Title

Integrating multi-omics data to decipher the cross talk between human tissues in obesity-related disorders

Author

Miao, Zong

Publication Date

2020

UNIVERSITY OF CALIFORNIA

Los Angeles

Integrating multi-omics data to decipher the cross talk between human tissues in obesity-related disorders

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Bioinformatics

by

Zong Miao

2020

ABSTRACT OF THE DISSERTATION

Integrating multi-omics data to decipher the cross talk between human tissues in obesity-related disorders

by

Zong Miao

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2020

Professor Päivi Pajukanta, Chair

Obesity and obesity-related diseases have become an increasing burden to the health systems worldwide. Since the obesity-related diseases, such as type 2 diabetes (T2D), non-alcoholic fatty liver disease (NAFLD), and cardiovascular disease, share many risk factors and affect multiple human tissues, the complexity of these diseases has prevented disentangling the underlying causal effects. To investigate the cross talk between various tissues and obesity-related complex diseases, we comprehensively analyzed single nucleus RNA sequence (sn-RNA-seq) data, bulk RNA-seq data in multiple tissue types, as well as genotype data in several independent cohorts. First, in Chapter III we utilized the adipose sn-RNA-seq data as a reference to estimate the cell-type composition in human subcutaneous adipose tissue. Using body mass index (BMI), adipose mitochondrial (MT) gene expression, and the estimated cell-type proportions as predictors, we explained ~40% of variance in systemic insulin resistance and accurately estimated insulin resistance in human cohorts with adipose RNA-seq data. Our analysis discovered the important role of adipose transcriptional activity and MT activity in the

development of systemic insulin resistance. Moreover, in Chapter IV we developed another prediction model that utilized BMI, waist circumference, age, sex, and serum lipid, liver enzyme, and glucose levels to accurately predict non-alcoholic fatty liver (NAFLD) in the UK Biobank (UKB) cohort. The novel NAFLD score (NAFLDS) model achieved a high accuracy (AUC = 0.9) and outperformed the existing fatty liver index (FLI) in predicting the NAFLD status in UKB. Using NAFLDS as the surrogate of the NAFLD status, we utilized *cis* expression quantitative trait loci (*cis*-eQTLs) in liver and coronary arteries to refine the instrumental variables (IV) for our Mendelian randomization (MR) analyses and demonstrated a one-way causal effect of NAFLD on CAD (beta = 0.024, p-value = 9.4e-6).

While analyzing the RNA-seq cohorts, we found that allele-specific expression (ASE) analysis is a powerful tool that improves the accuracy and power in identifying *cis* gene regulation in RNA-seq cohorts. However, the reference alignment bias remains a major obstacle in ASE analysis. The existing methods are either inaccurate or relatively slow, which makes it impractical to accurately estimate ASE events in larger cohorts. To address this issue, we developed ASElux that uses personal genotype data as the reference to fast and accurately align the ASE reads to both alleles. By applying ASElux into the GTEx lung samples, we showed that ASElux is at least ~5X faster than any existing method, while achieving a top accuracy.

In summary, we developed a novel method ASElux to resolve the reference alignment bias in the ASE analysis. Furthermore, we comprehensively combined multi-omics data from adipose, liver, and coronary artery tissues and established two causal effects regarding the obesity related diseases (obesity -> insulin resistance/pre-diabetes and NAFLD -> CAD).

The dissertation of Zong Miao is approved.

Eleazar Eskin

Jingyi Li

Yi Xing

Päivi Pajukanta, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Päivi Pajukanta for her guidance. She has been intensively involved in all the work presented in this thesis. I'm always amazed and inspired by her enthusiasm for science.

I would like to thank all my committee members, Professor Eleazar Eskin, Professor Jingyi Li, Professor Yi Xing, who have helped guide my dissertation and demonstrated commitment in this regard.

I would like to thank all my collaborators. Among all my collaborators and friends, I would particularly like to thank Marcus Alvarez, Kristina Garske, and David Pan. We have collaborated on many projects and it's always been a great pleasure to working with them.

I would like to thank my family. Six years is a long journey and they are nothing but supportive. Thanks to my lovely parents, they taught me to stay curious and never give up. My wife Shan Jiang is always able to cheers me up when I encounter setbacks.

This thesis was supported by American Heart Association (AHA) Predoctoral Fellowship 19PRE34430112.

Chapter II is a reprint of “ASElux: an ultra-fast and accurate allelic reads counter” by Zong Miao, Marcus Alvarez, Päivi Pajukanta, and Arthur Ko. Bioinformatics, 34(8), 2018, 1313–1320.

Chapter III is a submitted article entitled “The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance” by Zong Miao, Marcus Alvarez, Arthur Ko, Yash Bhagat, Elior Rahmani, Brandon Jew, Sini Heinonen, Karen L Mohlke, Markku Laakso, Kirsi H. Pietiläinen, Eran Halperin, and Päivi Pajukanta

Chapter IV is a version of an article in preparation entitled “Establishing a causal effect of non-alcoholic fatty liver disease on coronary artery disease” by Zong Miao, Kristina M. Garske, Dorota Kaminska, Janet S. Sinsheimer, Jussi Pihlajamäki, and Päivi Pajukanta

CURRICULUM VITAE

2013	B.S. (Biotechnology), Sun Yat-Sen University
2014-2020	Ph.D. student, Bioinformatics Program, UCLA
2017	Teaching Assistant, Human Genetics, UCLA

MANUSCRIPTS

Zong Miao, Marcus Alvarez, Arthur Ko, Yash Bhagat, Elior Rahmani, Brandon Jew, Sini Heinonen, Karen L Mohlke, Markku Laakso, Kirsi H. Pietiläinen, Eran Halperin, Päivi Pajukanta. The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance, Manuscript submitted, 2019

Zong Miao, Kristina M. Garske, Dorota Kaminska, Janet S. Sinsheimer, Jussi Pihlajamäki, Päivi Pajukanta. Establishing a causal effect of non-alcoholic fatty liver disease on coronary artery disease, Manuscript in preparation, 2019

Brandon Jew, Marcus Alvarez, Elior Rahmani, **Zong Miao**, Arthur Ko, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information, bioRxiv, 2019

Marcus Alvarez, Elior Rahmani, Brandon Jew, Kristina M. Garske, **Zong Miao**, Jihane N. Benhammou, Chun Jimmie Ye, Joseph R. Pisegna, Kirsi H. Pietiläinen, Eran Halperin, Päivi Pajukanta. Enhancing droplet-based single nucleus RNA-seq resolution using an unsupervised machine learning classifier DIEM, bioRxiv, 2019

Kristina M Garske, David Z Pan, **Zong Miao**, Yash V Bhagat, Caroline Comenho, Christopher R Robles, Jihane N Benhammou, Marcus Alvarez, Arthur Ko, Chun Jimmie Ye, Joseph R Pisegna, Karen L Mohlke, Janet S Sinsheimer, Markku Laakso, Päivi Pajukanta. Reverse gene–environment interaction approach to identify variants influencing body-mass index in humans, Nature Metabolism, 2019

David Z. Pan, Kristina M. Garske, Marcus Alvarez, Yash V. Bhagat, James Boocock, Elina Nikkola, **Zong Miao**, Chelsea K. Raulerson, Rita M. Cantor, Mete Civelek, Craig A. Glastonbury, Kerrin S. Small, Michael Boehnke, Aldons J. Lusis, Janet S. Sinsheimer, Karen L. Mohlke, Markku Laakso, Päivi Pajukanta, Arthur Ko. Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS, Nature Communication, 2018

Malika Kumar Freund, Kathryn S Burch, Huwenbo Shi, Nicholas Mancuso, Gleb Kichaev, Kristina M Garske, David Z Pan, **Zong Miao**, Karen L Mohlke, Markku Laakso, Päivi Pajukanta, Bogdan Pasaniuc, Valerie A Arboleda. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. AJHG, 2018.

Zong Miao, Marcus Alvarez, Päivi Pajukanta, Arthur Ko. ASElux: an ultra-fast and accurate allelic reads aligner. Bioinformatics, 2018.

Alejandra Rodríguez, Luis Gonzalez, Arthur Ko, Marcus Alvarez, **Zong Miao**, Yash Bhagat, Elina Nikkola, Ivette Cruz-Bautista, Olimpia Arellano-Campos, Linda L. Muñoz-Hernández, Maria-Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, Karen L. Mohlke, Markku Laakso, Teresa Tusie-Luna, Carlos A. Aguilar-Salinas, Päivi Pajukanta. Molecular characterization of the lipid genome-wide association study signal on chromosome 18q11.2 implicates HNF4A-mediated regulation of the TMEM241 gene. Arterioscler Thromb Vasc Biol. 2016.

Chapter I

Introduction

This thesis focuses on integrating multi-omics data from different human tissue types to investigate the cross talk between cardiometabolic tissues in the obesity-related complex diseases. During the past decades, the prevalence of obesity has constantly increased. Obesity has become a serious health problem and an increasing burden to the healthcare systems worldwide[1,2,3,4]. On average, the obese individuals' medical cost is 30% higher than the normal weight individuals' medical cost[2]. Moreover, the childhood obesity rates are also increasing fast globally[5]. Obese individuals have an increased risk of T2D, stroke, NAFLD, cardiovascular disease, hypertension, and several cancers[2,6,7,8]. Moreover, obesity is not just associated with these complex diseases because a high BMI has been shown to be a causal risk factor for T2D, NAFLD, and CAD in the previous studies[9,10,11]. Thus, obesity has become a greater health problem than smoking and drinking, and it is linked to multiple severe chronic medical conditions and a reduced quality of life[2].

With the rapid development of RNA-sequencing technology, the key tissues for obesity-related diseases, such as liver and adipose, have been sequenced to investigate their transcriptomes. For example, the GTEx V8 has collected 17,382 RNA-seq samples from 49 different tissue types[12,13]. Among these samples, 581 subcutaneous adipose samples and 208 liver samples were sequenced. Tissue-enriched expression quantitative traits loci (eQTLs) can provide a systematic understanding of how DNA variants affect gene expression in diverse sets of human tissues[13]. Moreover, some RNA-seq cohorts that are specifically designed for investigating obesity-related diseases not only collect RNA-seq samples from the relevant tissues, but also collect refined cardiometabolic phenotypes to decipher how the key tissues affect cardiometabolic diseases. The METabolic Syndrome In Men (METSIM) study contains ~10k middle aged Finnish men to investigate the risk factors (genetic and non-genetic) associated with

T2D and cardiovascular disease[14]. Combining the comprehensive cardiometabolic phenotype data, such as fasting serum glucose, Matsuda index, and serum lipid levels with RNA-seq and genetic data provides a great opportunity to disentangle these complex diseases and pin point the main risk factors associated with cardiometabolic disorders.

While analyzing RNA-seq data, a *cis*-eQTL analysis is a common method to identify causal variants that regulate local gene expression. However the power of *cis*-eQTL analysis is often constrained by the relatively small sample sizes. One way to improve the power in *cis*-eQTL analysis is to utilize the personal genetic data and identify the allele-specific gene expression (ASE)[15]. Since ASE detects imbalanced expression and may indicate a *cis*-regulation of the target gene, combining ASE and *cis*-eQTL data can improve the power and accuracy in detecting *cis*-regulation of the target genes[16,17,18,19,20]. However, the preference to align reads to the reference allele (i.e. reference bias) introduces a bias in ASE analysis in identifying the ASE sites. Several tools have been developed to address this problem. For example, GSNAP builds an allele-aware reference to align reads equally to two alleles[21]. In 2015, Geijn et al. reported a new tool called WASP to exclude the potential biased reads using a simulation-based method[15]. However, GSNAP and WASP are both computationally intensive and require a relatively long time to process the data. The extra long processing time is prohibitive for large cohorts, such as GTEx and METSIM. Thus, the lack of an ideal tool to detect ASE in large cohorts remains an obstacle for the use of ASE in *cis*-eQTL analysis. In Chapter II, we developed a novel tool, ASElux, that uses personal genetic variants as the reference to fast and accurately align the reads to both alleles. Using ASElux to count the allelic reads in RNA-seq data is at least 4X faster than with any of the existing methods while keeping the top accuracy, which makes it an ideal tool to investigate ASE in large RNA-seq cohorts.

T2D is a complex disease that can be diagnosed by high fasting glucose levels or the A1C test[22]. Among the different subtypes of diabetes, T2D is the most common one, and it is postulated to be caused by the loss of insulin secretion on the background of insulin resistance[22]. During the development of T2D, insulin secretion initially increases to compensate the insulin resistance and the disease occurs when β cell compensation fails[23]. Duret et al. have also shown that insulin resistance is a consistent feature of T2D risk in early onset populations, while insulin secretion is not as prevalent as insulin resistance[24]. This suggests the essential role of insulin resistance in the development of prediabetes and early stage of T2D. Randle et al. reported the "glucose fatty acid circle" that shows a substrate competition between fat and glucose oxidation[25]. This central concept that explains the development of insulin resistance has been verified in several previous studies[26,27,28,29]. Although muscle is a key tissue that is shown to be associated with insulin resistance[30,31,32,33], the adipose tissue mass has also been suggested to be significantly associated with insulin resistance[34]. Since the key functions of adipose tissue, i.e. lipogenesis (storing fat) and lipolysis (mobilizing the stored fat), make it one of the most important tissues contributing to obesity, it is important to thoroughly investigate the association between adipose tissue and insulin resistance. In Chapter III, we first utilized the UK Biobank (UKB) and METSIM cohort to explore the causal relationship between obesity and insulin resistance/prediabetes. Then we employed single nuclei RNA-seq data to decompose cell-type proportions of the subcutaneous adipose tissue in several bulk RNA-seq cohorts. The decomposed adipose tissue demonstrated that proportions of certain cell-types, i.e. adipocytes and macrophages, are significantly associated with obesity and insulin resistance. We also showed that adipose tissue together with BMI explain a large proportion of variance in systemic

insulin resistance. This supports the important role of adipose tissue in the obesity-related T2D development.

Non-alcoholic fatty liver disease (NAFLD) is a fast increasing obesity-related complex disease that causes a large burden to the health care system[35,36].NAFLD has a worldwide prevalence of 25% [35,36]. Moreover, ~8-19% of Asians, who have a normal weight (BMI < 25), have NAFLD[36]. Although obesity is recognized as a main risk factor for NAFLD[37], other factors, such as central adiposity, insulin resistance, and certain genetic variants (such as some PNPLA3 SNPs[38,39,40,41]) are also important risk factors for NAFLD that require further investigation. To measure the degree of liver steatosis, imaging techniques, such as abdominal magnetic resonance imaging (MRI), are being used. However, MRI is relatively expensive and not readily available. Thus, compared to obesity and T2D, the NAFLD GWAS studies are usually restricted by a small sample size that limits the power to identify risk variants[42,43,44,45]. Moreover, the serious, advanced forms of NAFLD, i.e. non-alcoholic steatohepatitis (NASH) and fibrosis, can lead to catastrophic health consequences, such as liver failure, and require histological assessment of the liver by an invasive biopsy. To assess NAFLD utilizing a non-invasive approach without imaging, several NAFLD scoring systems have been developed. For example, Bedogni et al. developed a fatty liver index (FLI) that uses serum gamma-glutamyl transferase (GGT), BMI, waist circumference, and serum triglycerides to predict the risk of NAFLD[46]. FLI has been employed and verified in several independent studies [47,48,49]. However, FLI was developed using a relatively small sample size and only limited traits were selected as the predictors. In Chapter IV, we improved the prediction of NALFD using a similar set of predictors while utilizing 2,181 NAFLD patients and 2,444 healthy controls that are verified by ICD codes and MRI data in the

UKB cohort. Our NAFLD score (NAFLDS) model outperformed the existing NAFLD models and was employed to impute the NAFLD status in the whole UKB cohort.

Noteworthy, the leading cause of death from NAFLD is coronary artery disease (CAD) instead of liver carcinoma or NASH [50,51]. An estimated 5-10% of people with NAFLD are dying from CAD[51]. On the other hand, among the people who have a high risk of CAD, the prevalence of NAFLD is also increased[52]. The association between NAFLD and CAD can be caused by the fact that both NAFLD and CAD share several risk factors, such as obesity, dyslipidemia, and T2D[53,54,55,56,57]. However, there can also be a direct causal link between NAFLD and CAD. For example, lipoprotein (a) is a known causal risk factor for CAD[58,59,60]. Apolipoprotein (a) is encoded by the LPA gene, which is almost solely expressed in the liver[13]. Thus, the serum LP(a) level is heavily influenced by the liver health status and it causally affects the risk of CAD. Tsimikas et al. recently reported that among CAD patients, the hepatocyte-directed antisense oligonucleotide AKCEA-APO(a)-LRx (APO(a)-LRx) effectively reduced the serum LP(a) level[61]. This verified that the liver LPA gene expression has a significant effect on serum LP(a) level, and the gene expression regulation therapy can become a possible new treatment for the CAD patients. However, serum LP(a) has a negative association[62,63] with NAFLD, which is opposite to the association between NAFLD and CAD. We also confirmed this negative association between NAFLD and LP(a) in the UKB cohort. Thus, LP(a) cannot mediate a potential causal effect from NAFLD on CAD. Due to the complicated risk factors shared by NAFLD and CAD, the causality between NAFLD and CAD remains elusive [64,65,66]. To resolve the causal relationship between NAFLD and CAD, we used the imputed NAFLDS in the UKB cohort to identify novel variants that are associated with NAFLD. Then we overlapped the GWAS variants with liver and coronary artery *cis*-eQTLs to identify GWAS variants that

directly affect the disease by regulating gene expression in the related tissues. Using these tissue-enriched *cis*-eQTL GWAS variant as the instrumental variables (IV), we reduced a potential horizontal pleiotropy and established a one-way causal effect of NAFLD on CAD.

In summary, we investigated transcriptional regulation of cardiometabolic tissues to identify novel causal factors for obesity-related common diseases, such as T2D, NAFLD, and CAD. In this thesis, we integrated multi-tissue RNA-seq data and sn-RNA-seq data with genome-wide variant and deep cardiometabolic phenotype data from several independent cohorts and UKB to discover new mechanisms that affect the development of these obesity-related diseases. We 1) developed a novel tool to fast and reliably count allelic expression in large bulk RNA-seq cohorts; 2) explored the potential critical role of adipose in systemic insulin resistance; and 3) built a novel NAFLD score to accurately estimate the NAFLD status from serum traits and used the imputed NALFDS to resolve the so far elusive causal relationship between NAFLD and CAD in the UKB.

References:

[1] Soares, E. M. K. V. K., Smith, D., & Grossi Porto, L. G. (2020). Worldwide prevalence of obesity among firefighters: A systematic review protocol. *BMJ Open*, *10*(1), 1–5. https://doi.org/10.1136/bmjopen-2019-031282

[2] Withrow, D., & Alter, D. A. (2011). The economic burden of obesity worldwide: A systematic review of the direct costs of obesity. *Obesity Reviews*, *12*(2), 131–141. https://doi.org/10.1111/j.1467-789X.2009.00712.x

[3] Bentham, J., Di Cesare, M., Bilano, V., Bixby, H., Zhou, B., Stevens, G. A., … Cisneros, J. Z. (2017). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *The Lancet*, *390*(10113), 2627–2642. https://doi.org/10.1016/S0140-6736(17)32129-3

[4] Reilly, J. J., El-Hamdouchi, A., Diouf, A., Monyeki, A., & Somda, S. A. (2018). Determining the worldwide prevalence of obesity. *The Lancet*, *391*(10132), 1773–1774. https://doi.org/10.1016/S0140-6736(18)30794-3

[5] Wang, Y., & Lobstein, T. (2006). Worldwide trends in childhood overweight and obesity. *International Journal of Pediatric Obesity*, *1*(1), 11–25. https://doi.org/10.1080/17477160600586747

[6] Folsom AR, Kaye SA, Potter JD, Prineas RJ. Association of incident carcinoma of the endometrium with body weight and fat distribution in older women: early findings of the Iowa Women's Health Study. Cancer Res 1989; 49: 6828–6831.

[7] Lee IM, Paffenbarger RS Jr. Quetelet's index and risk of colon cancer in college alumni. J Natl Cancer Inst 1992; 84: 1326–1331.

[8] 15. Sellers TA, Kushi LH, Potter JD, Kaye SA, Nelson CL, McGovern PG, Folsom AR. Effect of family history, body-fat distribution, and reproductive factors on the risk of postmeno- pausal breast cancer. NEnglJMed 1992; 326: 1323–1329

[9] Holmes, M. V., Lange, L. A., Palmer, T., Lanktree, M. B., North, K. E., Almoguera, B., … Keating, B. J. (2014). Causal effects of body mass index on cardiometabolic traits and events: A Mendelian randomization analysis. *American Journal of Human Genetics*, *94*(2), 198–208. https://doi.org/10.1016/j.ajhg.2013.12.014

[10] Cangeri Di Naso, F., Rosa Porto, R., Sarubbi Fillmann, H., Maggioni, L., Vontobel Padoin, A., Jacques Ramos, R., … Ivo Homem De Bittencourt, P. (2015). Obesity depresses the anti-inflammatory HSP70 pathway, contributing to NAFLD progression. *Obesity*, *23*(1), 120–129. https://doi.org/10.1002/oby.20919

[11] Parekh, S., & Anania, F. A. (2007). Abnormal Lipid and Glucose Metabolism in Obesity: Implications for Nonalcoholic Fatty Liver Disease. *Gastroenterology*, *132*(6), 2191–2207. https://doi.org/10.1053/j.gastro.2007.03.055

[12] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

[13] Ardlie, K. G., Deluca, D. S., Segre, a. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., … Lockhart, N. C. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

[14] Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A. J., … Boehnke, M. (2017). The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *Journal of Lipid Research*, *58*(3), 481–493. https://doi.org/10.1194/jlr.o072629

[15] van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. (2015). WASP: allele-specific software for robust discovery of molecular quantitative trait loci. *Nature Methods*, *12*(11), 1061–1063. https://doi.org/10.1101/011221

[16] Li, G., Bahn, J. H., Lee, J. H., Peng, G., Chen, Z., Nelson, S. F., & Xiao, X. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, *40*(13), 1–13. https://doi.org/10.1093/nar/gks280

[17] Heap, G. a., Yang, J. H. M., Downes, K., Healy, B. C., Hunt, K. a., Bockett, N., … Plagnol, V. (2009). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human Molecular Genetics*, *19*(1), 122–134. https://doi.org/10.1093/hmg/ddp473

[18] Hasin-Brumshtein, Y., Hormozdiari, F., Martin, L., van Nas, A., Eskin, E., Lusis, A. J., & Drake, T. a. (2014). Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics*, *15*(1), 471. https://doi.org/10.1186/1471-2164-15-471

[19] Knowles, D. A., Davis, J. R., Raj, A., Zhu, X., Potash, J. B., Weissman, M. M., … Battle, A. (2015). Allele-specific expression reveals interactions between genetic variation and environment. *BioRxiv*, 025874. https://doi.org/10.1101/025874

[20] Kukurba, K. R., Zhang, R., Li, X., Smith, K. S., Knowles, D. a., How Tan, M., … Montgomery, S. B. (2014). Allelic Expression of Deleterious Protein-Coding Variants across Human Tissues. *PLoS Genetics*, *10*(5). https://doi.org/10.1371/journal.pgen.1004304

[21] Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads (Vol. 26, pp. 873–881). https://doi.org/10.1093/bioinformatics/btq057

[22] American Diabetes Association, (2016). Classification and diagnosis of diabetes. *Diabetes Care*, *39*(January), S13–S22. https://doi.org/10.2337/dc16-S005

[23] Kahn, B. (1998). Type 2 Diabetes: When Insulin Secretion Fails to Compensate for Insulin Resistance. *Cell*, 92, 593-596

[24] Druet, C., Tubiana-Rufi, N., Chevenne, D., Rigal, O., Polak, M., & Levy-Marchal, C. (2006). Rapid Communication: Characterization of insulin secretion and resistance in type 2 diabetes of adolescents. *Journal of Clinical Endocrinology and Metabolism*, *91*(2), 401–404. https://doi.org/10.1210/jc.2005-1672

[25] Randle PJ, Garland PB, Hales CN & Newsholme EA. The glucose fatty-acid cycle. Its role in insulin sensitivity and the metabolic disturbances of diabetes mellitus. Lancet 1963 1 785–789.

[26] Bonadonna, R. C., Groop, L. C., Simonson, D. C., & DeFronzo, R. A. (1994). Free fatty acid and glucose metabolism in human aging: Evidence for operation of the Randle cycle. *American Journal of Physiology - Endocrinology and Metabolism*, *266*(3 29-3). https://doi.org/10.1152/ajpendo.1994.266.3.e501

[27] Hue, L., & Taegtmeyer, H. (2009). The Randle cycle revisited: A new head for an old hat. *American Journal of Physiology - Endocrinology and Metabolism*, *297*(3), 578–591. https://doi.org/10.1152/ajpendo.00093.2009

[28] Bevilacqua, S., Buzzigoli, G., Bonadonna, R., Brandi, L. S., Oleggini, M., Boni, C., … Ferrannini, E. (1990). Operation of randle's cycle in patients with NIDDM. *Diabetes*, *39*(3), 383–389. https://doi.org/10.2337/diab.39.3.383

[29] Massao Hirabara, S., De Oliveira Carvalho, C. R., Mendonça, J. R., Piltcher Haber, E., Fernandes, L. C., & Curi, R. (2003). Palmitate acutely raises glycogen synthesis in rat soleus muscle by a mechanism that requires its metabolization (Randle cycle). *FEBS Letters*, *541*(1–3), 109–114. https://doi.org/10.1016/S0014-5793(03)00316-8

[30] Kraegen, E. W., Clark, P. W., Jenkins, A. B., Daley, E. A., Chisholm, D. J., & Storlien, L. H. (1991). Development of muscle insulin resistance after liver insulin resistance in high-fat-fed rats. *Diabetes*, *40*(11), 1397–1403. https://doi.org/10.2337/diab.40.11.1397

[31] Koves, T. R., Ussher, J. R., Noland, R. C., Slentz, D., Mosedale, M., Ilkayeva, O., … Muoio, D. M. (2008). Mitochondrial Overload and Incomplete Fatty Acid Oxidation Contribute to Skeletal Muscle Insulin Resistance. *Cell Metabolism*, *7*(1), 45–56. https://doi.org/10.1016/j.cmet.2007.10.013

[32] DeFronzo, R. A., & Tripathy, D. (2009). Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes Care*, *32 Suppl 2*. https://doi.org/10.2337/dc09-s302

[33] Petersen, K. F., Dufour, S., Savage, D. B., Bilz, S., Solomon, G., Yonemitsu, S., … Shulman, G. I. (2007). The role of skeletal muscle insulin resistance in the pathogenesis of the metabolic

syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(31), 12587–12594. https://doi.org/10.1073/pnas.0705408104

[34] Newgard, C. B., An, J., Bain, J. R., Muehlbauer, M. J., Stevens, R. D., Lien, L. F., Haqq, A. M., …, Svetkey, L. P. A (2009) Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. Cell Metab., 9 (4), 311− 326; 9 (6), 565−566 (correction)

[35] Araújo, A. R., Rosso, N., Bedogni, G., Tiribelli, C., & Bellentani, S. (2018). Global epidemiology of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: What we need in the future. Liver International, 38(November 2017), 47–51. https://doi.org/10.1111/liv.13643

[36] Fan, J. G., Kim, S. U., & Wong, V. W. S. (2017). New trends on obesity and NAFLD in Asia. *Journal of Hepatology*, *67*(4), 862–873. https://doi.org/10.1016/j.jhep.2017.06.003

[37]Machado, M. V., & Cortez-Pinto, H. (2016). Diet, microbiota, obesity, and NAFLD: A dangerous quartet. *International Journal of Molecular Sciences*, *17*(4), 1–20. https://doi.org/10.3390/ijms17040481

[38] Scorletti, E., West, A. L., Bhatia, L., Hoile, S. P., McCormick, K. G., Burdge, G. C., … Byrne, C. D. (2015). Treating liver fat and serum triglyceride levels in NAFLD, effects of PNPLA3 and TM6SF2 genotypes: Results from the WELCOME trial. *Journal of Hepatology*, *63*(6), 1476–1483. https://doi.org/10.1016/j.jhep.2015.07.036

[39] Liu, Y. L., Patman, G. L., Leathart, J. B. S., Piguet, A. C., Burt, A. D., Dufour, J. F., … Anstee, Q. M. (2014). Carriage of the PNPLA3 rs738409 C >g polymorphism confers an increased risk of non-alcoholic fatty liver disease associated hepatocellular carcinoma. *Journal of Hepatology*, *61*(1), 75–81. https://doi.org/10.1016/j.jhep.2014.02.030

[40] Lallukka, S., Sevastianova, K., Perttilä, J., Hakkarainen, A., Orho-Melander, M., Lundbom, N., … Yki-Järvinen, H. (2013). Adipose tissue is inflamed in NAFLD due to obesity but not in NAFLD due to genetic variation in PNPLA3. *Diabetologia*, *56*(4), 886–892. https://doi.org/10.1007/s00125-013-2829-9

[41] Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L. A., … Hobbs, H. H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature Genetics*, *40*(12), 1461–1465. https://doi.org/10.1038/ng.257

[42] Di Costanzo, A., Belardinilli, F., Bailetti, D., Sponziello, M., D'Erasmo, L., Polimeni, L., … Arca, M. (2018). Evaluation of Polygenic Determinants of Non-Alcoholic Fatty Liver Disease (NAFLD) By a Candidate Genes Resequencing Strategy. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-21939-0

[43] Chen, Q. R., Braun, R., Hu, Y., Yan, C., Brunt, E. M., Meerzaman, D., … Buetow, K. (2013). Multi-SNP Analysis of GWAS Data Identifies Pathways Associated with Nonalcoholic Fatty Liver Disease. *PLoS ONE*, *8*(7), 1–11. https://doi.org/10.1371/journal.pone.0065982

[44] Shang, X. R., Song, J. Y., Liu, F. H., Ma, J., & Wang, H. J. (2015). GWAS-Identified Common Variants With Nonalcoholic Fatty Liver Disease in Chinese Children. *Journal of Pediatric Gastroenterology and Nutrition*, *60*(5), 669–674. https://doi.org/10.1097/MPG.0000000000000662

[45] Eslam, M., Valenti, L., & Romeo, S. (2018). Genetics and epigenetics of NAFLD and NASH: Clinical impact. *Journal of Hepatology*, *68*(2), 268–279. https://doi.org/10.1016/j.jhep.2017.09.003

[46] Bedogni, G., Bellentani, S., Miglioli, L., Masutti, F., Passalacqua, M., Castiglione, A., & Tiribelli, C. (2006). The fatty liver index: A simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterology*, *6*, 1–7. https://doi.org/10.1186/1471-230X-6-33

[47] Cuthbertson, D. J., Weickert, M. O., Lythgoe, D., Sprung, V. S., Dobson, R., Shoajee-Moradie, F., … Kemp, G. J. (2014). External validation of the fatty liver index and lipid accumulation product indices, using 1H-magnetic resonance spectroscopy, to identify hepatic steatosis in healthy controls and obese, insulin-resistant individuals. *European Journal of Endocrinology*, *171*(5), 561–569. https://doi.org/10.1530/EJE-14-0112

[48] Koehler, E. M., Schouten, J. N. L., Hansen, B. E., Hofman, A., Stricker, B. H., & Janssen, H. L. A. (2013). External Validation of the Fatty Liver Index for Identifying Nonalcoholic Fatty Liver Disease in a Population-based Study. *Clinical Gastroenterology and Hepatology*, *11*(9), 1201–1204. https://doi.org/10.1016/j.cgh.2012.12.031

[49] Koehler, E. M., Schouten, J. N. L., Hansen, B. E., Hofman, A., Stricker, B. H., & Janssen, H. L. A. (2013). External Validation of the Fatty Liver Index for Identifying Nonalcoholic Fatty Liver Disease in a Population-based Study. *Clinical Gastroenterology and Hepatology*, *11*(9), 1201–1204. https://doi.org/10.1016/j.cgh.2012.12.031

[50] Brouwers, M. C. G. J., Simons, N., Stehouwer, C. D. A., Koek, G. H., Schaper, N. C., & Isaacs, A. (2019). Relationship Between Nonalcoholic Fatty Liver Disease Susceptibility Genes and Coronary Artery Disease. *Hepatology Communications*, *3*(4), 587–596. https://doi.org/10.1002/hep4.1319

[51] Wong, V. W. S., Wong, G. L. H., Yip, G. W. K., Lo, A. O. S., Limquiaco, J., Chu, W. C. W., … Chan, H. L. Y. (2011). Coronary artery disease and cardiovascular outcomes in patients with non-alcoholic fatty liver disease. *Gut*, *60*(12), 1721–1727. https://doi.org/10.1136/gut.2011.242016

[52] Adibi, A., Jaberzadeh-Ansari, M., Dalili, A.-R., Omidifar, N., & Sadeghi, M. (2013). Association between Nonalcoholic Fatty Liver Disease (NAFLD) and Coronary Artery Disease (CAD) in Patients with Angina Pectoris. *Open Journal of Medical Imaging*, *03*(03), 97–101. https://doi.org/10.4236/ojmi.2013.33015

[53] Wong, V. W. S., Wong, G. L. H., Yip, G. W. K., Lo, A. O. S., Limquiaco, J., Chu, W. C. W., … Chan, H. L. Y. (2011). Coronary artery disease and cardiovascular outcomes in patients

with non-alcoholic fatty liver disease. *Gut*, *60*(12), 1721–1727. https://doi.org/10.1136/gut.2011.242016

[54] Araújo, A. R., Rosso, N., Bedogni, G., Tiribelli, C., & Bellentani, S. (2018). Global epidemiology of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: What we need in the future. *Liver International*, *38*(November 2017), 47–51. https://doi.org/10.1111/liv.13643

[55] Chhabra, R., O'Keefe, J. H., Patil, H., O'Keefe, E., Thompson, R. C., Ansari, S., … Helzberg, J. H. (2013). Association of coronary artery calcification with hepatic steatosis in asymptomatic individuals. *Mayo Clinic Proceedings*, *88*(11), 1259–1265. https://doi.org/10.1016/j.mayocp.2013.06.025

[56] Patil, R., & Sood, G. K. (2017). Non-alcoholic fatty liver disease and cardiovascular risk. *World Journal of Gastrointestinal Pathophysiology*, *8*(2), 51. https://doi.org/10.4291/wjgp.v8.i2.51

[57] Adibi, A., Jaberzadeh-Ansari, M., Dalili, A.-R., Omidifar, N., & Sadeghi, M. (2013). Association between Nonalcoholic Fatty Liver Disease (NAFLD) and Coronary Artery Disease (CAD) in Patients with Angina Pectoris. *Open Journal of Medical Imaging*, *03*(03), 97–101. https://doi.org/10.4236/ojmi.2013.33015

[58] Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., … Kathiresan, S. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*, *45*(11), 1345–1353. https://doi.org/10.1038/ng.2795

[59] Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., … Samani, N. J. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*, *45*(1), 25–33. https://doi.org/10.1038/ng.2480

[60] Kronenberg, F. (2016). Human Genetics and the Causal Role of Lipoprotein(a) for Various Diseases. *Cardiovascular Drugs and Therapy*, *30*(1), 87–100. https://doi.org/10.1007/s10557-016-6648-3

[61] Tsimikas, S., Karwatowska-Prokopczuk, E., Gouni-Berthold, I., Tardif, J. C., Baum, S. J., Steinhagen-Thiessen, E., … Witztum, J. L. (2020). Lipoprotein(a) Reduction in Persons with Cardiovascular Disease. *The New England Journal of Medicine*, *382*(3), 244–255. https://doi.org/10.1056/NEJMoa1905239

[62] Nam, J. S., Jo, S., Kang, S., Ahn, C. W., Kim, K. R., & Park, J. S. (2016). Association between lipoprotein(a) and nonalcoholic fatty liver disease among Korean adults. *Clinica Chimica Acta*, *461*, 14–18. https://doi.org/10.1016/j.cca.2016.07.003

[63] Zhang, J., Zhao, Y., Xu, C., Hong, Y., Lu, H., Wu, J., & Chen, Y. (2014). Association between serum free fatty acid levels and nonalcoholic fatty liver disease: A cross-sectional study. *Scientific Reports*, *4*, 1–6. https://doi.org/10.1038/srep05832

[64] Targher, G., Marra, F., & Marchesini, G. (2008). Increased risk of cardiovascular disease in non-alcoholic fatty liver disease: Causal effect or epiphenomenon? *Diabetologia*, *51*(11), 1947–1953. https://doi.org/10.1007/s00125-008-1135-4

[65] Santos, R., Valentic, L., Romeod, S. (2019). Does nonalcoholic fatty liver disease cause cardiovascular disease? Current knowledge and gaps. Atherosclerosis, Vol. 282, 110–120

[66] Francque, S. M., van der Graaff, D., & Kwanten, W. J. (2016). Non-alcoholic fatty liver disease and cardiovascular risk: Pathophysiological mechanisms and implications. *Journal of Hepatology*, *65*(2), 425–443. https://doi.org/10.1016/j.jhep.2016.04.005

Chapter II

ASElux: an ultra-fast and accurate allelic reads counter

Bioinformatics, 34(8), 2018, 1313–1320
doi: 10.1093/bioinformatics/btx762
Advance Access Publication Date: 23 November 2017
Original Paper

OXFORD

Gene expression

ASElux: an ultra-fast and accurate allelic reads counter

Zong Miao[1,2], Marcus Alvarez[1], Päivi Pajukanta[1,2,3] and Arthur Ko[1,3,*]

[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, [2]Bioinformatics Interdepartmental Program and [3]Molecular Biology Institute, UCLA, Los Angeles, CA 90024, USA

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on August 17, 2017; revised on October 25, 2017; editorial decision on November 18, 2017; accepted on November 22, 2017

Abstract

Motivation: Mapping bias causes preferential alignment to the reference allele, forming a major obstacle in allele-specific expression (ASE) analysis. The existing methods, such as simulation and SNP-aware alignment, are either inaccurate or relatively slow. To fast and accurately count allelic reads for ASE analysis, we developed a novel approach, ASElux, which utilizes the personal SNP information and counts allelic reads directly from unmapped RNA-sequence (RNA-seq) data. ASElux significantly reduces runtime by disregarding reads outside single nucleotide polymorphisms (SNPs) during the alignment.

Results: When compared to other tools on simulated and experimental data, ASElux achieves a higher accuracy on ASE estimation than non-SNP-aware aligners and requires a much shorter time than the benchmark SNP-aware aligner, GSNAP with just a slight loss in performance. ASElux can process 40 million read-pairs from an RNA-sequence (RNA-seq) sample and count allelic reads within 10 min, which is comparable to directly counting the allelic reads from alignments based on other tools. Furthermore, processing an RNA-seq sample using ASElux in conjunction with a general aligner, such as STAR, is more accurate and still $\sim 4\times$ faster than STAR + WASP, and $\sim 33\times$ faster than the lead SNP-aware aligner, GSNAP, making ASElux ideal for ASE analysis of large-scale transcriptomic studies. We applied ASElux to 273 lung RNA-seq samples from GTEx and identified a splice-QTL rs11078928 in lung which explains the mechanism underlying an asthma GWAS SNP rs11078927. Thus, our analysis demonstrated ASE as a highly powerful complementary tool to cis-expression quantitative trait locus (eQTL) analysis.

Availability and implementation: The software can be downloaded from https://github.com/abl0719/ASElux.

Contact: zmiao@ucla.edu or a5ko@ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Allele specific expression (ASE) denotes the preferential allelic expression of a gene in the diploid genome. Integrating ASE with expression quantitative trait locus (eQTL) analysis improves fine-mapping accuracy and sensitivity (Kumasaka *et al.*, 2015), thus helping identify biologically meaningful regulatory signals such as imprinting and cis regulation. Although several methods have been developed to identify ASE events from RNA-sequencing (RNA-seq) data (Castel *et al.*, 2015; León-novelo *et al.*, 2014; Liu *et al.*, 2014; Li *et al*, 2012), mapping bias remains a major obstacle in ASE analysis (Degner *et al.*, 2009; Panousis *et al.*, 2014; Stevenson *et al.*, 2013). Therefore, there is an important scientific knowledge gap that motivates the development of fast and accurate allelic expression analysis tools.

Previously, simulations have been used to identify variant sites showing bias towards one allele (Buil *et al.*, 2015). However,

simulations perform sub-optimally in practice since they are largely based on single-end reads whereas most RNA-seq data are now paired-end reads. There are also methods that utilize available genotype information and build personal allelic reference genomes for an allele-aware alignment that are implemented in programs such as SNP-o-matic (Manske and Kwiatkowski, 2009) and GSNAP (Wu and Nacu, 2010). In these approaches, the aligners are aware of single nucleotide polymorphisms (SNP) and align reads against both alleles. Even though the SNP-aware methods are more accurate than simulation-based approaches, they are more time consuming and computationally intensive, which makes them impractical for large RNA-seq datasets. A recently developed allele-specific analysis method (WASP) (van de Geijn *et al.*, 2015) substitutes the SNP base with the alternative genotype in allelic reads and re-aligns those reads to correct for the reference bias. By excluding the allelic reads that are affected by different genotypes, WASP obtains extremely low false positive rate when identifying ASE SNPs. However, the process of generating reads with alternative genotypes in WASP takes a relatively long time (~3.5 h) and many reads are excluded due to its stringent requirements.

To this end, we developed a new and more efficient approach, ASElux, which focuses on SNP-overlapping reads and combines the alignment and estimation of allelic expression into one step. Since accurate genotyping is essential for ASE analysis, the genotype information is usually obtained separately from the RNA-seq data using SNP array or genome/exome sequencing (Lonsdale *et al.*, 2013). ASElux builds a personal allelic reference genome by using the individual's existing genotype information to generate all possible ASE reads and pre-screen the RNA-seq data. This allows us to perform SNP-aware alignment and to efficiently identify only the reads that cover the unique set of SNPs present in each individual. Compared to all of the tested tools, ASElux is ultra-fast while achieving the closest allelic mapping accuracy to the benchmark SNP-aware aligner, GSNAP. Adding the time consumption to analyze an RNA-seq sample using a general-purpose aligner, such as STAR, the overall runtime of ASE analysis using ASElux is still ~4 times faster than STAR (Dobin *et al.*, 2013) followed by WASP (STAR + WASP), which re-aligns the reads with SNPs to decrease the reference bias (van de Geijn *et al.*, 2015). We applied ASElux to 273 lung transcriptomes from the Genotype-Tissue Expression Project (GTEx) (Lonsdale *et al.*, 2013) to demonstrate the increased power of ASE analysis in detecting local gene regulation. The high speed and accuracy of this novel ASE software makes it possible to analyze ASE in large datasets, helping efficient transformative interrogation of variants.

2 Materials and methods

2.1 Workflow of ASElux

Since only ~10% of sequencing reads can be identified as SNP-overlapping, ASElux saves time by focusing on aligning reads that overlap with an individual's SNPs obtained either from a genotype array, imputed SNPs based on a reference panel, or DNA sequencing. To implement this new alignment, we designed a hybrid index system that performs both genome-wide alignment and personal SNP-aware alignment (Supplementary Fig. S1A). The hybrid index system contains a static index that is built once for each reference genome. ASElux aggregates the genic regions in the reference genome to form a trimmed genome and uses a suffix array (Nong *et al.*, 2009; Manber and Myers, 1990) as the static index for a fast alignment. The other part of our hybrid index system is the dynamic index. We extract the flanking sequence on both sides of the exonic SNP and store that in the dynamic index. The dynamic index is generated before alignment and it takes only ~3 min to build it for each individual. Supplementary Figure S1B shows the workflow of aligning paired-end reads. For a pair of reads, we follow the workflow twice to treat each read first as the main read and then as the mate read. To accommodate sequencing errors, ASElux by default allows up to two mismatches elsewhere than at the SNP site. The user can set the number of allowed mismatches to fit various read lengths. We first use the dynamic index to identify the allelic reads during the alignment. Only the reads that match the dynamic index would be mapped to the genome with the static index to locate their multi-alignment loci. Then we try to align the other read, known as the mate read, near the identified multi-alignment loci. Thus, we only align the mate read if the main read matches the dynamic index. If both reads are uniquely aligned to one gene, we count the reads for the allele they originated from.

2.2 Filtering candidate SNPs

Since exonic reads can provide the best estimation of gene expression, ASElux disregards non-exonic SNPs and alignments for ASE analysis. Within ASElux, we provide a fast and useful tool to select exonic SNPs using genome annotation. A standard genome annotation contains overlapped exons and transcripts due to alternative splicing, and the overlapping information is redundant for pruning SNPs. To facilitate the pruning process, we merge all overlapping exons from different transcripts within the same gene into one. Small indels are another mechanism of allelic expression, but they tend to cause an alignment error leading to bias in ASE estimation (Heap *et al.*, 2010; Stevenson *et al.*, 2013). Thus, most ASE analyses focus on SNPs alone rather than the combination of SNPs and indels for the better accuracy (David *et al.*, 2017). Therefore, we load the SNP and indel information from the vcf file and disregard all SNPs within one read length of an indel. The distance allowed between SNPs and indels varies according to the read length of the particular set of RNA-seq data. For example, if the read length is 50 bp, all SNPs within 50 bp of any indels in each individual would be disregarded from the further alignment. As shown in the 20 GTEx samples, only ~0.9% of the SNPs were excluded by this process (Supplementary Table S1).

2.3 Hybrid index system

To perform a personalized SNP-aware alignment and maintain a high speed, we designed a hybrid index system that contains both static and dynamic indices. The static indices are built only once for each reference genome. Since only a small proportion of RNA-seq reads consist of intergenic reads (Mortazavi *et al.*, 2008), ASElux uses the genic regions as the reference genome to achieve the least compromised balance between the alignment accuracy and speed and relatively low memory usage. We locate the start and the end of each gene so that the sequence between them covers all the components of a gene (exons, introns, UTRs etc.). Then we aggregate the sequences of all genes to form a trimmed genome. In the human genome (hg19), ASElux generates a new genome that contains ~1.5 billion bp out of the ~3 billion bp. For a genome that contains N genes, we construct N suffix arrays for the N genes and 1 more general suffix array for the trimmed genome as the static index using the sais algorithm (Nong *et al.*, 2011). Although in theory searching globally in one suffix array is faster than using N suffix arrays for N genes, in practice combining local and global indices is faster due to the low-level memory management strategy in a modern computer

(Kim *et al.*, 2015). Briefly, the static index is built based on the trimmed reference genome and accordingly, the global alignment is not allele-specific. The suffix array indices of the trimmed genome and genes costs ~30 GB of RAM (10 bytes for each base). We only use ~15 GB with the trimmed genome, thus keeping our overall RAM usage at ~ 20GB.

For each individual, we build a personalized dynamic index for SNP-aware alignment. We first prune the non-exonic SNPs to make sure that ASElux focuses on aligning only the expression-related reads. For each exonic SNP, we extract N-1 bp flanking sequence on both sides of the exonic SNP from the reference transcripts, where N is the read length, and replace the allele at the SNP location to generate reference sequences for all possible exonic reads that overlap with the SNP. To cover the SNPs adjacent to various splicing junctions, we extract the SNP flanking regions from all transcripts in each gene. Thus, each SNP has two 2N-1 bp long sequences for the reference and alternative alleles from each transcript. If the individual has additional known SNPs within the flanking sequence, we generate all possible haplotypes with alternative alleles of these adjacent SNPs to avoid misaligning reads with multiple variants. As there are regions with extremely high SNP density, ASElux only counts the first 10 heterozygous SNPs in each read. Noteworthy, as most indices are unique, we do not expect ambiguous indices to substantially bias the alignment of the ASE reads. To quickly locate SNP-overlapping reads, we aggregate all of the generated sequences as the dynamic index and build a suffix array for it. Then we save the generated sequences, SNPs and gene names for the dynamic index to query.

2.4 Alignment

Aligning only to the SNP-overlapping regions of the genome to identify the allelic reads is the key to the high speed of ASElux. For paired-end reads, we treat one read as the main read and the other as the mate read to help alignment. As shown in Supplementary Figure S1B and Algorithm 1 of Supplementary Methods, we check if the main read can be identified as an allelic read and use the mate read to properly align the whole read fragment. Only the ASE reads that are aligned to the dynamic index with up to two mismatches by default (not counting the SNP locus) will be aligned against the static index built on the trimmed genome (global alignment) to identify all of the multi-alignment loci. During the local alignment step, ASElux tries to locally align each main read's mate read to the static index of the same gene. Thus, both the global alignment and the local alignment are against the static index. Since the read fragment should come from the same gene, we require the read mates to be aligned to the same gene. In the case where the major read is multi-aligned, we count the major read towards the ASE estimate only if both the main and mate reads are aligned to the same gene and at least one of them is uniquely aligned. Finally, we exchange the roles of the main and mate read and align the main read again to identify all possible alignments for the read pair. Thus, each read is treated once as the main read of the paired end reads.

2.4.1 Alignment against the dynamic index

Similarly to STAR, ASElux uses a binary search strategy to identify the Maximal Mappable Prefix (MMP) of a read in a suffix array index. The alignment of the main read starts from the left end of the read and identifies the longest common sequence with the dynamic index. Since the suffix array is built in the forward genome direction, we also align the reverse complement of the read to cover both directions. Algorithm 2 in the Supplementary Methods shows the process of aligning reads against the dynamic index. For the reads with mismatches, the alignment process stops at the mismatched locus and restarts at the base after the mismatched locus. Thus, several regions divided by the mismatched loci in the main reads are aligned to different loci in the dynamic index. For the regions aligned by no less than 20 bp, ASElux compares the whole read against the sequence around the mapped loci to check if the main read can be aligned to the locus with no more than 2 mismatches (using the ASElux default setting) while not counting indels (Supplementary Fig. S2). Since ASElux aligns the main read while being aware of the individual's SNP loci, the mismatches are mainly caused by sequencing errors or unknown adjacent variants. Furthermore, we calculated that for a 100-bp read, allowing for up to 2 mismatches (using the default setting) covers 99.985% of the reads with the typical sequencing error rate of 0.1% per base expected for the Illumina platform (Schirmer *et al.*, 2016). Although ASElux allows 2 mismatches for ASE reads by default, users can adjust the number of allowed mismatches to fit for the various read lengths.

2.4.2 Local alignment

Using the static index, ASElux aligns the mate read against the same gene region that the main read aligns to. Therefore, the reads without mismatches, indels, or splice junctions are perfectly mapped to the reference genome in this step. Supplementary Figure S3 shows an example of aligning a junction read. For reads that are not identical to the reference genome, the MMP is a substring of the read that stops before a variant or splice site. As shown in Algorithm 3 of the Supplementary Methods, we skip eight bases to avoid mapping indels or SNPs and search the MMP again for the unmapped part of the read. We chose to skip 8 bp in line with STAR because in practice most indels would safely be skipped with this set-up and it still allows us to utilize the remaining read for alignment. Separate MMPs of a read indicate that mismatches or splicing occurs between MMPs. We repeatedly search for the MMP until all parts of the read are mapped or we have searched more than the default of four times, indicating that the read should not be mapped to the reference due to too many mismatches or splicing loci. We selected the default of four times since it provided the best balance between the alignment accuracy and speed. After identifying all MMPs, we reassemble the read and only accept the read alignment if the read was properly reconstructed such that the MMPs are in the same order in the query read and the reference.

2.4.3 Global alignment

The global alignment is similar to the local alignment but extends to the trimmed genome in the static index. Hence, the MMP can originate from multiple local indices, indicating that the read is aligned to multiple genes. Since the lengths of the perfectly aligned prefixes in multi-aligned reads vary and searching for the MMP requires a perfect alignment, the multi-aligned reads will only be aligned to the locus that has the longest prefix shared with the reference genome. Thus, if we only align a read once to the trimmed genome, the multi-aligned loci that have the shorter perfectly aligned prefixes would be missed. Since it is crucial to find all possible multi-alignment targets for the ASE reads, we developed a masked binary search strategy to align the read to additional possible loci by masking off the known alignment results (Algorithm 4 of the Supplementary Methods). To globally fast align the ASE read, we utilize the fact that the information about the one perfectly mapped locus is available for the ASE reads. To find all possible genes where

the read may be mapped to, we skip the locus that the read is already aligned to when searching for the MMP for the read. Since smaller MMPs have too many matches to the trimmed genome by chance alone, we only use MMPs longer than 20 bases and record the genes they reside in. Then we locally align the main read and the mate read in those genes to finish the alignment. ASElux can repeatedly align the reads with more and more masked genes. Therefore, the loci with smaller MMPs will not be missed due to the existence of the other loci with longer MMPs. In more detail, to find all MMPs within the read, we will start from the beginning of the read and search for the longest shared sequence between the particular read and the trimmed genome. We move along the read to find all MMPs longer than 20 bp in the read. Accordingly, there can be several MMPs which all must be longer than 21 bp. After the global alignment, we still locally align the mate read (Supplementary Fig. S1B), which ensures that a locus with only 20 bp match will not be identified as a properly aligned locus. As the next step, since the static index contains no SNP information, we align not only the ASE read but also the read that resides in the same locus with a different genotype. ASElux combines the alignment results of the two reads to make sure that we have the most comprehensive multi-alignment result.

The details of alignment with existing methods as well as the simulation data are described in the Supplementary Methods.

2.5 ASE and splice-QTL analyses in the GTEx project

We processed 273 RNA-seq samples from the GTEx project (Lonsdale *et al.*, 2013) with ASElux. We downloaded the RNA-seq data and the imputed genotype data from the dbGaP accession phs000424.v6.p1. We randomly selected 20 samples for the comparisons in this study. The samples have on average 40 million 50-bp paired-end reads. Reads were aligned to the human genome (hg19) with the four tested aligners. We used the default alignment parameters of all the tested methods. The uniquely aligned reads were then kept for the subsequent analyses.

The results of the cis-eQTL analysis (version 6) were obtained from the GTEx portal (Ardlie *et al.*, 2015). For each individual, we pruned out all SNPs aligned with less than 30 reads or less than 6 reads from one allele. To identify ASE SNPs across the population, we picked all SNPs that were heterozygous and passed the read count threshold in at least 30 individuals. We performed a paired *t*-test with the read counts of the reference allele and alternative allele from all individuals. The SNPs with Bonferroni corrected *P*-values less than 0.05 were identified as ASE SNPs. The GWAS SNPs ($P \leq 5 \times 10^{-8}$, two-sided) were obtained from the NHGRI GWAS Catalog (Welter *et al.*, 2014). We calculated the linkage disequilibrium (LD) between the ASE SNPs and GWAS SNPs within 1 Mb distance and obtained all of the SNPs in LD ($R2 \geq 0.8$) with the ASE SNPs using PLINK (Purcell *et al.*, 2007). Then we annotated the SNPs in LD with the ASE SNPs using ANNOVAR (Wang *et al.*, 2010).

For the splice-QTL analysis, we aligned the 273 GTEx lung samples with STAR and identified all the splice events using LeafCutter. Following the analytical guideline of LeafCutter, we then used MatrixeQTL (Shabalin, 2012) to identify whether rs11078928 is a significant splice-QTL of GSDMB. For the isoform level eQTL analysis, we used RSEM (Li and Dewey, 2011) to estimate the isoform expression of the 273 GTEx lung samples and calculated the proportional transcript expression as the transcript expression level over the total gene expression as the phenotype in the eQTL analysis performed by MatrixeQTL (Shabalin, 2012).

3 Results

3.1 Test on simulated RNA-seq dataset

We first tested ASElux and other alignment methods on a simulated RNA-seq dataset with ~180 M 2× 50 bp paired-end reads (see Supplementary Methods). Since comprehensively testing the alignment bias is important, we generated a high coverage simulated dataset. SNPs and junction reads were introduced to mimic real RNA-seq data. We added alternative alleles to the simulated reads based on imputed genotypes from a random GTEx sample and set both alleles to be equally expressed, which allowed us to accurately calculate the mapping bias of all methods. Besides ASElux, we also tested STAR 2.4.2a (Dobin *et al.*, 2013), GSNAP 2015-6-23 (Wu and Nacu, 2010), HISAT2 2.0.4 (Kim *et al.*, 2015) and WASP (van de Geijn *et al.*, 2015) on the simulated dataset using the default parameters during the alignment (see Supplementary Methods). Since we focus on the alignment bias, we only tested the mapping function of WASP. We used the reference genome hg19 for all aligners and the GENCODE v19 annotation if the gene annotation could be supplied. To utilize the power of SNP-aware alignment, we used GSNAP to build a SNP-integrated alignment index for GSNAP. The HISAT2 alignment index was downloaded from its website along with the SNPs and transcript information. We used the default parameters for each aligner.

Using the genome-wide SNP data (genotyped and imputed) from GTEx (Lonsdale *et al.*, 2013), we calculated read counts of each allele at exonic SNP sites to estimate ASE. The proportion of reference allele read counts when compared to the total read counts indicates the imbalance of allelic expression. Since the two alleles were equally expressed in the simulated dataset, the expected RACR of each SNP is 0.5. Accordingly, we measured the reference bias as the deviation of RACR from 0.5. Since each method aligns allelic reads differently, we performed the reference bias analysis using SNPs with enough aligned reads in all methods. ASElux, GSNAP, STAR and HISAT2 uniquely aligned ~10 M allelic reads whereas WASP aligned ~24% less reads than the other tested methods using the same simulated dataset (Supplementary Table S2). Figure 1 shows the proportion of SNPs in different bias categories. Although the majority of the SNPs displayed a bias less than 5% using all methods, ASElux achieved the highest accuracy by properly accounting for allele imbalance for ~90% of the SNPs. Among the biased SNPs identified by each method, ASElux and the SNP-aware GSNAP showed substantially fewer SNPs with reference allele bias (~70%) when compared to HISAT2 and STAR (~99%). Even though STAR + WASP identified the fewest SNPs with a bias more than 5%, still the majority (88%) of the SNPs identified by WASP showed a bias in the range of more than 0% but less than 5% (Fig. 1). STAR alone performed worst since no SNP information was used during the alignment. Even though HISAT2 considers all common SNPs (MAF > 1%), it performs better than STAR but not as well as WASP, GSNAP and ASElux.

To test the ability of identifying ASE SNPs by ASElux and other methods, we generated another simulated dataset with 20% of genes exhibiting imbalanced allelic expression. These imbalanced genes were randomly selected and one random allele from the selected genes was overexpressed. Compared to the less expressed allele, we generated 1.5–3.5× more reads from the overexpressed allele. To mimic real RNA-seq data, we introduced sequencing error in addition to SNPs and junction reads. To ensure a 50× coverage for each allele, we overexpress one allele by generating more reads when simulating the imbalanced allelic expression. Using the binomial test, we identified a SNP as an ASE SNP if the Bonferroni corrected

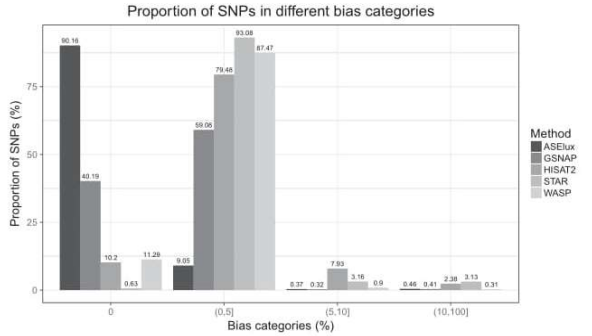


Fig. 1. Proportions of SNPs in different bias categories show that ASElux performs better than general RNA-seq aligners. Y axis shows the proportion of SNPs, and X axis shows the different bias categories. The bias is the absolute difference between the predicted proportion of the reference allele reads and the real proportion of reference allele reads (0.5)

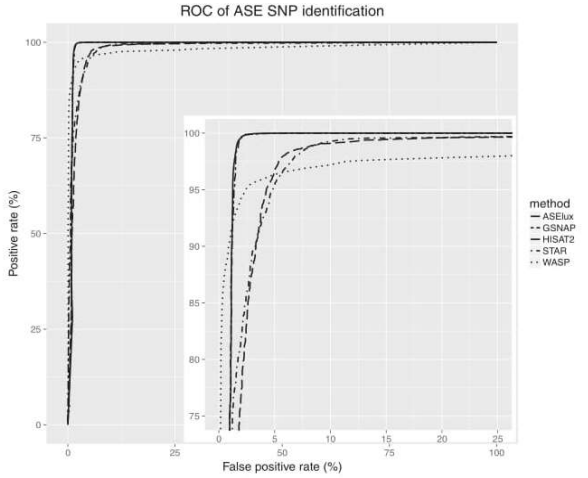


Fig. 2. The receiver operating characteristic (ROC) of ASE SNP identification shows that ASElux performs as well as GSNAP, and outperforms HISAT2 and STAR in a simulated dataset. The X axis is the false positive rate and the Y axis is the positive rate

P-value is less than 0.05. To ensure that the tested methods are fairly compared, we used the intersection of the SNPs identified by all of the tested methods. The receiver operating characteristic (ROC) curve (Fig. 2) indicates that ASElux outperforms HISAT2 and STAR alone on identifying ASE SNPs. Since GSNAP and ASElux both utilize personal SNP information, they identified all of the ASE SNPs (true positive rate = 100%) while maintaining a low false positive rate of ~5%. Although ASElux performed better than GSNAP in the first simulation test (Fig. 1), ASElux and GSNAP showed a comparable number of SNPs showing more than 5% bias. Thus, GSNAP and ASElux performed similarly based on the ROC curve (Fig. 2). The false positive rate of WASP is the lowest among all the tested methods while the true positive is below 92%. In the first simulation test under the null condition (Fig. 1), WASP showed the smallest number of SNPs that have bias more than 5%, suggesting that WASP tends to be highly conservative in order to achieve a low false positive rate. However, since WASP filters out potentially falsely aligned reads by STAR, some SNPs might have insufficient coverage to pass a stringent threshold, which may contribute to the low positive rate of WASP.

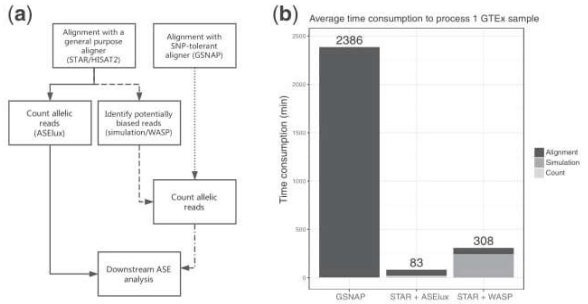


Fig. 3. ASElux is faster than the other tested methods (WASP and GSNAP). (**a**) The workflow of ASE analysis using different programs. (**b**) The estimated average time consumptions to count allelic reads from 10 GTEx RNA-seq samples. The X axis shows the tested method. The Y axis is the time needed for processing the dataset

3.2 Speed benchmarks

We performed the speed benchmark on a server with 64-bit Intel CPUs @2.66 GHz with ~95GB RAM. Figure 3A shows the common workflow of ASE analysis using different methods. Researchers can first map reads using a RNA-seq aligner (STAR/HISAT2) and then count allelic reads with specialized tools, such as ASElux or WASP, or alternatively use a SNP-aware aligner (GSNAP) and count allelic reads directly based on the alignment. Figure 3B shows the average time consumption of a single thread to perform ASE analysis on 10 samples from the GTEx project using STAR + ASElux, STAR + WASP and GSNAP, respectively. Among the tested methods, GSNAP used ~12 GB RAM, HISAT2 ~8 GB RAM and ASElux ~22 GB of RAM, respectively. WASP itself requires no more than 1GB of RAM but the actual RAM requirement of WASP depends on the alignment tool it uses, e.g. STAR would need additional ~30 GB RAM. Counting allelic reads with ASElux is, however, ultra-fast since it only takes ~20 min to process a GTEx RNA-seq sample. Therefore, STAR + ASElux can has a ~33× faster processing speed than GSNAP. WASP requires ~4 CPU hours for each GTEx sample, which makes STAR + WASP ~4× slower than STAR + ASElux. The tests shown in Figure 3 were based on single thread mode. ASElux, HISAT2, STAR and GSNAP all have a multi-thread mode, however, WASP does not support multi-thread computing. Thus, we also tested the multithread mode of each tool except for WASP (Supplementary Fig. S4), which resulted in similar relative alignment speeds across the tools as in the single thread mode. As I/O often plays a significant factor in runtime, the system cache was cleared before each alignment run to avoid any bias due to pre-loaded reference index in the memory during the benchmarking. On average, index loading contributes up to 25% of the overall runtime of ASElux without caching.

3.3 Comparing GSNAP, ASElux, HISAT2 and STAR on 20 experimental samples

To evaluate the performance of ASElux, GSNAP, STAR, HISAT2 and WASP on real RNA-seq data, we processed 20 lung RNA-seq samples from the GTEx (Lonsdale *et al.*, 2013) cohort with the five methods. Each sample contains ~40 million pairs of 76 bp reads. The genotype data of each sample consists of the imputed and genome-wide SNP array data from the GTEx study. For each sample, ~120 000 exonic SNPs were obtained from the VCF file. We

built a personalized index for each sample for ASElux (see methods) and GSNAP, and provided the same indices as in the simulated analysis to STAR and HISAT2. After alignment, we extracted the allelic read counts on each heterozygous SNP for further analyses.

The level of imbalanced allelic expression represented by the RACR provides more information on ASE than the statistics by the binomial test. In an ASE analysis, the proportion of reference allelic reads closer to 0 or 1 often indicates stronger allele-specific gene expression. Therefore, we compared the allelic imbalance (AIB), which is the difference between 0.5 and RACR, derived by ASElux to AIBs by the other methods. Under the null hypothesis that most SNPs will not have an ASE effect, we expect equal expression from both the reference and alternative haplotypes. Consequently, the theoretical distribution of AIB should be centered at zero with a few outliers towards the two tails. If the reference bias hampers the alignment, the mean and median of AIB of all SNPs would shift up from 0, which is shown in Figure 4. GSNAP shows a minimal reference bias in the test. Although ASElux shows a higher average AIB when compared to GSNAP (Fig. 4), its average AIB is significantly lower than the AIBs obtained using WASP, HISAT2 and STAR. WASP, HISAT2 and STAR aligned significantly more reads to the reference allele, indicating a higher reference bias. Although WASP showed the lowest false positive rate in the simulation test, the majority of the WASP SNPs still had a bias more than 0% and less than 5%, which is similar to HISAT2 (Fig. 1). The AIBs derived from the 20 GTEx samples confirmed this similarity between WASP and HISAT2.

ASElux uniquely aligned ~1.3 M allelic reads for each sample; whereas WASP aligned ~1.5 M allelic reads; GSNAP and HISAT2~1.7 M allelic reads; and STAR ~2.8 M allelic reads for each sample, respectively (Supplementary Fig. S5a, Table S2). ASElux identified ~15% fewer SNPs than GSNAP but ~37% more than WASP (Supplementary Fig. S5b). It is worth noting, however, that not all SNPs identified by STAR and HISAT2 are suitable for downstream ASE analysis. Previous studies show that more than 10% of the heterozygous SNPs would be excluded when employing a simulation procedure to correct for the reference alignment bias, (Kukurba *et al.*, 2014; Panousis *et al.*, 2014) while using a general purpose aligner. Thus, overall ASElux would identify a similar number of heterozygous SNPs that are suitable for the downstream ASE analysis when compared to STAR, and HISAT2.

Although STAR uniquely aligned more reads than the other tested tools, it identified a similar number of SNPs with a coverage of ≥30 reads when compared to HISAT2 and GSNAP (Supplementary Table S3, Fig. S4b). The extra allelic reads aligned by STAR mainly overlap with the low coverage SNPs that do not contribute to the ASE analysis (Supplementary Fig. S6). Since WASP depends on STAR for the alignment, a large amount of reads in WASP also overlap with the low coverage SNPs (Supplementary Fig. S6). Thus, WASP identified less SNPs than the other tested tools with similar number of reads aligned (Supplementary Fig. S5).

3.4 ASE analysis strengthens cis-eQTL analysis in identifying local regulation of gene expression

Utilizing the ultra-fast speed of ASElux, we applied ASElux to a dataset of 273 lung RNA-seq samples and imputed SNP array data from the GTEx study. Figure 5 shows that ASElux has significantly less reference bias when compared to the allelic read counts reported by GTEx using their ASE analysis pipeline (Ardlie *et al.*, 2015) (*P*-value$< 2.2 \times 10^{-16}$, two-sided t-test). The distribution of RACR from ASElux is centered at 0.5 whereas the distribution reported by GTEx displays an upward bias. To verify whether ASElux has identified enough heterozygous SNPs for the ASE analysis, we compared the number of SNPs identified by ASElux and the GTEx study. In both analysis by ASElux and the GTEx study a heterozygous SNP must be covered by ≥30 reads to be counted for the downstream ASE analysis. A median of 6385 SNPs passed the simulation correction in the GTEx study (Panousis *et al.*, 2014) for each sample which is ~20% less than the median SNP number identified by ASElux in the 273 GTEx lung samples, indicating that ASElux can identify more ASE SNPs than the GTEx ASE protocol.

In addition to ASE analysis, a cis-eQTL analysis is also widely used to detect allele-specific regulation of gene expression. We compared the ASE results from ASElux to the cis-eQTL results on the same set of SNPs publicly available at the GTEx website. Among the 273 lung samples we aligned with ASElux, we identified 21 550 heterozygous exonic SNPs covered by at least 30 reads in no less than 30 samples. Using a paired t-test, we identified 2765 SNPs residing in 1790 genes that showed ASE ($P < 2.32 \times 10^{-6}$, two-sided). Although not all ASE events are caused by exonic SNPs, the paired

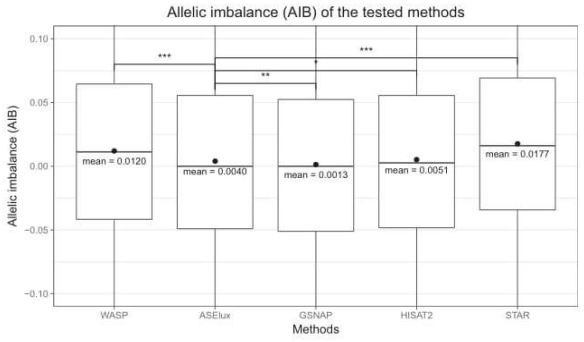


Fig. 4. For ASE analysis, ASElux has less reference bias than WASP, and the general aligners, STAR and HISAT, when testing the 20 real RNA-seq samples. * indicates a *P*-value of 1.05×10^{-3} (two-sided); ** indicates a *P*-value of 7.44×10^{-14} (two-sided); and *** indicates a *P*-value $< 2.2 \times 10^{-16}$ (two-sided). Y axis displays the allele frequency differences between the tested methods and 0.5 (i.e. the two alleles are expressed equally). The reference bias of ASElux is significantly smaller than that of HISAT2, STAR and WASP, but higher than GSNAP. Y axis was limited from -0.1 to 0.1 to show the distribution of most SNPs. The red dots indicate the mean values of each method

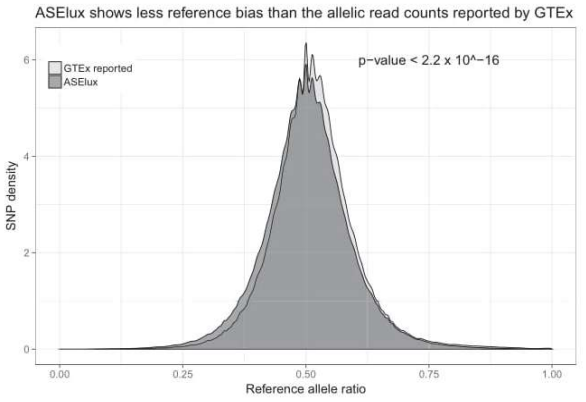


Fig. 5. Compared to the allelic read counts reported by GTEx, ASElux shows less reference bias in 273 lung samples (*P*-value $< 2.2 \times 10^{-16}$, two-sided). The X axis shows the reference allele ratio, and the y axis shows the density of all SNPs

t-test of the exonic SNPs should identify either the causal exonic ASE SNPs or the exonic SNPs tagged by non-coding causal variants. Next, we investigated whether these 2765 ASE SNPs were also identified as cis-eQTLs by GTEx. Using the gene-specific permutation threshold used by GTEx (Ardlie *et al.*, 2015), 1421 of the ASE SNPs were significant cis-eQTLs in lung for the genes they are located in. Overall, 1344 (48.61%) of the ASE SNPs were missed by the cis-eQTL analysis, and 965 (53.91%) of the ASE genes were not identified as cis-eQTL genes. Accordingly, the combination of ASE and cis-eQTL analysis increased the power to identify variants associated with local regulation of gene expression when compared to cis-eQTL analysis alone.

To further investigate the association between ASE and lung disorders, we calculated the linkage disequilibrium (LD) between ASE SNPs and 67 GWAS SNPs (Ardlie *et al.*, 2015) of lung disorders, such as smoking; asthma; lung cancer; chronic obstructive pulmonary disease (COPD); and pulmonary hypertension. There are 11 ASE SNPs in strong LD ($r2 > 0.8$) with the GWAS SNPs (Supplementary Table S4). Of the 11 ASE SNPs, 10 are identified as cis-eQTLs of the genes in which they reside. Both the cis-eQTL and ASE analysis indicate that the alternative genotype of the 10 ASE SNPs is associated with a lower gene expression. It is worth noting, however, that the ASE SNP rs2305480 located in Gasdermin B (GSDMB) is in LD ($R2 = 1$) with a GWAS SNP rs11078927 which is associated with the increased risk of asthma (Bouzigon *et al.*, 2008; Bønnelykke *et al.*, 2014). This SNP rs11078927 has never been identified as a significant cis-eQTL of GSDMB in the lung tissue before. Moreover, rs2305480 has also been identified as a GWAS SNP of another inflammatory disorder, ulcerative colitis (McGovern *et al.*, 2010), supporting the role of the GSDMB gene in several disorders with a known inflammatory component.

We further investigated the potential mechanism of the GWAS SNP rs11078927 and discovered rs11078928, which is a splice donor site variant previously identified in the whole blood and suggested to be involved in asthma (Morrison *et al.*, 2013). It is in tight LD ($R2 = 0.99$) with two asthma GWAS hits, rs2305480 (the ASE SNP) and rs11078927. To examine the splicing effect in a human tissue highly relevant for asthma, we performed a splice-QTL analysis in 273 GTEx lung RNA-seq samples using LeafCutter and identified rs11078928 as a significant splice-QTL of GSDMB in the lung (Fig. 6). The genotype of rs11078928 is significantly associated (P-value $= 4.63 \times 10^{-35}$, two-sided) with the proportional expression level of the junction reads overlapping exon 5 and exon 6 of GSDMB, which is consistent with the splicing event identified previously in the whole blood (Morrison *et al.*, 2013).

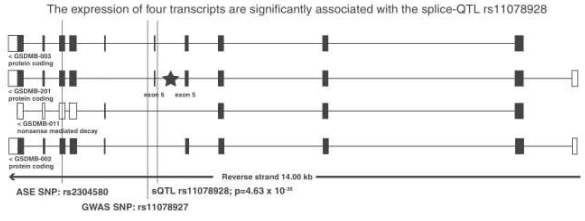


Fig. 6. Using the proportional transcript expression as the phenotype, four transcripts GSDMB-003, GSDMB-201, GSDMB-011 and GSDMB-002 are significantly associated with the genotype of the splice-QTL SNP, rs11078928 (P-value $< 9.43 \times 10^{-4}$, two-sided). The stars indicate that genotypes of rs11078928 are significantly associated with the splice junction reads between the exon 5 and exon 6 of GSDMB (P-value $= 4.63 \times 10^{-35}$, two-sided)

To determine which isoform expression of GSDMB is impacted by the splice variant, we used RSEM (Li and Dewey, 2011) to estimate the expression of isoforms in 273 GTEx lung samples and used the proportional transcript expression as the phenotype for an isoform eQTL analysis. The relative expression of four transcripts GSDMB-003, GSDMB-201, GSDMB-011 and GSDMB-002 are significantly associated with the genotype of the splice-QTL rs11078928 [P-value $< 9.43 \times 10^{-4}$ using linear regression via MatrixeQTL (Shabalin, 2012) with Bonferroni correction] (Fig. 6, Supplementary Table S5). Thus, the biological mechanism underlying the asthma risk SNPs, rs2305480 and rs11078927, is likely mediated by the SNP rs11078928 via splicing regulation on GSDMB in the human lungs.

We further functionally annotated the 52 460 SNPs ($R2 > 0.8$) tagged by the ASE SNPs, identified by ASElux, using ANNOVAR (Wang *et al.*, 2010) (Supplementary Fig. S7). There are 19 additional SNPs identified as splice variants by ANNOVAR and 7 of them were missed by the GTEx cis-eQTL analysis. Taken together, an ASE analysis provides substantially more power for analysis of local gene expression, complementing the regular cis-eQTL analysis.

4 Discussion

With growing interest in ASE analysis, mapping bias remains a critical barrier that hinders the accuracy of ASE analysis in RNA-seq. We provide a novel approach, ASElux that focuses solely on SNP-overlapping reads, allowing a fast and accurate SNP-aware alignment for ASE analysis. To ensure a high alignment accuracy, we used the whole gene body (50% of the reference genome) to build the alignment index. It is worth noting that this speed gain is largely due to the fact that ASElux first aligns all reads to the very small dynamic index to identify the allelic reads and then only aligns them to the large static index. The size of the static index will not affect the speed substantially because the time complexity of searching through suffix array is O(mlog(n)), where the n is the size of the reference and m is the size of the pattern. In addition, ASElux shows a minimal reference bias when compared with other methods based on both simulated and experimental RNA-seq data. ASElux aligns against both alleles by employing personal dynamic indices to minimize the reference bias. We demonstrated that ASElux works optimally with short reads currently generated by most RNA-seq studies.

Due to the complexity of RNA-seq alignment and variable expression of genes across tissues, SNP-calling from RNA-seq is often less accurate than from DNA-sequencing data (Quinn *et al.*, 2013). Thus, external genotype information from whole exome sequencing (WES), whole genome sequencing (WGS), or SNP-arrays are preferred for ASE or eQTL analysis (Ardlie *et al.*, 2015). ASElux and all of the tools tested here do not directly identify SNPs from RNA-seq reads and are therefore only applicable to RNA-seq cohorts that have genotype data available. Simultaneously calling SNPs and ASE from RNA-seq data will enable ASE analyses in additional RNA-seq cohorts, but it will require development of new methods in the future.

Multi-alignment also presents a serious challenge in ASE analysis. Reads generated from different regions might be falsely identified as ASE reads due to their similar sequences. ASElux tries to find all possible multi-alignment loci in addition to the optimal alignment even if the read has the best alignment quality as an ASE read to stringently remove possible false ASE reads. As ambiguously aligned reads are more stringently excluded, ASElux tends to align

less allelic reads than the other tested tools. However, not all SNPs are reliable for the ASE analysis due to the reference alignment bias when using a general-purpose aligner such as STAR and HISAT2, and in fact, the previous studies show a ~10% loss in the number of SNPs during the simulation correction (Kukurba *et al.*, 2014; Panousis *et al.*, 2014). We have shown here that the high accuracy of ASElux has provided more reliable SNPs for the downstream ASE analysis than STAR did in the analyzed GTEx lung samples.

As an alignment tool exclusively designed for ASE analysis, ASElux outperforms most existing methods in speed and provides a better accuracy than the existing non-SNP-aware aligners for correcting the reference bias in alignment while also achieving the closest accuracy to GSNAP. ASElux is ultra-fast: it is able to process 40 million 2× 50 bp reads in 16 min. Combined with a general purpose aligner, such as STAR, STAR + ASElux is ~33 times faster than the golden standard SNP-aware aligner GSNAP, and ~4 times faster than the popular combination of STAR + WASP. The high speed and accuracy make ASElux an ideal tool to perform ASE analysis in large-scale RNA-seq studies. We demonstrated the usefulness of ASElux by performing the ASE analysis in lung RNA-seq data from 273 individuals of the GTEx project in two days (~70 CPU hours using multi-CPU). By comparing the ASE SNPs and eQTLs from the same dataset, we also demonstrated that the combination of ASE and cis-eQTL analysis provides more power to detect local regulation of gene expression.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on 08/14/2016 and dbGaP accession number phs000424.v6.p1 on 08/11/2016.

Funding

This work was supported by the National Institutes of Health (NIH) [grant numbers HL-095056, HL-28481]. A.K. was supported by the NIH [grant number F31HL127921] and M.A. was supported by the NIH [grant number T32HG002536].

Conflict of Interest: none declared.

References

Ardlie,K.G. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Bønnelykke,K. *et al.* (2014) A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.*, **46**, 51–55.

Bouzigon,E. *et al.* (2008) Effect of 17q21 variants and smoking exposure in early-onset asthma. *N. Engl. J. Med.*, **359**, 1985–1994.

Buil,A. *et al.* (2015) Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.*, **47**, 88–91.

Castel,S.E. *et al.* (2015) Tools and best practices for allelic expression analysis. *Genome Biol.*, **16**, 195.

David,A.K. *et al.* (2017) Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods*, **14**, 699–702.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Heap,G.A. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Kukurba,K. *et al.* (2014) Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet.*, **10**, e1004304.

Kumasaka,N. *et al.* (2015) Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.*, **48**, 206–213.

León-Novelo,L.G. *et al.* (2014) A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics*, **15**, 920.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,G. *et al.* (2012) Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.*, **40**, 1–13.

Liu,Z. *et al.* (2014) Comparing computational methods for identification of allele-specific expression based on next generation sequencing data. *Genet. Epidemiol.*, **38**, 591–598.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Manber,U. and Myers,G. (1990) Suffix string arrays: a new searches method for on-line. *Proc. first Annu. ACM-SIAM Symp. Discret. Algorithms*, 319–327.

Manske,H.M. and Kwiatkowski,D.P. (2009) SNP-o-matic. *Bioinformatics*, **25**, 2434–2435.

McGovern,D. *et al.* (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.*, **42**, 332–337.

Morrison,F.S. *et al.* (2013) The splice site variant rs11078928 may be associated with a genotype-dependent alteration in expression of GSDMB transcripts. *BMC Genomics*, **14**, 627.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nong,G. *et al.* (2009) Linear suffix array construction by almost pure induced-sorting. In: 2009 Data Compression Conference, pp. 193–202.

Nong,G. *et al.* (2011) Two efficient algorithms for linear time suffix array construction. *IEEE Trans. Comput.*, **60**, 1471–1484.

Panousis,N.I. *et al.* (2014) Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. **15**, 467.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Quinn,E.M. *et al.* (2013) Development of strategies for SNP detection in RNA-Seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One*, **8**, e58815.

Schirmer,M. *et al.* (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**, 125.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Stevenson,K.R. *et al.* (2013) Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics*, **14**, 536.

van de Geijn,B. *et al.* (2015) WASP: allele-specific software for robust discovery of molecular quantitative trait loci. *Nat. Methods*, **12**, 1061–1063.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, 1001–1006.

Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Supplementary tables:

Supplemental Table 1. Less than 1% of the SNPs are excluded in the ASElux analysis because they are adjacent to INDELs.

Sample name	Number of the excluded SNPs	Number of the remaining SNPs	Proportion of the excluded SNPs (%)
GTEX-13N1W	885	107339	0.82
GTEX-11DXZ	1164	120636	0.96
GTEX-X4XX	888	107369	0.82
GTEX-13OW6	962	105505	0.90
GTEX-OOBJ	1089	119545	0.90
GTEX-XQ8I	905	106793	0.84
GTEX-X585	989	120302	0.82
GTEX-QEG4	908	105506	0.85
GTEX-13JVG	851	105251	0.80
GTEX-Y5V6	942	106175	0.88
GTEX-13D11	933	105622	0.88
GTEX-X5EB	946	104997	0.89
GTEX-12ZZW	885	105707	0.83
GTEX-ZLV1	827	105661	0.78
GTEX-13QJ3	917	105864	0.86
GTEX-XUJ4	828	107092	0.77
GTEX-X4XY	891	105546	0.84
GTEX-XBEC	964	105250	0.91
GTEX-YFC4	899	105595	0.84
GTEX-139YR	904	104773	0.86

Supplemental Table 2. Number of the allelic reads aligned by each method.

Methods	Median of 20 samples	Mean of 20 samples	GTEX-13N1W	GTEX-11DXZ
ASElux	1411680	1427785	1448457	1583544
WASP	1596755	1607712	1615524	1707106
STAR	2928688	2903249	3035030	3022777
HISAT	1841112	1789573	1907193	1940462
GSNAP	1778662	1751781	1813082	1847666
	GTEX-13D11	GTEX-X5EB	GTEX-12ZZW	GTEX-ZLV1
ASElux	1360661	1284373	1374903	1514770
WASP	1420672	1490861	1577986	1834123
STAR	2834599	2581049	2747308	3198729
HISAT	1809144	1549335	1658227	1873079
GSNAP	1744242	1534365	1657635	1913695
	GTEX-X4XX	GTEX-13OW6	GTEX-OOBJ	GTEX-XQ8I
ASElux	1187930	1448550	1322975	1231999
WASP	1030483	1474473	1668605	1231519
STAR	3158780	2711660	2755387	2810048
HISAT	1945811	1804526	1512018	1868316
GSNAP	1794169	1801006	1572970	1837019
	GTEX-13QJ3	GTEX-XUJ4	GTEX-X4XY	GTEX-XBEC
ASElux	1035310	721168	1080764	1022613
WASP	1176780	775960	1292241	953487
STAR	2617469	2049129	2232504	2841092
HISAT	1353661	1137901	1295088	1737545
GSNAP	1465169	1179572	1274297	1543226
	GTEX-X585	GTEX-QEG4	GTEX-13JVG	GTEX-Y5V6
ASElux	1550535	1383654	1545208	1790518
WASP	1871069	1551725	1903785	1844965
STAR	3387669	2719446	3327758	3514072
HISAT	1838873	1767707	1951838	2285574
GSNAP	1832315	1681174	1906546	2198372
	GTEX-YFC4	GTEX-139YR	sim_A	sim_B
ASElux	1395430	1307405	10170552	5824938
WASP	1576968	1328522	7636477	4028250
STAR	2689094	2720051	10615486	6085666
HISAT	1695771	1698869	10726304	6146009
GSNAP	1656754	1555761	11087068	6362983

sim_A	Both alleles are equally expressed in the simulation data A.
sim_B	About 20% of genes exhibit imbalanced allelic expression in the simulation data B.

Supplemental Table 3. Number of the SNPs identified by each method.

Methods	Median of 20 samples	Mean of 20 samples	GTEX-13N1W	GTEX-11DXZ
ASElux	7882	8357	7242	9050
WASP	5931	6393	5357	6816
STAR	9115	9781	8311	10583
HISAT	8842	9437	8233	10291
GSNAP	9035	9602	8298	10365
	GTEX-13D11	GTEX-X5EB	GTEX-12ZZW	GTEX-ZLV1
ASElux	7401	8010	7753	8616
WASP	5192	6012	5849	6894
STAR	8506	9258	8972	10310
HISAT	8377	9010	8674	9773
GSNAP	8481	9219	8851	9974
	GTEX-X4XX	GTEX-13OW6	GTEX-OOBJ	GTEX-XQ8I
ASElux	7140	7574	7837	6542
WASP	4359	5618	6128	4497
STAR	10007	8723	9047	8643
HISAT	9411	8468	8656	8406
GSNAP	9551	8663	8852	8495
	GTEX-13QJ3	GTEX-XUJ4	GTEX-X4XY	GTEX-XBEC
ASElux	6739	4144	7141	6150
WASP	4576	2220	5397	3835
STAR	8539	6313	8408	8798
HISAT	8078	6232	8117	8375
GSNAP	8408	6352	8272	8486
	GTEX-X585	GTEX-QEG4	GTEX-13JVG	GTEX-Y5V6
ASElux	8956	7533	7720	9240
WASP	6861	5737	6274	6949
STAR	10513	8691	9279	10687
HISAT	9949	8550	8814	10362
GSNAP	10151	8537	8922	10432
	GTEX-YFC4	GTEX-139YR	sim_A	sim_B
ASElux	7174	6641	60228	57307
WASP	5483	4703	55186	50645
STAR	8333	7846	61721	60658
HISAT	8068	7549	61960	60861
GSNAP	8117	7662	61918	61324

sim_A	Both alleles are equally expressed in the simulation data A.
sim_B	About 20% of genes exhibit imbalanced allelic expression in the simulation data B.

Supplemental Table 4. The ASE SNPs that are in LD (r^2>0.8) with lung disease GWAS SNPs (refs 16 and 19).

chr	position	type	gene	AB	ASE_p	cis_eQTL_beta	cis_eQTL_p	GWAS_SNP	GWAS_trait	LD_r2
6	32605207	5_prime_UTR_variant	HLA-DQA1	0.69	1.95E-18	-1.18	6.66E-84	rs9272346	asthma	0.99
6	32605257	missense_variant	HLA-DQA1	0.73	2.82E-20	-1.06	1.33E-53	rs9272346	asthma	0.82
6	32605274	synonymous_variant	HLA-DQA1	0.70	3.94E-26	-1.19	1.53E-86	rs9272346	asthma	0.99
6	32609126	missense_variant	HLA-DQA1	0.71	6.19E-13	-1.19	4.00E-88	rs9272346	asthma	1.00
6	32610009	missense_variant	HLA-DQA1	0.56	1.29E-16	-1.13	1.61E-66	rs9272346	asthma	0.91
17	38062196	missense_variant	GSDMB	0.59	5.49E-11	NA	NA	rs11078927	asthma	1.00
17	38062217	non_coding_transcript_exon_variant	GSDMB	0.58	1.88E-08	-0.18	3.65E-07	rs7216389	asthma	0.95
17	38063381	synonymous_variant	GSDMB	0.63	3.77E-09	-0.18	3.65E-07	rs7216389	asthma	0.95
6	32605295	synonymous_variant	HLA-DQA1	0.61	5.68E-17	-0.53	5.20E-09	rs2395185	lung_cancer	0.85
6	32609855	synonymous_variant	HLA-DQA1	0.62	1.48E-11	-0.57	3.83E-10	rs2395185	lung_cancer	0.91
6	32610008	missense_variant/synonymous_variant	HLA-DQA1	0.51	1.23E-15	-0.58	2.41E-10	rs2395185	lung_cancer	0.91

Column names

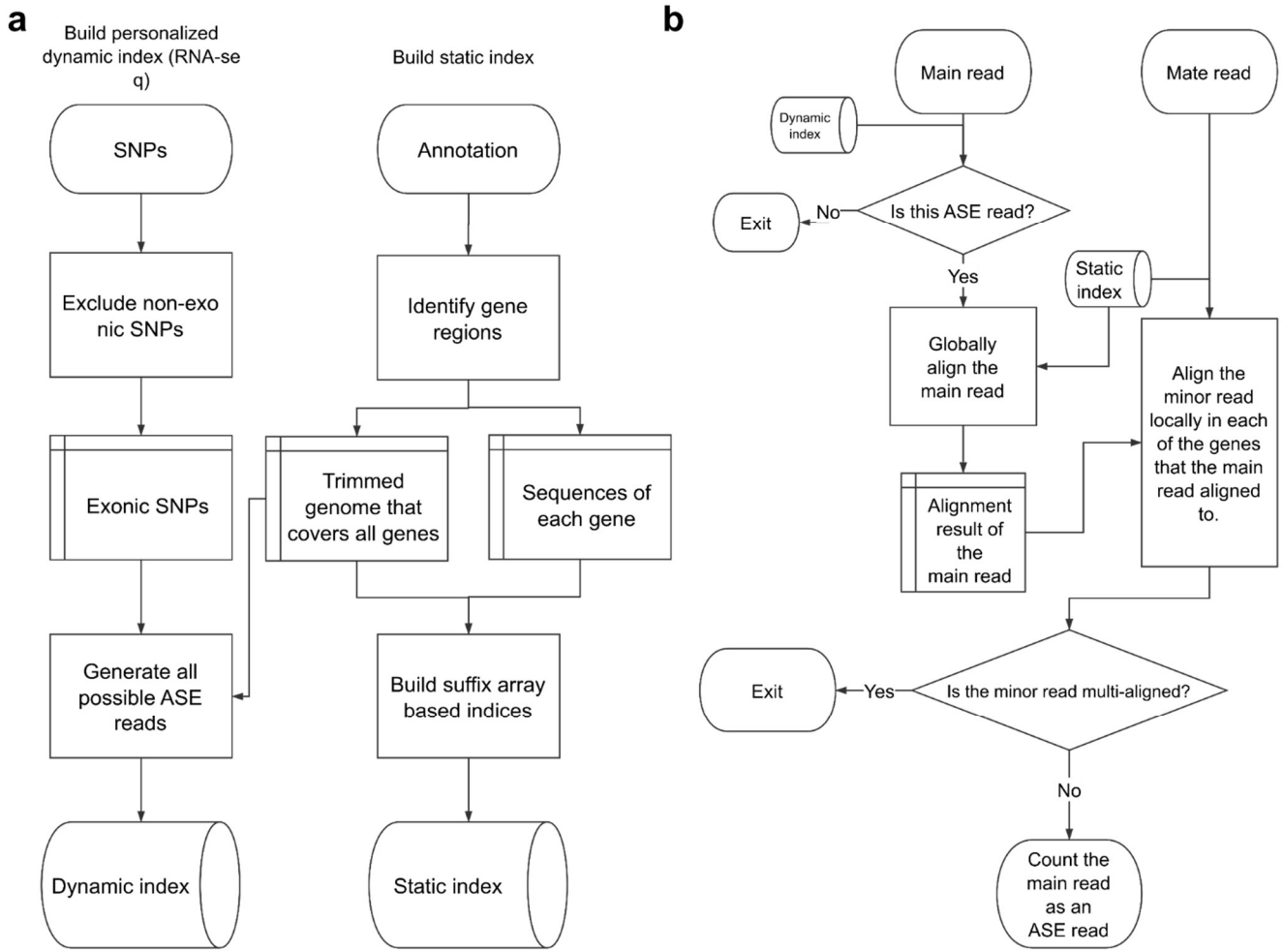
chr	The chromosome of the ASE SNP
position	The chromosomal position of the ASE SNP
type	The type of the ASE SNP
gene	The gene that the ASE SNPs reside in
AB	Allelic balance of the ASE SNP
ASE_p	The p value of the paired t test for the ASE SNPs
cis_eQTL_beta	The beta value of the cis-eQTL analysis
cis_eQTL_p	The p value of the cis-eQTL analysis
GWAS_SNP	The rs ID of the GWAS SNP
GWAS_trait	The trait of the GWAS SNP
LD_r2	The LD (r2) between the ASE SNP and the GWAS SNP

Supplemental Table 5. Transcripts which are significantly associated with rs11078928.

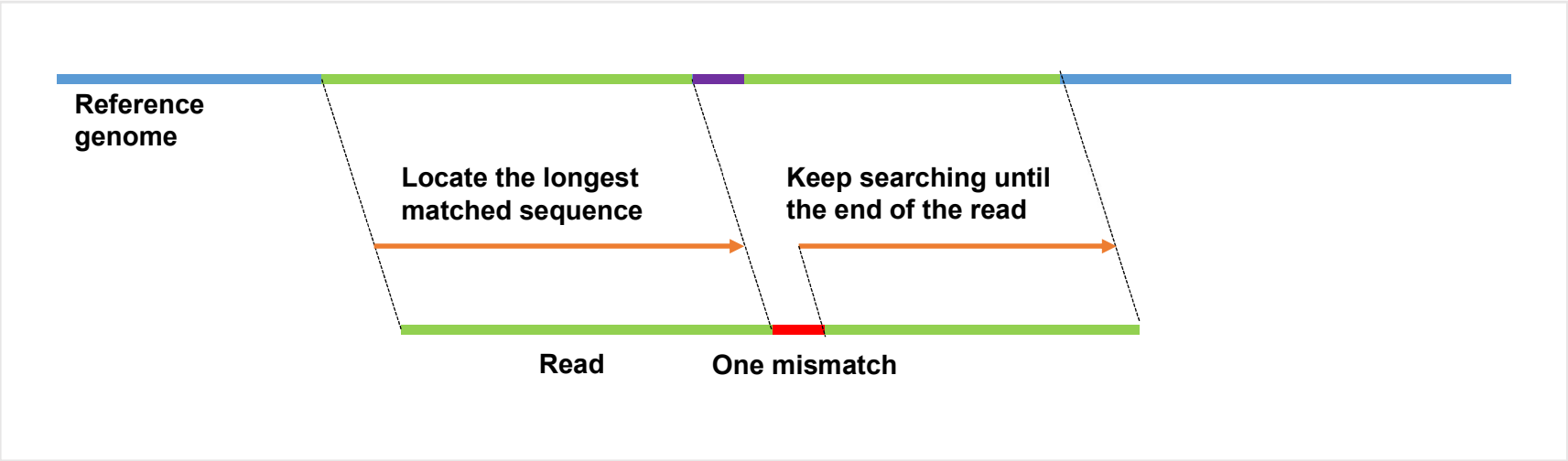
SNP	transcript	beta	t-stat	p-value	FDR
17_38064469	ENST00000523371.1	1.05E-02	3.62E+00	3.48E-04	1.84E-03
17_38064469	ENST00000360317.3	-5.53E-02	-8.93E+00	7.00E-17	1.24E-15
17_38064469	ENST00000394179.1	9.81E-03	7.67E+00	3.09E-13	4.10E-12
17_38064469	ENST00000309481.7	2.08E-02	4.88E+00	1.82E-06	1.93E-05

We used the proportional transcript expression compared to the gene level expression instead of TPM as the phenotype.

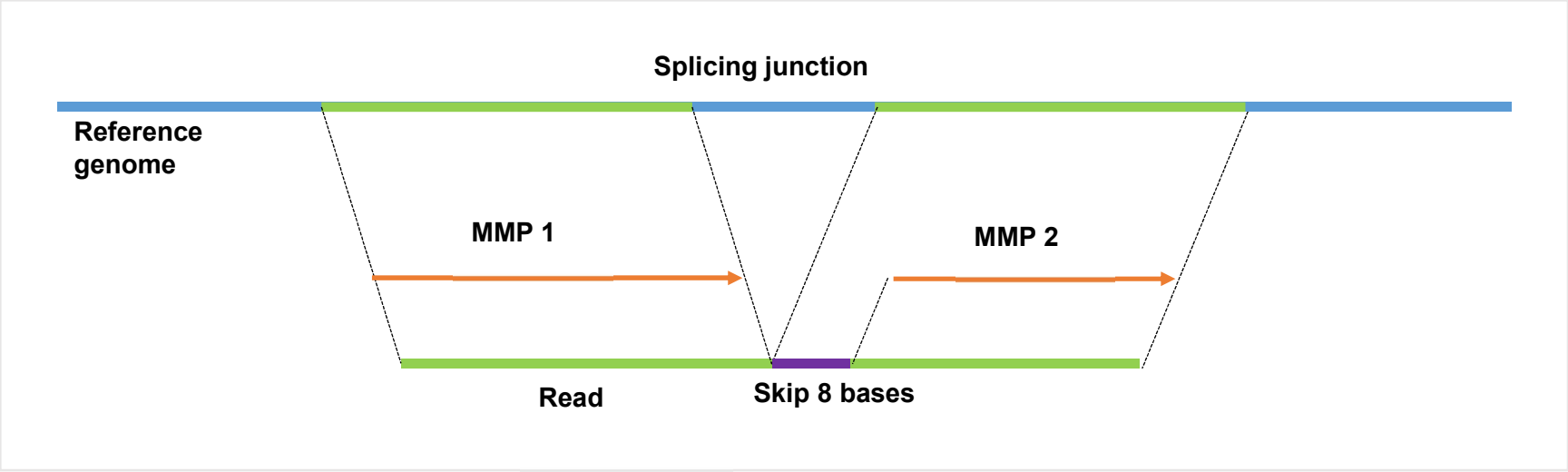
Supplementary Figures:



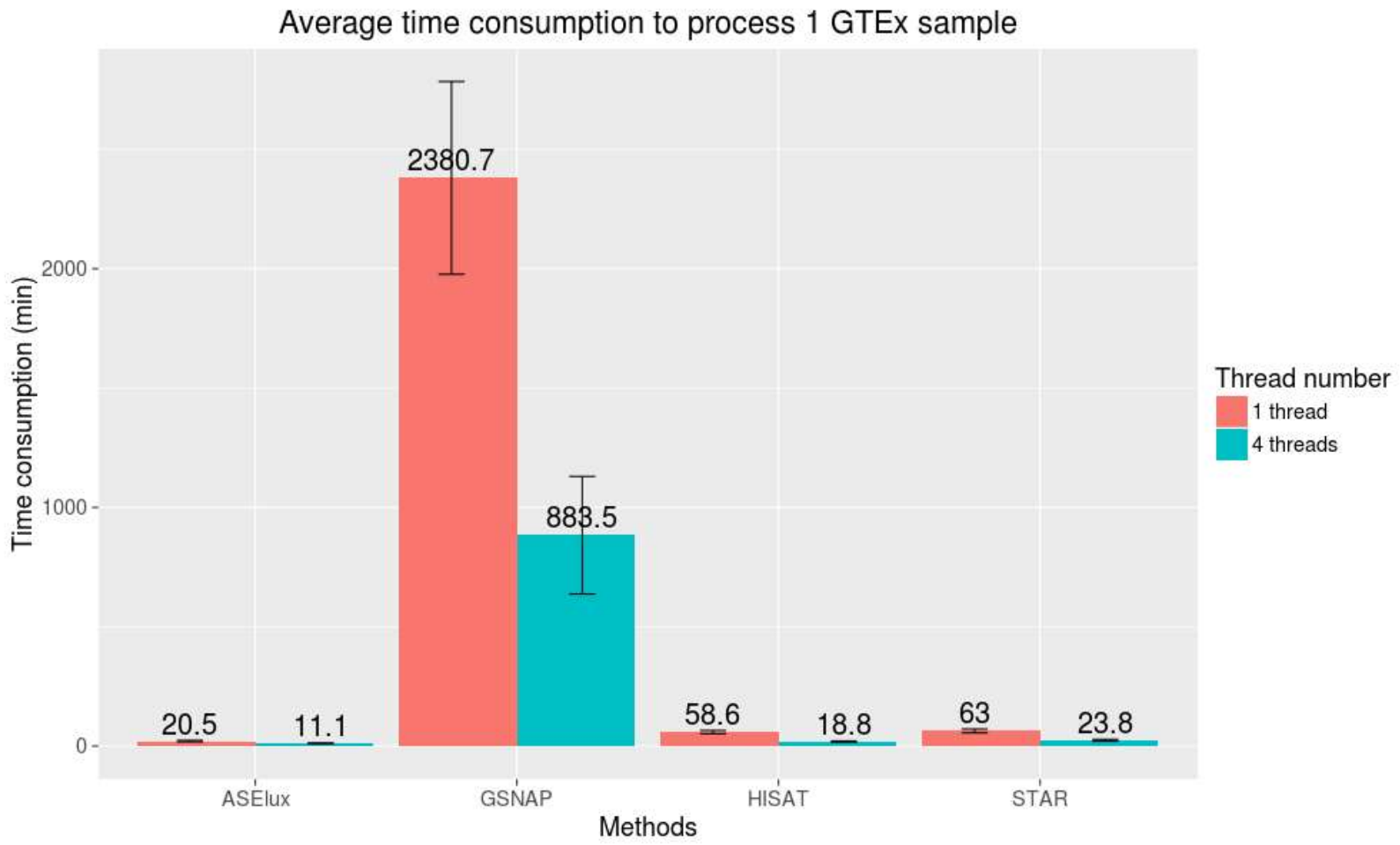
Supplemental Figure 1. The workflow of ASElux. (a). The process of building the hybrid index system. (b). The process of aligning ASE reads with a personalized index.



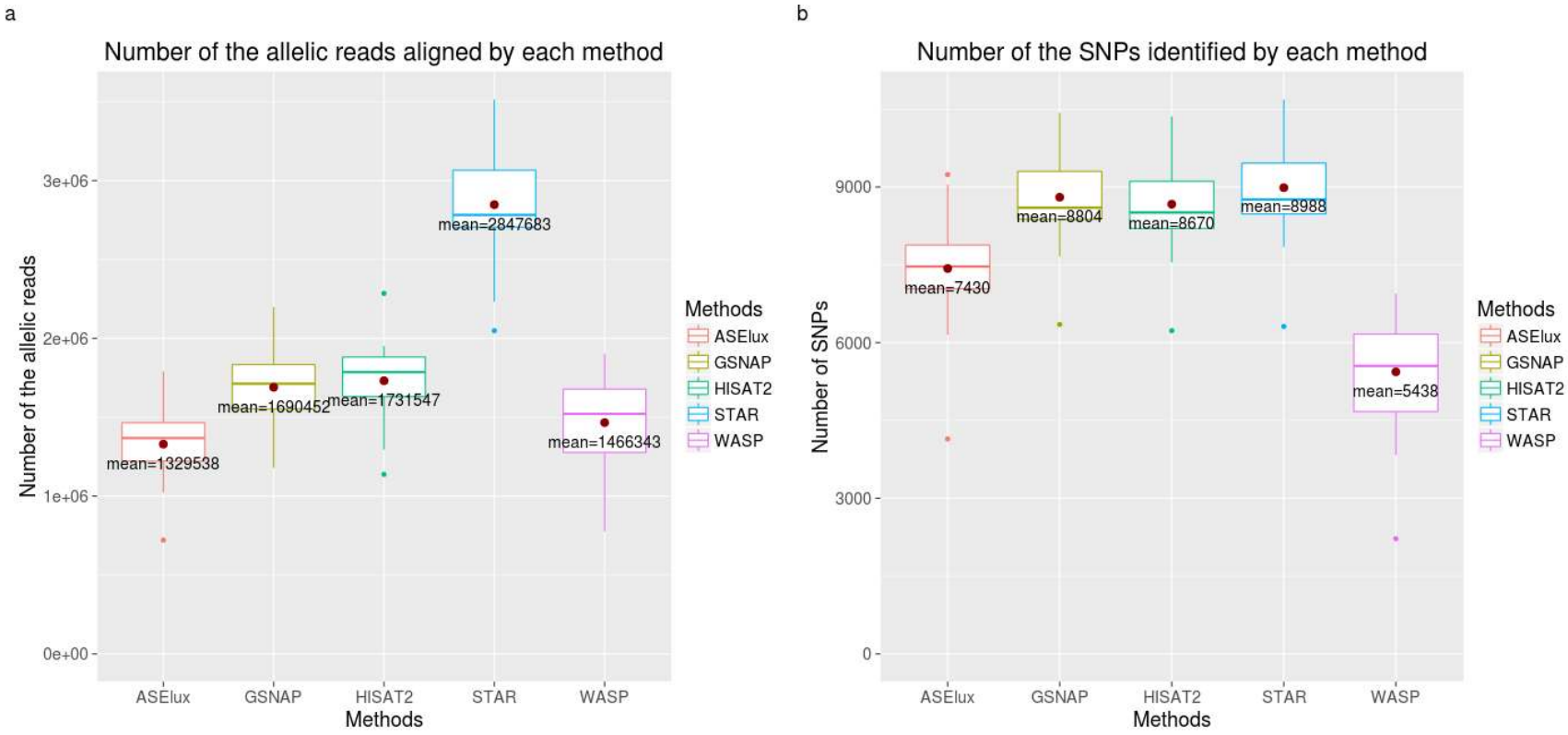
Supplemental Figure 2. An example of aligning a read with one mismatch when aligning the main read to the dynamic index. Step 1: find the longest matched sequence. Step 2: skip 1 base and continue searching until reaches the end of the read. Step 3: check if the whole read can be aligned with up to 2 mismatches.



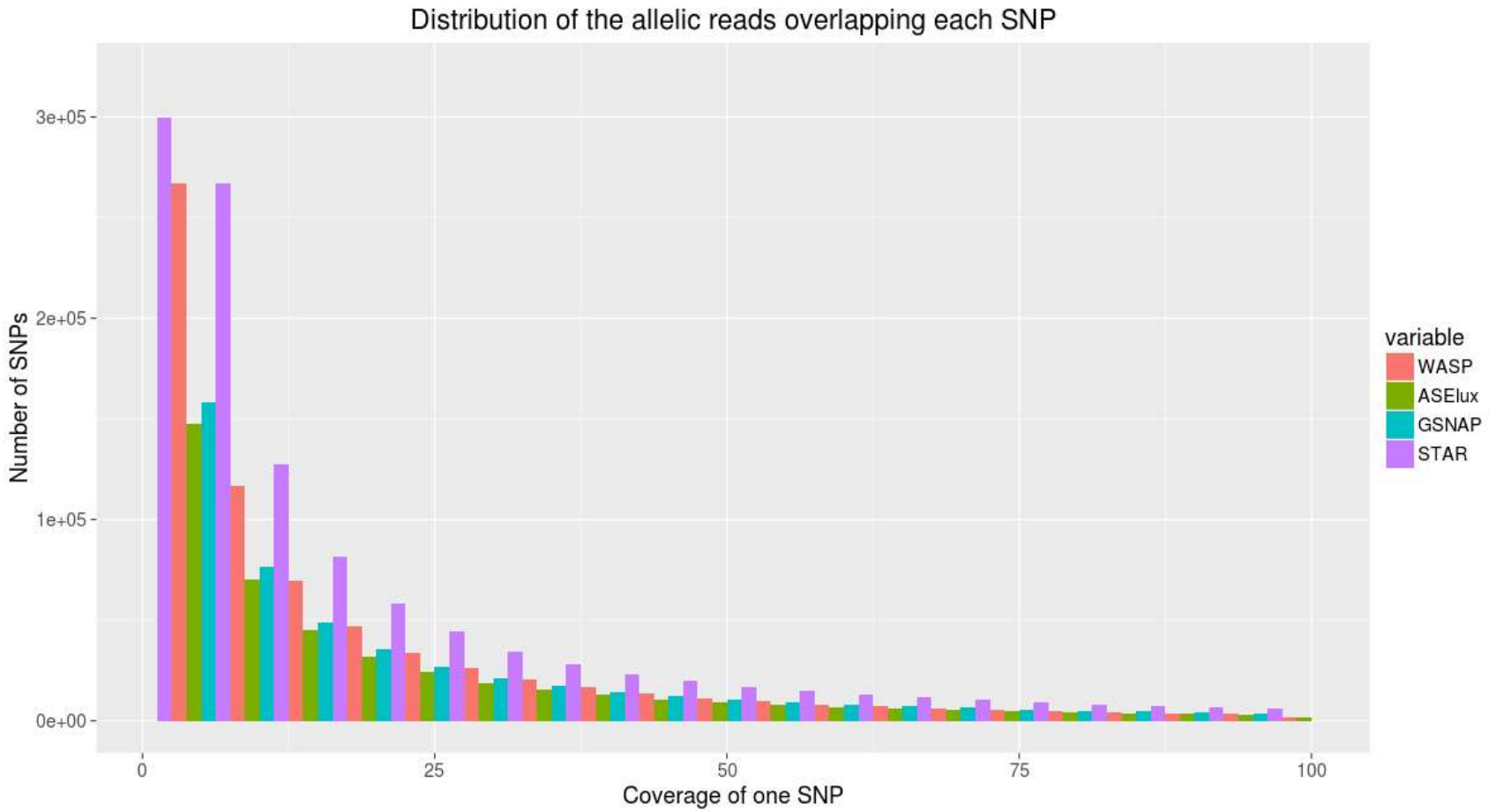
Supplemental Figure 3. An example of aligning a junction read. Step 1: find the MMP for the read. Step 2: skip 8 bases and find the MMP for the rest of the read. Step 3: reassemble the read on the reference genome.



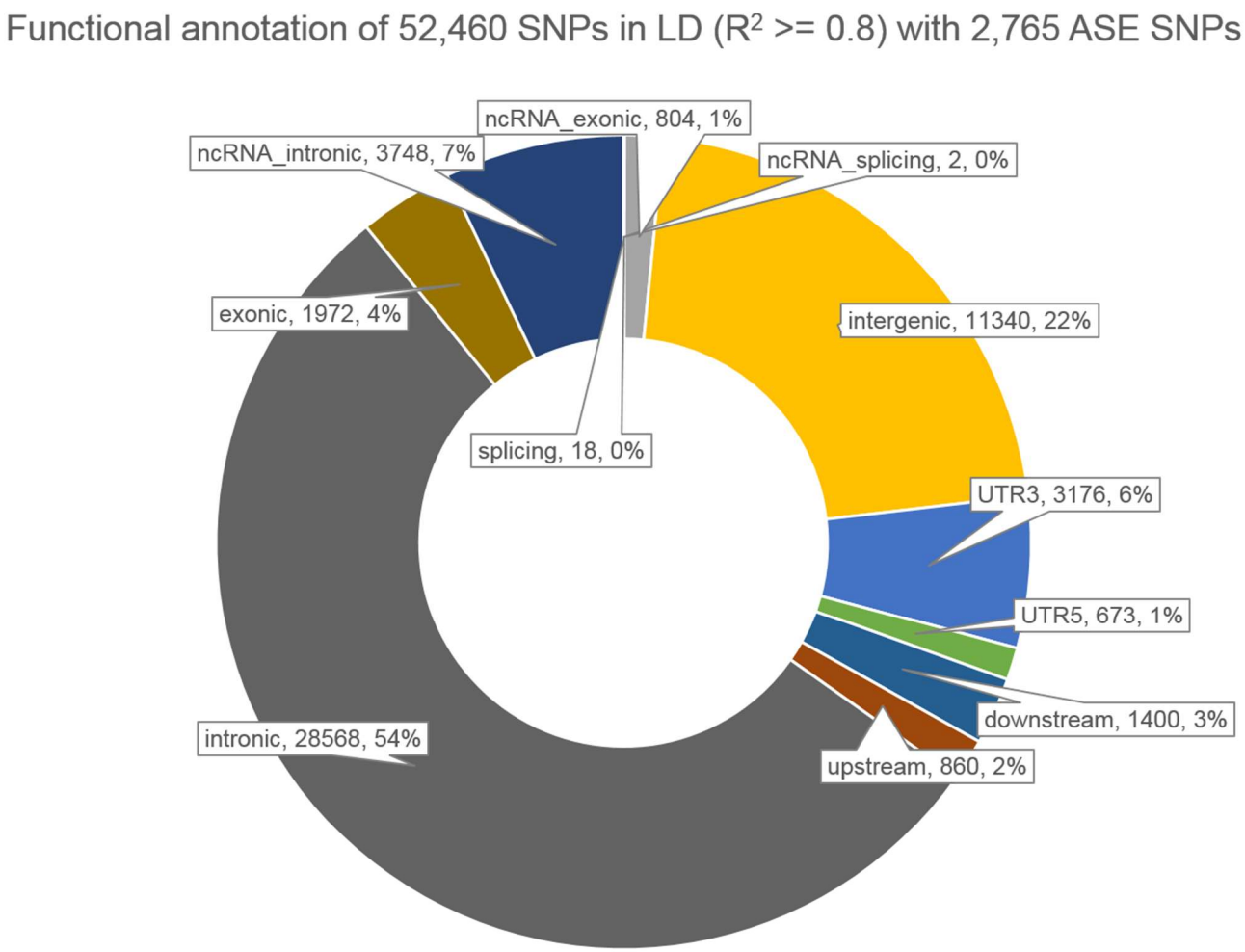
Supplemental Figure 4. The alignment speeds of each tool in both the single and multi-thread modes are displayed and ASElux is faster than all other tested methods in both modes. The X axis shows the names of the tools. The Y axis is the time needed for processing the data set.



Supplemental Figure 5. (a). The number of all uniquely aligned allelic reads aligned by each method. (b). The number of SNPs that have a coverage of =>30 reads identified by each method.



Supplemental Figure 6. The number of allelic reads overlapping each SNP shows that STAR and WASP identified a lot more SNPs with a low coverage (<30 reads) than GSNAP and ASElux.



Supplemental Figure 7. The functional annotation of the 52,460 SNPs in LD (R^2>=0.8) with the identified 2,765 ASE SNPs in 273 GTEx lung samples.

Chapter III

The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance

The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance

Zong Miao[1,2], Marcus Alvarez[1], Arthur Ko[3], Yash Bhagat[1], Elior Rahmani[4], Brandon Jew[2,4], Sini Heinonen[5], Karen L Mohlke[6], Markku Laakso[7], Kirsi H. Pietiläinen[5,8], Eran Halperin[1,4,9,10], Päivi Pajukanta[1,2,11*]

1. Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA, 90095
2. Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA, 90095.
3. Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA, 90095.
4. Computer Science Department in the School of Engineering, UCLA, Los Angeles, CA, USA, 90095
5. Obesity Research Unit, Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland, 00014
6. Department of Genetics, University of North Carolina, Chapel Hill, NC, USA, 27599
7. Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland, 70210
8. Obesity Center, Endocrinology, Abdominal Center, Helsinki University Central Hospital and University of Helsinki, Helsinki, Finland, 00260
9. Department of Computational Medicine, UCLA, Los Angeles, CA, USA, 90095
10. Department of Anesthesiology and Perioperative Medicine, UCLA, Los Angeles, CA, USA, 90095
11. Institute for Precision Health, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA, 90095

* Corresponding author:
Päivi Pajukanta, MD, PhD
Professor
Vice Chair
Department of Human Genetics
David Geffen School of Medicine at UCLA
Gonda Center, Room 6335B
695 Charles E. Young Drive South
Los Angeles, California 90095-7088, USA
Email: ppajukanta@mednet.ucla.edu

Abstract:

Reverse causality has made it difficult to establish the causal directions between obesity and prediabetes and obesity and insulin resistance. To disentangle whether obesity causally drives prediabetes and insulin resistance already in non-diabetic individuals, we utilized the UK Biobank and METSIM cohort to perform a Mendelian randomization (MR) analyses in the non-diabetic individuals. Our results show that both prediabetes and systemic insulin resistance are caused by obesity ($p=3.50\times10^{-80}$ and $p=8.96\times10^{-25}$). As obesity reflects the amount of body fat, we next studied how adipose tissue affects insulin resistance. We performed both bulk RNA-sequencing and single nucleus RNA sequencing on frozen human subcutaneous adipose biopsies to assess adipose cell-type heterogeneity and mitochondrial (MT) gene expression in insulin resistance. We discovered that the adipose MT gene expression and body mass index (BMI) are both independently associated with insulin resistance ($p\leq0.05$ for each) when adjusting for the decomposed adipose cell-type proportions. Next, we showed that these 3 factors, adipose MT gene expression, BMI, and adipose cell types, explain a substantial amount (44.39%) of variance in insulin resistance and can be used to predict it ($p\leq2.64\times10^{-5}$ in 3 independent human cohorts). In summary, we demonstrated that obesity causes both prediabetes and insulin resistance, and discovered that individuals' adipose cell-type composition, adipose MT gene expression, and BMI predict their insulin resistance, emphasizing the critical role of adipose tissue in systemic insulin resistance.

Author Summary

Obesity is a global health epidemic predisposing to type 2 diabetes (T2D) and other cardiometabolic disorders. Previous studies have shown that obesity has a causal effect on T2D;

however, it remains unknown whether obesity causes prediabetes and insulin resistance already in non-diabetic individuals. By utilizing almost half a million individuals from the UK Biobank and the Finnish METSIM cohort, we identified a significant causal effect of obesity on prediabetes and insulin resistance among the non-diabetic individuals. Next, we investigated the role of subcutaneous adipose tissue in these obesogenic effects. We discovered that the adipose mitochondrial gene expression and body mass index (BMI) are independently associated with insulin resistance after adjusting for the tissue heterogeneity. For the latter, we estimated the adipose cell type proportions by utilizing single-nucleus RNA sequencing of frozen adipose tissue biopsies. Moreover, we established a prediction model to estimate insulin resistance using BMI and adipose RNA-sequencing data, which enlightens the importance of adipose tissue in insulin resistance and provides a helpful tool to impute the insulin resistance for existing adipose RNA-sequencing cohorts. Overall, we discover the causal effect of obesity on prediabetes and insulin resistance and the key role of adipose tissue in insulin resistance.

Introduction:

The global obesity epidemic is driving the concomitant rapid increase in the prevalence of cardiometabolic disorders, including type 2 diabetes (T2D)[1,2]. It is well established that obesity, prediabetes, and insulin resistance are tightly associated[3,4,5,6,7,8]. Moreover, inflammation has been identified as the link between obesity and insulin resistance[9,10,11,12,13,14]. For example, Roberts-Toler et al. showed that diet-induced obesity can cause insulin resistance in mouse brown adipose tissue[15]. Wensveen et al. showed that natural killer cells can mediate the association between obesity and insulin resistance in mice[16]. However, the direction of the causal effect between obesity and insulin resistance remains elusive in humans[17,18]. Thus, direct evidence of obesity causing systemic insulin resistance in humans is still lacking. To this end, we performed a Mendelian randomization (MR) analysis using the genotype and metabolic traits of unrelated non-diabetic individuals from both UK Biobank (UKB)[19] and the Finnish METabolic Syndrome In Men (METSIM) cohort[20]. Our MR analysis provides strong evidence for the first time that obesity causes both prediabetes and insulin resistance in non-diabetic humans.

The key functions of adipose tissue, i.e. lipogenesis (storing fat) and lipolysis (mobilizing the stored fat), make it one of the most important tissues contributing to obesity. Thus, it would be important to better understand how much this endocrine tissue contributes to insulin resistance and T2D. Since adipose tissue is complex and contains multiple cell-types, adipose cell-type composition may be affected by obesity. Weisberg et al. showed that the number of macrophages increases in the adipose tissue of obese mice[21]. Furthermore, in human adipose tissue, BMI was reported to be negatively correlated with the number of adipocytes[22] and positively correlated with the size of the adipocytes[21,22,23]. However, the effects of different adipose cell-type proportions on insulin resistance have not been systematically assessed in humans previously.

Fluorescence-activated cell sorting (FACS) has been used for characterizing and defining some of the cell types in human adipose samples [24,25,26]. Even though the number of marker proteins that can be simultaneously measured in FACS has progressively increased[27], FACS relies on predetermined cell-type specific marker proteins to isolate different cell types and is thus unable to discover new cell types or sub-cell types. Moreover, since FACS requires a large starting number of cells (more than 10,000) in suspension[28], it is unable to isolate single cells from a low quantity cell population. Overall, it is highly challenging to evaluate all cell types of solid tissues, such as adipose tissue, using FACS. Thus, to thoroughly investigate the tissue heterogeneity in human adipose tissue, we performed single nucleus RNA sequencing (sn-RNA-seq) of all adipose cell types using frozen human subcutaneous adipose biopsies. We then utilized the sn-RNA-seq data to define expression profiles of signature genes in different adipose cell-types to decompose cell-type proportions in the bulk adipose RNA-seq cohorts. This helped us leverage the gene expression information available in the adipose bulk RNA-seq data to assess whether adipose cell-type composition influences systemic insulin resistance.

Previous studies have shown that the biogenesis and metabolic activities of the mitochondria (MT) are impaired in the adipose tissue of obese individuals[2,29,30,31,32]. Experimental evidence also shows that declined MT function can elicit insulin resistance in mice[33,34]. Paglialunga et al. further demonstrated that elevated MT reactive oxygen species (ROS) emission in murine white adipose tissue contributes to insulin resistance[34]. Furthermore, dysfunction of MT in muscle and liver associates with insulin resistance in humans[35,36]. However, since the MT activity, BMI, and systemic insulin resistance are associated with each other, it is unclear whether these associations are caused by independent mechanisms or confounded by a shared trait. To this end, we investigated whether MT gene expression in human adipose tissue is independently associated

with systemic insulin resistance and BMI. Furthermore, we built a prediction model to investigate whether systemic insulin resistance (i.e. Matsuda index) in humans can be predicted using adipose cell-type proportions, adipose MT gene expression, and BMI as an input. Overall our studies helped determine the causal role of obesity in human insulin resistance, of which a major portion is driven by adipose tissue.

Results:

Obesity causes prediabetes and systemic insulin resistance in non-diabetic individuals:

Although obesity and prediabetes are known to be associated[7,8], there is no previous evidence about the causal direction between them in non-diabetic individuals. To this end, we performed an MR analysis to investigate whether prediabetes (assessed by serum HbA1C level between 5.7-6.4[37]) is caused by obesity. Figure 1A shows the MR models we used to explore causal associations between prediabetes and BMI. For this MR analysis, we first utilized the UKB (n=~380k) to identify 962 non-redundant SNPs (R^2<0.01) significantly associated with BMI ($p<5\times10^{-8}$) as the genetic instrumental variable (IV) in the MR analysis (see Methods). Then we used MR-PRESSO[38] that adjusts for the potential horizontal pleiotropy and identified a significant positive causal effect of BMI on prediabetes in the UKB (estimate effect = 0.053; p-value = 3.0×10^{-73}). Figure 1B shows the effects of the 962 IVs on the exposure variable and outcome in the UKB. To test the possibility of the reverse causal path, we utilized 284 independent prediabetes GWAS SNPs as IVs in the UKB and explored the potential causal effect of prediabetes on obesity. Figure 1C shows that there is no causal effect of prediabetes on BMI (estimated effect = 0.065, p-value = 0.096). Thus, using MR in the UKB, we established a one-way causal effect of obesity on prediabetes.

To further investigate this finding, we next explored the causal relationship between obesity and systemic insulin resistance (i.e. decreased insulin sensitivity assessed by the Matsuda index) in the METSIM cohort. Of the 962 independent significant BMI GWAS SNPs identified in the UKB, we selected the ones that were nominally (p-value<0.05) associated with BMI in METSIM as IVs. Including 84 such BMI associated SNPs as IVs and using MR-PRESSO, we discovered a negative causal effect of BMI on the Matsuda index (i.e. insulin sensitivity) in METSIM

(estimate effect = -0.51, p-value = 4.9e-23). It is worth noting that MR-PRESSO did not find any evidence of pleiotropy (p-value = 1.0) when testing the causal effect of BMI on the Matsuda index in METSIM. This suggests that our MR model fulfills the second and third assumptions of MR analysis (i.e. the IV is not associated with the hidden confounders or with the outcome variable (i.e. Matsuda index) when conditioning on the exposure variable (i.e. BMI)). When investigating the opposite direction of causality (i.e. insulin resistance -> obesity) using a similar pipeline, we found no genome-wide significant SNPs associated with the Matsuda index in METSIM or other cohorts of previous studies. Furthermore, no insulin resistance parameters are measured in the UKB. Therefore, the first assumption of MR (i.e. IV is associated with the exposure variable) cannot be fulfilled. As METSIM may be underpowered to identify genome-wide significant SNPs for the Matsuda index, the current sample size of the Matsuda index GWAS does not allow a reliable MR analysis in this opposite direction. Accordingly, assessment of this direction using MR warrants further investigation in larger GWAS cohorts with the Matsuda index available for study.

In summary, we established a one-way causal effect of obesity on prediabetes among the non-diabetic individuals from the UKB. In contrast, prediabetes did not show any evidence for a causal effect on BMI. We then followed up this finding by identifying a negative causal effect of BMI on insulin sensitivity (i.e. Matsuda index) in the non-diabetic individuals from METSIM. Although METSIM is underpowered to investigate the reverse-causal effect (i.e. insulin resistance -> BMI), the MR analyses performed in both UKB and METSIM show that obesity causes prediabetes and insulin resistance before the development of diabetes and thus, prediabetes is less likely to cause obesity among the non-diabetic population.

Adipose mitochondrial (MT) gene expression plays a key role in insulin resistance:

We have shown that obesity leads to increased insulin resistance in human using MR analysis. Since obesity reflects the amount of body fat, adipose tissue may play an important role in obesity-related insulin resistance. Thus, we further investigated how adipose tissue affects the Matsuda index in the METSIM cohort. Among the 4k unrelated individuals in METSIM, 335 had bulk RNA-seq data from the subcutaneous adipose tissue biopsies. To estimate MT gene expression, we estimated the transcripts per million (TPM) values of each gene in the RNA-seq data and used the sum of TPMs from all 37 MT encoded genes to represent the MT gene expression. We also included the first 3 genetic PCs as covariates when correcting the MT expression to adjust for the potential population stratification (see Methods). We corrected Matsuda index for age, age^2 and excluded people who have type 2 diabetes (T2D) (n=11). The Matsuda index and MT gene expression were inverse normal transformed to obtain normal distribution. Using these data, we discovered that the MT gene expression is significantly associated with Matsuda index (p-value = 9.60×10^{-15}, n= 324) (Figure 2A).

To further replicate and validate this finding, we tested the association between MT gene expression and insulin resistance in the RNA-seq data from 5 different tissues in the GTEx cohort. Since the GTEx cohort does not have the Matsuda index measured, we tested the MT gene expression difference between the patients with T2D and non-diabetic individuals. Figure 2B shows that, as in METSIM, the patients with T2D in GTEx have significantly lower MT gene expression in the subcutaneous and visceral adipose tissue than the non-diabetic individuals (Supplementary Figure 1A). However, in three other non-adipose tissues from GTEx, only muscle MT gene expression (n=305) showed the significant difference between T2D patients and non-diabetics ($P=3.42\times10^{-2}$). In the liver and whole blood (total n=412), no significant difference was observed (Supplementary Figure 1). These results support the important role of adipose

tissue and muscle in the insulin resistance related metabolic process; however, due to the limited sample size of the GTEx liver cohort (n=123), we cannot exclude the potential role of liver in insulin resistance. Taken together, the adipose MT gene expression is significantly associated with insulin resistance in the METSIM cohort and T2D in the GTEx cohort. In both subcutaneous and visceral adipose tissue, the MT gene expression is significantly lower in insulin resistant individuals.

Assessing tissue heterogeneity and adipose cell type proportions using single nucleus RNA sequencing:

Single nucleus RNA sequencing reveals 8 cell-types in human adipose tissue:

Adipose tissue is a complex tissue that consists of multiple cell-types, such as adipocytes, preadipocytes, macrophages, fibroblasts, and vascular cells. Even though adipocytes comprise ~90% of the total volume in the human adipose tissue, they only take ~50% of the total cell count[39,40,41]. We hypothesized that the metabolic processes in different contexts and adipose cell-types associated with obesity may substantially affect systemic insulin resistance. To investigate how adipose cell-type heterogeneity affects insulin resistance, we performed single-nucleus RNA sequencing (sn-RNA-seq) on 6 frozen human subcutaneous adipose tissue biopsies (see the Methods) and used cell-type-specific gene expression data as a reference to identify signature genes for each adipose cell-type in order to estimate cell-type proportions from the bulk adipose RNA-seq profiles.

Using the 10x Genomics platform, we sequenced on average ~2,600 nuclei for each sample and obtained the non-zero expression of ~500 genes per cell (for sample-specific metrics, see Supplementary table 1). Next, we used Seurat[55] to cluster the sn-RNA-seq data and identified 8 adipose cell-type clusters based on the gene expression profiles of the adipose nuclei. It is worth

noting that the adipocyte cluster comprises 44.0% of the total cell number, which is in line with the previous findings[41]. Figure 3A shows the tSNE plots of the 8 adipose cell-type clusters in 15,623 nuclei. Supplementary Figure 2A shows the tSNE plot that is colored by the sample IDs. These data show that clustering is largely driven by distinct gene expression profiles from different adipose cell-types rather than by the differences between individuals.

Estimating adipose cell-type proportions using the sn-RNA-seq as the reference:

Next, we used MuSiC to estimate the proportions of each cell-type from the bulk adipose RNA-seq data. Utilizing both bulk RNA-seq and sn-RNA-seq data from these 6 individuals, we estimated the cell-type proportions of the 6 individuals from bulk RNA-seq data and compared the decomposition results with the true cell-type proportions from the sn-RNA-seq data to verify our decomposition method. We employed a leave-one-out approach to decompose the cell-type proportions of each sample while using the sn-RNA-seq data of the other 5 samples as the reference. Figure 3B shows that the estimated adipose cell-type proportions have a high concordance with the true adipose cell-type proportions. Thus, our decomposition method provides reliable estimated adipose cell-type proportions then we used the 6 sn-RNA-seq samples as reference.

After verifying the accuracy of cell-type decomposition by MuSiC, we applied the method to the 335 subcutaneous adipose bulk RNA-seq samples from the METSIM cohort to estimate the proportions of the 8 adipose cell-types. We first checked the associations between the Matsuda index and the 8 estimated cell-type proportions using linear regression. In the association tests, we inverse normal transformed the Matsuda index and included age as a covariant. Four of 8 estimated cell-type proportions showed a significant association with the Matsuda index (Supplementary table 2). It is worth noting that both dendritic cells and macrophages exhibited a

strong association with the Matsuda index (p-values < 5×10^{-5}), suggesting that immune cell types contribute to insulin resistance in human adipose tissue.

BMI and adipose MT gene expression are independently associated with systemic insulin resistance after adjusting for tissue heterogeneity:

It has been shown previously that BMI associates with adipose MT activity[42]. In line with this we observed that MT gene expression is significantly associated with BMI in the adipose RNA-seq data from METSIM (n=324) (p-value = 2.94×10^{-10}). However, it is unknown if BMI or adipose tissue heterogeneity causes the association between the adipose MT expression and insulin resistance. To investigate this, we explored the associations between the Matsuda index and age, BMI, adipose MT gene expression, and adipose cell-type proportions using a multi-variable linear model (see model 1 in Methods). In our multi-variable linear model, BMI, MT expression, and the estimated proportions of dendritic and fibroblasts cells in adipose all showed significant associations with the Matsuda index ($p<0.05$) (Supplementary table 3). This result suggests that the adipose tissue heterogeneity, MT gene expression, and BMI all have independent contributions to the variance in the Matsuda index. Noteworthy, this model explained 44.39% of the variance (R^2) in the Matsuda index, which is higher than using any trait alone: BMI (R^2 = 30.89%); MT expression (R^2=14.64%); and estimated cell-type proportions (R^2=29.24%). Moreover, when we excluded BMI from model 1, the estimated adipose cell-type proportions and MT expression together explained a substantial amount (R^2=35.42%) of the variance in the Matsuda index. The high variance explained by model 1 makes it possible to predict the Matsuda index, i.e. systemic insulin resistance, using BMI, adipose MT gene expression and the estimated adipose cell-type proportions.

Utilizing adipose RNA-seq data to predict systemic insulin resistance:

Although the Matsuda index is an important biomarker for glucose metabolism, its measurement requires an oral glucose tolerance test, which is not available in many human metabolic cohorts. To this end, we developed a prediction model using elastic net regularization[43] that combines BMI, MT gene expression, age, and cell-type proportion information to predict the Matsuda index using adipose RNA-seq data. The prediction model was verified in three different cohorts: METSIM, GTEx, and FTC. In METSIM, we performed a 100-fold cross validation on the prediction model. Figure 4A shows that the predicted Matsuda index has a high concordance with the true Matsuda index ($r = 0.65$, p-value=7.22×10^{-40}). To further confirm this promising prediction of insulin resistance, we used the METSIM cohort to train a model and then estimated the Matsuda index in two independent adipose RNA-seq cohorts: GTEx and FTC (see Methods). Figure 4B shows that in the GTEx subcutaneous adipose samples, the T2D patients have significant lower predicted Matsuda index when compared to the non-diabetic GTEx individuals (p-value=2.58×10^{-5}). Since the monozygotic twin participants share the identical genetic background, we tested the association between the predicted Matsuda index and the true Matsuda index both in the full FTC cohort and the unrelated individuals by randomly selecting one individual from each twin pair. Figure 4C-D shows that the predicted Matsuda index is similarly well concordant with the true Matsuda index in the full FTC cohort ($r = 0.51$, p-value=1.21×10^{-7}) and in the unrelated FTC individuals ($r = 0.46$, p-value=1.20×10^{-3}).

Furthermore, the predicted Matsuda index had the best concordance with the true Matsuda index when compared to any of the tested traits alone in METSIM and FTC. Supplementary Figure 3 shows that neither MT gene expression nor BMI can predict the Matsuda index as accurately as our prediction model in the METSIM and FTC cohorts. Therefore, this prediction model can potentially be used to impute systemic insulin resistance into other adipose RNA-seq

cohorts, in which this key glucose metabolism trait has not been measured. Supplementary table 4 shows the betas in the trained prediction model that can be applied to other cohorts. In summary, we discovered that a substantial amount (44.39%) of the variance in the systemic insulin resistance, measured using the Matsuda index, can be explained by adipose MT gene expression, adipose cell-type proportions, and BMI. By combining the information from these traits, we were repeatedly able to predict the Matsuda index with a great accuracy when compared to the prediction results with any single trait alone. Since the Matsuda index is an important biomarker for glucose metabolism in humans, our prediction model can be utilized to impute the Matsuda index into adipose RNA-seq cohorts where this key metabolic trait is missing.

Discussion:

Even though the previous MR studies have shown that obesity has a causal effect on T2D[44,45], causal effects of obesity on prediabetes or insulin resistance among non-diabetic individuals are unknown. In the present study, we utilized the extensive UK Biobank cohort[19] and carefully phenotyped Finnish METSIM cohort[20] to show that obesity causes prediabetes and causally increases insulin resistance in the non-diabetic population using the MR analysis. Our MR result sheds new light on the long-standing reverse causality question between obesity and insulin resistance by establishing its directionality. Stancakova et al. have showed earlier that the Matsuda index is the best index of insulin sensitivity when compared to other surrogate indexes of insulin resistance using an M value from the euglycemic hyperinsulinemic clamp as the gold standard.[46] Therefore, the Matsuda index is largely a measure of systemic rather than adipose-based insulin resistance. However, when we examined the role of adipose cell-type heterogeneity, adipose MT gene expression, and BMI in systemic insulin resistance, we discovered that even when excluding BMI from the calculation, the estimated adipose cell-type proportions and adipose MT gene expression together still explain a substantial amount (R^2=35.42%) of the variance in the Matsuda index. When we included BMI into this analysis, all three factors are independently associated with insulin resistance ($p<0.05$ for each) and the R^2 increased to 44.39% which is higher than using any trait alone (R^2<=30.89%). This surprisingly high proportion of variance explained by adipose tissue (i.e. adipose cell types and MT gene expression) and BMI suggests that adipose tissue has an important role in the systemic insulin resistance. Based on this novel finding, we built a prediction model using adipose cell-types, adipose MT gene expression, and BMI that accurately predicted insulin resistance across multiple cohorts.

To investigate how adipose tissue heterogeneity affects systemic insulin resistance, we performed sn-RNA-seq using 6 frozen human subcutaneous adipose tissue samples. Noteworthy, the previous studies investigating adipose cell-type heterogeneity used FACS[24,25,26]; however this application is limited to a small number of well identified non-adipocyte cell-types and is unable to detect refined new subtypes. There are no previous publications performing sn-RNA-seq from frozen human adipose tissue, and thus the role of cell-type heterogeneity in insulin resistance has not been investigated for all main adipose cell types before. After careful quality control, the sn-RNA-seq generated 15,623 nuclei from the 6 adipose tissue biopsies and identified 8 adipose cell-type clusters based on their gene expression. We then used the sn-RNA-seq data as the reference data to detect cell-type specific signature genes in each adipose cell type cluster and decomposed the cell-type proportions in the METSIM, FTC, and GTEx adipose bulk RNA-seq cohorts, leveraging thus substantially the information contained in these existing bulk RNA-seq cohorts. Notably, the estimated cell-type proportions of macrophages and dendritic cells exhibited a significant association with insulin resistance, demonstrating the key role of the obesity-related low-grade inflammation process in systemic insulin resistance[9,10,12,13,14].

Even though insulin resistance is an essential clinical metabolic trait in obesity-related cardiometabolic diseases, it is often not measured in the existing adipose RNA-seq cohorts, such as the GTEx cohort[48]. Moreover, although the adipose tissue is suggested to be relevant in the development of insulin resistance[11,15,33], to the best of our knowledge, the variance in insulin resistance parameters that can be explained by adipose tissue has not been reported previously. Strikingly, we found that 44.39% of the variance in systemic insulin resistance (i.e. the Matsuda index) can be explained by the adipose cell-types, adipose MT expression, and BMI using the METSIM cohort. Thus, we developed an elastic net prediction model to predict the Matsuda

index using these traits. The prediction model was trained in a subset of the METSIM cohort. The model not only successfully predicted the Matsuda index in the METSIM test cohort but also predicted well the Matsuda index in 2 independent cohorts, the GTEx and FTC, indicating that we can predict the missing systemic insulin resistance estimates to cohorts lacking metabolic phenotype data, such as GTEx. Since the prediction model is based on adipose RNA-seq data, the predicted Matsuda index can potentially be used to proportionally estimate how much of insulin resistance is driven by adipose tissue versus other metabolic tissues. This would help subtype different forms of insulin resistance states underlying the development of type 2 diabetes.

In summary, we have shown that obesity has a significant causal effect on prediabetes and insulin resistance using the MR analysis. By leveraging bulk RNA-seq data in large adipose RNA-seq cohorts using a small amount of adipose sn-RNA-seq data to decompose adipose cell-types, we show that a substantial proportion (44.39%) of systemic insulin resistance can be explained by certain adipose cell-type proportions, MT gene expression, and BMI. This new finding not only establishes the key role of adipose tissue in regulating insulin resistance but also provides a useful method to impute insulin resistance estimates to human transcriptome cohorts.

Research Design and Methods:

Study cohorts:

We analyzed the genotype and phenotype data of ~510k individuals from two cohorts: METSIM cohort (n=10,198)[20], and UK Biobank cohort[19] (n=391,816) In the METSIM cohort, middle-aged Finnish males were recruited at the University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland, and the biochemical lipid, glucose, and other clinical and metabolic phenotypes were measured as described previously[20]. Briefly, a 2-h oral glucose tolerance test (OGTT) (75 g of glucose) was performed in the METSIM cohort, and samples for plasma glucose and insulin were drawn at 0, 30, and 120 min[20]. We evaluated insulin resistance in the non-related, non-diabetic METSIM participants using the Matsuda index that was calculated based on the OGTT values, as described in detail previously[47]. The METSIM study design was approved by local ethics committee and all participants gave a written informed consent. This research has been conducted using the UK Biobank Resource under Application Number 33934. The UK Biobank data was downloaded from the UK Biobank data repository on 08/23/2018.

We analyzed the RNA-seq data in 751 human subcutaneous adipose samples from 3 different cohorts: METSIM cohort (n=335)[20], Genotype-Tissue Expression (GTEx) cohort (n=308)[48,49], and 54 monozygotic Finnish monozygotic twin cohort (FTC) (n=108)[50,51,30]. The GTEx adipose RNA-seq data were downloaded from dbGaP (accession number phs000424.v6.p1) on 08/11/2016. In addition to subcutaneous adipose tissue, we also analyzed the GTEx visceral adipose tissue, blood, liver, and muscle RNA-seq data (v7). We used the FTC cohort to verify the prediction model of the Matsuda index. In this cohort, we generated RNA-seq data from subcutaneous adipose tissue of 54 MZ twin pairs (n=108). Supplementary Table 5 shows the

clinical characteristics of the participants in the 3 cohorts. We also selected the adipose biopsies of 6 individuals from FTC for the sn-RNA-seq experiment. Supplementary Table 1 shows the phenotypic characteristics of the 6 Finnish individuals whose adipose biopsies were processed for sn-RNA-seq. The 6 individuals have roughly similar ages, 3 of the 6 are males, and 3 of the 6 have a normal BMI (BMI<25).

GWAS and Mendelian randomization (MR) analysis:

To identify candidate instrumental variables (IVs) for the MR analysis in UKB, we first performed GWAS analyses of BMI and prediabetes in the UKB. The prediabetes cases were identified by serum HbA1c level between 5.7-6.4[37]. We excluded the individuals who had HbA1c > 6.4 or had been diagnosed as diabetic to ensure that only non-diabetic individuals were included in the GWAS analyses. We used BOLT-LMM[52] to explore the associations between the genotypes and the target phenotype, while accounting for the population stratification. We inverse normal transformed BMI to ensure a normal distribution and included age, age^2, sex, array type, center ID, and 20 genotype PCs as covariates. To fulfill the first assumption of MR (i.e. IVs should be significantly associated with the exposure variable), we selected the independent (R2<0.01) GWAS SNPs (p-value<5e-8) of BMI and prediabetes as candidate IVs for the MR analysis. Then we used MR-PRESSO[38] to identify causal associations between BMI and prediabetes while controlling for the potential pleiotropy.

When searching for a causal effect of BMI on insulin resistance, we performed GWAS analyses of BMI and Matsuda index using the non-diabetic individuals METSIM using BOLT-LMM[52] following the same pipeline as in UKB. For these GWAS analyses, we inverse normal transformed BMI and Matsuda index and included age, age^2, and 10 genotype PCs as the covariates. In the GWAS, we did not identify any Matsuda index-associated SNPs in METSIM

using a genome-wide significant cut point of p-value<5.0×10^{-8}. For BMI, we utilized the BMI GWAS results from the UK Biobank atlas of genetic associations[53], which also employed an LMM based regression method. To avoid pleiotropy in the MR analysis, we only selected the SNPs that are associated with BMI in the UK Biobank (p-value<$5.0x10^{-8}$) but not with the Matsuda index (p-value>0.05) in METSIM as the candidate IVs. To identify non-redundant SNPs, we LD pruned (R^2=0.0) the candidate SNPs, which resulted in 398 BMI-associated SNPs shared by the UK Biobank and METSIM cohort. We used these 398 BMI-associated SNPs as IVs and utilized MR-PRESSO to correct for potential pleiotropy and test for the causal effect between BMI and the Matsuda index in the direction obesity -> insulin resistance. The opposite causal direction, insulin resistance -> obesity, could not reliably be assessed using IVs due to the lack of genome-wide significant Matsuda index SNPs in the METSIM cohort (see Results for details).

Single nucleus RNA-sequencing and clustering:

Frozen subcutaneous adipose tissue was minced over dry ice and transferred into ice cold lysis buffer consisting of 0.1% NP-40, 10mM Tris-Hcl, 10 mM NaCl, and 3 mM MgCl2. After a 10-minute incubation period, the lysate was gently homogenized using a dounce and filtered through a 70 μm MACS smart strainer (Miltenyi Biotec #130-098-462) to remove debris. Nuclei were centrifuged at 500 g for 5 minutes at 4°C and re-suspended in wash buffer consisting of 1X PBS, 1.0% BSA, and 0.2 U/μl RNase inhibitor. We further filtered nuclei using a 40 μm Flowmi cell strainer (Sigma Aldrich # BAH136800040) and centrifuged at 500 g for 5 minutes at 4°C. Pelleted nuclei were re-suspended in wash buffer and immediately processed with the 10X Chromium platform following the Single Cell 3′ v2 protocol.

We used Cell Ranger[54] to build a pre-mRNA alignment reference based on the reference gencode 19 and estimate the UMIs in each cell. As the quality control, we excluded the cells that had <300 genes expressed and kept only the genes that were expressed in at least 3 cells. Then we used Seurat[55] to simultaneously cluster all the qualified cells from the 6 individuals. We identified 8 clusters and 697 signature genes (Supplementary Table 6) that have a higher expression in one of the clusters over the others.

Decomposition of adipose cell-type proportions from bulk RNA-seq data:

We first used Cell Ranger to re-align the single nucleus reads to a mature mRNA reference (gencode 19) and then estimated the pseudo-bulk gene expression in the 6 individuals. Next, treating the candidate gene expression of the sn-RNA-seq data as the reference, we used MuSiC[56] to estimate the cell-type proportions from the bulk RNA-seq data. To validate the accuracy of our decomposition method, we performed both sn-RNA-seq and bulk RNA-seq using the subcutaneous adipose biopsies from the same 6 individuals. Then we predicted the cell-type proportions using the bulk RNA-seq data and compared the decomposition results to the cell-type proportions estimated from the sn-RNA-seq data of the same individuals. To ensure the independence of the test data, we used the leave-one-out strategy. In more detail, when we estimated cell-type proportions of one individual, we used the sn-RNA-seq data from the other 5 individuals as the reference. Thus, all of the estimated cell-type proportions of each individual are based on an unrelated data set.

Because sn-RNA-seq captures not only mature mRNAs but also pre-mRNAs, the expression patterns of some genes are expected to be different between the nuclei and bulk RNA-seq data. For example, the MALAT1 gene (ENSG00000251562) exhibits an average TPM of 254 in the bulk adipose tissue in METSIM while its average TPM in the sn-RNA-seq data is 391,375.

Accordingly, we observed that the decomposition results were biased by these different expression patterns when we used all the ~16,000 expressed genes as suggested by MuSiC (Supplementary Figure 2B). To improve the accuracy of the decomposition, we calculated the difference in mean of the log-transformed gene expression across all genes from the target bulk RNA-seq samples and the 6 pseudo-bulk samples. Then we normalized the expression differences and kept the genes that have chi square statistic <=1. After this filtering process, we kept ~4,000 genes that have similar expression in both the single nucleus and bulk RNA-seq data. When estimating the cell-type proportions in the bulk RNA-seq data, we used the sn-RNA-seq data from all of the 6 samples as the reference. Since cell-type proportions are estimated from RNA-seq data that is affected by technical factors, we also adjusted the proportion of each cell-type for RNA-seq technical factors when testing for the association between cell-type proportions and traits in the METSIM cohort.

QC for estimating MT gene expression:

We used the same pipeline to estimate MT expression in all cohorts. First, we used FastQC[57] to verify the sequence quality of the RNA-seq data. Then, we performed a 2-pass alignment using STAR[58] (reference genome: gencode 19, hg19) and subsequently used featureCounts[59] to estimate the TPM of each gene. Only uniquely mapped reads were counted for gene expression. MT gene expression was defined as the sum of TPMs of all MT encoded genes. Since gene expression estimates from RNA-seq data are affected by multiple technical factors[60], we corrected MT gene expression for 11 known technical factors (Supplementary Table 7) and 3 genotype PCs. We chose to correct for 3 genotype PCs to follow a similar pipeline as implemented in the GTEx project[49]. Since the GTEx RNA-seq samples were collected from

deceased individuals, we also adjusted MT gene expression for the post-mortem sample collection time in the GTEx cohort.

However, it is worth noting that the MT genome is small and has a simple structure when compared to the autosomal chromosomes. Thus, RNA metrics estimated from MT reads have a different pattern compared to that of autosomal reads. Since MT reads comprise a relatively large proportion of total reads (Supplementary Table 7), we discovered that the technical factors estimated from the RNA-seq data, such as intergenic read percent and exonic read percent, are heavily correlated with the MT read percent of each sample. Supplementary Figure 4A-B shows that in the METSIM cohort, almost all the RNA metrics estimated by Picard Tools[61] show a significant association with the MT read percent, with the percent of intergenic reads exhibiting the strongest association with the MT read percent ($R=0.86$, p-value$=7.54 \times 10^{-103}$). Since the MT read percent reflects MT gene expression, these correlated RNA metrics cannot well represent the true technical covariation. Correcting for these factors when estimating MT gene expression would thus remove signals from MT gene expression. To address this issue, we first excluded the MT reads from the RNA-seq data and then estimated these technical factors from the reads aligned to the nuclear genome using Picard Tools. The new unbiased technical factors showed much weaker associations with the MT read percent (Supplementary Figure 4C-D).

Disentangling the associations between MT expression, BMI, insulin resistance (i.e. the Matsuda index), and tissue heterogeneity:

Since individuals with T2D are insulin resistant and the antidiabetic medication may influence the outcome, we removed them from all analyses involving the Matsuda index. We built a multi-variable linear model treating the Matsuda index as the dependent variable to

identify the associations between MT expression, BMI, estimated adipose cell-type proportions and Matsuda index in the METSIM cohort:

$$Matsuda \sim \beta_b * BMI + \sum \beta_{ci} * CT_i + \beta_M * MT \quad \text{(Model 1)}$$

MT indicates the corrected MT gene expression. Matsuda indicates the Matsuda index. CT_i is the estimated cell-type proportion in adipose tissue. The βs are the estimated parameters from the multi-variable linear models. Since the sum of the 8 estimated cell-type proportions equals to 1, the degree of freedom of the cell-type proportions is 7 instead of 8. We excluded the proportion of endothelial cells from the model due to its less accurate prediction when compared to the other cell-types.

To predict the Matsuda index using the other traits, we employed the following elastic net regularization[62] to predict the β for each variable in model 1:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|Matsuda - X\beta\| + \lambda\|\beta\| + \lambda\|\beta\|^2)$$

We used the 'glmnet' package[63] to obtain the λ that has the minimum mean cross-validated error in the training data set, and then used the specified λ and βs to predict the Matsuda index. To evaluate the prediction accuracy in both models, we performed a 100-fold verification in the METSIM cohort. In more detail, we randomly split the individuals into 100 groups, and then for each group, we predicted its value based on the model that we trained with the other 99 groups. We also verified this model in 2 independent cohorts: FTC and GTEx. For building the final prediction model for these 2 cohorts, we used all individuals in the METSIM cohort as the training set and predicted the Matsuda index in GTEx and FTC as verification. Using the predicted Matsuda index, we performed a Pearson correlation test to check the association between the estimated and true Matsuda index. Since GTEx do not have the Matsuda index

available, we compared the predicted Matsuda index values between the GTEx individuals with and without T2D as the verification.

Data and Resource Availability:

This research has been conducted using the UK Biobank Resource under Application Number 33934. The UK Biobank data is available from the UK Biobank data repository, but restrictions apply to the availability of these data, which were used under license for the current study and therefore are not publicly available. Data are however available from the authors upon reasonable request and with permission of UK Biobank. The GTEx dataset analyzed during the current study are available in the dbGAP repository, phs000424.v6.p1. The data on the METSIM cohort are available through METSIM data access committee (http://www.nationalbiobanks.fi/index.php/studies2/10-metsim). The FTC dataset and sn-RNA-seq data is available from the corresponding author upon reasonable request.

Acknowledgements:

We thank the individuals who participated in the METSIM, GTEx, FTC, and UK Biobank as well as Jaakko Kaprio and Aila Rissanen for the contributions to the FTC study. We also thank the UNGC sequencing core at UCLA for performing RNA sequencing. This study was funded by National Institutes of Health (NIH) grants HL-095056, HL-28481, and U01 DK105561. Z.M was supported by the AHA grant 19PRE34430112, M.A. was supported by the HHMI Gilliam grant, and A.K. by the NIH grant F31HL127921. E.H., E.R. and B.J. were supported by the National Science Foundation grant 1705197. K.H.P. was supported by the Academy of Finland (272376, 266286, 314383, 315035), Finnish Medical Foundation, Finnish Diabetes Research Foundation, Novo Nordisk Foundation, Gyllenberg Foundation, Sigrid Juselius Foundation, Helsinki University Hospital Research Funds, Government Research Funds and University of Helsinki. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the article. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx data used for the analyses described in this manuscript were obtained from the GTEx Portal on 08/14/2016 and dbGAP (accession number phs000424.v6.p1) on 08/11/2016. This research has been conducted using the UK Biobank Resource under application number 33934.

Author Contributions:

Z.M. and P.P. designed the study. Z.M., M.A, A.K., B.J., E.R., E.H., and P.P. performed methods development and statistical analysis. M.A. performed the single nuclei RNA sequencing experiments. Z.M., M.A., A.K. performed computational analysis of the data. M.L., K.M., and

P.P. produced the METSIM RNA-seq data. K.H.P. and P.P. produced the FTC RNA-seq data.

Z.M. and P.P. wrote the manuscript and all authors read, reviewed, and/or edited the manuscript.

Conflict of interest:

All authors declare no potential conflicts of interest relevant to this article.

Figures:

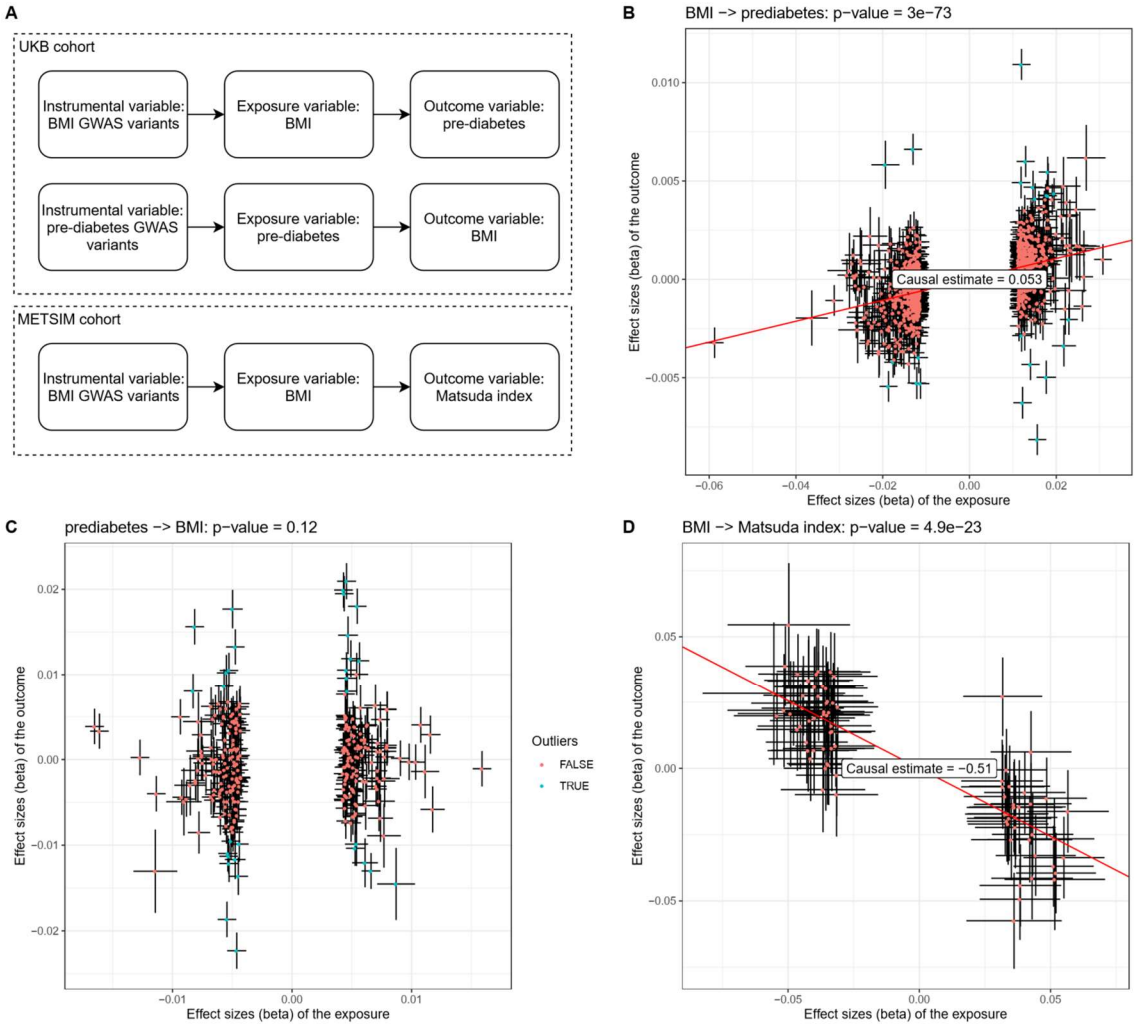


Figure III-1. MR analysis shows the causal relationship between BMI and Matsuda index, i.e. obesity leads to insulin resistance. (A) Variables used in the MR analysis. (B) The variant effect sizes on the exposure (BMI) are associated with the variant effect sizes on the outcome (i.e. the Matsuda index). The slope indicates the estimated causal effect of the exposure on the outcome.

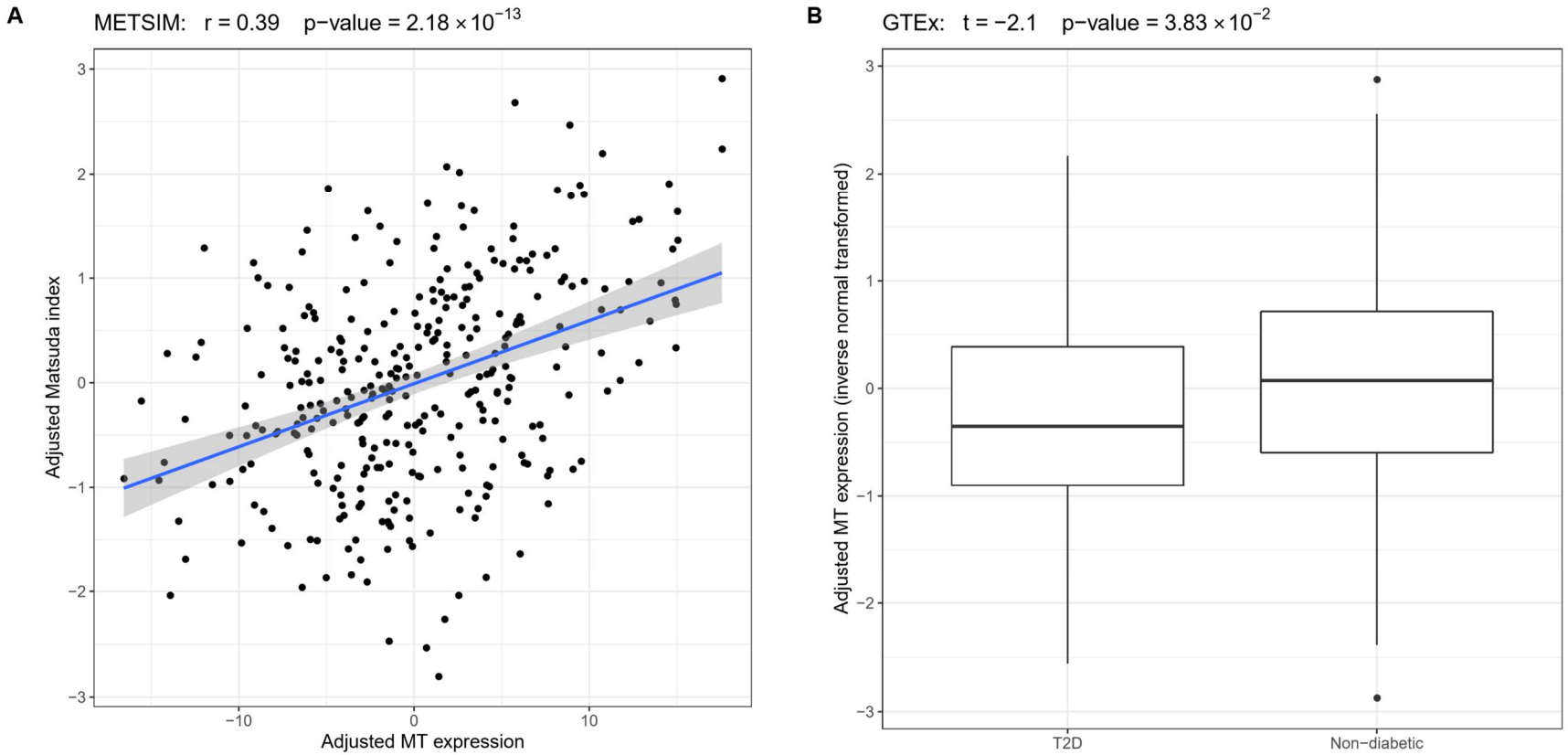


Figure III-2. A low adipose MT gene expression is associated with insulin resistance. (A) In METSIM, the adjusted adipose MT gene expression is significantly associated with the adjusted Matsuda index. (B) In GTEx, the non-diabetic individuals have significantly higher adjusted MT gene expression (inverse normal transformed) than the T2D patients.

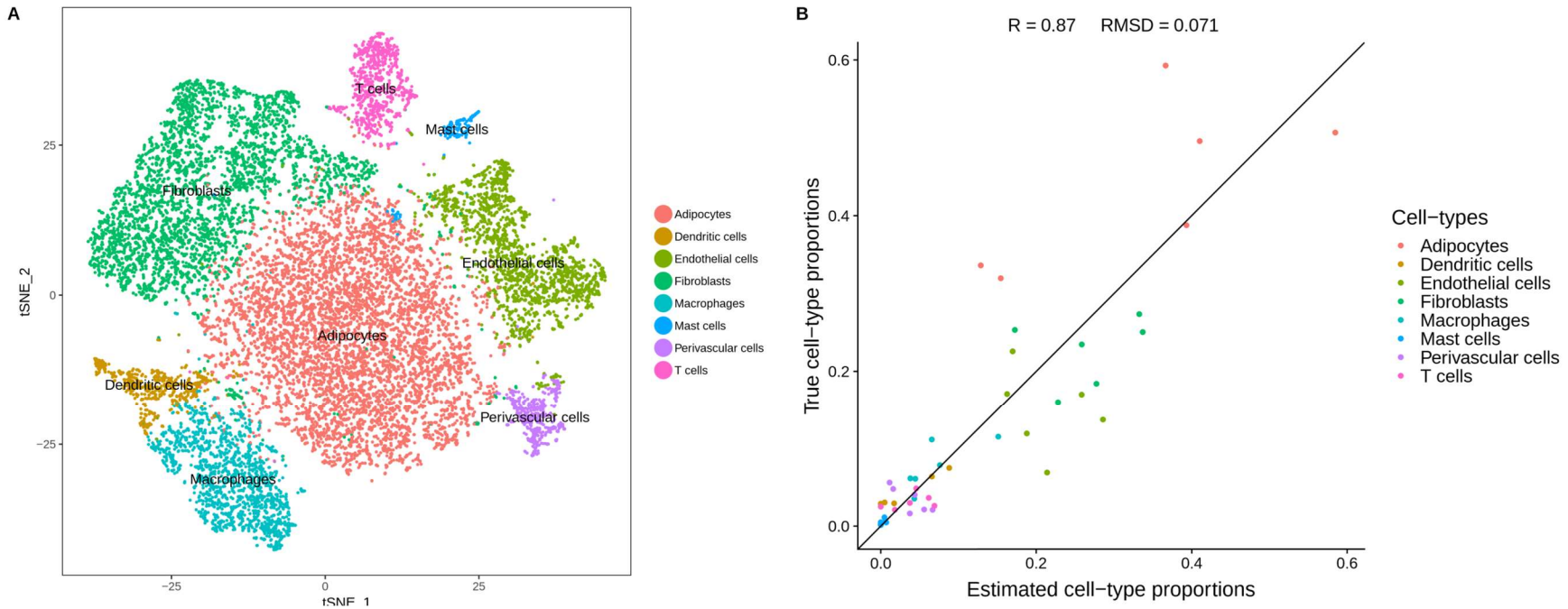


Figure III-3. Analysis of sn-RNA-seq data reveals tissue heterogeneity in human subcutaneous adipose tissue. (A) We identified 8 cell-type clusters in 15,623 nuclei from frozen human adipose tissue from 6 Finnish individuals. The t-SNP plot is colored by the identified cell types. (B) Using sn-RNA-seq as reference, the estimated adipose cell-type proportions from bulk adipose RNA-seq data are well concordant with the true cell-type proportions. (C) In METSIM, 4 of the estimated cell-type proportions showed significant associations with the Matsuda index. Asterisks indicate significant p-values after the Bonferroni correction. *: adjusted $p < 5\times10^{-2}$, **: adjusted $p < 5\times10^{-5}$, *: adjusted $p < 5\times10^{-8}$.

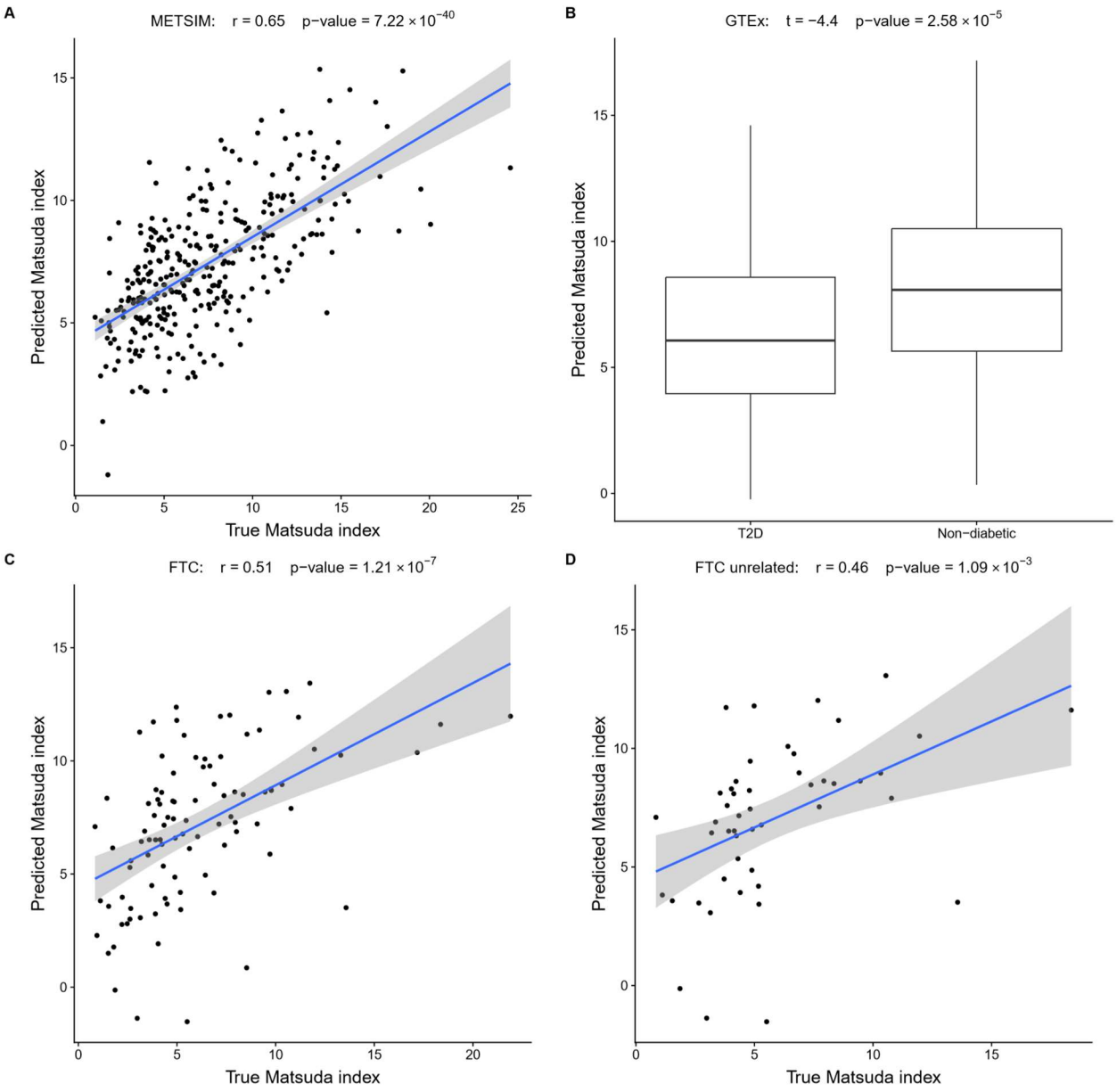


Figure III-4. The predicted Matsuda index is well concordant with the true Matsuda index values in 3 different cohorts: METSIM, GTEx, and FTC. (A) In METSIM, the estimated and true Matsuda index are significantly associated. (B) In GTEx, the predicted Matsuda index is significantly higher in non-diabetic individuals when compared to the T2D patients. (C) In FTC, the estimated and true Matsuda index are significantly associated. (D) We randomly choose one individual from each twin pair to select the unrelated individuals from FTC. Among the unrelated individuals, the estimated and true Matsuda index are also significantly associated, indicating that the twin status did not bias the prediction of the Matsuda index.

Supplementary tables:

Supplementary table III-1: Characteristics of SN-RNA-seq samples.

Sample ID	Age	Sex	BMI	Cell number	Number of genes/cell
1	59.2	1	25.9	2202	550
2	66.6	2	28.9	2809	641
3	58.64	1	22.1	1794	419
4	69.3	2	24.6	3405	430
5	68.2	1	31.4	3385	501
6	66.6	2	24.1	2028	527

Sex: male = 1, female = 2

Supplementary table III-2: The estimated adipose cell-type proportions are associated with the Matsuda index.

Cell type	R	P-value
Adipocytes	0.343	2.18E-10
Dentritic cells	-0.450	1.56E-17
Endothelial cells	0.026	NS
Fibroblasts	-0.246	7.28E-06
Macrophages	-0.288	1.33E-07
Mast cells	0.112	NS
Perivascular cells	0.087	NS
T cells	-0.009	NS

NS: Non-significant p-value

Supplementary table III-3: A multi-linear model shows the significant associations between the Matsuda index and the other traits.

Dependent variable:	Matsuda index		
	Beta	SE	P-value
BMI	-0.099	0.013	2.93E-13
MT	0.121	0.048	1.19E-02
Adipocytes	-0.769	1.032	NS
Dentritic cells	-6.474	1.964	1.09E-03
Fibroblasts	-2.925	1.186	1.42E-02
Macrophages	-2.523	1.466	NS
Mast cells	15.505	11.083	NS
Perivascular cells	2.667	1.805	NS
T cells	-1.037	1.385	NS
Age	-0.016	0.156	NS
Age2	0.000	0.001	NS

NS: Non-significant p-value

Supplementary table III-4: The prediction model of Matsuda index.

Trait	Beta
BMI	-0.3447276
age	-0.03886025
age2*	-0.00064973
MT **	0.5996329
GRS	0.02179454
Adipocytes	2.797682
Dendritic cells	-17.27797
Endothelial cells	3.042322
Fibroblasts	-7.809582
Macrophages	-3.779918
Mast cells	123.4385
Perivascular cells	12.85419
Intercept	20.82579

*: age square

**: MT expression is adjusted for technical factors and then normalized using Z score.

Supplementary table III-5: Phenotypes of the human adipose RNA-seq cohorts.

	GTEx	METSIM	FTC
Gender (female/male)	122 (39.61%)/186(60.39%)	0/335(100%)	64 (59.26%) / 44 (40.74%)
T2D	61 (19.81%)	11 (3.28%)	19 (17.59%)
Age	51.97 (12.52)	54.15 (4.93)	46.26 (17.80)
BMI	27.08 (4.20)	26.82 (3.69)	29.07 (5.96)

Supplementary table III-6: Signature genes identified in 8 adipose cell-types.

Gene	p-value	avg_logFC	pct.1	pct.2	p_val_adj	cluster
PDE3B	0	1.529582897	0.743	0.201	0	Adipocyte
SLC19A3	0	1.275624365	0.437	0.084	0	Adipocyte
GPAM	0	1.221092173	0.54	0.156	0	Adipocyte
SORBS1	0	1.220938965	0.624	0.179	0	Adipocyte
CLSTN2	0	1.212084332	0.382	0.089	0	Adipocyte
ITGA7	0	1.20413228	0.414	0.091	0	Adipocyte
TRHDE-AS1	0	1.193259819	0.469	0.103	0	Adipocyte
GYG2	0	1.187016967	0.39	0.075	0	Adipocyte
HOOK2	0	1.183383533	0.384	0.075	0	Adipocyte
ACACB	0	1.182436557	0.691	0.253	0	Adipocyte
WDPCP	0	1.172724797	0.552	0.201	0	Adipocyte
RP11-236F9.2	0	1.167185028	0.564	0.171	0	Adipocyte
AC002066.1	0	1.130780201	0.314	0.055	0	Adipocyte
AC004538.3	0	1.124465835	0.468	0.119	0	Adipocyte
DMD	0	1.111543233	0.697	0.258	0	Adipocyte
RP11-125B21.2	0	1.087217657	0.266	0.043	0	Adipocyte
TENM3	0	1.086309837	0.349	0.077	0	Adipocyte
ERBB4	0	1.080648559	0.352	0.088	0	Adipocyte
MLXIPL	0	1.064005503	0.397	0.096	0	Adipocyte
TNS1	0	1.057270537	0.418	0.107	0	Adipocyte
GHR	0	1.046183298	0.684	0.262	0	Adipocyte
DIRC3	0	1.042296757	0.365	0.086	0	Adipocyte
EGFEM1P	0	1.041280584	0.287	0.052	0	Adipocyte
BCL2	0	1.034813546	0.54	0.16	0	Adipocyte
CACNA2D1	0	1.001483496	0.452	0.158	0	Adipocyte
PTPRF	0	1.000247299	0.312	0.067	0	Adipocyte
NTM	0	0.991678401	0.436	0.119	0	Adipocyte
AQP7	0	0.983107303	0.373	0.091	0	Adipocyte
SIK2	0	0.978307874	0.523	0.185	0	Adipocyte
SLC7A6	0	0.974049837	0.346	0.079	0	Adipocyte
ELMOD3	0	0.970048775	0.345	0.084	0	Adipocyte
CTIF	0	0.966901484	0.327	0.073	0	Adipocyte
PPP2R1B	0	0.966449689	0.307	0.069	0	Adipocyte
PNPLA2	0	0.961403137	0.462	0.149	0	Adipocyte
ITIH5	0	0.954169563	0.492	0.169	0	Adipocyte
LIPE-AS1	0	0.953746905	0.479	0.153	0	Adipocyte
TEAD1	0	0.951418561	0.472	0.144	0	Adipocyte
ACSS2	0	0.948104028	0.313	0.07	0	Adipocyte
PTGER3	0	0.936144988	0.318	0.077	0	Adipocyte
PCDH9	0	0.932196098	0.625	0.239	0	Adipocyte
PLIN1	0	0.921351944	0.683	0.267	0	Adipocyte

NEAT1	0	0.884069756	0.914	0.664	0	Adipocyte
MAST4	0	0.881247303	0.429	0.134	0	Adipocyte
MGEA5	0	0.870253752	0.539	0.209	0	Adipocyte
PPP1R12B	0	0.852387408	0.432	0.143	0	Adipocyte
PLIN4	0	0.847154152	0.58	0.23	0	Adipocyte
RP11-444D3.1	0	0.834094982	0.498	0.178	0	Adipocyte
COBLL1	0	0.828266098	0.427	0.151	0	Adipocyte
TLN2	0	0.785297097	0.473	0.183	0	Adipocyte
FRMD4A	0	0.778822842	0.482	0.178	0	Adipocyte
EBF1	0	0.746505631	0.82	0.481	0	Adipocyte
MALAT1	0	0.599117598	1	1	0	Adipocyte
ACSS3	6.82E-305	0.8895013	0.313	0.079	1.56E-300	Adipocyte
LIMA1	1.93E-302	0.700130192	0.507	0.207	4.40E-298	Adipocyte
PFKFB3	5.13E-294	0.841045933	0.384	0.125	1.17E-289	Adipocyte
TMEM132C	2.54E-293	0.867131236	0.345	0.102	5.81E-289	Adipocyte
SOX5	1.02E-290	0.774854403	0.396	0.134	2.33E-286	Adipocyte
LAMA4	1.06E-290	0.717849464	0.526	0.232	2.43E-286	Adipocyte
ASPH	2.09E-275	0.710076297	0.483	0.203	4.76E-271	Adipocyte
PBX1	1.34E-273	0.699168405	0.461	0.185	3.05E-269	Adipocyte
PRKAR2B	3.04E-272	0.830445969	0.292	0.076	6.94E-268	Adipocyte
ADIPOQ	1.07E-271	0.730080736	0.406	0.145	2.45E-267	Adipocyte
LIMCH1	1.06E-267	0.805433759	0.381	0.133	2.41E-263	Adipocyte
LIPE	8.37E-264	0.66594134	0.454	0.182	1.91E-259	Adipocyte
SOS1	2.74E-255	0.70964169	0.432	0.174	6.26E-251	Adipocyte
GABRE	8.20E-255	0.854788939	0.253	0.059	1.87E-250	Adipocyte
RNF150	1.86E-251	0.765114236	0.302	0.087	4.25E-247	Adipocyte
MDFIC	2.85E-250	0.693947305	0.403	0.154	6.51E-246	Adipocyte
CLMP	2.16E-239	0.767736858	0.286	0.082	4.93E-235	Adipocyte
CIDEC	7.65E-239	0.707713136	0.33	0.108	1.75E-234	Adipocyte
PDZD2	2.81E-237	0.683734404	0.408	0.164	6.42E-233	Adipocyte
MGST1	3.06E-237	0.559994622	0.576	0.288	6.99E-233	Adipocyte
UGP2	1.08E-235	0.718832185	0.398	0.157	2.46E-231	Adipocyte
RP11-399D6.2	1.06E-229	0.765502189	0.304	0.097	2.42E-225	Adipocyte
PDZRN3	3.57E-226	0.750382729	0.327	0.113	8.15E-222	Adipocyte
PLA2G16	1.10E-223	0.626236095	0.357	0.131	2.52E-219	Adipocyte
GPD1	9.11E-219	0.679442253	0.322	0.11	2.08E-214	Adipocyte
ECHDC2	4.38E-218	0.674421411	0.309	0.102	1.00E-213	Adipocyte
PHLDB2	1.00E-217	0.746665652	0.279	0.085	2.28E-213	Adipocyte
G0S2	1.40E-215	0.604836131	0.618	0.356	3.19E-211	Adipocyte
AGPAT2	2.04E-208	0.648249842	0.354	0.135	4.67E-204	Adipocyte
GBE1	4.49E-208	0.583134352	0.426	0.187	1.03E-203	Adipocyte
KANK1	1.02E-206	0.657369948	0.269	0.082	2.32E-202	Adipocyte
ANO6	1.16E-201	0.655017777	0.334	0.126	2.64E-197	Adipocyte

PLXNA4	1.27E-201	0.694699006	0.278	0.09	2.90E-197	Adipocyte
ADAMTS12	5.74E-200	0.669814282	0.275	0.088	1.31E-195	Adipocyte
NRIP1	3.71E-198	0.644744228	0.302	0.105	8.47E-194	Adipocyte
AFF3	8.67E-196	0.58343795	0.33	0.124	1.98E-191	Adipocyte
EFNA5	6.09E-193	0.630015641	0.278	0.092	1.39E-188	Adipocyte
VKORC1L1	7.34E-192	0.65960032	0.251	0.077	1.68E-187	Adipocyte
EHBP1	1.10E-188	0.539337976	0.442	0.207	2.50E-184	Adipocyte
GPC6	1.42E-188	0.66784395	0.265	0.086	3.24E-184	Adipocyte
CRIM1	4.78E-187	0.649589011	0.346	0.141	1.09E-182	Adipocyte
FIGN	1.00E-185	0.654404124	0.254	0.08	2.29E-181	Adipocyte
SLTM	2.57E-180	0.57674137	0.311	0.118	5.86E-176	Adipocyte
YAP1	7.43E-180	0.564158657	0.297	0.108	1.70E-175	Adipocyte
MAPK10	6.23E-179	0.633979846	0.276	0.096	1.42E-174	Adipocyte
RBPMS	1.82E-178	0.527427065	0.442	0.211	4.15E-174	Adipocyte
ITSN1	1.14E-175	0.448650635	0.436	0.203	2.60E-171	Adipocyte
ADRBK2	6.51E-174	0.605180546	0.298	0.113	1.49E-169	Adipocyte
APBB1IP	1.63E-173	0.560952279	0.272	0.095	3.72E-169	Adipocyte
RTN3	3.35E-172	0.575237299	0.285	0.104	7.65E-168	Adipocyte
PRICKLE2	5.08E-171	0.578309439	0.261	0.089	1.16E-166	Adipocyte
UBE2E2	5.67E-169	0.460472853	0.422	0.2	1.29E-164	Adipocyte
ACSL1	2.55E-168	0.621609981	0.251	0.085	5.82E-164	Adipocyte
PHLDB1	1.17E-167	0.642194891	0.282	0.106	2.67E-163	Adipocyte
FERMT2	4.31E-167	0.5991449	0.255	0.088	9.83E-163	Adipocyte
FASN	1.26E-163	0.724726641	0.264	0.096	2.89E-159	Adipocyte
PALMD	4.74E-163	0.520804883	0.318	0.13	1.08E-158	Adipocyte
PRKAG2	1.03E-162	0.573353591	0.278	0.104	2.35E-158	Adipocyte
ADH1B	4.28E-160	0.495467473	0.453	0.23	9.78E-156	Adipocyte
COL4A2	1.23E-157	0.499563669	0.412	0.2	2.81E-153	Adipocyte
TNS3	4.67E-155	0.495152437	0.256	0.091	1.07E-150	Adipocyte
MAGI2	1.89E-148	0.46487481	0.329	0.143	4.31E-144	Adipocyte
FOXO1	2.18E-146	0.485131825	0.37	0.175	4.97E-142	Adipocyte
PPARG	5.29E-145	0.427357589	0.457	0.24	1.21E-140	Adipocyte
PTPRS	7.82E-144	0.475696744	0.35	0.16	1.78E-139	Adipocyte
FAM13A	8.82E-143	0.493100009	0.385	0.189	2.01E-138	Adipocyte
RHOBTB3	1.64E-142	0.48037428	0.305	0.129	3.73E-138	Adipocyte
ZBTB16	1.93E-142	0.554251885	0.329	0.15	4.41E-138	Adipocyte
FHIT	7.22E-140	0.416769596	0.333	0.148	1.65E-135	Adipocyte
RP11-736K20.5	4.48E-138	0.518624079	0.257	0.101	1.02E-133	Adipocyte
EMP1	1.32E-132	0.487443766	0.338	0.16	3.00E-128	Adipocyte
LPL	1.41E-130	0.495012431	0.326	0.152	3.22E-126	Adipocyte
IMMP2L	4.19E-128	0.395064074	0.463	0.255	9.57E-124	Adipocyte
CAT	1.04E-125	0.438714881	0.315	0.145	2.37E-121	Adipocyte
DLG1	4.26E-125	0.431403633	0.27	0.114	9.72E-121	Adipocyte

MAGI1	2.10E-124	0.330919746	0.384	0.193	4.78E-120	Adipocyte
PCED1B	4.46E-123	0.413431449	0.26	0.108	1.02E-118	Adipocyte
ABCA1	6.79E-122	0.478028212	0.27	0.117	1.55E-117	Adipocyte
APBB2	9.18E-122	0.39330311	0.409	0.216	2.10E-117	Adipocyte
SH3D19	3.95E-121	0.398860987	0.426	0.23	9.02E-117	Adipocyte
THRB	1.65E-119	0.39858423	0.313	0.147	3.77E-115	Adipocyte
TMEM135	2.38E-119	0.411121832	0.293	0.133	5.42E-115	Adipocyte
CSMD1	6.76E-119	0.598067988	0.272	0.123	1.54E-114	Adipocyte
RP11-1000B6.3	1.02E-117	0.39898297	0.268	0.116	2.34E-113	Adipocyte
TCF7L2	3.68E-117	0.397299817	0.36	0.182	8.41E-113	Adipocyte
USP33	9.30E-112	0.384847586	0.272	0.121	2.12E-107	Adipocyte
SEMA3A	8.50E-109	0.333645265	0.285	0.13	1.94E-104	Adipocyte
GOLGA4	9.76E-105	0.366520915	0.313	0.154	2.23E-100	Adipocyte
CPM	3.06E-104	0.321261419	0.34	0.174	6.98E-100	Adipocyte
LENG8	7.18E-103	0.467180816	0.264	0.123	1.64E-98	Adipocyte
11-Sep	1.35E-102	0.377903043	0.268	0.125	3.07E-98	Adipocyte
TNRC6A	2.50E-102	0.341203705	0.357	0.187	5.70E-98	Adipocyte
UVRAG	1.77E-99	0.360030075	0.34	0.179	4.05E-95	Adipocyte
INSR	3.99E-98	0.397673064	0.255	0.118	9.12E-94	Adipocyte
PDE4DIP	7.01E-93	0.311985061	0.276	0.134	1.60E-88	Adipocyte
ADK	9.30E-93	0.328729375	0.272	0.132	2.12E-88	Adipocyte
UBR3	1.40E-90	0.325001589	0.268	0.13	3.19E-86	Adipocyte
FNDC3B	2.91E-89	0.269423843	0.435	0.253	6.64E-85	Adipocyte
FOXP2	7.73E-89	0.302324152	0.258	0.123	1.76E-84	Adipocyte
SESTD1	7.15E-81	0.35062778	0.269	0.139	1.63E-76	Adipocyte
BNC2	1.55E-80	0.259599498	0.272	0.138	3.55E-76	Adipocyte
PARD3	6.82E-80	0.26522838	0.296	0.156	1.56E-75	Adipocyte
ACYP2	4.93E-74	0.250900362	0.267	0.139	1.13E-69	Adipocyte
COL4A1	8.53E-74	0.308712734	0.269	0.143	1.95E-69	Adipocyte
ALCAM	0	2.269577233	0.384	0.021	0	Dendritic Cells
CCDC88A	4.18E-192	1.566360428	0.393	0.063	9.54E-188	Dendritic Cells
RBM47	2.12E-187	1.422987583	0.38	0.059	4.84E-183	Dendritic Cells
CHST11	1.88E-180	1.435111951	0.303	0.038	4.28E-176	Dendritic Cells
PTPRC	3.04E-160	1.251057242	0.328	0.049	6.94E-156	Dendritic Cells
MYO1F	3.82E-159	1.291099409	0.301	0.042	8.71E-155	Dendritic Cells
MSR1	6.73E-148	1.358588414	0.332	0.055	1.54E-143	Dendritic Cells
MYO9B	9.82E-128	1.295386649	0.337	0.065	2.24E-123	Dendritic Cells
SLC8A1	5.03E-127	1.296624098	0.379	0.082	1.15E-122	Dendritic Cells
SAMHD1	1.14E-120	1.252426819	0.313	0.059	2.60E-116	Dendritic Cells
GAS7	4.91E-98	1.178945332	0.342	0.082	1.12E-93	Dendritic Cells
MITF	7.64E-80	1.226954203	0.38	0.119	1.74E-75	Dendritic Cells
IQGAP2	1.72E-79	0.876720118	0.286	0.066	3.93E-75	Dendritic Cells
ETV6	5.47E-79	1.13789594	0.293	0.073	1.25E-74	Dendritic Cells

ATG7	1.40E-74	1.045184476	0.4	0.134	3.19E-70	Dendritic Cells
HDAC9	2.27E-67	1.082466092	0.362	0.119	5.17E-63	Dendritic Cells
FRMD4B	7.85E-60	0.688343015	0.31	0.091	1.79E-55	Dendritic Cells
HLA-DRA	5.21E-59	1.375880128	0.297	0.094	1.19E-54	Dendritic Cells
TPRG1	1.83E-57	1.499100499	0.33	0.12	4.19E-53	Dendritic Cells
CTSB	1.13E-55	1.12632528	0.29	0.091	2.59E-51	Dendritic Cells
ZEB21	1.41E-55	0.92542531	0.505	0.241	3.22E-51	Dendritic Cells
ANKRD44	3.00E-55	0.930838217	0.268	0.079	6.84E-51	Dendritic Cells
CD74	1.12E-45	1.075264014	0.368	0.151	2.56E-41	Dendritic Cells
DPYD1	2.53E-43	0.749994781	0.464	0.224	5.77E-39	Dendritic Cells
RNF130	1.22E-37	0.839074289	0.313	0.13	2.78E-33	Dendritic Cells
ARL15	4.39E-37	0.664221584	0.31	0.123	1.00E-32	Dendritic Cells
FTL1	7.59E-36	1.083337827	0.75	0.577	1.73E-31	Dendritic Cells
ASAP1	5.04E-34	0.814141175	0.346	0.16	1.15E-29	Dendritic Cells
FAM49B	9.67E-34	0.813853798	0.324	0.145	2.21E-29	Dendritic Cells
FTLP3	1.60E-31	1.113377369	0.288	0.125	3.64E-27	Dendritic Cells
DENND4C	4.76E-31	0.835881348	0.306	0.14	1.09E-26	Dendritic Cells
ELMO1	3.91E-30	0.632387731	0.255	0.101	8.92E-26	Dendritic Cells
DOCK8	2.27E-27	0.761421578	0.272	0.121	5.18E-23	Dendritic Cells
DLEU2	1.02E-25	0.55732156	0.272	0.12	2.33E-21	Dendritic Cells
RASAL2	1.46E-25	0.660847221	0.29	0.136	3.32E-21	Dendritic Cells
SRGAP2B	3.86E-25	0.529061742	0.259	0.112	8.80E-21	Dendritic Cells
NUMB	4.62E-25	0.814550877	0.303	0.151	1.05E-20	Dendritic Cells
ARHGAP26	1.32E-24	0.673312986	0.284	0.135	3.02E-20	Dendritic Cells
PABPC1	2.80E-24	0.668089107	0.321	0.163	6.40E-20	Dendritic Cells
SAT1	1.37E-23	0.661291715	0.337	0.177	3.13E-19	Dendritic Cells
TAOK3	4.25E-23	0.675103592	0.268	0.128	9.71E-19	Dendritic Cells
PSAP1	1.19E-21	0.645194694	0.395	0.228	2.71E-17	Dendritic Cells
FTH11	2.85E-21	0.759722787	0.611	0.44	6.50E-17	Dendritic Cells
PLXDC21	2.23E-20	0.472383671	0.366	0.203	5.09E-16	Dendritic Cells
ZSWIM6	6.26E-20	0.512110903	0.254	0.122	1.43E-15	Dendritic Cells
SPIDR	4.15E-19	0.589775863	0.295	0.158	9.48E-15	Dendritic Cells
TMSB4X1	2.86E-17	0.432853548	0.707	0.55	6.53E-13	Dendritic Cells
ACTB1	8.81E-17	0.587878024	0.467	0.31	2.01E-12	Dendritic Cells
C10orf11	3.21E-14	0.431920182	0.426	0.285	7.33E-10	Dendritic Cells
TRPS1	4.35E-14	0.484295363	0.304	0.181	9.93E-10	Dendritic Cells
GNAQ	9.91E-14	0.457929594	0.266	0.151	2.26E-09	Dendritic Cells
SNX29P21	7.13E-11	0.304528894	0.611	0.484	1.63E-06	Dendritic Cells
CELF1	2.78E-10	0.416719774	0.281	0.178	6.34E-06	Dendritic Cells
TMSB4XP81	7.02E-10	0.365118626	0.317	0.207	1.60E-05	Dendritic Cells
TMSB101	7.81E-10	0.287806907	0.533	0.397	1.78E-05	Dendritic Cells
AKAP13	1.27E-09	0.390400788	0.266	0.168	2.91E-05	Dendritic Cells
CST31	2.88E-09	0.527134116	0.366	0.258	6.58E-05	Dendritic Cells

ARHGAP10	9.84E-09	0.434079077	0.252	0.163	0.000224644	Dendritic Cells
ADRBK21	3.11E-08	0.419664573	0.272	0.185	0.00071002	Dendritic Cells
RPS9	3.92E-08	0.289295543	0.301	0.202	0.000894327	Dendritic Cells
EPB41L2	4.84E-08	0.504208372	0.29	0.208	0.001104726	Dendritic Cells
APOE	1.03E-07	0.427988724	0.277	0.187	0.002354319	Dendritic Cells
PAN3	1.72E-06	0.323918227	0.255	0.179	0.039258458	Dendritic Cells
ZFAND3	1.08E-05	0.319000794	0.299	0.228	0.245850647	Dendritic Cells
RPS191	1.41E-05	0.262949636	0.284	0.207	0.320917464	Dendritic Cells
MEF2A	5.01E-05	0.253458797	0.259	0.192	1	Dendritic Cells
UBE2E21	6.19E-05	0.273657079	0.362	0.288	1	Dendritic Cells
PICALM	0.000227374	0.276821528	0.272	0.211	1	Dendritic Cells
SSH2	0.000267885	0.257596572	0.27	0.21	1	Dendritic Cells
KIAA1217	0	2.195558766	0.611	0.101	0	Endothelial
BTNL9	0	2.104606529	0.512	0.041	0	Endothelial
CLDN5	0	2.059441284	0.455	0.046	0	Endothelial
ST6GALNAC3	0	2.043887541	0.387	0.02	0	Endothelial
CADM2	0	2.020472526	0.427	0.059	0	Endothelial
MECOM	0	1.988349793	0.365	0.022	0	Endothelial
LDB2	0	1.795059088	0.557	0.095	0	Endothelial
ABLIM3	0	1.788427328	0.364	0.034	0	Endothelial
VWF	0	1.766945412	0.463	0.064	0	Endothelial
RBP7	0	1.72670259	0.34	0.047	0	Endothelial
CDH13	0	1.669612314	0.337	0.041	0	Endothelial
IFI27	0	1.639419307	0.327	0.049	0	Endothelial
EGFL7	0	1.636854216	0.339	0.028	0	Endothelial
SPARCL1	0	1.635045878	0.488	0.076	0	Endothelial
ELTD1	0	1.628297942	0.309	0.02	0	Endothelial
PTPRB	0	1.617432287	0.252	0.011	0	Endothelial
GPR116	0	1.539778947	0.294	0.023	0	Endothelial
A2M	0	1.535889272	0.449	0.069	0	Endothelial
KALRN	0	1.506534418	0.366	0.073	0	Endothelial
ID1	0	1.438405292	0.257	0.027	0	Endothelial
EMCN	0	1.412001765	0.26	0.024	0	Endothelial
BST2	0	1.362453807	0.295	0.044	0	Endothelial
HLA-C	0	1.356590743	0.364	0.074	0	Endothelial
HLA-E	0	1.321191229	0.351	0.07	0	Endothelial
HLA-B	4.82E-297	1.297668467	0.428	0.11	1.10E-292	Endothelial
GNG11	1.62E-286	1.369689832	0.36	0.081	3.70E-282	Endothelial
FABP41	8.14E-276	1.240802421	0.888	0.733	1.86E-271	Endothelial
EPAS1	4.17E-272	1.222019045	0.323	0.067	9.53E-268	Endothelial
HLA-A	1.05E-262	1.239485944	0.41	0.112	2.39E-258	Endothelial
CD36	1.85E-244	0.835887429	0.861	0.626	4.23E-240	Endothelial
CCSER1	4.52E-237	1.136723247	0.252	0.045	1.03E-232	Endothelial

B2M1	6.81E-237	1.243095709	0.71	0.397	1.55E-232	Endothelial
IGFBP7	2.75E-220	1.138087363	0.382	0.111	6.27E-216	Endothelial
ENG	3.13E-219	1.095417481	0.257	0.051	7.13E-215	Endothelial
CD59	9.51E-217	1.150009025	0.264	0.055	2.17E-212	Endothelial
MEF2C	4.18E-212	0.97502785	0.403	0.124	9.54E-208	Endothelial
ABLIM1	1.20E-202	1.165155297	0.361	0.109	2.75E-198	Endothelial
H19	8.03E-201	1.076839693	0.257	0.055	1.83E-196	Endothelial
TCF41	1.96E-195	0.93894737	0.475	0.178	4.48E-191	Endothelial
IFITM3	2.11E-192	1.138816406	0.327	0.093	4.81E-188	Endothelial
CD741	2.90E-189	0.980812067	0.391	0.127	6.63E-185	Endothelial
TMTC1	5.34E-179	1.240782477	0.312	0.093	1.22E-174	Endothelial
TMSB4X2	8.23E-171	0.849203033	0.769	0.527	1.88E-166	Endothelial
FABP5	5.72E-167	1.030336996	0.296	0.085	1.31E-162	Endothelial
TMSB102	4.64E-157	0.843104495	0.645	0.369	1.06E-152	Endothelial
MAGI12	7.64E-157	1.040711809	0.498	0.24	1.74E-152	Endothelial
RALGAPA2	9.45E-155	1.143792913	0.306	0.099	2.16E-150	Endothelial
DOCK4	9.54E-142	0.934468055	0.286	0.089	2.18E-137	Endothelial
PTMA	1.85E-132	0.856328159	0.351	0.132	4.23E-128	Endothelial
TMSB4XP82	2.34E-132	0.867650415	0.424	0.183	5.33E-128	Endothelial
SYNE2	1.57E-131	0.906357617	0.292	0.098	3.59E-127	Endothelial
TXNIP1	2.03E-127	0.81593346	0.532	0.281	4.63E-123	Endothelial
PITPNC1	7.32E-124	0.883607566	0.256	0.08	1.67E-119	Endothelial
PTRF	4.59E-122	0.771052485	0.337	0.129	1.05E-117	Endothelial
CALM1	1.27E-120	0.841287684	0.259	0.083	2.91E-116	Endothelial
MGLL	9.25E-119	0.874077124	0.326	0.128	2.11E-114	Endothelial
HSPB1	2.82E-115	0.887413677	0.266	0.091	6.44E-111	Endothelial
TACC1	1.11E-112	0.862277577	0.4	0.186	2.53E-108	Endothelial
RPL3P41	7.11E-112	0.717558041	0.406	0.181	1.62E-107	Endothelial
ARHGAP29	4.17E-110	0.937931951	0.292	0.113	9.51E-106	Endothelial
RASAL21	1.66E-97	0.863839502	0.296	0.121	3.80E-93	Endothelial
SASH1	2.72E-97	0.721297401	0.302	0.121	6.20E-93	Endothelial
AC016739.21	3.35E-93	0.564810811	0.468	0.239	7.65E-89	Endothelial
TIMP31	7.79E-92	0.611389944	0.537	0.32	1.78E-87	Endothelial
RPLP11	2.46E-88	0.537992073	0.508	0.28	5.62E-84	Endothelial
RP11-742N3.11	2.20E-87	0.606911973	0.457	0.243	5.01E-83	Endothelial
MYL61	2.84E-87	0.617739987	0.541	0.329	6.48E-83	Endothelial
RPS7	2.61E-85	0.71965108	0.256	0.099	5.97E-81	Endothelial
CTD-2192J16.151	1.96E-84	0.611896972	0.381	0.184	4.47E-80	Endothelial
H3F3B	2.16E-84	0.666853469	0.312	0.137	4.92E-80	Endothelial
PLCB1	3.56E-84	0.834345665	0.264	0.108	8.12E-80	Endothelial
RPL111	3.30E-83	0.568372126	0.408	0.205	7.53E-79	Endothelial
RPL8	8.24E-81	0.584864087	0.355	0.168	1.88E-76	Endothelial

RP11-234A1.11	3.91E-80	0.55302004	0.467	0.257	8.92E-76	Endothelial
ZFP36L1	1.29E-79	0.629009807	0.277	0.116	2.93E-75	Endothelial
RP13-258O15.1	1.39E-78	0.580306843	0.261	0.104	3.17E-74	Endothelial
PTPRM1	2.03E-78	0.756544061	0.465	0.284	4.63E-74	Endothelial
RPL34P181	5.71E-78	0.581787515	0.408	0.211	1.30E-73	Endothelial
FAU	6.24E-78	0.589804756	0.282	0.119	1.42E-73	Endothelial
AC022431.1	3.08E-75	0.561638538	0.278	0.118	7.02E-71	Endothelial
RPS181	3.55E-74	0.550643546	0.382	0.195	8.10E-70	Endothelial
AKT3	2.60E-72	0.654261665	0.289	0.129	5.93E-68	Endothelial
RPL15	3.77E-70	0.601003736	0.299	0.138	8.60E-66	Endothelial
RPL321	1.48E-69	0.522711203	0.436	0.244	3.39E-65	Endothelial
RP11-122C9.1	5.10E-69	0.542181052	0.325	0.156	1.16E-64	Endothelial
RPL341	5.53E-69	0.527811638	0.483	0.286	1.26E-64	Endothelial
RPL10	2.23E-68	0.568039658	0.349	0.176	5.09E-64	Endothelial
AC004453.81	3.00E-67	0.526103774	0.401	0.217	6.86E-63	Endothelial
RP11-367G18.2	5.46E-67	0.548365225	0.267	0.117	1.25E-62	Endothelial
RPS21	9.68E-67	0.515893195	0.357	0.181	2.21E-62	Endothelial
NRP1	1.41E-65	0.60362119	0.355	0.188	3.22E-61	Endothelial
RPL13AP51	6.32E-65	0.472443864	0.359	0.182	1.44E-60	Endothelial
RPL31	1.49E-64	0.520187011	0.325	0.16	3.41E-60	Endothelial
RPL131	7.44E-64	0.491894389	0.455	0.268	1.70E-59	Endothelial
RPL35P5	1.58E-63	0.497047873	0.328	0.163	3.61E-59	Endothelial
AC007969.51	3.04E-63	0.485074827	0.355	0.183	6.95E-59	Endothelial
RPL13P121	1.66E-62	0.474327711	0.369	0.195	3.78E-58	Endothelial
RPL27A	7.79E-62	0.502304805	0.337	0.172	1.78E-57	Endothelial
RPS25	3.73E-61	0.522464772	0.293	0.14	8.51E-57	Endothelial
RPS61	1.49E-60	0.535656658	0.357	0.191	3.39E-56	Endothelial
UBC	5.75E-60	0.579187189	0.311	0.157	1.31E-55	Endothelial
PALMD1	2.32E-59	0.697463922	0.339	0.189	5.30E-55	Endothelial
PDE4D	1.42E-58	0.268028483	0.255	0.113	3.25E-54	Endothelial
RPS8	1.57E-58	0.538154981	0.299	0.149	3.59E-54	Endothelial
ACTB2	2.12E-58	0.482925093	0.476	0.294	4.84E-54	Endothelial
RPS151	7.19E-58	0.463351042	0.368	0.201	1.64E-53	Endothelial
RPS27A1	4.72E-57	0.504037907	0.372	0.206	1.08E-52	Endothelial
PLEKHA7	5.21E-57	0.53166357	0.266	0.127	1.19E-52	Endothelial
RP11-889L3.1	1.72E-56	0.517672339	0.265	0.126	3.91E-52	Endothelial
RPL351	1.81E-56	0.451854405	0.383	0.212	4.12E-52	Endothelial
RPL38	1.18E-55	0.486388923	0.263	0.124	2.70E-51	Endothelial
RPL36	6.32E-55	0.464402635	0.338	0.179	1.44E-50	Endothelial
RPL19	1.34E-54	0.482165482	0.327	0.172	3.05E-50	Endothelial
TPT1	2.10E-54	0.45819254	0.323	0.169	4.79E-50	Endothelial
SYNE1	2.24E-53	0.622584661	0.272	0.138	5.11E-49	Endothelial
RPS192	5.19E-53	0.472753138	0.349	0.191	1.18E-48	Endothelial

RPL23A	5.85E-53	0.441536684	0.262	0.126	1.34E-48	Endothelial
RPS141	2.04E-52	0.470571926	0.42	0.252	4.66E-48	Endothelial
RP11-51O6.1	2.23E-52	0.502016109	0.259	0.125	5.10E-48	Endothelial
RPL411	2.35E-51	0.466094572	0.383	0.221	5.36E-47	Endothelial
RPS3	5.97E-51	0.471518098	0.263	0.129	1.36E-46	Endothelial
RPS23	8.27E-51	0.449338725	0.281	0.141	1.89E-46	Endothelial
RPS23P8	1.03E-50	0.448930801	0.262	0.128	2.35E-46	Endothelial
RPS4X	2.49E-50	0.46535623	0.313	0.167	5.69E-46	Endothelial
RPL28	2.49E-49	0.413776339	0.285	0.145	5.69E-45	Endothelial
VIM1	7.09E-49	0.396563277	0.496	0.327	1.62E-44	Endothelial
LGALS11	1.19E-48	0.47250654	0.547	0.397	2.72E-44	Endothelial
RPS24	3.79E-48	0.427074828	0.283	0.145	8.64E-44	Endothelial
RPS13	1.99E-47	0.410018759	0.261	0.13	4.55E-43	Endothelial
ADIRF1	2.06E-47	0.411019148	0.622	0.476	4.71E-43	Endothelial
RPS16	3.12E-47	0.419093743	0.283	0.147	7.11E-43	Endothelial
RPL35A	3.47E-47	0.434353354	0.301	0.16	7.93E-43	Endothelial
OOEP1	4.61E-47	0.389309824	0.345	0.191	1.05E-42	Endothelial
CAV11	3.28E-46	0.483188312	0.386	0.239	7.49E-42	Endothelial
EIF1	4.05E-46	0.480354646	0.327	0.186	9.25E-42	Endothelial
RPLP21	5.11E-46	0.40306727	0.525	0.368	1.17E-41	Endothelial
RPL30	7.03E-43	0.433498422	0.259	0.135	1.60E-38	Endothelial
RPS12	7.62E-43	0.439445023	0.28	0.15	1.74E-38	Endothelial
ACTG1	1.07E-42	0.449825645	0.251	0.129	2.43E-38	Endothelial
MGST3	1.97E-42	0.541278382	0.254	0.136	4.49E-38	Endothelial
FTH12	5.89E-41	0.293087869	0.583	0.428	1.35E-36	Endothelial
RPL37A1	7.56E-41	0.378113995	0.342	0.2	1.73E-36	Endothelial
RPL29	1.29E-40	0.426476003	0.269	0.145	2.95E-36	Endothelial
UBA52	8.40E-39	0.388609717	0.259	0.138	1.92E-34	Endothelial
GPX3	2.08E-38	0.338750575	0.325	0.189	4.75E-34	Endothelial
RP11-543P15.1	1.56E-37	0.361010863	0.286	0.16	3.55E-33	Endothelial
RPS91	1.85E-37	0.378904595	0.322	0.19	4.22E-33	Endothelial
RP11-864N7.2	3.25E-37	0.43685896	0.256	0.139	7.42E-33	Endothelial
RPS201	4.69E-37	0.376863686	0.358	0.222	1.07E-32	Endothelial
ATP5E	3.55E-36	0.359656182	0.258	0.141	8.10E-32	Endothelial
RPS29	8.05E-36	0.353822011	0.326	0.194	1.84E-31	Endothelial
IGFBP5	2.76E-35	0.324127797	0.258	0.142	6.30E-31	Endothelial
HSPG2	2.79E-33	0.412255568	0.354	0.23	6.37E-29	Endothelial
CD63	5.03E-32	0.326832398	0.309	0.188	1.15E-27	Endothelial
S100A10	8.51E-32	0.336763531	0.253	0.144	1.94E-27	Endothelial
SERF21	3.42E-26	0.300835989	0.325	0.211	7.80E-22	Endothelial
PPARG2	9.92E-25	0.413437887	0.421	0.316	2.26E-20	Endothelial
POLR2L	5.22E-24	0.321161919	0.268	0.169	1.19E-19	Endothelial
ITM2B	6.30E-24	0.319826177	0.268	0.168	1.44E-19	Endothelial

FBXL7	2.30E-23	0.314436651	0.307	0.207	5.25E-19	Endothelial
LHFP	2.97E-23	0.302334931	0.264	0.166	6.78E-19	Endothelial
SPARC1	2.67E-22	0.300072825	0.447	0.346	6.09E-18	Endothelial
PLXDC22	1.27E-21	0.336012631	0.294	0.198	2.89E-17	Endothelial
RP11-236F9.22	3.20E-16	0.359412531	0.413	0.321	7.29E-12	Endothelial
HMGB1	3.98E-12	0.272417032	0.282	0.213	9.08E-08	Endothelial
SNX29P22	1.50E-11	0.324655561	0.539	0.481	3.42E-07	Endothelial
NEGR11	0	2.380213188	0.561	0.065	0	Fibroblast
DCLK1	0	2.140848065	0.466	0.056	0	Fibroblast
LAMA2	0	2.021476135	0.481	0.092	0	Fibroblast
NOVA13	0	1.839205864	0.588	0.183	0	Fibroblast
DCN3	0	1.779722478	0.591	0.163	0	Fibroblast
TNXB	0	1.603384084	0.348	0.064	0	Fibroblast
COL1A21	0	1.563082899	0.453	0.111	0	Fibroblast
COL1A1	0	1.559131625	0.261	0.043	0	Fibroblast
ABCA9	0	1.433585716	0.347	0.088	0	Fibroblast
COL6A21	0	1.36291633	0.452	0.155	0	Fibroblast
CFD3	0	1.359334337	0.728	0.5	0	Fibroblast
ROBO2	2.62E-304	1.477827364	0.296	0.073	5.99E-300	Fibroblast
RORA3	3.70E-303	1.195172859	0.526	0.263	8.45E-299	Fibroblast
COL6A3	3.35E-296	1.28508574	0.267	0.058	7.66E-292	Fibroblast
ABCA10	2.22E-285	1.4487257	0.328	0.101	5.07E-281	Fibroblast
GSN2	5.34E-285	1.044329559	0.719	0.537	1.22E-280	Fibroblast
TSHZ22	1.34E-269	1.189682511	0.444	0.195	3.05E-265	Fibroblast
ABCA6	3.12E-249	1.140279018	0.366	0.135	7.13E-245	Fibroblast
PRKG1	1.07E-235	1.05149854	0.296	0.09	2.43E-231	Fibroblast
ABCA8	3.68E-232	1.275724767	0.255	0.069	8.40E-228	Fibroblast
KAZN	2.91E-228	1.272017751	0.271	0.082	6.65E-224	Fibroblast
APOD1	5.69E-221	1.516118397	0.391	0.173	1.30E-216	Fibroblast
ANK2	1.77E-202	1.212550021	0.253	0.078	4.03E-198	Fibroblast
CCDC80	1.23E-193	1.123362557	0.326	0.13	2.80E-189	Fibroblast
COL6A1	4.20E-177	1.03091918	0.338	0.148	9.58E-173	Fibroblast
MGP	1.49E-169	1.244313348	0.298	0.12	3.41E-165	Fibroblast
FBXL71	9.59E-165	1.104511133	0.356	0.174	2.19E-160	Fibroblast
SLIT3	8.56E-162	1.047956936	0.361	0.18	1.95E-157	Fibroblast
DLC12	1.00E-156	0.862602632	0.484	0.304	2.29E-152	Fibroblast
COL3A12	5.10E-146	1.016523199	0.375	0.202	1.17E-141	Fibroblast
PID1	8.84E-130	0.861759626	0.276	0.119	2.02E-125	Fibroblast
RUNX1T1	1.88E-115	0.98710318	0.265	0.124	4.30E-111	Fibroblast
IGFBP51	1.47E-107	0.930361329	0.261	0.121	3.35E-103	Fibroblast
LAMC1	9.86E-104	0.908158977	0.28	0.143	2.25E-99	Fibroblast
LINC004782	2.84E-93	0.811685167	0.379	0.248	6.47E-89	Fibroblast
S100A41	1.17E-85	0.772408422	0.319	0.184	2.68E-81	Fibroblast

ZEB1	2.18E-83	0.852323396	0.261	0.141	4.97E-79	Fibroblast
GPX31	1.38E-78	0.73715415	0.302	0.174	3.15E-74	Fibroblast
S100A61	3.35E-76	0.73516689	0.41	0.288	7.65E-72	Fibroblast
LHFP1	1.60E-75	0.732395162	0.268	0.149	3.65E-71	Fibroblast
TCF42	1.16E-70	0.603363148	0.308	0.182	2.64E-66	Fibroblast
CALD1	1.78E-65	0.717082757	0.266	0.158	4.06E-61	Fibroblast
AUTS23	9.97E-60	0.586825351	0.403	0.299	2.28E-55	Fibroblast
PLXDC23	5.43E-55	0.6440814	0.288	0.184	1.24E-50	Fibroblast
PTPRG2	4.29E-53	0.608550191	0.397	0.306	9.80E-49	Fibroblast
CST32	2.22E-46	0.521717724	0.336	0.238	5.06E-42	Fibroblast
SVEP1	1.37E-44	0.69151274	0.283	0.198	3.12E-40	Fibroblast
REV3L	1.45E-43	0.671838786	0.296	0.212	3.31E-39	Fibroblast
VIM2	3.29E-37	0.487508045	0.407	0.328	7.50E-33	Fibroblast
RPL412	6.64E-29	0.436123854	0.297	0.221	1.52E-24	Fibroblast
NFIB1	2.03E-27	0.472174752	0.332	0.269	4.62E-23	Fibroblast
PARD3B2	1.64E-23	0.494502839	0.341	0.29	3.74E-19	Fibroblast
RPL132	1.48E-22	0.386379307	0.338	0.275	3.38E-18	Fibroblast
RPL342	8.21E-21	0.352318416	0.356	0.295	1.87E-16	Fibroblast
RPLP12	8.86E-21	0.338277308	0.355	0.291	2.02E-16	Fibroblast
NFIA1	1.11E-20	0.412208373	0.381	0.338	2.54E-16	Fibroblast
RPL13P122	1.39E-20	0.412333174	0.261	0.201	3.17E-16	Fibroblast
FTH13	9.66E-18	0.310474388	0.474	0.437	2.20E-13	Fibroblast
ZBTB201	2.38E-17	0.2944115	0.525	0.516	5.44E-13	Fibroblast
TIMP32	3.73E-17	0.356552909	0.38	0.335	8.52E-13	Fibroblast
RPL34P182	1.20E-16	0.357897782	0.276	0.221	2.74E-12	Fibroblast
RP11-234A1.12	2.65E-15	0.336022431	0.32	0.269	6.04E-11	Fibroblast
RBMS31	5.92E-15	0.288796883	0.478	0.453	1.35E-10	Fibroblast
RPS142	1.47E-14	0.339409029	0.308	0.261	3.35E-10	Fibroblast
AC016739.22	1.62E-14	0.314914704	0.304	0.254	3.69E-10	Fibroblast
AC004453.82	5.81E-14	0.35391874	0.274	0.228	1.33E-09	Fibroblast
RPS152	1.08E-13	0.317665686	0.258	0.209	2.46E-09	Fibroblast
RPS182	1.39E-13	0.345703015	0.253	0.205	3.17E-09	Fibroblast
RPL112	1.39E-13	0.330508835	0.266	0.217	3.17E-09	Fibroblast
RPS27A2	8.60E-13	0.287083543	0.263	0.214	1.96E-08	Fibroblast
RPL322	2.39E-12	0.335301311	0.299	0.257	5.46E-08	Fibroblast
CELF21	2.55E-12	0.322701156	0.39	0.366	5.81E-08	Fibroblast
NAALADL22	3.09E-11	0.422839335	0.303	0.276	7.05E-07	Fibroblast
ANXA2	1.28E-10	0.37648008	0.289	0.26	2.93E-06	Fibroblast
RPS202	1.43E-10	0.330832978	0.267	0.228	3.27E-06	Fibroblast
WSB1	3.62E-10	0.373065558	0.26	0.23	8.26E-06	Fibroblast
LPP1	7.61E-10	0.283732465	0.406	0.392	1.74E-05	Fibroblast
RPL352	1.46E-08	0.274929213	0.259	0.224	0.00033282	Fibroblast
RPLP22	1.55E-08	0.255108767	0.402	0.382	0.000353187	Fibroblast

HSPG21	1.09E-07	0.317426859	0.264	0.239	0.002488599	Fibroblast
CHD91	2.59E-05	0.309026522	0.277	0.264	0.591006044	Fibroblast
F13A1	0	2.588030853	0.469	0.022	0	Macrophage
FRMD4B1	0	2.412446231	0.542	0.049	0	Macrophage
PDE4D1	0	2.310055954	0.541	0.084	0	Macrophage
RBPJ	0	2.146269665	0.523	0.143	0	Macrophage
LGMN	0	2.120988174	0.45	0.052	0	Macrophage
MS4A6A	0	2.067403019	0.36	0.021	0	Macrophage
MAMDC2	0	2.016148564	0.387	0.048	0	Macrophage
SEPP11	0	1.972323916	0.647	0.163	0	Macrophage
SCN9A	0	1.957267118	0.268	0.012	0	Macrophage
SLC9A9	0	1.810956572	0.421	0.082	0	Macrophage
IQGAP21	0	1.805184796	0.368	0.041	0	Macrophage
COLEC12	0	1.800661573	0.279	0.03	0	Macrophage
STAB1	0	1.713315643	0.276	0.035	0	Macrophage
RBM471	3.92E-276	1.604938132	0.281	0.046	8.95E-272	Macrophage
MYO5A	4.98E-269	1.569030674	0.343	0.075	1.14E-264	Macrophage
NAV2	6.15E-254	1.621616281	0.353	0.084	1.40E-249	Macrophage
RNASE1	1.86E-249	1.870590699	0.366	0.092	4.24E-245	Macrophage
WWP1	2.33E-236	1.583481587	0.346	0.086	5.32E-232	Macrophage
MEF2C1	2.43E-209	1.330708266	0.408	0.128	5.54E-205	Macrophage
C20orf194	7.93E-203	1.439855731	0.353	0.102	1.81E-198	Macrophage
SLC8A11	6.05E-195	1.443076952	0.29	0.07	1.38E-190	Macrophage
MTSS1	4.14E-190	1.518813658	0.314	0.085	9.44E-186	Macrophage
SRGAP2B1	3.59E-171	1.336474453	0.319	0.094	8.19E-167	Macrophage
MAN1A1	2.47E-167	1.383474566	0.41	0.159	5.64E-163	Macrophage
PDGFC	1.24E-151	1.429010474	0.293	0.091	2.83E-147	Macrophage
HDAC91	1.05E-141	1.272411303	0.316	0.106	2.39E-137	Macrophage
ATG71	7.26E-136	1.252126748	0.333	0.122	1.66E-131	Macrophage
ME1	4.62E-135	1.396201429	0.284	0.093	1.05E-130	Macrophage
ZEB23	4.30E-130	1.152195166	0.458	0.227	9.82E-126	Macrophage
ELMO11	2.64E-124	1.105191847	0.275	0.088	6.02E-120	Macrophage
PID11	3.15E-100	1.079334032	0.327	0.138	7.20E-96	Macrophage
TRPS11	4.13E-91	1.075164727	0.348	0.167	9.42E-87	Macrophage
CD742	2.43E-88	1.033952532	0.321	0.14	5.56E-84	Macrophage
ITPR2	5.68E-79	1.104234955	0.303	0.142	1.30E-74	Macrophage
FOXO3	8.10E-74	1.096635807	0.274	0.124	1.85E-69	Macrophage
FCHSD2	1.42E-73	1.015722612	0.258	0.11	3.24E-69	Macrophage
LDLRAD4	5.17E-73	1.052598702	0.296	0.141	1.18E-68	Macrophage
ZSWIM61	1.20E-67	1.086978421	0.251	0.112	2.75E-63	Macrophage
C10orf112	2.02E-64	0.857162654	0.43	0.274	4.61E-60	Macrophage
ABCA61	1.77E-63	0.868518497	0.332	0.175	4.03E-59	Macrophage
ARHGAP242	1.09E-60	0.888142026	0.413	0.264	2.48E-56	Macrophage

SAT11	3.48E-59	0.901498047	0.312	0.168	7.94E-55	Macrophage
SNX29	6.09E-59	0.898955389	0.267	0.131	1.39E-54	Macrophage
FOXP11	3.64E-58	0.812889968	0.381	0.232	8.31E-54	Macrophage
PAPD4	6.63E-56	0.925427142	0.292	0.156	1.51E-51	Macrophage
PEAK1	1.55E-54	0.922235068	0.305	0.169	3.53E-50	Macrophage
ITSN14	2.98E-53	0.810283458	0.419	0.285	6.80E-49	Macrophage
NRP11	1.12E-51	0.875226905	0.331	0.194	2.56E-47	Macrophage
STARD13	2.90E-49	0.94080233	0.283	0.158	6.62E-45	Macrophage
MEF2A1	9.21E-49	0.865800243	0.31	0.181	2.10E-44	Macrophage
CPM3	1.80E-44	0.848126353	0.354	0.229	4.10E-40	Macrophage
AKAP131	3.01E-40	0.760417157	0.276	0.159	6.88E-36	Macrophage
GNAQ1	1.09E-37	0.751556645	0.253	0.144	2.48E-33	Macrophage
FTL2	1.38E-29	0.570013308	0.636	0.578	3.15E-25	Macrophage
TCF12	1.14E-23	0.687308075	0.291	0.206	2.59E-19	Macrophage
CST33	1.52E-20	0.596920759	0.338	0.253	3.46E-16	Macrophage
PSAP2	1.79E-17	0.556336037	0.3	0.226	4.09E-13	Macrophage
AFF31	2.64E-13	0.634035568	0.259	0.202	6.03E-09	Macrophage
QKI1	4.57E-12	0.555226117	0.302	0.254	1.04E-07	Macrophage
DPYD2	3.56E-11	0.439576936	0.28	0.227	8.13E-07	Macrophage
MED13L	1.17E-09	0.450189659	0.314	0.271	2.66E-05	Macrophage
TXNIP2	1.30E-09	0.288891078	0.358	0.305	2.97E-05	Macrophage
FHIT2	8.47E-08	0.479746595	0.257	0.22	0.001932672	Macrophage
MBNL11	7.14E-06	0.313159136	0.382	0.366	0.162984709	Macrophage
AUTS24	0.002701629	0.322511806	0.327	0.324	1	Macrophage
TPSB2	0	3.963803176	0.838	0.013	0	Mast Cells
TPSAB1	0	3.741246083	0.754	0.007	0	Mast Cells
AC004791.2	0	3.538027141	0.681	0.002	0	Mast Cells
HPGD	0	2.326799269	0.309	0.008	0	Mast Cells
HDC	0	2.248647987	0.298	0.002	0	Mast Cells
KIT	0	2.156220816	0.267	0.003	0	Mast Cells
RAB27B	0	2.08870495	0.262	0.004	0	Mast Cells
VWA5A	5.39E-239	2.023174892	0.251	0.008	1.23E-234	Mast Cells
HPGDS	1.77E-188	1.843250505	0.251	0.01	4.04E-184	Mast Cells
SYTL3	4.64E-60	1.775407858	0.277	0.041	1.06E-55	Mast Cells
SLC24A3	7.49E-47	1.761742422	0.298	0.059	1.71E-42	Mast Cells
NTM5	8.49E-37	1.60684885	0.565	0.245	1.94E-32	Mast Cells
SMYD3	1.43E-32	1.667716678	0.424	0.15	3.26E-28	Mast Cells
ELMO12	9.58E-20	1.207605675	0.298	0.104	2.19E-15	Mast Cells
AGAP1	8.38E-13	1.084375701	0.283	0.125	1.91E-08	Mast Cells
FER	1.14E-10	1.04291185	0.298	0.154	2.60E-06	Mast Cells
SNX29P23	5.75E-07	0.538296669	0.586	0.487	0.013118044	Mast Cells
FOXP12	7.98E-07	0.588121748	0.377	0.246	0.018215318	Mast Cells
AKAP132	0.000259101	0.667239934	0.257	0.17	1	Mast Cells

ANXA1	0.00065147	0.696135504	0.283	0.206	1	Mast Cells
EXOC6B	0.001014123	0.604649873	0.293	0.22	1	Mast Cells
PPP3CA	0.00693087	0.593169469	0.267	0.207	1	Mast Cells
C10orf113	0.008359834	0.445713942	0.351	0.289	1	Mast Cells
COL25A1	0	2.524444573	0.601	0.031	0	Perivascular
KCNAB1	0	2.357137001	0.555	0.033	0	Perivascular
RGS6	0	2.324522556	0.634	0.048	0	Perivascular
PRKG11	0	2.0882734	0.762	0.121	0	Perivascular
POSTN	0	2.072924253	0.456	0.027	0	Perivascular
MYO1B	0	1.931623983	0.467	0.028	0	Perivascular
STEAP4	5.53E-291	1.568671148	0.286	0.017	1.26E-286	Perivascular
EBF2	1.35E-282	1.832289532	0.425	0.042	3.08E-278	Perivascular
FRMD3	2.50E-244	1.660091459	0.381	0.039	5.70E-240	Perivascular
THBS4	4.20E-238	1.500251199	0.271	0.019	9.60E-234	Perivascular
NOTCH3	6.48E-218	1.386349498	0.28	0.023	1.48E-213	Perivascular
GUCY1A2	2.58E-215	1.508843095	0.258	0.019	5.89E-211	Perivascular
ABCC9	7.68E-186	1.621988806	0.399	0.058	1.75E-181	Perivascular
NDUFA4L2	9.59E-179	1.506033679	0.253	0.023	2.19E-174	Perivascular
PDGFRB	1.09E-178	1.487964363	0.396	0.057	2.49E-174	Perivascular
ENOX1	2.82E-171	1.450340433	0.289	0.032	6.45E-167	Perivascular
NR2F2-AS1	2.41E-153	1.462714026	0.355	0.054	5.50E-149	Perivascular
COL18A1	3.24E-152	1.330782062	0.291	0.036	7.39E-148	Perivascular
IGFBP71	1.00E-148	1.497788519	0.551	0.131	2.29E-144	Perivascular
RGS5	1.37E-132	1.301391095	0.319	0.049	3.12E-128	Perivascular
MIR4435-1HG	2.15E-125	1.394942336	0.37	0.07	4.90E-121	Perivascular
DLC14	4.09E-114	1.259800322	0.767	0.335	9.33E-110	Perivascular
TINAGL1	6.34E-114	1.177989563	0.253	0.036	1.45E-109	Perivascular
COL5A3	3.18E-110	1.385599289	0.421	0.101	7.26E-106	Perivascular
GRK5	5.00E-100	1.190188226	0.328	0.066	1.14E-95	Perivascular
BTNL91	5.18E-84	0.924851688	0.368	0.089	1.18E-79	Perivascular
TPM1	4.30E-70	0.986836521	0.286	0.067	9.81E-66	Perivascular
SLCO3A1	1.02E-64	0.988689841	0.267	0.061	2.32E-60	Perivascular
SH3RF1	2.58E-63	1.062150399	0.284	0.072	5.88E-59	Perivascular
IGFBP52	1.63E-62	0.880706651	0.438	0.147	3.73E-58	Perivascular
MT2A	7.58E-62	1.137365757	0.328	0.094	1.73E-57	Perivascular
SPARCL11	2.74E-57	0.762246015	0.377	0.118	6.26E-53	Perivascular
PLCL1	6.97E-57	1.161419429	0.282	0.077	1.59E-52	Perivascular
THSD7B	3.82E-54	1.250913615	0.253	0.068	8.71E-50	Perivascular
KALRN1	1.36E-51	0.901829402	0.324	0.102	3.11E-47	Perivascular
CD364	4.69E-50	0.583564667	0.868	0.647	1.07E-45	Perivascular
LHFP2	2.09E-47	0.841818946	0.432	0.17	4.77E-43	Perivascular
ASAP11	6.81E-47	0.906414779	0.407	0.159	1.55E-42	Perivascular
TSHZ25	1.47E-43	0.691892234	0.533	0.247	3.35E-39	Perivascular

DPYSL2	6.95E-43	0.760575332	0.262	0.079	1.59E-38	Perivascular
EPS8	7.83E-39	0.865349072	0.416	0.186	1.79E-34	Perivascular
MEF2C2	1.12E-38	0.60655098	0.385	0.15	2.55E-34	Perivascular
7-Sep	1.12E-35	0.775212328	0.264	0.09	2.56E-31	Perivascular
CALD11	2.60E-32	0.783257681	0.392	0.178	5.94E-28	Perivascular
RASAL22	1.98E-31	0.762307949	0.326	0.136	4.52E-27	Perivascular
APBB24	3.07E-31	0.687825249	0.52	0.288	7.00E-27	Perivascular
TCF7L1	6.09E-31	0.733613165	0.256	0.093	1.39E-26	Perivascular
COL6A22	7.12E-31	0.474283024	0.467	0.22	1.62E-26	Perivascular
RAPGEF2	9.73E-27	0.68532357	0.258	0.102	2.22E-22	Perivascular
PPARG6	1.24E-26	0.648322151	0.54	0.322	2.83E-22	Perivascular
RBPMS5	4.72E-26	0.619388637	0.518	0.299	1.08E-21	Perivascular
A2M1	8.47E-26	0.429714332	0.278	0.109	1.93E-21	Perivascular
PDZD24	1.77E-25	0.677760192	0.463	0.258	4.03E-21	Perivascular
RP11-1101K5.1	2.14E-25	0.779216898	0.295	0.132	4.88E-21	Perivascular
TIMP35	4.16E-25	0.50902582	0.568	0.339	9.48E-21	Perivascular
AC016739.23	1.08E-24	0.462962216	0.487	0.26	2.46E-20	Perivascular
ADAMTS121	7.71E-24	0.774121412	0.33	0.16	1.76E-19	Perivascular
PTK2	8.75E-24	0.720989481	0.302	0.138	2.00E-19	Perivascular
PTEN	6.92E-23	0.622256299	0.324	0.154	1.58E-18	Perivascular
RPLP13	6.56E-22	0.43496779	0.52	0.3	1.50E-17	Perivascular
AGAP11	6.97E-22	0.677669845	0.273	0.123	1.59E-17	Perivascular
FYN	1.18E-21	0.572574313	0.289	0.131	2.68E-17	Perivascular
MAML21	4.71E-21	0.546097258	0.537	0.344	1.07E-16	Perivascular
CDC42BPA	1.35E-20	0.676225208	0.308	0.154	3.08E-16	Perivascular
FABP44	9.78E-20	0.553067747	0.85	0.749	2.23E-15	Perivascular
LPP3	1.86E-19	0.480283179	0.579	0.39	4.24E-15	Perivascular
NBEAL11	3.84E-18	0.48008447	0.419	0.243	8.77E-14	Perivascular
SASH11	6.80E-18	0.499337388	0.284	0.138	1.55E-13	Perivascular
IFITM31	1.24E-17	0.490040457	0.253	0.117	2.83E-13	Perivascular
PTPRG5	2.76E-17	0.521606582	0.502	0.323	6.30E-13	Perivascular
SYNE21	3.14E-17	0.430167222	0.251	0.117	7.16E-13	Perivascular
MKL2	3.29E-17	0.623773278	0.251	0.122	7.51E-13	Perivascular
DLEU21	1.11E-16	0.594555197	0.251	0.121	2.53E-12	Perivascular
ZEB24	7.00E-15	0.37989211	0.405	0.246	1.60E-10	Perivascular
MYL9	8.87E-15	0.634039698	0.267	0.141	2.02E-10	Perivascular
COL4A24	4.80E-14	0.400486724	0.445	0.282	1.09E-09	Perivascular
RPL3P42	1.51E-13	0.419506453	0.35	0.204	3.44E-09	Perivascular
CTD-2192J16.152	4.64E-13	0.393943817	0.35	0.203	1.06E-08	Perivascular
MAPK101	5.52E-13	0.556994263	0.289	0.166	1.26E-08	Perivascular
UTRN1	8.30E-13	0.45340725	0.449	0.301	1.89E-08	Perivascular
ACTG11	2.49E-12	0.269144295	0.267	0.14	5.68E-08	Perivascular

TCF43	4.99E-12	0.272991151	0.357	0.209	1.14E-07	Perivascular
GPX32	1.18E-11	0.263118416	0.341	0.201	2.69E-07	Perivascular
MGST31	2.71E-11	0.357427239	0.262	0.146	6.19E-07	Perivascular
PTMA1	2.77E-11	0.262205157	0.278	0.154	6.33E-07	Perivascular
PABPC11	4.27E-11	0.287450486	0.289	0.165	9.74E-07	Perivascular
RPL101	6.98E-11	0.384205929	0.322	0.193	1.59E-06	Perivascular
COL5A2	1.38E-10	0.472940589	0.302	0.183	3.14E-06	Perivascular
RPS11	1.73E-10	0.305406812	0.251	0.139	3.96E-06	Perivascular
PTRF1	2.02E-10	0.262900119	0.267	0.15	4.62E-06	Perivascular
MKLN1	2.26E-10	0.445809223	0.337	0.215	5.15E-06	Perivascular
CHD92	4.76E-10	0.425293641	0.385	0.264	1.09E-05	Perivascular
RBMS34	5.69E-10	0.308656132	0.586	0.455	1.30E-05	Perivascular
LDB21	5.74E-10	0.256793799	0.26	0.146	1.31E-05	Perivascular
SPARC3	6.51E-10	0.385476944	0.485	0.354	1.49E-05	Perivascular
ANGPTL4	8.23E-10	0.52381429	0.256	0.154	1.88E-05	Perivascular
COL6A11	1.83E-09	0.327012405	0.308	0.191	4.19E-05	Perivascular
NAV1	2.33E-09	0.322372627	0.275	0.166	5.32E-05	Perivascular
COL4A11	2.87E-09	0.364914754	0.302	0.191	6.55E-05	Perivascular
COL3A14	3.80E-09	0.26622972	0.368	0.24	8.67E-05	Perivascular
PDE8A	4.86E-09	0.424945869	0.26	0.159	0.000110855	Perivascular
HSPG23	7.46E-09	0.275353215	0.363	0.241	0.000170162	Perivascular
TACC11	1.30E-08	0.27981542	0.324	0.208	0.000295703	Perivascular
RBMS1	1.87E-08	0.286088078	0.328	0.215	0.000425684	Perivascular
ACTB3	2.69E-08	0.293792783	0.441	0.312	0.000613867	Perivascular
HIP1	3.29E-08	0.407455586	0.256	0.16	0.000751105	Perivascular
RPS81	7.59E-08	0.362066337	0.26	0.164	0.001731461	Perivascular
ADIRF2	7.79E-08	0.267201536	0.61	0.49	0.001778395	Perivascular
CD631	1.05E-07	0.261107712	0.308	0.199	0.002393536	Perivascular
PARD31	1.51E-07	0.353662053	0.308	0.21	0.003448334	Perivascular
ZEB11	9.09E-07	0.260189484	0.258	0.168	0.020738328	Perivascular
LRP4	3.75E-06	0.309534853	0.392	0.295	0.085620141	Perivascular
UVRAG2	0.000266257	0.272035082	0.317	0.242	1	Perivascular
DIAPH21	0.000559781	0.258208635	0.33	0.263	1	Perivascular
SKAP1	0	2.272873288	0.338	0.013	0	T Cells
FAM65B	0	2.029144115	0.293	0.019	0	T Cells
ARHGAP15	0	1.952743262	0.394	0.053	0	T Cells
PTPRC1	8.90E-304	1.820928274	0.358	0.043	2.03E-299	T Cells
MYO1F1	2.68E-164	1.558022942	0.257	0.041	6.12E-160	T Cells
CCND3	9.41E-110	1.438551535	0.26	0.06	2.15E-105	T Cells
PRKCH	5.42E-81	1.326248118	0.293	0.095	1.24E-76	T Cells
B2M2	1.10E-49	0.764624806	0.619	0.424	2.51E-45	T Cells
HLA-B1	1.82E-45	0.887142488	0.312	0.139	4.16E-41	T Cells
RABGAP1L	1.95E-38	1.071448797	0.29	0.141	4.46E-34	T Cells

HLA-A1	9.63E-27	0.759406982	0.268	0.141	2.20E-22	T Cells
MBNL13	2.23E-24	0.647678136	0.473	0.362	5.09E-20	T Cells
SNX29P24	4.63E-16	0.487872253	0.55	0.485	1.06E-11	T Cells
TMSB4X3	5.68E-15	0.33005517	0.628	0.552	1.30E-10	T Cells
RPS22	2.66E-09	0.551989292	0.269	0.199	6.06E-05	T Cells
RPS193	2.34E-06	0.488725134	0.26	0.207	0.053415349	T Cells
RPL3P43	3.33E-05	0.394878407	0.255	0.206	0.760600345	T Cells
RPL343	0.000107206	0.278307736	0.351	0.307	1	T Cells
RPS27A3	0.000132024	0.31455855	0.267	0.224	1	T Cells
S100A42	0.000137144	0.301640805	0.262	0.215	1	T Cells
ACTB4	0.000763594	0.30373759	0.343	0.315	1	T Cells
RP11-742N3.12	0.001004715	0.2999617	0.301	0.267	1	T Cells
FOXP13	0.003461248	0.390951084	0.271	0.246	1	T Cells

avg_logFC indicates average log fold change; p_val_adj indicates adjusted p-value

pct.1 indicates the percentage of cells where the gene is detected in the first group

pct.2 indicates the percentage of cells where the gene is detected in the second group

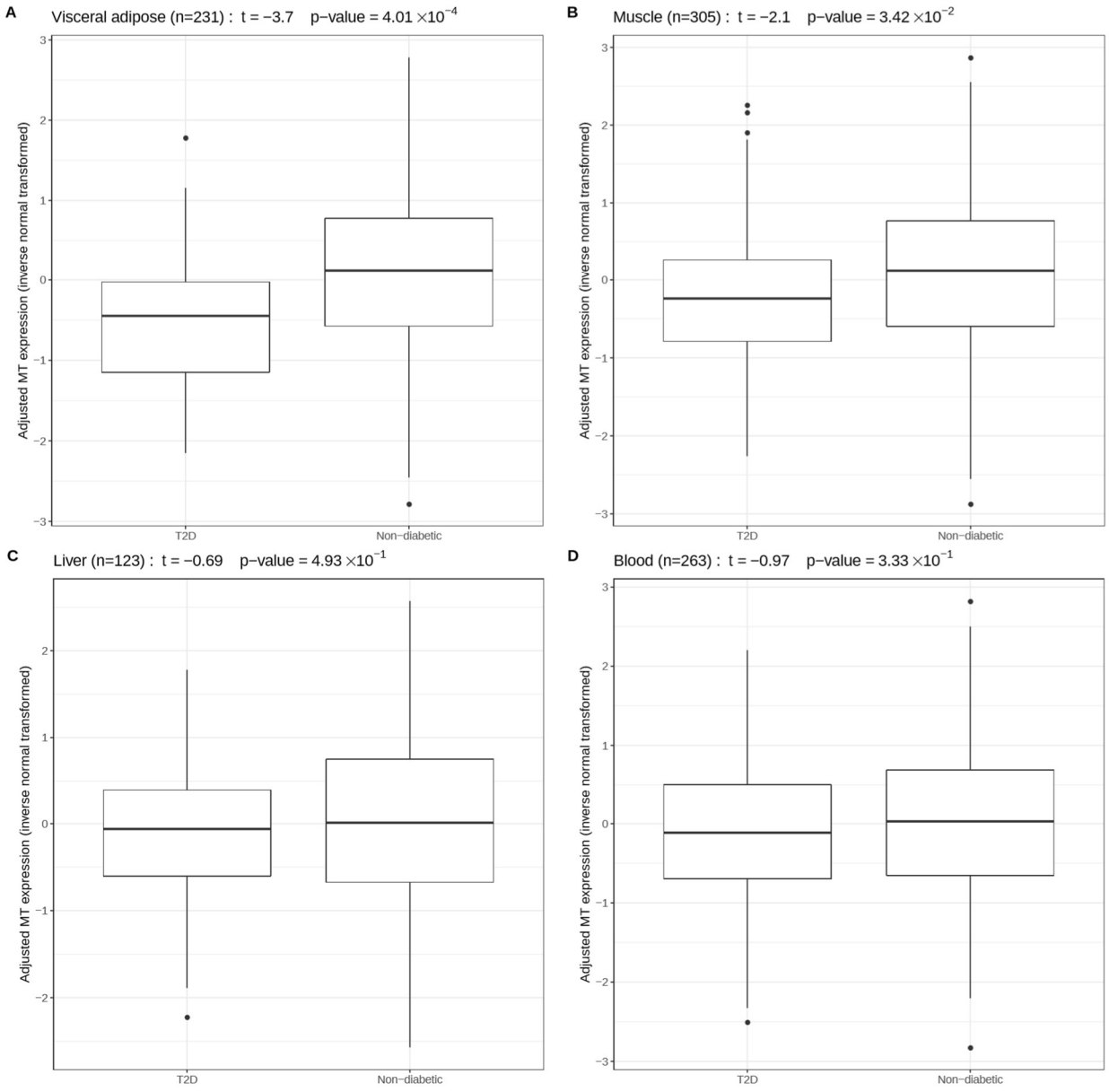
Supplementary table III-7: Technical factors observed in the METSIM adipose RNA-seq data.

	Mean	SD	Min	Max
PCT_CODING_BASES	43.93%	2.94%	31.90%	50.27%
PCT_UTR_BASES	45.75%	4.01%	34.36%	56.38%
PCT_INTRONIC_BASES	6.69%	2.21%	2.62%	15.92%
PCT_INTERGENIC_BASES	3.63%	2.34%	2.08%	24.62%
PCT_MRNA_BASES	89.68%	3.16%	72.03%	95.24%
MEDIAN_3PRIME_BIAS	0.81	0.20	0.58	1.88
MEDIAN_5PRIME_TO_3PRIME_BIAS	0.47	0.19	0.02	0.89
MEDIAN_CV_COVERAGE	0.56	0.12	0.38	1.11
RIN	7.74	0.64	6.00	8.90
TOTAL_MPREADS	3.6E+07	7878216	1.6E+07	6E+07
Batch	*	*	*	*
MT read percent **	10.52%	3.60%	3.65%	20.87%

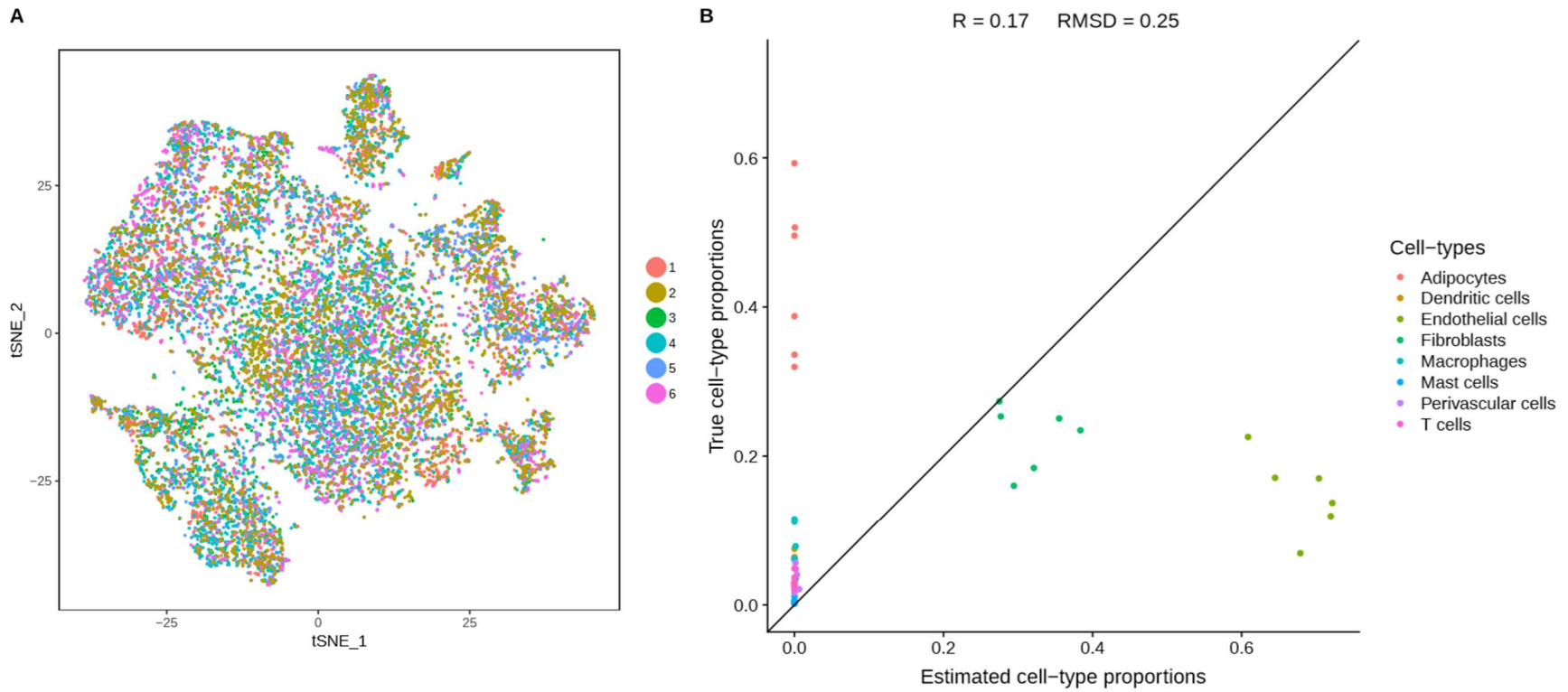
*: We adjusted for the batches and thus, the summary statistics, such as mean and SD, are not applicable.

**: MT expression was adjusted for 11 technical factors, excluding the MT read percent.

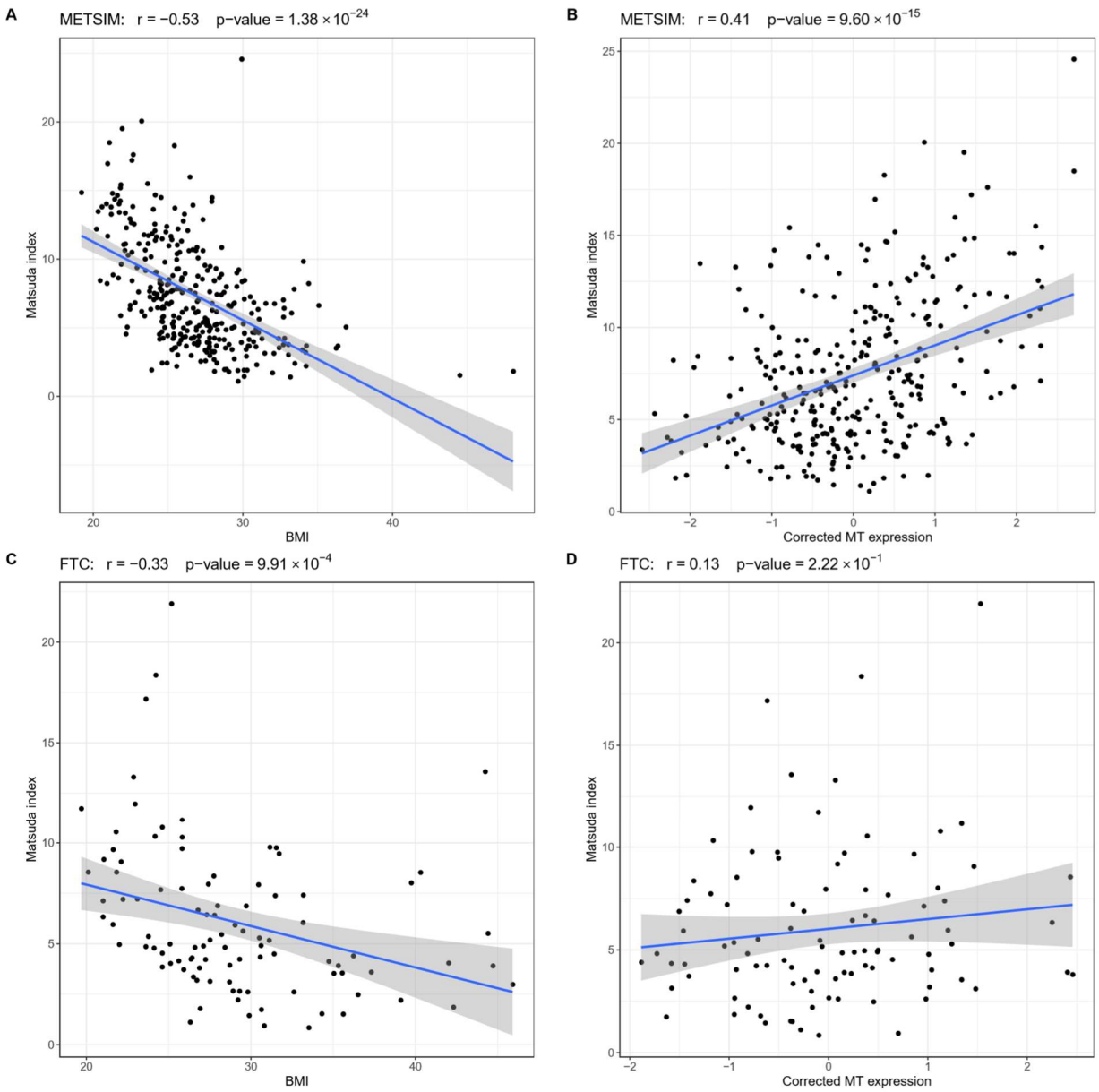
Supplementary Figures:



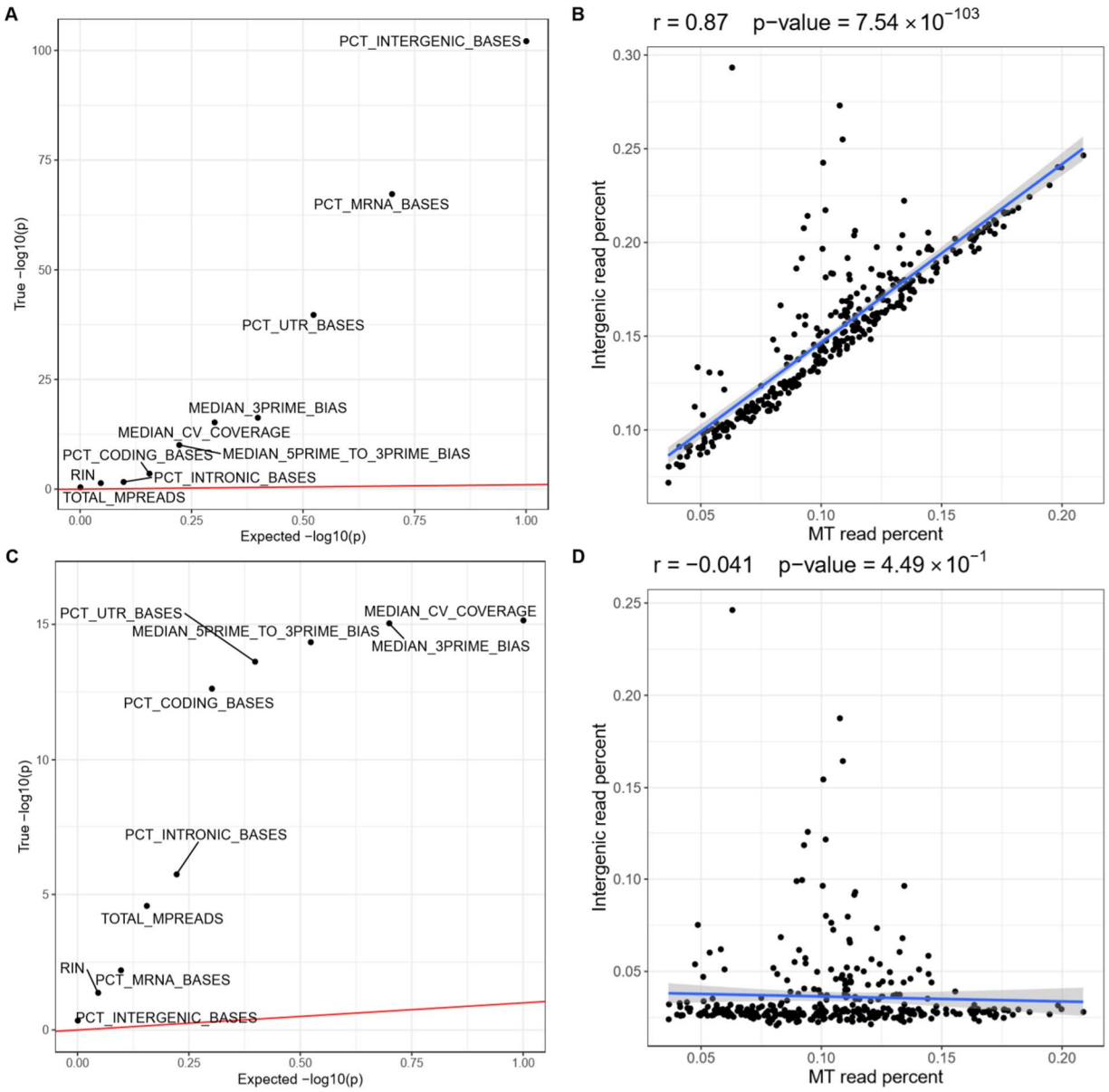
Supplementary Figure III-1. In GTEx, we used a t-test to show the association between the adjusted MT gene expression (inverse normal transformed) and T2D status in 4 different tissues. In (A) visceral adipose and (B) muscle, the adjusted MT expression is significantly higher in the non-diabetic individuals than in the T2D patients. In (C) liver and (D) blood, there is no evidence of differential MT gene expression between the non-diabetic individuals and the T2D patients.



Supplementary Figure 2. Our QC process ensures that the SN-RNA-seq accurately estimated the cell-type proportions in subcutaneous adipose tissue. (A) The t-SNE plot shows no evidence of a batch effect in SN-RNA-seq clustering. The dots are colored by sample IDs. (B) When using all genes in the SN-RNA-seq data without any filtering, the estimated cell-type proportions are not concordant with the true cell-type proportions. Thus, using the selected genes (Figure 2B) performs much better than using all genes as reference in the decomposition process.



Supplementary Figure 3. Pearson correlations show the associations between the predictors (BMI, corrected MT expression) and Matsuda index in the METSIM and FTC cohorts. The predicted Matsuda index is always more strongly associated with the true Matsuda index than any of the predictors (Figure 4). (A) The correlation between raw BMI and the Matsuda index in METSIM. (B) The correlation between the corrected MT gene expression and Matsuda index in METSIM. (C) The correlation between raw BMI and the Matsuda index in FTC. (D) The correlation between the corrected MT gene expression and Matsuda index in FTC.



Supplementary Figure 4. Using the METSIM cohort as an example, we demonstrate that when estimating RNA metrics, including the MT reads, the RNA metrics are heavily biased by the MT read percent. When excluding the MT reads to estimate the RNA metrics, the correlations between the MT read percent and other RNA metrics are reduced. (A). The qq-plot shows the correlations between the MT read percent and estimated RNA metrics, including MT reads. (B). The intergenic read percent is dominated by the MT read percent, including the MT reads. (C). The qq-plot shows the correlations between the MT read percent and estimated RNA metrics, excluding the MT reads. (C). The intergenic read percent is not correlated with the MT read percent, excluding the MT reads.

References:

[1] Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX, et al. Obesity and cardiovascular disease: Pathophysiology, evaluation, and effect of weight loss. Circulation. 2006;113(6):898–918.

[2] Zamora-Mendoza R, Rosas-Vargas H, Ramos-Cervantes MT, Garcia-Zuniga P, Perez-Lorenzana H, Mendoza-Lorenzo P, et al. Dysregulation of mitochondrial function and biogenesis modulators in adipose tissue of obese children. Int J Obes. 2018;42(4):618–24.

[3] Martinez KE, Tucker LA, Bailey BW, LeCheminant JD. Expanded normal weight obesity and insulin resistance in US adults of the national health and nutrition examination survey. J Diabetes Res. 2017;2017.

[4] Chung JO, Cho DH, Chung DJ, Chung MY. Associations among Body Mass Index, Insulin Resistance, and Pancreatic β-Cell Function in Korean Patients with New-Onset Type 2 Diabetes FAU - Chung, Jin Ook FAU Cho, Dong Hyeok FAU - Chung, Dong Jin FAU - Chung, Min Young. Korean J Intern Med. 2012;27(1):66–71.

[5] Meah FA, DiMeglio LA, Greenbaum CJ, Blum JS, Sosenko JM, Pugliese A, et al. The relationship between BMI and insulin resistance and progression from single to multiple autoantibody positivity and type 1 diabetes among TrialNet Pathway to Prevention participants. Diabetologia. 2016;59(6):1186–95.

[6] Cheng YH, Tsao YC, Tzeng IS, Chuang HH, Li WC, Tung TH, et al. Body mass index and waist circumference are better predictors of insulin resistance than total body fat percentage in middle-aged and elderly Taiwanese. Med (United States). 2017;96(39):1–6.

[7] Neeland, I. J., Turer, A. T., Ayers, C. R., Powell-Wiley, T. M., Vega, G. L., Farzaneh-Far, R., … De Lemos, J. A. Dysfunctional adiposity and the risk of prediabetes and type 2 diabetes in obese adults. JAMA. 2012; 308(11), 1150–1159.

[8] Goran, M. I., Lane, C., Toledo-Corral, C., & Weigensberg, M. J. Persistence of pre-diabetes in overweight and obese hispanic children; Association with progressive insulin resistance, Poor β-cell function, and increasing visceral fat. Diabetes. 2008; 57(11), 3007–3012.

[9] Dandona P, Aljada A, Bandyopadhyay A. Inflammation the link between insulin resistance,. Trends Immunol. 2004;25(1):4–7.

[10] Yang WM, Jeong HJ, Park SW, Lee W. Obesity-induced miR-15b is linked causally to the development of insulin resistance through the repression of the insulin receptor in hepatocytes. Mol Nutr Food Res. 2015;59(11):2303–14.

[11] Pedersen DJ, Guilherme A, Danai L V., Heyda L, Matevossian A, Cohen J, et al. A major role of insulin in promoting obesity-associated adipose tissue inflammation. Mol Metab. 2015;4(7):507–18.

[12] Adabimohazab R, Garfinkel A, Milam EC, Frosch O, Mangone A, Convit A. Does Inflammation Mediate the Association Between Obesity and Insulin Resistance? Inflammation. 2016;39(3):994–1003. z

[13] Saltiel AR, Olefsky JM, Saltiel AR, Olefsky JM. Inflammatory mechanisms linking obesity and metabolic disease Find the latest version : Inflammatory mechanisms linking obesity and metabolic disease. J Clin Invest. 2017;127(1):1–4.

[14] 5. Li P, Oh DY, Bandyopadhyay G, Lagakos WS, Talukdar S, Osborn O, et al. LTB4 promotes insulin resistance in obese mice by acting on macrophages, hepatocytes and myocytes. Nat Med. 2015;21(3):239–47.

[15] Roberts-Toler C, O'Neill BT, Cypess AM. Diet-induced obesity causes insulin resistance in mouse brown adipose tissue. Obesity. 2015;23(9):1765–70.

[16] Wensveen FM, Jelenčić V, Valentić S, Šestan M, Wensveen TT, Theurich S, et al. NK cells link obesity-induced adipose stress to inflammation and insulin resistance. Nat Immunol. 2015;16(4):376–85.

[17] Noakes TD. So What Comes First: The Obesity or the Insulin Resistance? And Which Is More Important? Clin Chem. 2018;64(1):7–9.

[18] Bluher M. Adipose tissue inflammation: a cause or consequence of obesity-related insulin resistance? Clin Sci. 2016;130(18):1603–14.

[19] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9.

[20]Laakso M, Kuusisto J, Stančáková A, Kuulasmaa T, Pajukanta P, Lusis AJ, et al. The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. J Lipid Res. 2017;58(3):481–93.

[21] Weisberg SP, Leibel RL, Anthony W, Jr F, Weisberg SP, Mccann D, et al. Obesity is associated with macrophage accumulation in adipose tissue Find the latest version : Obesity is associated with. J Clin Invest. 2003;112(12):1796–808.

[22] Van Harmelen V, Skurk T, Röhrig K, Lee YM, Halbleib M, Aprath-Husmann I, et al. Effect of BMI and age on adipose tissue cellularity and differentiation capacity in women. Int J Obes. 2003;

[23] Reilly SM, Saltiel AR. Adapting to obesity with adipose tissue inflammation. Nat Rev Endocrinol. 2017;13(11):633–43.

[24] Majka SM, Miller HL, Helm KM, Acosta AS, Childs CR, Kong R, et al. Analysis and isolation of adipocytes by flow cytometry. Methods in Enzymology 2014; 537:281–296 p.

[25] Ehrlund A, Acosta JR, Björk C, Hedén P, Douagi I, Arner P, et al. The cell-type specific transcriptome in human adipose tissue and influence of obesity on adipocyte progenitors. Sci Data. 2017;4:1–11.

[26] Hagberg CE, Li Q, Kutschke M, Bhowmick D, Kiss E, Shabalina IG, et al. Flow Cytometry of Mouse and Human Adipocytes for the Analysis of Browning and Cellular Heterogeneity. Cell Rep. 2018;24(10):2746-2756.e5.

[27] Wu M, Singh AK. Single-cell protein analysis. Curr Opin Biotechnol. 2012;23(1):83–8.

[28] Hu P, Zhang W, Xin H, Deng G. Single Cell Isolation and Analysis. Front Cell Dev Biol. 2016 Oct 25;4(October):135–82.

[29] Vink RG, Roumans NJ, Fazelzadeh P, Tareen SHK, Boekschoten M V., Van Baak MA, et al. Adipose tissue gene expression is differentially regulated with different rates of weight loss in overweight and obese humans. Int J Obes. 2017;41(2):309–16.

[30] Heinonen S, Buzkova J, Muniandy M, Kaksonen R, Ollikainen M, Ismail K, et al. Impaired mitochondrial biogenesis in adipose tissue in acquired obesity. Diabetes. 2015;64(9):3135–45.

[31] Lindinger PW, Christe M, Eberle AN, Kern B, Peterli R, Peters T, et al. Important mitochondrial proteins in human omental adipose tissue show reduced expression in obesity. J Proteomics. 2015;124:79–87.

[32] Mardinoglu A, Kampf C, Asplund A, Fagerberg L, Hallström BM, Edlund K, et al. Defining the human adipose tissue proteome to reveal metabolic alterations in obesity. J Proteome Res. 2014;13(11):5106–19.

[33] Vernochet C, Damilano F, Mourier A, Bezy O, Mori MA, Smyth G, et al. Adipose tissue mitochondrial dysfunction triggers a lipodystrophic syndrome with insulin resistance, hepatosteatosis, and cardiovascular complications. FASEB J. 2014;28(10):4408–19.

[34] Paglialunga S, Ludzki A, Root-McCaig J, Holloway GP. In adipose tissue, increased mitochondrial emission of reactive oxygen species is important for short-term high-fat diet-induced insulin resistance in mice. Diabetologia. 2015;58(5):1071–80.

[35] Petersen KF, Petersen KF, Befroy D, Dufour S, Dipietro L, Cline GW, et al. Mitochondrial Dysfunction in the Elderly : Possible Role in Insulin. Science (80-). 2007;1140(2003):1140–3.

[36] Sanyal AJ, Campbell-Sargent C, Mirshahi F, Rizzo WB, Contos MJ, Sterling RK, et al. Nonalcoholic steatohepatitis: Association of insulin resistance and mitochondrial abnormalities. Gastroenterology. 2001;120(5):1183–92.

[37] American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 2010; Jan; 33(Supplement 1): S62-S69.

[38] Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50(5):693–8.

[39] Lee MJ, Wu Y, Fried SK. Adipose tissue heterogeneity: Implication of depot differences in adipose tissue for obesity complications. Mol Aspects Med. 2013;34(1):1–11.

[40] Lynes MD, Tseng YH. Deciphering adipose tissue heterogeneity. Ann N Y Acad Sci. 2018;1411(1):5–20.

[41] Kaaman M, Sparks LM, Van Harmelen V, Smith SR, Sjölin E, Dahlman I, et al. Strong association between mitochondrial DNA copy number and lipogenesis in human white adipose tissue. Diabetologia. 2007;50(12):2526–33.

[42] Heinonen S, Buzkova J, Muniandy M, Kaksonen R, Ollikainen M, Ismail K, et al. Impaired mitochondrial biogenesis in adipose tissue in acquired obesity. Diabetes. 2015;64(9):3135–45.

[43] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(5):768.

[44] Holmes M V., Lange LA, Palmer T, Lanktree MB, North KE, Almoguera B, et al. Causal effects of body mass index on cardiometabolic traits and events: A Mendelian randomization analysis. Am J Hum Genet. 2014;94(2):198–208.

[45] Khera A V., Chaffin M, Wade KH, Zahid S, Brancale J, Xia R, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell. 2019;177(3):587-596.e9.

[46] Stancáková A, Javorsky M, Kuulasmaa T, Haffner SM, Kuusisto J, Stančáková A, et al. Changes in Insulin Sensitivity and Insulin Release in Relation to Glycemia and Glucose Tolerance in 6,414 Finnish Men. Diabetes. 2009;58(5):1212–21.

[47] Matsuda M, DeFronzo R. Insulin Sensitivity Indices Obtained Fro m Comparison with the euglycemic insulin clamp. Diabetes Care. 1999;22(9):1462–70.

[48] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.

[49] Ardlie KG, Deluca DS, Segre a. V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 2015;348(6235):648–60.

[50] Granér M, Seppälä-Lindroos A, Rissanen A, Hakkarainen A, Lundbom N, Kaprio J, et al. Epicardial fat, cardiac dimensions, and low-grade inflammation in young adult monozygotic twins discordant for obesity. Am J Cardiol. 2012;109(9):1295–302.

[51] Jukarainen S, Heinonen S, Rämö JT, Rinnankoski-Tuikka R, Rappou E, Tummers M, et al. Obesity is associated with low nad+/sirt pathway expression in adipose tissue of BMI-discordant monozygotic twins. J Clin Endocrinol Metab. 2016;101(1):275–83.

[52] Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Price, A. L. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics. 2015; 47(3), 284–290.

[53] Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50(11):1593–9.

[54] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:1–12.

[55] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.

[56] Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1).

[57] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[58] Dobin A, Davis C a., Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

[59] Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.

[60] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13.

[61] http://broadinstitute.github.io/picard

[62] Zou H, Hastie T. Erratum: Regularization and variable selection via the elastic net (Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320)). J R Stat Soc Ser B Stat Methodol. 2005;67(5):768.

[63] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models. J Stat Software. 2010;33(1):1–3.

Chapter IV

Establishing a causal effect of non-alcoholic fatty liver disease on coronary artery disease

Establishing a causal effect of non-alcoholic fatty liver disease on coronary artery disease

Zong Miao[1,2], Kristina M. Garske[1], Dorota Kaminska[1,3,4], Janet S. Sinsheimer[2,5], Jussi Pihlajamäki[3,6], Päivi Pajukanta[1,2,7]*

[1]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California, USA.

[2]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA, 90095.

3Institute of Public Health and Clinical Nutrition University of Eastern Finland, Kuopio, Finland.

[4]Turku PET Centre, Turku University Hospital, Turku, Finland.

[5]Department of Computational Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California, USA.

[6]Clinical Nutrition and Obesity Center, Kuopio University Hospital, Kuopio, Finland.

[7]Institute for Precision Health, David Geffen School of Medicine at UCLA, Los Angeles, California, USA.

*: Corresponding author:

Päivi Pajukanta, MD, PhD

Professor

Vice Chair

Department of Human Genetics

David Geffen School of Medicine at UCLA

Gonda Center, Room 6335B

695 Charles E. Young Drive South

Los Angeles, California 90095-7088, USA

Email: ppajukanta@mednet.ucla.edu

Abstract:

The prevalence of non-alcoholic fatty liver disease (NAFLD) is rapidly increasing worldwide due to the ongoing obesity epidemic. NAFLD is also emerging as an important risk factor for poor cardiovascular outcomes. Yet, the causal direction between NAFLD and coronary artery disease (CAD) has not been established conclusively. Moreover, steatosis (fat in the liver) and NAFLD have remained grossly underdiagnosed because currently their diagnosis requires cumbersome imaging technologies, such as abdominal magnetic resonance imaging (MRI), and/or invasive liver biopsy. These difficulties in diagnosing NAFLD limit the sample sizes of NAFLD studies and obstruct identifying the causal relationships between NAFLD and CAD. Here we utilized the UK Biobank (UKB) cohort to establish a prediction model that estimates the NAFLD status based on common serum traits and anthropometric measures. After training and testing in two independent groups, we verified that our NAFLD score (NAFLDS) outperformed an established fatty liver index (FLI) and serum ALT in predicting the NAFLD status. Moreover, we imputed the NAFLDS for all individuals in the UKB cohort and used the NAFLDS as a surrogate for the NAFLD status to explore the causal relationship between NAFLD and CAD. We selected the GWAS SNPs of NAFLDS that are also liver-specific *cis*-eQTLs as the candidate instrumental variables (IVs) for Mendelian randomization analysis. By selecting the NAFLDS GWAS SNPs that specifically regulate liver gene expression, we 1) reinforce the possibility of a valid regulatory effect of the selected NAFLDS VIs on NAFLD; and 2) reduce the potential horizontal pleiotropy in which the NAFLDS IVs affect CAD through other pathways not mediated by NAFLD. Using these tissue of origin enriched eQTL IVs, we identified a significant causal effect of NAFLDS on CAD using MR-PRESSO (Effect size = 0.024, p-value = 9.4e-6). We also employed a similar reverse workflow but did not see a causal effect of CAD on NALFD

(p-value = 0.26). In summary, we established a prediction model (NALFDS) that outperforms the currently available fatty liver index and ALT as a predictor of NAFLD in UKB. Using the predicted NAFLDS and tissue-enriched eQTL IVs, we established the significant causal effect of NAFLD on CAD while controlling for potential horizontal pleiotropy in a bi-directional MR analysis.

Introduction:

It is estimated that over 25% of adults worldwide have non-alcoholic fatty liver disease (NAFLD), and an increase in its prevalence has paralleled that of other cardiometabolic disorders, such as obesity and type 2 diabetes (T2D)[1]. The degree of steatosis (fat in the liver) can be measured through different imaging techniques, the most accurate of which is abdominal magnetic resonance imaging (MRI)[2,3]. However, unlike anthropometric measures, such as body mass index (BMI), or biochemical measures, such as serum liver enzymes and lipids levels, abdominal MRI is not readily available, and thus NAFLD may go undiagnosed for years. Thus, NAFLD is likely under-diagnosed due to the relative difficulty in obtaining reliable measures of liver fat. Moreover, NAFLD may progress to non-alcoholic steatohepatitis (NASH) and fibrosis[4,5,6]. However, abdominal MRI cannot identify inflammation, ballooning, or fibrosis, which can only be diagnosed through histological assessment of liver biopsy. These measures are the hallmarks of NASH and considered to be especially problematic for liver disease and poor cardiovascular outcomes[7].

Statistical models that use serum markers as predictors have been employed to predict the NAFLD risk. For example, Giorgio Bedogni et al[8] reported a fatty liver index (FLI) that has been widely applied and verified in various studies[9,10,11]. However, the existing prediction models are usually built on a limited sample size, which restricts the robustness/accuracy of the prediction model. A validation study has shown that FLI did not outperform the simple waist circumference in predicting NAFLD among the general population of northern Iran[12]. Although novel machine learning (ML) methods have also been used in predicting NAFLD in some recent studies[13,14,15], they are still limited by the small sample size and suffer from a potential overfitting problem in certain small population groups. Moreover, the necessity of complex ML methods over simple

regression models is under discussion[16]. The complexity of machine learning models also makes them difficult to be applied to different heterogeneous cohorts. To improve the prediction of NAFLD using serum traits, we utilized the individuals with ICD9 and ICD10 -based NAFLD diagnoses in the extensive UK Biobank (UKB) cohort[17] as the ground truth for the NAFLD status in our modeling. Accordingly, using the large training cohort (n=4,625), we built a robust prediction model of NAFLD and imputed the NAFLD scores (NAFLDS) in the full UKB cohort.

Genome-wide association studies (GWAS) and subsequent fine mapping have been performed for NAFLD to identify causal variants and genes. One study was recently done in 9,677 individuals of European ancestry to identify the GWAS loci associated with NALFD status[18]. However, due to the previously mentioned relative scarcity of abdominal MRI and liver biopsy, identifying risk loci for NAFLD has been slower than with other cardio-metabolic diseases, such as obesity, T2D, or hypercholesterolemia. Given that diagnosing NAFLD and NASH by either imaging or liver histology is not readily available, one alternative method for identifying patients with likely NAFLD is to establish the risk of NAFLD from the correlated clinical traits, such as serum liver enzyme, glucose, and lipid levels. Accordingly, we built the NAFLDS in UKB, used the NAFLDS as the surrogate of NAFLD, and performed a GWAS to powerfully identify the variants associated with NAFLDS in UKB.

The leading cause of death from NAFLD is coronary artery disease (CAD), with an estimated 5-10% of people with NAFLD dying from CAD[19,20]. It is unclear whether the increased risk of CAD mortality in NAFLD patients is due to other metabolic traits known to be linked to CAD and correlated with NAFLD (e.g. dyslipidemia, T2D, or obesity), and thus the causal direction between NAFLD and CAD has remained elusive[19,21,22,23]. For example, two large studies with long-term follow-up failed to confirm the causal link between NAFLD and

CAD[24,25]. It is important to establish which CAD risk factors are causal because therapeutic interventions should be targeted to the causal risk factors. This is highlighted by the previous Mendelian Randomization (MR)-based discoveries, demonstrating that increased LDL cholesterol (LDL-C) is causal for CAD, whereas decreased HDL cholesterol (HDL-C) is not[26]. Here, to disentangle the causal relationship between NAFLD and CAD, we present evidence that genetically determined liver health (measured through the imputed NAFLD scores) is causal for CAD in a bi-directional MR analysis.

Results:

Estimating NALFD score in UKB cohort:

To impute the NAFLD status using available traits in UKB, we first identified the NAFLD patients and healthy individuals using the same ICD9 and ICD10 codes as employed in several previous large administrative data-based studies of NAFLD prevalence and incidence[27,28,29,30,31,32], and MRI data as the ground truth. We identified 2,181 NAFLD patients by their ICD9/10 codes (see details in Methods) and treated 2,444 individuals with a non-steatotic liver, verified by MRI (liver fat percent < 5%), as healthy controls. Then we randomly selected 3,700 individuals (80%) from the identified individuals as the training group and the remaining 925 individuals as the testing group. We selected 14 NAFLD related traits, including age, BMI, liver enzymes, and blood glucose/lipid traits as predictors using LASSO[33] to build a prediction model that can best predict the NAFLD status in the training data. Then we applied the prediction model to the testing group and compared the predicted NAFLD score (NAFLDS) to ALT and fatty liver index (FLI)[8] on their performance of predicting NAFLD. Figure 1A shows that NAFLDS outperformed FLI in predicting the NAFLD status in the testing group and achieved the highest AUC in a ROC curve.

To investigate the relative importance of the different predictors, we also applied a random forest model to the same training/testing groups. GGT, waist circumference, and BMI ranked high and were identified as the most important predictors. On the contrary, the diabetic traits, such as HbA1c and T2D, were less important predictors. Thus, we trained another prediction model that only relies on the liver enzymes and anthropometric measures, i.e. ALT, AST, GGT, AST/ALT, waist circumference, sex, age, age^2, and BMI. The simplified prediction model (NAFLDS_simple) also outperformed FLI and any predictor alone in the testing group (Figure

1B). Thus, when all predictors in the NAFLDS model are not available, the NAFLDS_simple can be employed to obtain a similar prediction power on the NAFLD status. Table 1 shows the estimated betas of both NAFLDS and NAFLDS_simple. Since the NAFLDS model is shown to accurately predict NAFLD in the testing group, we trained the prediction model using all the 4,625 individuals who have the ground truth and applied the trained model to the full UKB cohort. The predicted NAFLDS was then used as a surrogate for the NAFLD status in our downstream analysis.

NAFLD exhibits a causal effect on CAD:

To determine whether there is a causal effect between NAFLD and CAD risk, we performed a bi-directional MR analysis using the UKB. MR requires the use of proper instrumental variables (IVs), which are often SNPs that are known to significantly contribute to the exposure (GWAS SNPs). In the UKB cohort, we first predicted the NAFLDS score as a surrogate of the NAFLD status. Then we identified 40,918 NAFLDS GWAS variants, which were treated as candidate IVs for the causal analysis of NAFLD on CAD. Similarly, we performed a separate GWAS for CAD to identify the candidate IVs for the causal analysis of CAD on NAFLDS. With only 17,188 CAD individuals in UKB, we identified fewer significant CAD GWAS SNPs (n=841) than in the NAFLDS GWAS. Therefore, we also included the reported known CAD GWAS SNPs from the large Cardiogram meta-study[34] to expand our CAD GWAS SNP pool.

Moreover, the IVs used in an MR analysis should preferably have a known function to provide evidence of lack of horizontal pleiotropy, which is a violation of the MR assumptions[35]. To refine the NAFLDS and CAD GWAS SNPs to those with a plausible function in the liver and coronary arteries, respectively, we determined which of the NAFLDS and CAD GWAS SNPs are *cis* expression quantitative trait loci (*cis*-eQTLs) in their respective tissues. We used RNA-

sequence (RNA-seq) data of 259 liver biopsies from the Kuopio obesity surgery (KOBS) cohort[36] to identify the liver *cis*-eQTLs. We also downloaded the *cis*-eQTLs identified in liver and coronary artery tissue from GTEx v8[37] and excluded *cis*-eQTL SNPs that overlapped between the liver and coronary arteries to avoid using IVs that function in both tissues. In total, 162,293 shared *cis*-eQTLs were identified in both KOBS and GTEx liver cohorts and 464,236 *cis*-eQTLs were identified in the GTEx coronary artery samples. We then obtained our final list of candidate IVs for NAFLDS and CAD by overlapping the respective *cis*-eQTLs with the significant (p<5e-8) NAFLDS or CAD GWAS SNPs.

Figure 2A shows the framework of our MR models. Using our approach described above to obtain the tissue of origin enriched eQTL NAFLDS IVs, we first selected 68 independent SNPs ($R^2 \leq 0.2$) that are associated with NAFLDS in the UKB and are liver, but not coronary artery, *cis*-eQTLs (FDR<0.05). We identified a significant positive causal effect (beta = 0.024, p-value=9.4e-6) of NAFLDS on CAD in the UKB using MR-PRESSO[35] that corrects for potential horizontal pleiotropy in the MR analysis. Figure 2B,C shows the association between the estimated effects (reported by MR-PRESSO) of the IVs on NAFLDS and CAD when treating NAFLDS as the exposure variable. We also tested the potential reverse causal effect of CAD on NAFLDS. We identified 18 independent SNPs ($R^2 \leq 0.2$) that are both CAD GWAS SNPs and coronary artery, but not liver, *cis*-eQTLs. Figure 2C shows the effects of the IVs on CAD and NAFLDS when treating CAD as the exposure variable. Using MR-PRESSO to correct for the potential horizontal pleiotropy, we did not find a significant causal effect of CAD on NAFLDS (beta=0.33, p-value=0.26). The smaller number of IVs (n=18) in the CAD->NAFLDS MR analysis when compared to the NAFLDS->CAD MR analysis (n=68) indicates that we had less power to detect the potential reverse causal effect. However, when we limited the IVs of the

NAFLDS -> CAD MR analysis to the same number of IVs (n=18), we still observed a significant causal effect of NAFLDS on CAD (beta = 0.021, p-value = 0.0064). In summary, we identified the IVs for the MR analyses using the large UKB cohort for GWAS SNPs of both NAFLDS and CAD, and refined these IVs to those with functional evidence in their respective tissues. Our bi-directional MR analysis suggests that NAFLD causally increases the risk of CAD and did not identify evidence of reverse causality (CAD causing increased NAFLDS).

Discussion:

We used the extensive UKB cohort to develop a reliable prediction model of NAFLD. By combining the relevant serum traits (i.e. liver enzymes, lipids (triglycerides, cholesterol), diabetes-related traits (HbA1c, T2D status), age, sex, waist circumference, and BMI), our NAFLDS model achieved a high prediction accuracy on NAFLD (AUC = 0.9) and outperformed the existing FLI[8] index and serum GGT levels. Since the predictors are non-independent traits, the estimated betas cannot be directly used to infer the importance of the predictors in NALFD. Thus, we also employed a random forest method to predict NAFLD with the same predictors, and identified that GGT, waist circumference, and BMI are the most important predictors for NAFLD. In the NAFLD model, we did not include serum glucose levels since the UKB participants are not under fasting, which can heavily bias the serum glucose level. It is suggested that serum glucose and lipid levels are independent predictors for NAFLD[8,38], and that GGT is the only liver enzyme that is an independent predictor for NAFLD[8]. However, using only the anthropometric measures and liver enzymes, our NAFLDS_simple model achieved a similar power in predicting NAFLD status as our NAFLDS model that also utilized the lipid and glucose traits. Thus, despite the studies that shows weak associations between ALT/AST and NAFLD[8,39], our NAFLDS_simple model reinstates the important role of liver enzymes in predicting the NAFLD status.

It is difficult to distinguish the specific contribution of NAFLD on CAD from the other risk factors that are shared by NAFLD and CAD[23]. For example, obesity is a known risk factor for both NAFLD[40] and CAD[41]. Thus, it is important to avoid the potential horizontal pleiotropy in the Mendelian randomization analysis when investigating the causal relationships between

NAFLD and CAD. Here, we included BMI as a covariate to identify the GWAS variants that are associated with NAFLDS/CAD without being mediated by the obesity status (BMI). Furthermore, we combined the GWAS variants and the tissue-enriched *cis*-eQTLs to identify the GWAS SNPs that affect gene expression specifically in the liver or coronary arteries. The overlapped tissue-enriched *cis*-eQTL GWAS variants could thus exhibit a direct causal role in the development of NAFLD/CAD. Using the tissue-enriched *cis*-eQTL GWAS SNPs as IVs and applying MR-PRESSO will reduce the potential pleiotropy and thus improve the reliability of the MR analysis.

In summary, we used key clinical metabolic measurements to build a novel NAFLDS model that is easy to employ and outperforms the existing NAFLD prediction model in the UKB cohort. When some serum traits, such as HbA1c, triglycerides, and cholesterol, are not available, our NAFLDS_simple model can be used to predict the NAFLD status in a similar accuracy as the full NAFLDS model. Furthermore, we combined the GWAS variants and tissue-enriched *cis*-eQTLs to identify the GWAS SNPs that affect gene expression specifically in the liver or coronary arteries. Using these tissue-enriched *cis*-eQTL GWAS SNPs as IVs and applying MR-PRESSO to avoid the potential pleiotropy, we identified the one-way causal role of NAFLD on CAD.

Methods:

Materials and data cohorts:

This research has been conducted using the UK Biobank Resource under Application Number 33934[17]. The GTEx coronary artery *cis*-eQTL results were obtained from the GTEx portal in the version of dbGaP Accession phs000424.v8.p2[37]. The KOBS cohort was recruited at the University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland. We collected the liver biopsies from 259 participants who underwent bariatric surgery. Details about the KOBS sample collection and clinical measurements have been described previously[36,42].

Predicting the NAFLD score in UKB:

In the UKB cohort, the NAFLD cases were identified using the following ICD9/10 codes: 571.5, 571.8, 571.9, K74.0, K74.6, K75.8, K76.0, similarly as in previous large administrative data-based studies of NAFLD prevalence and incidence[27-32]. We selected the individuals who have a liver fat percent < 5%, assessed by abdominal MRI, as the healthy controls. To predict the NAFLD status using the other traits, we employed the elastic net regularization[43] to predict the effect sizes (β) for each variable. We used the 'glmnet' package to obtain the λ that has the minimum mean cross-validated error in the training data set, and then used the specified λ and βs to predict the NAFLDS. Table 1 shows the predictors that were included in the NAFLDS prediction model. To evaluate the prediction accuracy, we performed an 80/20 verification in the UKB cohort. In more detail, we randomly split the individuals into 2 groups so that the training group contained 80% of the individuals and the remaining 20% of the individuals were included in the testing group. In the testing group, we compared the predicted NAFLDS, FLI, and GGT in

predicting the NAFLD status using a ROC curve. To impute the final NAFLDS in the full UKB cohort, we trained the NAFLDS model using all the individuals who have a ground truth (combining both the training and testing group) and then applied the prediction model to the full UKB cohort. The predicted NAFLDS was used as the surrogate for NAFLD status in our following GWAS analysis.

GWAS analysis:

We used a linear mixed model implemented by BOLT-LMM[44] to identify the associations between the genetic variants and the selected traits (NAFLDS, and CAD) while taking the population structure in the UKB into account. The CAD patients were identified using the ICD9 and ICD10 codes, as described by Amit et al[45]. The NALFDS were inverse normal transformed to maintain a normal distribution. We also included age, age^2, sex, BMI, top 20 genotype PCs, array type, and center ID as covariates. To avoid the potential bias from different population structures, only the unrelated Caucasian participants were included in the analysis. Moreover, we excluded the individuals with a liver disease other than NAFLD from the GWAS analysis using the following ICD9/ICD10 codes: 571.1-4, 571.6, 572.0, 572.8, 573.3, 573.8-9, K70.0-4, K70.9, K71.0-2, K71.5-9, K72.0-1, K72.9, K73.0-2, K73.8-9, K74.1-5, K75.0, K75.2-4, K75.9, K76.1-3, K76.6-9, and K77.0. Overall, 388,253 individuals were included in the CAD GWAS analysis and 316,216 individuals were included in the NAFLDS GWAS analysis. For any un-mentioned parameters, we followed the suggestions from the BOLT-LMM manual.

Mendelian randomization analysis:

Using the summary statistics that we obtained from our GWAS analysis, we explored the causal relationship of NAFLDS <-> CAD. We first overlapped the KOBS/GTEx liver *cis*-eQTLs and the GTEx coronary artery *cis*-eQTLs and filtered out the shared SNPs that might affect both the liver and coronary arteries. The *cis*-eQTLs that only exist in one of these tissues were identified as the tissue-enriched *cis*-eQTLs. When using NAFLDS as the exposure variable, we overlapped the significant NAFLDS GWAS variants with both the KOBS and GTEx liver-enriched *cis*-eQTL variants to identify the SNPs that most likely affect the liver health status, reflected by the NAFLDS status. Then we LD pruned (R2 = 0.2) the overlapped SNPs and treated the independent SNPs as instrumental variables (IV). When testing the causal effect of CAD on NAFLDS, we included both the UKB CAD GWAS SNPs and the CARDIoGRAMplusC4D CAD GWAS SNPs[34] as the candidate IVs and overlapped these GWAS SNPs with GTEx coronary artery-enriched *cis*-eQTLs. We LD pruned (R2 = 0.2) the GWAS *cis*-eQTLs and treated the independent GWAS SNPs as IVs. Next, we used MR-PRESSO[35] to correct for potential horizontal pleiotropy and test for the causal effects between NAFLDS and CAD in both directions.

Tables:

Table IV-1. Betas estimated in NAFLDS model.

Predictors	NAFLDS	NAFLDS_simple
GGT	0.013788	0.014916
BMI	0.039514	0.200233
waist	0.06058	-
ALT	0.008904	0.014813
AST	0.037278	0.032677
HbA1c	0.035956	-
AST/ALT	-0.12989	-0.20661
TG	0.349888	-
Cholesterol	-0.28504	-
Albumin	-0.00351	-
age	-0.14701	-0.15243
age2	0.001531	0.001715
sex	-1.02516	-0.21388
T2D	0.412348	-

The predictors are ranked by their importance in the random forest prediction model.

Figures:

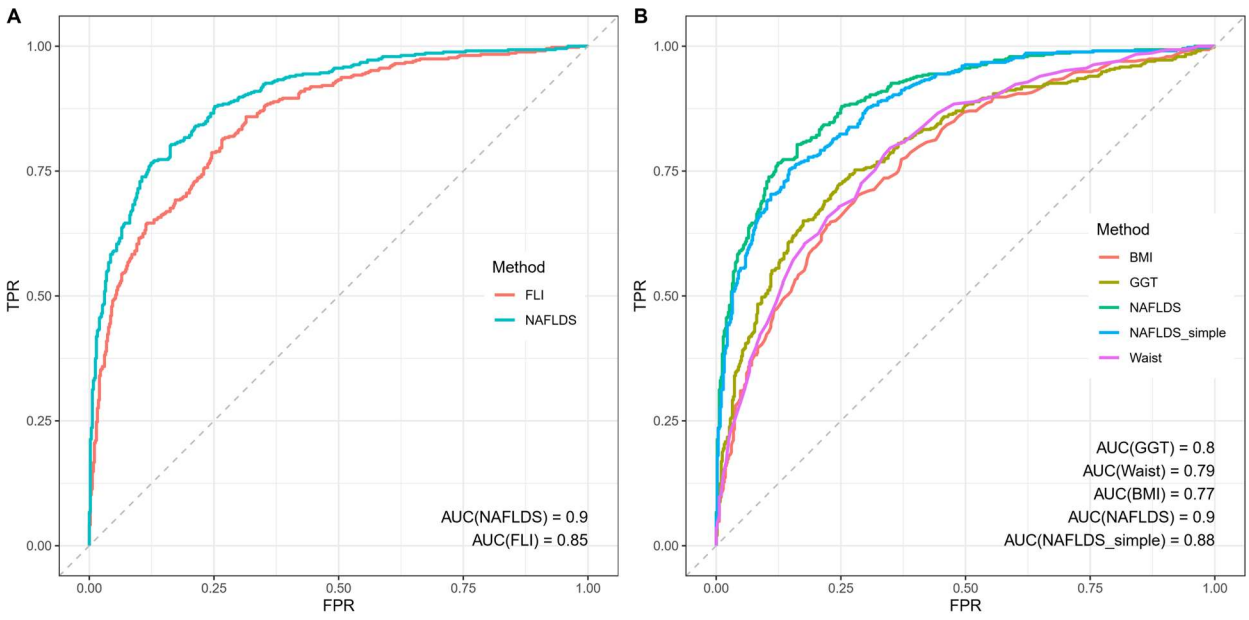


Figure IV-1. ROC plots show that NAFLDS outperforms the existing NAFLD predictors. A). Validated by a ROC curve, NAFLDS outperforms FLI and achieves a higher AUC. B) The NAFLDS_simple model obtains a similar prediction power on NALFD as NAFLDS. Both NAFLDS and NAFLDS_simple outperform any predictor alone. The top 3 predictors GGT, BMI, and waist do not have a large difference in predicting NAFLD.

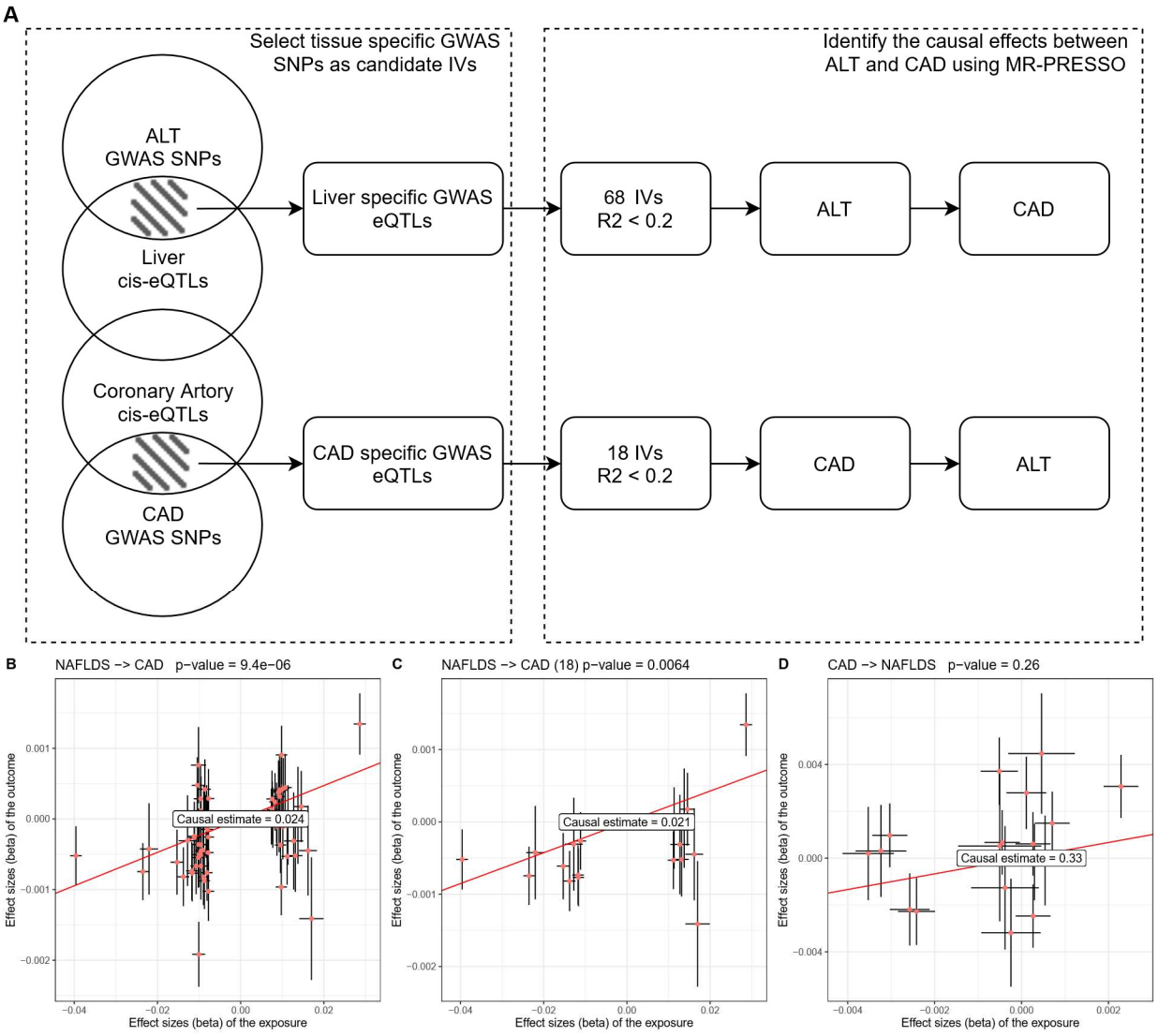


Figure IV-2. MR analysis shows the causal effect of NAFLDS on CAD. A) Workflow of combining liver/coronary artery *cis*-eQTL and UKB GWAS variants to perform a bi-directional MR between NAFLDS and CAD. B) The estimated effects of IVs on NAFLDS (exposure) and CAD (outcome). C) When we downsampled the NAFLDS -> CAD IVs to 18, we still observed a significant causal effect of NAFLDS on CAD. D) There is no evidence supporting the causal effect of CAD on NAFLDS.

References:

[1] Araújo, A. R., Rosso, N., Bedogni, G., Tiribelli, C., & Bellentani, S. (2018). Global epidemiology of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: What we need in the future. *Liver International*, *38*(November 2017), 47–51. https://doi.org/10.1111/liv.13643

[2] Cowin, G. J., Jonsson, J. R., Bauer, J. D., Ash, S., Ali, A., Osland, E. J., … Galloway, G. J. (2008). Magnetic resonance imaging and spectroscopy for monitoring liver steatosis. *Journal of Magnetic Resonance Imaging*, *28*(4), 937–945. https://doi.org/10.1002/jmri.21542

[3] Imajo, K., Kessoku, T., Honda, Y., Tomeno, W., Ogawa, Y., Mawatari, H., … Nakajima, A. (2016). Magnetic Resonance Imaging More Accurately Classifies Steatosis and Fibrosis in Patients with Nonalcoholic Fatty Liver Disease Than Transient Elastography. *Gastroenterology*, *150*(3), 626-637.e7. https://doi.org/10.1053/j.gastro.2015.11.048

[4] Younossi, Z., Anstee, Q. M., Marietti, M., Hardy, T., Henry, L., Eslam, M., … Bugianesi, E. (2017). Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nature Publishing Group*, *14*(1), 11–20. https://doi.org/10.1038/nrgastro.2017.109

[5] Ong, J. P., & Younossi, Z. M. (2007). Epidemiology and Natural History of NAFLD and NASH. *Clinics in Liver Disease*, *11*(1), 1–16. https://doi.org/10.1016/j.cld.2007.02.009

[6] Hashimoto, E., Taniai, M., & Tokushige, K. (2013). Characteristics and diagnosis of NAFLD/NASH. *Journal of Gastroenterology and Hepatology (Australia)*, *28*(S4), 64–70. https://doi.org/10.1111/jgh.12271

[7] Hagström, H., Nasr, P., Ekstedt, M., Hammar, U., Stål, P., Hultcrantz, R., & Kechagias, S. (2017). Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. *Journal of Hepatology*, *67*(6), 1265–1273. https://doi.org/10.1016/j.jhep.2017.07.027

[8] Bedogni, G., Bellentani, S., Miglioli, L., Masutti, F., Passalacqua, M., Castiglione, A., & Tiribelli, C. (2006). The fatty liver index: A simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterology*, *6*, 1–7. https://doi.org/10.1186/1471-230X-6-33

[9] Cuthbertson, D. J., Weickert, M. O., Lythgoe, D., Sprung, V. S., Dobson, R., Shoajee-Moradie, F., … Kemp, G. J. (2014). External validation of the fatty liver index and lipid accumulation product indices, using 1H-magnetic resonance spectroscopy, to identify hepatic steatosis in healthy controls and obese, insulin-resistant individuals. *European Journal of Endocrinology*, *171*(5), 561–569. https://doi.org/10.1530/EJE-14-0112

[10] Koehler, E. M., Schouten, J. N. L., Hansen, B. E., Hofman, A., Stricker, B. H., & Janssen, H. L. A. (2013). External Validation of the Fatty Liver Index for Identifying Nonalcoholic Fatty Liver Disease in a Population-based Study. *Clinical Gastroenterology and Hepatology*, *11*(9), 1201–1204. https://doi.org/10.1016/j.cgh.2012.12.031

[11] Koehler, E. M., Schouten, J. N. L., Hansen, B. E., Hofman, A., Stricker, B. H., & Janssen, H. L. A. (2013). External Validation of the Fatty Liver Index for Identifying Nonalcoholic Fatty Liver Disease in a Population-based Study. *Clinical Gastroenterology and Hepatology*, *11*(9), 1201–1204. https://doi.org/10.1016/j.cgh.2012.12.031

[12] Motamed, N., Sohrabi, M., Ajdarkosh, H., Hemmasi, G., Maadi, M., Sayeedian, F. S., … Zamani, F. (2016). Fatty liver index vs waist circumference for predicting non-alcoholic fatty liver disease. *World Journal of Gastroenterology*, *22*(10), 3023–3030. https://doi.org/10.3748/wjg.v22.i10.3023

[13] Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. (Alex), Poly, T. N., … (Jack) Li, Y. C. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*(March), 23–29. https://doi.org/10.1016/j.cmpb.2018.12.032

[14] Canbay, A., Kälsch, J., Neumann, U., Rau, M., Hohenester, S., Baba, H. A., … Sowa, J. P. (2019). Non-invasive assessment of NAFLD as systemic disease—A machine learning perspective. *PLoS ONE*, *14*(3), 1–15. https://doi.org/10.1371/journal.pone.0214436

[15] Sowa, J. P., Heider, D., Bechmann, L. P., Gerken, G., Hoffmann, D., & Canbay, A. (2013). Novel Algorithm for Non-Invasive Assessment of Fibrosis in NAFLD. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0062439

[16] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*(February), 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[17] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

[18] Namjou, B., Lingren, T., Huang, Y., Parameswaran, S., Cobb, B. L., Stanaway, I. B., … Harley, J. B. (2019). GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network. *BMC Medicine*, *17*(1), 1–19. https://doi.org/10.1186/s12916-019-1364-z

[19] Brouwers, M. C. G. J., Simons, N., Stehouwer, C. D. A., Koek, G. H., Schaper, N. C., & Isaacs, A. (2019). Relationship Between Nonalcoholic Fatty Liver Disease Susceptibility Genes and Coronary Artery Disease. *Hepatology Communications*, *3*(4), 587–596. https://doi.org/10.1002/hep4.1319

[20] Wong, V. W. S., Wong, G. L. H., Yip, G. W. K., Lo, A. O. S., Limquiaco, J., Chu, W. C. W., … Chan, H. L. Y. (2011). Coronary artery disease and cardiovascular outcomes in patients with non-alcoholic fatty liver disease. *Gut*, *60*(12), 1721–1727. https://doi.org/10.1136/gut.2011.242016

[21] Targher, G., Marra, F., & Marchesini, G. (2008). Increased risk of cardiovascular disease in non-alcoholic fatty liver disease: Causal effect or epiphenomenon? *Diabetologia*, *51*(11), 1947–1953. https://doi.org/10.1007/s00125-008-1135-4

[22] Santos, R., Valentic, L., Romeod, S. (2019). Does nonalcoholic fatty liver disease cause cardiovascular disease? Current knowledge and gaps. Atherosclerosis, Vol. 282, 110–120

[23] Francque, S. M., van der Graaff, D., & Kwanten, W. J. (2016). Non-alcoholic fatty liver disease and cardiovascular risk: Pathophysiological mechanisms and implications. *Journal of Hepatology*, *65*(2), 425–443. https://doi.org/10.1016/j.jhep.2016.04.005

[24] Stepanova, M., & Younossi, Z. M. (2012). Independent Association Between Nonalcoholic Fatty Liver Disease and Cardiovascular Disease in the US Population. *Clinical Gastroenterology and Hepatology*, *10*(6), 646–650. https://doi.org/10.1016/j.cgh.2011.12.039

[25] Lazo, M., Hernaez, R., Bonekamp, S., Kamel, I. R., Brancati, F. L., Guallar, E., & Clark, J. M. (2011). Non-alcoholic fatty liver disease and mortality among US adults: Prospective cohort study. *BMJ (Online)*, *343*(7836), 1245. https://doi.org/10.1136/bmj.d6891

[26] Jansen, H., Samani, N. J., & Schunkert, H. (2014). Mendelian randomization studies in coronary artery disease. *European Heart Journal*, *35*(29), 1917–1924. https://doi.org/10.1093/eurheartj/ehu208

[27] Williams VF, Taubman SB, Stahlman S. Non-alcoholic Fatty Liver Disease (NAFLD), Active Component, U.S. Armed Forces, 2000–2017.

[28] Allen AM, Van Houten HK, Sangaralingham LR, Talwalkar JA, McCoy RG. Healthcare cost and utilization in non-alcoholic fatty liver disease: realworld data from a large US claims database. Hepatology. 2018;68:2230-2238.

[29] Jablonski KL, Jovanovich A, Holmen J, et al. Low 25-hydroxyvitamin D level is independently associated with non-alcoholic fatty liver disease. Nutr Metab Cardiovasc Dis NMCD. 2013;23:792-798.

[30] Corey KE, Kartoun U, Zheng H, Shaw SY. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. Dig Dis Sci. 2016;61:913-919.

[31] Wild SH, Walker JJ, Morline JR, et al. Cardiovascular disease, cancer, and mortality among people with type 2 diabetes and alcoholic or nonalcoholic fatty liver disease hospital admission. Diabetes Care. 2018;41:341–347.

[32] Alexander M, Loomis AK, Fairburn-Beech J, et al. Real-world data reveal a diagnostic gap in non-alcoholic fatty liver disease. BMC Medicine. 2018;16:130.

[33] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models. J Stat Software. 2010;33(1):1–3.

[34] Brouwers, M. C. G. J., Simons, N., Stehouwer, C. D. A., Koek, G. H., Schaper, N. C., & Isaacs, A. (2019). Relationship Between Nonalcoholic Fatty Liver Disease Susceptibility Genes and Coronary Artery Disease. *Hepatology Communications*, *3*(4), 587–596. https://doi.org/10.1002/hep4.1319

[35] Verbanck, M., Chen, C. Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, *50*(5), 693–698. https://doi.org/10.1038/s41588-018-0099-7

[36] Männistö, V. T., Simonen, M., Hyysalo, J., Soininen, P., Kangas, A. J., Kaminska, D., … Pihlajamäki, J. (2015). Ketone body production is differentially altered in steatosis and non-alcoholic steatohepatitis in obese humans. *Liver International*, *35*(7), 1853–1861. https://doi.org/10.1111/liv.12769

[37] Ardlie, K. G., Deluca, D. S., Segre, a. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., … Lockhart, N. C. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

[38] Bugianesi, E., McCullough, A. J., & Marchesini, G. (2005). Insulin resistance: A metabolic pathway to chronic liver disease. *Hepatology*, *42*(5), 987–1000. https://doi.org/10.1002/hep.20920

[39] Bedogni, G., Miglioli, L., Masutti, F., Tiribelli, C., Marchesini, G., & Bellentani, S. (2005). Prevalence of and risk factors for nonalcoholic fatty liver disease: The dionysos nutrition and liver study. *Hepatology*, *42*(1), 44–52. https://doi.org/10.1002/hep.20734

[40] Gross, B., Pawlak, M., Lefebvre, P., & Staels, B. (2017). PPARs in obesity-induced T2DM, dyslipidaemia and NAFLD. *Nature Reviews Endocrinology*, *13*(1), 36–49. https://doi.org/10.1038/nrendo.2016.135

[41] Jahangir, E., De Schutter, A., & Lavie, C. J. (2014). The relationship between obesity and coronary artery disease. *Translational Research*, *164*(4), 336–344. https://doi.org/10.1016/j.trsl.2014.03.010

[42] Benhammou, J.H., Ko, A., Alvarez, M., Kaikkonen, M. U., Rankin, C., Garske, K. M., … Pajukanta P. (2019). Novel Lipid Long Intervening Noncoding RNA, Oligodendrocyte Maturation-Associated Long Intergenic Noncoding RNA, Regulates the Liver Steatosis Gene Stearoyl-Coenzyme A Desaturase As an Enhancer RNA. *Hepatology Communications*, Vol 3 (10), 1356-1372.

[43] Zou H, Hastie T. Erratum: Regularization and variable selection via the elastic net (Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320)). J R Stat Soc Ser B Stat Methodol. 2005;67(5):768.

[44] Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., … Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. https://doi.org/10.1038/ng.3190

[45] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., … Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224.

Chapter V

Discussion and future directions

RNA-sequencing technologies have been rapidly improving in recent years, resulting in an unprecedented opportunity to sequence transcriptomes from multiple human and model organism tissues. For example, the GTEx project[1,2] has been collecting RNA-seq samples from various human tissues, and the recently updated GTEx v8 has processed 17,382 RNA-seq samples from 49 tissue types of 979 individuals[2]. This comprehensive human RNA-seq atlas provides insight into how transcriptional activity acts differently across tissues. For example, LPA is a liver-specifically expressed gene, which encodes an important protein lipoprotein (a) (LP(a)) that has been identified as a risk factor of CAD[3]. Previous studies have shown that serum LP(a) level is associated with NAFLD[4,5]. Our preliminary analysis also showed that the LPA gene expression is significantly associated with the steatosis grade in the KOBS cohort. This suggests that the cross talk between the liver and heart plays an important role in the progression of NAFLD and CAD.

In this thesis, I focused on integrating multi-omics data, including bulk RNA-seq, single nuclei RNA-seq (sn-RNA-seq), and genome-wide genotype data with clinical phenotypes to identify essential variants that regulate obesity-related cardiometabolic disorders, and investigate how the cross talk between the different tissues affects these complex diseases. For example, in chapter III, we discovered that the MT gene expression of subcutaneous adipose tissue plays an important role in systemic insulin resistance, whereas the muscle, which has previously been considered to be the key tissue for insulin resistance[6,7,8,9], did not show a similar association between its MT expression and systemic insulin resistance. This analysis reveals that the multi-tissue cross talk in complex diseases forms a knowledge gap that needs to be more comprehensively and systematically addressed in future studies.

In **chapter II**, we developed a novel tool to fast and accurately count allelic reads in RNA-seq data. Before ASElux was developed, the existing methods in ASE analysis either suffered from severe reference alignment bias or were slow and could not be applied to large RNA-seq cohorts. By applying ASElux to the GTEx cohort, the allelic read count from ASElux showed a significantly better reference/alternative allele balance compared to the traditional pipeline employed by the GTEx project[2]. Due to the limited number of heterozygous exonic SNPs, ASE analysis is not able to identify as many SNPs as *cis*-eQTL analysis. However, since the reads from both alleles are processed equally in the RNA-seq data, ASE analysis is less likely to be biased by the usual RNA-seq bias sources, such as RIN values and technical factors. Thus, ASE analysis provides an optimal tool to fine map *cis*-eQTL and GWAS variants. For example, in chapter II, I used ASElux to identify a splice-QTL, rs11078928 that is in tight LD ($R2 > 0.99$) with an asthma GWAS variant rs11078927[10] and an ASE variant. The concordant results between ASE and splice-QTL indicate a potential mechanism of rs11078928 in regulating splicing of the GSDMB gene in the lungs, and thereby predisposing to asthma.

Based on the responses from the current users, I will improve ASElux in the following aspects:

Many users do not have personalized genotype data available for an ASE analysis. Therefore, some users have tried to treat all common SNPs as the reference and then employ ASElux. However, ASElux is not designed to handle all common SNPs as the alignment reference. It makes ASElux either failing or being slow when counting the allelic reads. To address this issue, I propose to combine ASElux with an RNA-seq variant calling tool to first verify the individual's common exonic SNPs and then employ the verified SNPs in calling allelic expression.

ASElux takes the individual's personal SNPs and builds a smaller annotated reference genome to achieve an unbiased alignment. However, this method is limited to SNP array data and cannot be applied to the whole genome sequence (WGS) data due to the limitation of RAM available. Using the current algorithm, ASElux takes ~20GB of RAM to accurately align the exonic reads. However, a large proportion of RNA-seq data consists of intronic reads. If we expand the unbiased reference genome of ASElux to both the intronic and exonic reads, ASElux would require much more RAM, which makes it unpractical to run. Thus, I developed a novel algorithm to directly update the suffix array (i.e. the index system employed by ASElux) based on the small changes (SNPs and INDELs) made to the reference genome. Currently, the new algorithm is able to update the suffix array of the human genome in ~5 minutes when updating the reference genome with ~3 million SNPs. By employing this new algorithm in ASElux, we would be able to detect allelic expression not only in exomes but also in the intronic regions.

ASElux is designed for RNA-seq data. However, since the coverage in the ATAC-seq/ChIP-seq peak regions is usually high, there would be sufficient reads to detect allele-specific binding/open chromatin in DNA-based sequencing data. Similar to ASE analysis, the reference alignment bias in ATAC-seq/ChIP-seq data is also an important issue that needs to be addressed. Thus, I propose to modify ASElux to fit the needs for the alignment of DNA-based sequencing data. The DNA version of ASElux would provide a fast and unbiased allele-specific read counter for additional sequence data types.

In **chapter III**, I first explored the causality between obesity and insulin resistance in the UKB cohort. A previous study has shown that obesity is a causal risk factor for T2D[11]. However, the direction of the causal effect between obesity and insulin resistance remains elusive in humans[12,13]. As the Matsuda index is considered a reliable indicator for systemic insulin

resistance[14], we utilized the Finnish METSIM cohort[15], which is one of the largest cohorts in which the Matsuda index has been measured, in order to identify the genetic variants that are associated with insulin resistance. Combining with the UKB cohort, we first identified the causal effect of obesity on insulin resistance and prediabetes. Next, we further explored the role of adipose tissue in systemic insulin resistance. Muscle, instead of the subcutaneous adipose tissue, is usually considered as the key tissue for systemic insulin resistance. However, in chapter III, we identified the important role of subcutaneous adipose tissue in systemic insulin resistance. We first verified that the overall adipose MT gene expression is significantly associated with BMI and the Matsuda index in the METSIM cohort. Compared to BMI, the Matsuda index has a stronger association with the overall adipose MT gene expression. When combining with the estimated adipose cell-type proportions, the adipose tissue together with BMI explained ~45% of the total variance in insulin resistance. Moreover, we built a prediction model that uses adipose RNA-seq data and BMI to impute the Matsuda index in 3 independent cohorts. Since the Matsuda index is often not available in large cohorts, our imputation model can be applied to estimate the insulin resistance in adipose RNA-seq cohorts, such as GTEx, in future studies.

While investigating the association between adipose tissue and systemic insulin resistance, we utilized single-nuclei RNA-sequencing (sn-RNA-seq) data to explore tissue composition in the subcutaneous adipose tissue. Using subcutaneous adipose sn-RNA-seq data as the reference, we decomposed cell-type proportions based on bulk adipose RNA-seq data. We verified the estimated cell-type proportions in 6 individuals who had both the sn-RNA-seq and bulk RNA-seq data sequenced. We discovered that a large proportion of genes have different expression patterns between bulk and sn-RNA-seq data. Thus, we proposed a simple strategy that first identifies the genes that have similar expression patterns by comparing the sn-RNA-seq and the

bulk RNA-seq data. Then, we only used the concordant genes to estimate the cell-type proportions. Using this straight forward design, we improved the accuracy of estimating adipose cell-type compositions. Furthermore, we developed a method Bisque[16] that first learns the transformation model between sn-RNA-seq and bulk RNA-seq data, and then uses the transformed bulk gene expression to estimate the cell-type proportions. The novel method is shown to outperform the existing decomposition method in both adipose and brain sn-RNA-seq data[16].

We have shown that in the METSIM cohort, the overall MT gene expression and estimated cell-type proportions together explained ~30% of the variance in insulin resistance (measured by a high value of the Matsuda index). The MT gene expression and several estimated cell-type proportions were significantly associated with the Matsuda index in a LASSO model. This suggests that adipose tissue has an important role in systemic insulin resistance. However, we did not observe the same association between the muscle MT gene expression and the Matsuda index in the GTEx cohort. Since muscle is considered to be an important tissue for systemic insulin resistance, the missing link between muscle MT gene expression and Matsuda index warrants further investigation.

Noteworthy, we have shown that the overall adipose MT expression is significantly associated with BMI and insulin resistance in adipose RNA-seq data. Usually, the MT read percent is considered as a technical factor that needs to be adjusted when estimating the gene expression; however, adjusting gene expression for MT read percent in adipose tissue will remove the true signals between certain genes and BMI/insulin resistance. Moreover, since the MT reads contribute to a large proportion of the adipose RNA-seq data, the gene expression and technical factors are heavily biased by BMI in the adipose RNA-seq data. We have shown that

the MT read percent is significantly associated with almost all technical factors in a usual RNA-seq pipeline. Thus, it is important to first exclude the MT reads before estimating gene expression or technical factors, such as 3 prime bias and exonic reads percent. Since we did not observe similar associations between MT reads and other tissues, such as muscle and liver, our study suggests a special caution regarding the MT reads in adipose RNA-seq data.

In **chapter IV**, we employed a new prediction model of NAFLD in the UKB cohort to examine the causality between NAFLD and CAD. First, we built a prediction model that uses multiple metabolic traits (lipids, liver enzymes, and glucose/T2D/HbA1c) and anthropometric measures (BMI and waist circumference). The predictors explained ~40% of the total variance of NAFLD in a logistic model and achieved a high prediction accuracy of AUC = 0.9. The NAFLD score (NAFLDS) outperforms the existing fatty liver predictor (FLI)[17] and any predictor alone. Due to the complex clinical procedures required for diagnosing NAFLD, previous NALFD GWASs have been restricted by a small sample size. To overcome this limitation, we imputed the NAFLDS in the UKB cohort and treated NAFLDS as a surrogate for the NAFLD status to perform a GWAS and identify genetic variants associated with NAFLD. In the UKB cohort, we identified 40,918 NAFLDS GWAS variants, most of which are novel.

To explore the causal relationships between NAFLD and CAD, we utilized tissue-enriched *cis*-eQTLs to reduce the potential horizontal pleiotropy in the MR analysis. Since NAFLD and CAD share many risk factors, such as obesity and hypertension, it is important to adjust for the potential effects of these shared risk factors. Thus, we first identified the NAFLDS GWAS variants that are also liver *cis*-eQTLs in two independent cohorts (GTEx and KOBS) and then excluded any *cis*-eQTLs in the coronary arteries in GTEx. In this way, we selected the *cis*-eQTL NAFLDS GWAS variants that most likely to have a direct effect on NAFLD in the liver. Using

these *cis*-eQTL NAFLDS GWAS variants as IVs, we showed a significant causal effect of NAFLDS on CAD (effect size = 0.024, p-value = 9.4e-6). On the contrary, there was no evidence for a reverse causal effect (CAD -> NAFLDS) using a similar pipeline. Since NAFLDS is a predicted score that consists of various traits, we cannot rule out the possibility that some of the predictors introduce horizontal pleiotropy to the MR. Thus, we also performed a similar bi-direction MR of the liver enzyme ALT <-> CAD. Although ALT has a weaker association with NAFLD than NAFLDS, it is widely used as a surrogate of liver health in many previous studies[18]. We identified the same one-way causal effect of liver health on the heart using ALT and CAD as the target traits. The two MR analysis shows that NAFLD has a significant causal effect on CAD, whereas no reverse causal effect (CAD -> NAFLD) was identified at the same significance level.

In summary, we integrated RNA-seq data from multiple independent human cohorts and tissues, including lung, adipose, liver, and coronary arteries, with deep cardiometabolic phenotype data and other omics data, such as sn-RNA-seq and genome-wide SNP data, to investigate the role of these tissues in obesity-related cardiometabolic disorders. We first developed a fast and reliable tool, ASElux that counts allele-specific expression by combining RNA-seq data and personal level genotype information. The fast speed makes ASElux ideal for performing ASE analysis in large scale RNA-seq cohorts, such as GTEx and METSIM. Then, we used sn-RNA-seq data as a reference and decomposed the cell-type proportions from adipose bulk RNA-seq samples. Using the estimated adipose composition, we showed that adipose tissue plays an important role in systemic insulin resistance. Next, we built a novel model that accurately predicts the NAFLD status in UKB and outperforms the existing NAFLD scoring system. Using the predicted NAFLDS, we discovered a one-way causal effect of NAFLD on CAD. The obesity-related cardiometabolic disorders have become a rapidly developing health

burden worldwide. Our studies suggest that by comprehensively integrating multi-omics data from various human tissues, we will be able to decipher the key causal mechanisms in the development of complex cardiometabolic disorders and save more lives in the near future.

Reference:

[1] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

[2] Ardlie, K. G., Deluca, D. S., Segre, a. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., … Lockhart, N. C. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

[3] Jansen, H., Samani, N. J., & Schunkert, H. (2014). Mendelian randomization studies in coronary artery disease. *European Heart Journal*, *35*(29), 1917–1924. https://doi.org/10.1093/eurheartj/ehu208

[4] Coassin, S., Schönherr, S., Weissensteiner, H., Erhart, G., Forer, L., Lee Losso, J., … Kronenberg, F. (2019). A comprehensive map of single-base polymorphisms in the hypervariable LPA kringle IV type 2 copy number variation region. *Journal of Lipid Research*, *60*(1), 186–199. https://doi.org/10.1194/jlr.M090381

[5] Coassin, S., Erhart, G., Weissensteiner, H., Eca Guimarães De Araújo, M., Lamina, C., Schönherr, S., … Kronenberg, F. (2017). A novel but frequent variant in LPA KIV-2 is associated with a pronounced Lp(a) and cardiovascular risk reduction. *European Heart Journal*, *38*(23), 1823–1831. https://doi.org/10.1093/eurheartj/ehx174

[6] Kraegen, E. W., Clark, P. W., Jenkins, A. B., Daley, E. A., Chisholm, D. J., & Storlien, L. H. (1991). Development of muscle insulin resistance after liver insulin resistance in high-fat-fed rats. *Diabetes*, *40*(11), 1397–1403. https://doi.org/10.2337/diab.40.11.1397

[7] Koves, T. R., Ussher, J. R., Noland, R. C., Slentz, D., Mosedale, M., Ilkayeva, O., … Muoio, D. M. (2008). Mitochondrial Overload and Incomplete Fatty Acid Oxidation Contribute to Skeletal Muscle Insulin Resistance. *Cell Metabolism*, *7*(1), 45–56. https://doi.org/10.1016/j.cmet.2007.10.013

[8] DeFronzo, R. A., & Tripathy, D. (2009). Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes Care*, *32 Suppl 2*. https://doi.org/10.2337/dc09-s302

[9] Petersen, K. F., Dufour, S., Savage, D. B., Bilz, S., Solomon, G., Yonemitsu, S., … Shulman, G. I. (2007). The role of skeletal muscle insulin resistance in the pathogenesis of the metabolic syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(31), 12587–12594. https://doi.org/10.1073/pnas.0705408104

[10] Morrison, F. S., Locke, J. M., Wood, A. R., Tuke, M., Pasko, D., Murray, A., … Harries, L. W. (2013). The splice site variant rs11078928 may be associated with a genotype-dependent alteration in expression of GSDMB transcripts. *BMC Genomics*, *14*(1), 627. https://doi.org/10.1186/1471-2164-14-627

[11] Holmes, M. V., Lange, L. A., Palmer, T., Lanktree, M. B., North, K. E., Almoguera, B., … Keating, B. J. (2014). Causal effects of body mass index on cardiometabolic traits and events: A Mendelian randomization analysis. *American Journal of Human Genetics*, *94*(2), 198–208. https://doi.org/10.1016/j.ajhg.2013.12.014

[12] Noakes TD. So What Comes First: The Obesity or the Insulin Resistance? And Which Is More Important? Clin Chem. 2018;64(1):7–9.

[13] Bluher M. Adipose tissue inflammation: a cause or consequence of obesity-related insulin resistance? Clin Sci. 2016;130(18):1603–14.

[14] Stancáková, A., Javorsky, M., Kuulasmaa, T., Haffner, S. M., Kuusisto, J., Stančáková, A., … Laakso, M. (2009). Changes in Insulin Sensitivity and Insulin Release in Relation to Glycemia and Glucose Tolerance in 6,414 Finnish Men. *Diabetes*, *58*(5), 1212–1221. https://doi.org/10.2337/db08-1607.A.S.

[15] Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A. J., … Boehnke, M. (2017). The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *Journal of Lipid Research*, *58*(3), 481–493. https://doi.org/10.1194/jlr.o072629

[16] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, Eran Halperin (2019). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. BioRxiv. https://doi.org/10.1101/669911

[17] Bedogni, G., Bellentani, S., Miglioli, L., Masutti, F., Passalacqua, M., Castiglione, A., & Tiribelli, C. (2006). The fatty liver index: A simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterology*, *6*, 1–7. https://doi.org/10.1186/1471-230X-6-33

[18] Bedogni, G., Miglioli, L., Masutti, F., Tiribelli, C., Marchesini, G., & Bellentani, S. (2005). Prevalence of and risk factors for nonalcoholic fatty liver disease: The dionysos nutrition and liver study. *Hepatology*, *42*(1), 44–52. https://doi.org/10.1002/hep.20734