB, we can detect exonic variants resulting in protein-coding changes which affect MetS. This will establish a series of protein-altering variants that provide insight into factors of MetS either for improvements in genetic testing or follow-ups for functional studies to investigate the biological mechanisms underlying the associations.

Aim 1: Determine Novel Gene-Based Associations with MetS.

Aim 1a: Identify genes associated with components of disease. Since MetS diagnosis depends on 5 different quantitative traits(waist circumference, diastolic & systolic blood pressure, triglycerides, HDL, and fasting blood glucose), we can extract each of those phenotypes from our discovery study cohort. We then run gene-based analysis techniques such as the SKAT-O to identify associations between genes and each of these MetS components.

Aim 1b: Integrate discovered signals across components to study the holistic effect on disease. After identifying significant genes across the 5 factors of MetS, we will conduct enrichment analysis on the gene-based results from Aim 1a genes associated with multiple MetS profiles. We will also perform gene-based association tests against a joint MetS diagnosis and severity score variables.

Based on the rare exonic variants we expect to identify and implicate specific genes that associate with components of MetS. We expect to construct a network of how these genes affect MetS as a whole, and if certain genes share similarities in effect on MetS providing further insights into its genetic basis.

Aim 2: Determine sequence-specific mechanisms of disease and progression.

Aim 2a: Within a gene, identify loci of individual variants affecting MetS. We will perform Leave One Variant Out (LOVO) studies on the effects of variants within an implicated gene to understand if the effects of the protein-altering variants are restricted to a specific region of the gene or not.

Aim 2b: Design CRISPR CAS9 KO guides based on exonic variants. Based on the specific type of rare variants identified (frameshift, splicing, missense, etc. . .) we will design a CAS9 guide library surrounding those regions to study the functional effects of the identified variants in vitro.

We expect to produce a series of specific genomic loci potentially responsible for changes in MetS along with a CAS9 library for future functional experiments. This provides a direct target for future studies regarding the effects of genetic modification on MetS.

By studying new next-generation sequencing data to understand the role of rare genetic variants on MetS, we expect to identify genes that affect MetS and also provide biological insight into how they relate to the components of the condition. This study would lay the groundwork for further functional tests to investigate the biology behind the genetics of MetS and provide a genetics-based informatory tool to understand how one's genome contributes to risks in MetS.

Background and Motivation

Metabolic syndrome (MetS) is a combination of 5 health risk factors, including abdominal obesity, high blood pressure, high triglycerides, low high-density lipoprotein (HDL) cholesterol, and high fasting blood sugar, that together increase the risk for cardiovascular disease, type 2 diabetes, stroke, and other chronic conditions. 33% of Americans currently satisfy the MetS conditions making it a significant public health challenge. The underlying causes or biological processes behind MetS are not fully understood but the disease is thought to have a genetic component with a heritability estimate of 20-30%.

Genome-wide association studies (GWAS) are a powerful approach used in genetics research to identify genetic variations associated with complex diseases or traits. In a GWAS, researchers analyze the genetic information of large groups of individuals to identify genetic markers, or locations along the genome, that are associated with the disease or trait of interest. By comparing the frequency of these genetic markers between individuals with the disease or trait and those without, researchers can identify specific genetic variations that may relate to the development of the disease or trait. GWAS for MetS have identified a couple of promising candidate genes such as APOB, LPL, CETP, APOA5, GCKR, and ZNF259 to be associated with MetS incidence in large European populations (¿300k individuals). These genes implicate lipid metabolism to be a major contributor of MetS. However, since 2 of the defining characteristics for MetS are defined by lipids (HDL and triglycerides) and these genes have a lesser impact on the other 3 characteristics of the disease, it suggests that these genes alone do not explain the entire genetic effect on MetS.

Another limitation of most GWAS on MetS is that they are unable to detect associations between rare variants, defined as having a frequency of less than 1% in the population. Previous studies tended to focus on common variants as this data is easier to acquire and study due to their increased imputation accuracy and prevalence in a population. However, one trend that has been identified through multiple GWAS is that rarer variants tend to have much larger phenotypic effects on an individual 1. There is potential that some of the large individual effects of genetic variants on MetS can be discovered through rare variant analyses. The primary problem of detecting rare variant associations lies within the rare nature of the variants. By definition, individuals with a selected rare variant make up less than 1% of the population leading to massive sample size imbalances across carriers of a mutation vs non-carriers during the statistical tests. An

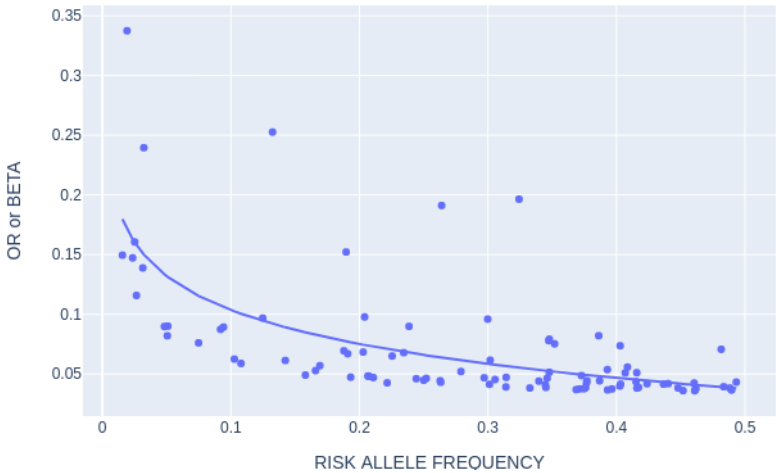Scatterplot of Risk Allele Frequency and Effect Size from Lind 2019 MetS GWAS results

Figure 1: **Scatterplot of Risk Allele Frquency and Effect Size** (reconstructed from Lind 2019 MetS GWAS results). Effect size of variant increases inversely to allele frequency.

alternative approach is to pool the effects of multiple rare variants, that we expect to behave similarly, together to combat the poor sample size imbalance. If we combine people that have any of multiple rare variants in an area we might have less imbalance of carrier vs non-carrier. One approach to this technique is known as the Optimal Sequence Kernal Association Test (SKAT-O). Most rare-variant associations use the SKAT-O in conjunction with variant annotation tools to test groups of variants belonging to the same gene or genomic region to directly implicate new genes with outcomes instead of specific point mutations.

Advancements in sequencing technology have resulted in a technique known as Whole exome sequencing (WES) which is used to sequence the protein-coding regions of an individual's genome, also known as the exome. WES offers additional insight particularly for rare variant GWAS as it is known to be more accurate than genotyping or imputation due to its ability to reconstruct exact sequences from an individual. There have been a couple of studies looking at the rare-variant associations in MetS using WES data. A study focusing on about 10,000 individuals with Korean ancestry identified MetS associations with rare variants in the CETP region. However, this study faced some limitations particularly due to the smaller sample size compared to previous studies. Due to the relatively small effects of genetic variants, it is recommended to run GWAS on large populations to increase the power of the tests and their ability to detect suspicious loci.

Large biobanks such as the UK Biobank (UKBB) have recently released their WES data for a sample of over 450k European individuals. This offers the opportunity to study the effects of rare variants across a large cohort of individuals and perhaps discover novel protein-coding variants associated with MetS. One of the benefits of the UKBB WES data is that mutations observed are exonic which means that their implication with genes and proteins will be easier to study and assess compared to variants identified between 2 genes or in a non-coding region of the genome. We hypothesize that by studying rare variant associations through SKAT-O in MetS using the UKBB WES data, we can discover new genes or loci implicated with disease and further increase our understanding of the genetic factors associated with MetS.

Significance

No large-scale (100k individuals) studies have been conducted on rare variants for MetS. Rare variants are important to study due to their relatively large phenotypic effects compared to common variants. An exome-wide analysis of rare variants associating with MetS can help identify protein-coding changes that are related to

disease providing further biological insight into the mechanisms at play. Furthermore, such a study will allow for directed functional follow-up and potentially guide new therapeutic approaches to MetS.

Datasets

UK Biobank (discovery): The UKBB large-scale population-based study consists of extensive health and lifestyle information from over 500,000 participants including various blood biochemistry data and sequence data. WES was performed on 454,787 participants across ¿26 million variants.

Michigan Genomics Initiative (validation): The MGI is a precision health research program based at the University of Michigan in Ann Arbor consisting of an opt-in study for hospital patients at the Michigan Medicine Hospital. This dataset consists of medical records for over 100,000 patients. Recently 70,439 individuals have been genotyped and imputed to the Trans-Omics for Precision Medicine (TopMed) rare variant reference panel consisting of 50 million variants.

Research Design and Methods

Aim 1: Determine Novel Gene-Based Associations with MetS.

Aim 1a: Identify genes associated with components of disease.

Rationale:

Due to the complex and multifaceted nature of MetS, it has been proposed in earlier studies that one consider the 5 distinct characteristics that lead to MetS diagnosis independently first. Most studies tend to analyze MetS across 2 steps: its individual risk factors and the aggregated diagnosis. This provides the added benefit of understanding how the different factors of MetS relate and differ from each other. Futhermore, if 2 associations affect different components of the disease in opposing fashion, a study focusing only on the aggregate variable of MetS will not be able to identify such relationships.

Approach:

First, the EWS data from the UKBB must be carefully assessed and standard GWAS Quality Control (QC) procedures should be performed. These include removing any sample missing more than 10% of the variants, removing any variant missing from more than 10% of the samples, variants outside Hardy-Weinberg equilibrium (HWE), variants with insufficient read-depth coverage, and variants with insufficient allelic balance. This QC process will primarily be completed through Plink and bcftools. This processed data will serve as the primary exonic dataset for discovery.

Next, the exonic variants will need to be annotated to a gene. Multiple annotation tools exist but we will use ANNOVAR and the RefGene database to implicate a variant with a gene. The refGene database also consists of functional annotations of which we will further label our exonic variants into 3 categories: loss-of-function (LoF), missense, or synonymous. Since the goal of this project is to identify rare exonic variants, any variant that does not contain an exonic annotation or has a population-level minor allele frequency ¿1% will be excluded. This step should produce 3 lists of variants for each gene (LoF, missense, synonymous).

Then, we will construct the phenotypes. The risk criteria across the 5 factors are blood pressure $\geq$ 130/85 mmHg or antihypertensive treatment, serum glucose $\geq$ 6.1 mmol/L or antidiabetic treatment, serum triglycerides $\geq$ 1.7 mmol/L, waist circumference 102 cm in men and 88 cm in women, HDL-cholesterol < 1.0 mmol/L in men and < 1.3 mmol/L in women for MetS diagnosis. In order to maintain the power of the GWAS we will opt to keep the traits as continuous variables. For traits involving controlling for medication, we will supplement medication values with the average value from the increased risk group. For example, for blood pressure, if an individual was taking antihypertensive treatment we will replace their blood pressure value with the average blood pressure across individuals with $\geq$ 130/85 mmHg. Triglycerides and glucose will be adjusted for fasting time in accordance with the approach from Lind 2019. Differences across sex will be controlled for in the genetic model. We will have 5 MetS characteristic phenotypes which will then each be rank-based inverse normal transformed to allow for comparison across each other.

Then, the SKAT-O tests will be conducted against each list of variants for each gene adjusting for 1st 10 genetic principle components (to deal with population stratification), age & squared age, and sex in accordance with the following null model:

$$y_{\text{risk factor}} \sim \sum_{i=1}^{10} \text{PC}_i + \text{age} + \text{age}^2 + \text{sex}$$

The SKAT-O tests will be performed using the SAIGE-GENE+ software to conduct a large number of tests quickly and efficiently over a large dataset such as the UKBB. Tests across the same gene (but different sets of variants) can be aggregated via the Cauchy Combination Test (CCT) in accordance with previous methods. We will attempt to obtain a list of exome-wide significant (EWS) genes defined as having a CCT SKAT-O p-value le 5e-6.

We will further assess that our implicated genes are free from LD with previously reported common variants by performing a conditional analysis on the 241 previously reported MetS associations from the GWAS catalog. The conditional analysis will involve recalculating the SKAT-O p-values for the previously implicated genes under a slightly changed altered null model:

$$y_{\text{risk factor}} \sim \sum_{i=1}^{10} \text{PC}_i + \text{age} + \text{age}^2 + \text{sex} + \sum_{j=0}^{M} \text{Dosage}_j$$

Where we are controlling for M markers and Dosage_j is the allele count (0,1,2) at marker j. By incorporating previous genetic signal into the null model we can assure that any associations identified are independent of previously implicated variants. We expect to test far fewer genes with the conditional models so we can use the standard p < 0.05 with Bonferroni correction as the significance threshold instead.

This should produce a list of novel genes significant for each of the 5 characteristics of MetS based on results from rare variant associations.

Potential Pitfalls & Alternative Approaches:

The low power for detecting rare-variant associations is a major concern, and it is possible that no genes pass the EWS threshold or the Bonferroni corrected conditional threshold for a given MetS trait.

One reason could arise could be due to a low proportion of causal variants within the variant lists for a given gene. It has been demonstrated that the rare-variant aggregation tests are extremely sensitive to proportion of causal variants within the variant lists for each gene. If a majority of variants within a variant list are not causal then the test will become underpowered. One way to address this concern would be to refine our lists of variants by subdividing variant lists. We can subdivide the variant lists by minor allele frequency category if we believe that only variants within a specific range of allele frequency are more predisposed to be causal. We can also subdivide the variant lists via a sliding window approach if we believe the causal variants to be unique to a specific region of the gene.

Aim 1b: Integrate discovered signals across components to study the holistic effect on disease.

Rationale:

In 1a we identify associated genes with components of MetS; however, previous studies have mentioned that studying the 5 factors independently neglects the joint and diverse nature of the disease. We can run a GWAS on the rare variants against MetS diagnosis; however, we can also use the component-wise results to inform and study some of the relationships of MetS as a whole. This will provide a deeper understanding of different disease subtypes and the genetic diversity of MetS.

Approach:

First, we can construct the MetS phenotypes for the individuals in the study cohort following the criteria from 1a. We will code it as a binary phenotype signifying MetS incidence.

Then we can then perform the gene-based rare variant analysis procedure similar to the approach outlined in the previous sub-aim. We can leverage the same variant annotation technique, but the null model will be slightly different to account for the binary nature of the phenotype:

$$ln(\frac{\pi_{\text{MetS incidence}}}{1 - \pi_{\text{MetS incidence}}}) \sim \sum_{i=1}^{10} \text{PC}_i + \text{age} + \text{age}^2 + \text{sex} + \sum_{j=0}^{M} \text{Dosage}_j$$

Where $ln(\frac{\pi_{\text{MetS incidence}}}{1-\pi_{\text{MetS incidence}}})$ is the log-odds of having MetS. We would still want to account for previously implicated MetS variants.

We can then start to implicate genes to associate with MetS. First we can consider any EWS genes to associate with MetS. Then we can leverage our component-wise discovery of MetS genes to potentially implicate more genes. We can restrict our genes of consideration to the group of genes associated from 1a and pay a less severe multiple testing penalty by looking at those component-wise significant genes with bonferroni corrected p-value threshold. This would give us a list of genes that potentially associate with MetS either strongly through EWS or association through component and disease (Bonferroni correction).

We can further validate this set of newly implicated MetS genes through the use of an independent external validation cohort. For the external validation cohort such as MGI we can construct the MetS phenotype. Since MGI is a hospital based cohort we would prefer to use MetS clinical diagnosis over blood biochemistry results; however, for patients with no history of disease we can default to MetS incidence constructed by blood test results and in patient visits. We will then fit an similar model to the above SKAT-O tests; however, since this is an independent cohort and we are only seeking validation on novel MetS hits we will not condition on the previous MetS results. We will only test the genes implicated by discovery cohort and we will define a gene association as replicating with a p $\leq$ 0.05 in the MGI cohort. This will result in a list of genes that we more certain of associating with the MetS disease as a whole and not just study specific associations or false positives.

To identify MetS disease subtypes we can also create a heatmap of genes against associated MetS features, then perform hierarchical clustering to identify genes that behave similarly across different traits of MetS. We can then conduct pathway analysis using tools like Depict to further offer biological insight or reasons why certain genes would implicate together. We can also identify if the different MetS gene clusters consist of individuals which are also similar via other phenotypes by looking at a Phenome Wide Association Study (PheWAS)

Then we can construct the classical MetS phenotype according to the criteria listed in 1a. This will be a binary incidence variable, however.

Next, we can conduct a rare variant GWAS on our MetS variable both with and without conditioning on the previously reported MetS GWAS common variants.

We will report any genes or variants implicated with the MetS phenotype conditional on the previously reported variants. Due to the binary nature of the MetS phenotype variable, we expect to implicate fewer genes during this process.

However, since we have a short list of genes that implicate the individual attributes of MetS we do not need to pay a severe multiple-testing burden. We can also just focus on the MetS unconditional results for the genes implicated in 1a.

Then we validate our MetS genes using an independent dataset such as MGI. The quality control pipeline follows similarly to 1a, however, the phenotypes we consider will be restricted to only MetS diagnosis. We will not condition on the previous variants and due to the nature that this is an independent dataset of individuals we will maintain significance at $p <= 0.05$. We will only test the genes implicated with MetS from UKBB, we will not conduct a GWAS.

We can also study how the newly implicated variants relate to MetS risk through the construction of Genetic Risk Scores (GRS). Following the equation outlined below:

$$GRS = \sum_{i=1}^{C_n} \beta_{C_i} X_{C_i} + \sum_{j=1}^{R_n} \beta_{R_j} X_{R_j}$$

Where C_n is the number of common markers, β_{C_i} is the effect size of the common variant extracted from GWAS catalog, X_M is the dosage of marker M, and R_n is the number of rare markers, and β_{R_i} is the effect size of the rare variant from the SAIGE single variant intermediate results. Previous studies have identified that the effects of rare variants on GRS are relatively small due to the low frequency of rare variants within a population, but it may be worth considering due to the potential to better distinguish between the ultra-high risk (top 1%) group and others.

We can look at other adverse outcomes implicated with MetS such as cardiovascular disease or type 2 diabetes and test to study if individuals with rare variants in MetS-implicated genes have different distributions of disease incidence compared to individuals without mutations in implicated genes.

Finally, we should have a list of genes whose rare variants are associated with MetS along with a more holistic understanding of how individual factors contribute to MetS disease and risk.

One thing to note is that GWAS of dichotomous traits tend to be lower powered

Aim 2: Determine sequence-specific mechanisms of disease and progression.

Using the implicated genes from Aim 1 we wish to construct work on functional follow-ups for in vitro experiments. It is known that we can study the effect of the removal of the entire gene using CRISPR-associated protein 9 (CAS9) knockouts however to further guide potential therapies we can try to isolate the gene-specific likely sequences or mutations causing the observed phenotype association.

Aim 2a: Within a gene, identify loci of individual variants affecting MetS.

For each implicated gene from 1, we will perform Leave One Variant Out (LOVO) studies on the effects of variants within that gene This involves fitting the gene-based test on multiple combinations of omitting a variant from a gene-based test. We expect to receive a plot that looks like Fig 2 with the significance of the association test on the y and the position of the omitted variant on the x.
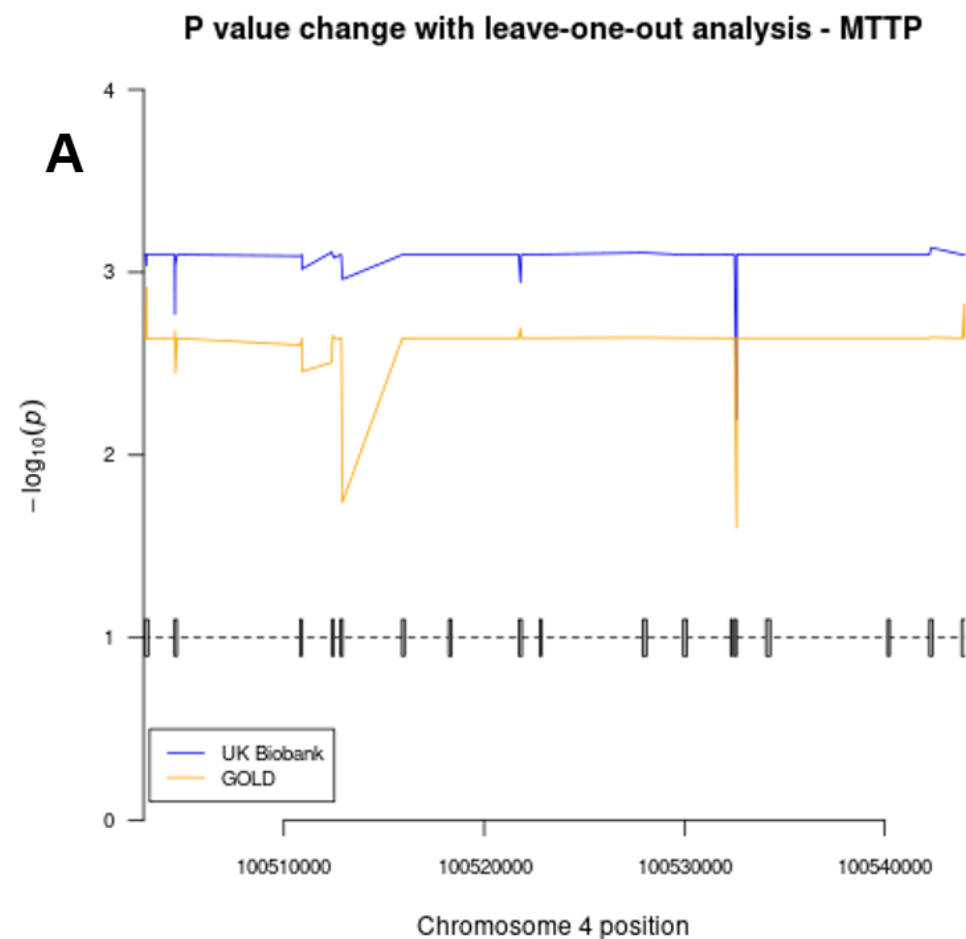

Figure 2: **Leave One Variant Out plot** (from Du et al 2021). Vertical drops signify associated single variants. V's signify regions of associating rare variants

We can then apply any peak-finding algorithm (like golden search or derivative approaches) to identify which variants resulted in large peaks of association signal change. This will help identify if the effects of the protein-altering variants are restricted to a specific region of the gene or not. we can then isolate specific regions of a gene and potential polypeptide sequences to study for follow-up analysis.

To verify that we have identified specific causal locations within a gene we can construct a gene-based test using only our implicated causal locations and identifying if the association improves or gets worse. If we do identify a singular causal region we expect the gene-based association p-value to improve. If the causal region of the gene cannot be well defined then we expect the gene-based association to perform worse.

5

We expect to have a list of genomic endpoints for each gene implicating the deleterious or important parts of a gene relating to MetS. If we cannot identify a single causal region then the endpoints should encapsulate the entire gene.

Aim 2b: Design CRISPR CAS9 KO guides based on exonic variants.

If the variants lie in the exons of a gene and result in a missense or loss-of-function mutation, we can even simulate the folded mutated protein using AlphaFold2 to assess hydrophobic and hydrophilic region changes across the proteins helping inform future functional studies for potential assays and tests to run with these results.

There are a plethora of tools for single-guide RNA (sgRNA) design for gene knockout experiments such as CRISPOR or CHOPCHOP, and we can use them to generate a list of potential gene knockout guides for each implicated gene.

Then for the genes from 2a of whom we were able to identify specific within gene sequences, we can further filter and assess the recommended guides to help build a more focused MetS gene knockout library.

The benefit of this library compared to a standard gene knockout library is the added ability to conduct partial knockouts to study not only the effect of gene omission but a targeted region-specific mutation of a MetS-implicated gene.

Finally, we expect to receive a set of sequences for each MetS-implicated gene that can directly be applied to a functional study of MetS for an in-vitro model.

Potential Outcomes and Conclusions

We expect to have identified a few new genes implicated to MetS through exonic rare variants along with a suite of sgRNA guides for follow-up CAS9 KO functional experiments.

However, due to the feed-forward nature of the study in which current results depend on previous results, we could fail to implicate any genes with MetS from Aim 1. In the case this happens we can Meta-Analysze UKBB and MGI together omitting the validation in an independent dataset step. We can make use of newer techniques such as MetaSTAAR which allows for gene-based meta-analysis.

Another limitation of the study is that the annotation of the variants greatly influences the gene-based results. If a variant is annotated to be loss-of-function, we assume that those types of variants have stronger effects compared to missense or synonymous variants. However, there may be regulatory functions in the body that can identify deleterious mutations and suppress gene expression of them negating their potential phenotypic effect. RefGene is a reputable resource however due to the many gaps in our understanding of biology its annotations may not be complete. If our implicated MetS genes are strange or don't validate with previous studies we might try different annotation methods for variants. A recent annotation method tool called FAVOR which makes use of more variant features might be able to address some of those weaknesses.

Background and Motivation
Significance of Research
Preliminary Data
Research Strategy
Rationale

Design & Methods

Potential Pitfalls & Alternative Approaches

Expected/Alternative outcomes

References

[1] M. G. Saklayen, "The Global Epidemic of the Metabolic Syndrome," *Curr Hypertens Rep*, vol. 20, p. 12, Feb 2018.

[2] F. Graziano, G. Biino, M. T. Bonati, B. M. Neale, R. Do, M. P. Concas, S. Vaccargiu, M. Pirastu, O. Terradura-Vagnarelli, M. Cirillo, M. Laurenzi, M. Mancini, A. Zanchetti, and M. Grassi, "Estimation of metabolic syndrome heritability in three large populations including full pedigree and genomic information," *Human Genetics*, vol. 138, pp. 739–748, Jul 2019.

[3] L. Lind, "Genome-Wide Association Study of the Metabolic Syndrome in UK Biobank," *Metab Syndr Relat Disord*, vol. 17, pp. 505–511, Dec 2019.

[4] H.-S. Lee, Y. Kim, and T. Park, "New common and rare variants influencing metabolic syndrome and its individual components in a korean population," *Scientific Reports*, vol. 8, p. 5701, Apr 2018.

[5] J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, A. Yadav, N. Banerjee, C. E. Gillies, A. Damask, S. Liu, X. Bai, A. Hawes, E. Maxwell, L. Gurski, K. Watanabe, J. A. Kosmicki, V. Rajagopal, J. Mighty, M. Jones, L. Mitnaul, E. Stahl, G. Coppola, E. Jorgenson, L. Habegger, W. J. Salerno, A. R. Shuldiner, L. A. Lotta, J. D. Overton, M. N. Cantor, J. G. Reid, G. Yancopoulos, H. M. Kang, J. Marchini, A. Baras, G. R. Abecasis, and M. A. R. Ferreira, "Exome sequencing and analysis of 454,787 UK Biobank participants," *Nature*, vol. 599, pp. 628–634, Nov 2021.

[6] N. Verweij, M. E. Haas, J. B. Nielsen, O. A. Sosina, M. Kim, P. Akbari, T. De, G. Hindy, J. Bovijn, T. Persaud, L. Miloscio, M. Germino, L. Panagis, K. Watanabe, J. Mbatchou, M. Jones, M. LeBlanc, S. Balasubramanian, C. Lammert, S. Enhörning, O. Melander, D. J. Carey, C. D. Still, T. Mirshahi, D. J. Rader, P. Parasoglou, J. R. Walls, J. D. Overton, J. G. Reid, A. Economides, M. N. Cantor, B. Zambrowicz, A. J. Murphy, G. R. Abecasis, M. A. Ferreira, E. Smagris, V. Gusarova, M. Sleeman, G. D. Yancopoulos, J. Marchini, H. M. Kang, K. Karalis, A. R. Shuldiner, G. Della Gatta, A. E. Locke, A. Baras, and L. A. Lotta, "Germline mutations in cideb and protection against liver disease," *New England Journal of Medicine*, vol. 387, no. 4, pp. 332–344, 2022. PMID: 35939579.

[7] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin, "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *Am J Hum Genet*, vol. 91, pp. 224–237, Aug 2012.

[8] X. Liu, S. White, B. Peng, A. D. Johnson, J. A. Brody, A. H. Li, Z. Huang, A. Carroll, P. Wei, R. Gibbs, R. J. Klein, and E. Boerwinkle, "WGSA: an annotation pipeline for human genome sequencing studies," *J Med Genet*, vol. 53, pp. 111–112, Feb 2016.