THE UNIVERSITY OF CHICAGO

GENETIC ANALYSES OF mRNA N^6-METHYLADENOSINE (m^6A) MODIFICATION AND ITS CONTRIBUTION TO HUMAN DISEASES HERITABILITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY
ZIJIE ZHANG

CHICAGO, ILLINOIS
AUGUST 2020

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF SUPPLEMENTARY FIGURES vii

LIST OF TABLES viii

LIST OF SUPPLEMENTARY TABLES ix

ACKNOWLEDGMENTS x

ABSTRACT xiii

1 CHAPTER 1: INTRODUCTION 1
- 1.1 Overview of the message RNA (mRNA) N^6-methyladenosine (m^6A) modification 1
  - 1.1.1 Regulatory proteins of mammalian m^6A pathway 1
  - 1.1.2 The role of m^6A pathway in biological processes 3
  - 1.1.3 Transcriptome-wide mapping of m^6A 4
- 1.2 Overview of quantitative trait loci (QTL) mapping 5
  - 1.2.1 Biological systems for QTL study 6
  - 1.2.2 Leverage regulatory QTLs to understand genetic architecture of complex traits 8
- 1.3 Dissertation overview 9

2 CHAPTER 2: SYSTEMATIC GENETIC ANALYSIS OF mRNA m^6A METHYLATION 11
- 2.1 Introduction 11
- 2.2 Material and methods 12
  - 2.2.1 Human lymphoblastoid cell line 12
  - 2.2.2 HeLa cell line 13
  - 2.2.3 RNA extraction and m^6A-sequencing 13
  - 2.2.4 Sequence alignment and joint peak calling 14
  - 2.2.5 Genotype data and imputation 15
  - 2.2.6 m^6A QTLs mapping 15
  - 2.2.7 Spatial distribution and genomic annotation of the m^6A QTLs 20
  - 2.2.8 Comparisons between m^6A QTLs vs. eQTLs and sQTLs 20
  - 2.2.9 Annotation of the m^6A QTLs and enrichment analysis 20
  - 2.2.10 Fine mapping m^6A QTLs, eQTLs and sQTLs 22
  - 2.2.11 Evaluating the role of RBP binding in regulating m^6A level 23
  - 2.2.12 Estimating contribution of RNA features and TFs to the m^6A QTLs 24
  - 2.2.13 Joint analysis of transcription-rate QTLs and m^6A QTLs 25
  - 2.2.14 Validating m^6A methyltransferase interactions with TFs 25
  - 2.2.15 Joint analysis of m^6A QTLs and other molecular traits QTLs 26

2.2.16 Re-analysis of ribosome-profiling data of METTL3 and YTHDF1 knockdown in HeLa cells . . . 28
2.2.17 Validating YBX3 function in repressing translation efficiency . . . 29
2.2.18 Heritability and enrichment analysis of GWAS summary statistics using stratified LD score regression (S-LDSC) . . . 30
2.2.19 TWAS and heritability analysis of m^6A peaks . . . 38
2.2.20 Colocalization analysis of m^6A QTLs and GWAS variants . . . 39
2.3 Results . . . 39
2.3.1 Analysis of m^6A peaks in LCLs . . . 39
2.3.2 Mapping genetic variants associated with m^6A . . . 40
2.3.3 m^6A QTLs are enriched in RNA-related features . . . 42
2.3.4 m^6A QTLs' effects on RBP motifs are correlated with their effects on m^6A level . . . 44
2.3.5 m^6A QTLs are enriched in transcription features . . . 45
2.3.6 QTL sharing between m^6A and related molecular traits . . . 48
2.3.7 m^6A-QTLs effect sizes are correlated with related molecular trait QTLs effect size in a context-dependent manner . . . 50
2.3.8 Re-analysis of ribosome-profiling data in METTL3 depleted HeLa cells 50
2.3.9 Validation of YBX3 as a translation repressor for YBX3-bound m^6A-modified transcripts . . . 51
2.3.10 m^6A-QTLs are enriched for GWAS signals . . . 53
2.3.11 Partition of GWAS trait heritability among functional features by S-LDSC . . . 53
2.3.12 Transcriptome-wide association studies (TWAS) using m^6A QTLs as a molecular trait . . . 55
2.4 Supplementary figures . . . 59
2.5 Supplementary tables . . . 72

3 CHAPTER 3: R PACKAGE RADAR FOR DIFFERENTIAL ANALYSIS OF MERIP-SEQ DATA . . . 80
3.1 Introduction . . . 80
3.2 Material and methods . . . 82
3.2.1 Biological samples . . . 82
3.2.2 RNA extraction and m^6A-MeRIP-seq . . . 83
3.2.3 Data preparation . . . 84
3.2.4 Read count pre-processing . . . 84
3.2.5 Adjust IP read count for pre-IP RNA expression level . . . 85
3.2.6 Data filtering . . . 85
3.2.7 Model for DM test . . . 85
3.2.8 Post-processing . . . 87
3.2.9 Simulation analysis . . . 87
3.2.10 Sample size analysis . . . 92
3.3 Results . . . 92
3.3.1 RADAR overcomes challenges in modeling MeRIP-seq data and accommodates complex study designs . . . 92

3.3.2 Comparative benchmarks of different methods using simulated datasets 95
3.3.3 Comparative benchmarks of different methods using four real m^{6}A-seq datasets . . . 99
3.3.4 RADAR analyses of m^{6}A-seq data connect phenotype with m^{6}A-modulated molecular mechanisms . . . 103
3.3.5 Validation of RADAR-detected DM sites by the SELECT method . . 105
3.4 Supplementary figures . . . 108

4 CHAPTER 4: CONCLUSION . . . 120
4.1 Summary and significance . . . 120
4.2 Limitations and future directions . . . 122

APPENDICES . . . 127

A ADDITIONAL m^{6}A-RELATED WORK DURING MY THESIS RESEARCH . . 128

REFERENCES . . . 131

LIST OF FIGURES

2.1 Fitting the overall IP efficiency correction term . . . 17
2.2 Fitting the GC bias correction term . . . 19
2.3 Re-normalize ribosome profiling libraries . . . 29
2.4 Overview of m^{6}A QTL mapping . . . 40
2.5 Mapping common genetic variants associated with m^{6}A . . . 42
2.6 Functional features enriched in m^{6}A QTLs . . . 43
2.7 m^{6}A QTLs may affect m^{6}A level via disrupting RRACH motif . . . 44
2.8 m^{6}A QTL effects on m^{6}A levels correlate with their effects on RBP motifs . . . 45
2.9 m^{6}A installation is coupled with transcriptional processes . . . 46
2.10 Test two models of transcriptional regulation of m^{6}A . . . 47
2.11 Joint analysis of m^{6}A QTLs and other molecular QTLs . . . 49
2.12 Re-analysis of ribosome profiling data in METTL3 knockdown HeLa cells . . . 52
2.13 Integrated analysis of m^{6}A QTLs and GWAS data of human complex traits . . . 54
2.14 m^{6}A-TWAS result summary and comparison with expression-TWAS and splicing-TWAS . . . 55
2.15 m^{6}A-TWAS identify m^{6}A peaks associated with lymphocyte count . . . 56
2.16 Colocalization of m^{6}A QTL and lymphocyte count variant at *DDX55* locus . . . 57

3.1 Unique features of m^{6}A-seq (MeRIP-seq) data . . . 93
3.2 Benchmarking RADAR on two simulation models . . . 96
3.3 Sensitivity of benchmarked methods on real m^{6}A-seq data . . . 100
3.4 Benchmarking false positive signals using permutation analysis on real m^{6}A-seq data. . . . 101
3.5 Pathways enriched in differential methylated genes identified in ovarian cancer and T2D datasets . . . 104
3.6 Experimental validation of RADAR-detected DM sites using SELECT method . 106

4.1 m^{6}A modification mediates the impact of genetic variation on human complex traits . . . 121
4.2 The influence of sample size on the statistical power of differential methylation analysis . . . 124
4.3 Analysis of statistical power and the number of replicates . . . 125

LIST OF SUPPLEMENTARY FIGURES

2.1 Joint m^6A peak calling and QTL mapping 59
2.2 Heritability of m^6A peaks and independence of m^6A QTLs, eQTL and sQTLs . 61
2.3 Contribution of RNA features and transcriptional features to m^6A variation . . 63
2.4 Downstream effects of m^6A are context dependent 65
2.6 Enrichment of GWAS signal in m^6A QTL 67
2.7 Enrichment of complex trait heritability in m^6A QTNs using RNA-features-informed priors 69
2.8 Partitioning complex trait heritability by m^6A QTLs, eQTLs and sQTLs 70
2.9 m^6A-TWAS identifies putative risk genes in human complex traits 71

3.1 Scatter plot of read count 108
3.2 Read count distribution of Input data 109
3.3 Sequencing depth distribution of m^6A-seq in published literatures 110
3.4 Variability distribution comparing RNA-seq and m^6A-seq (MeRIP-seq) 111
3.5 Evaluating performances of benchmarked methods on simulated data using sliding thresholds 112
3.6 P value and effect size estimates on simulated data 113
3.7 Coverage plot of individual samples for example DM sites and bogus sites in the T2D dataset 114
3.8 Compare the methods to adjust for gene expression level 115
3.9 Compare results obtained from shallower sequence depth with that from original depth 116
3.10 Motif analysis and topological distribution of putative DM sites 117
3.11 Coverage plot to visualize differential m^6A peaks in ovarian cancer 118
3.12 Representation of Insulin/IGF1-AKT-PDX1 pathway 119

LIST OF TABLES

2.1 Antibodies used in Co-IP experiment 26
2.2 Summary of GWAS data 30
2.3 Coloc result for m^6A-TWAS peaks in lymphocyte count 58

3.1 Oligo probes sequences and qPCR primer sequences 91
3.2 Selected DM sites for experimental validation 107

LIST OF SUPPLEMENTARY TABLES

2.1 Summary of Torus result for RBP binding sites 72
2.2 Summary of Torus result for microRNA binding sites 77

ACKNOWLEDGMENTS

First of all, I would like to thank my PhD advisor Prof. Chuan He for his great support since I joined the lab. Chuan has provided me with tremendous amount of opportunities to touch base on both technology development as well as basic science discovery. He gave me a lot of flexibility to explore and choose the direction that mostly fits me and provided all necessary resources I needed to accomplish my goals. He suggested me with projects of both high risk ones with more challenges and learning opportunities as well as low risk ones that can cover my back. More importantly, his enthusiasm and sparkling ideas influenced and motivated me to pursue innovative scientific questions. I would not have had such a fruitful PhD experience without the inspiring guidance and support from Chuan.

I would also like to thank Prof. Xin He, who provided a lot of support and guidance to my thesis research. Xin has taught me a lot statistical knowledge and provided me with many ideas to mine the gold from the data. He also encouraged me to not give up when we encountered difficulties during paper submission and this persistence has finally got rewarded. This thesis work would not be accomplished without the support from Xin. Prof. Matthew Stephens, another advisor of this work, has also helped and taught me a lot through out this project. Matthew's statistical expertise are crucial to ensure our analytical inferences are rigorous. When I write up my paper for publication, Matthew has provided a lot of help to convert my draft into a confluent essay with accurate expression of ideas. This process greatly advanced my scientific writing skill and would be extremely useful through out my career. I would also like to thank Prof. Yoav Gilad, a member of my thesis committee. Yoav is an expert in molecular QTL studies and has provided insightful suggestions about my thesis work. He helped me with many important considerations in QTL studies by asking questions and challenging me and thereby lead me to think through key considerations. This method, although made me nervous at the beginning, really helped me to build critical thinking. Additionally, I would like to thank Prof. Mengjie Chen, who has provided phenomenon mentorship in the computational method development project.

With her help, I learned not only useful statistics knowledge, but also turning my codes for data analysis into a well-wrapped package that can be easily reused by everyone else. She also guided me through my first first-author publication, which greatly enhanced my ability to turn research results into a publication.

I would like to give a big thank to Dr. Kaixuan Luo, who worked closely with me on this thesis work. Kaixuan's expertise in statistics is essential for the accomplishment of getting my thesis research published. He has put a lot of effort to help test and re-build our QTL mapping pipeline when our first version of manuscript was rejected in the initial submission. This process greatly improved our manuscript. It was our joint endeavor that paved the way to the final publication of my thesis work. I also need to thank many fellows in our lab and collaborator's labs. Dr. Ian Roundtree worked with me as a senior graduate student when I rotate in the He lab. He has trained me many basic experimental techniques with great patient and trust. Dr. Kai Chen worked with me when I officially joined the He lab. He also provided me with training of basic experimental skills and demonstrated a rigorous and attentive attitude toward research, which helped me to build a good habit in conducting experiment. I also benefited a lot from daily discussion with Dr. Guangzhen Luo, my desk neighbor in the lab. Guanzhen has helped me a lot in problem-solving in my early stages of learning bioinformatics. I need to thank Zhongyu Zou, and Jiakun Tian who provided crucial support in experiments for current work. Dr. Gao Wang has provided valuable support to the statistical fine-mapping part of our work, which is greatly appreciated. Dr. Jean Morrison and Laura Sieh have helped with text editing, which greatly improved the confluence and clarity of our manuscript. There are many other fellows who have lent me great support and provided valuable suggestions, I sincerely thank all of them.

Finally, there is my family. I need to thank my father, who is also a great scientist. It was him who cultivated my curiosity to the unknown world and led me into the door of scientific exploration in biology. He has provided me with many useful suggestions in career planning and encouraged me to overcome difficulties in research activities. I have to also thank my

wife for her great care and support during these years of graduate school. It wasn't an easy decision for her to put away her job in hometown and came over to U.S. to accompany me for the graduate school. She also helped me with experiments in this work when I need extra hands, particularly when dozens of cell lines need to be cultured and hundreds of sequencing libraries need to be processed simultaneously. Of course, I wouldn't came this far without the love and encourage from my mother, grandparents and all members of my extended family. I'm really grateful for all their support, in the past and in the future.

ABSTRACT

N^6-methyladenosine (m^6A) plays a critical role in regulating various aspects of mRNA metabolism and translation in eukaryotes. Despite rapid progress in this field, gaps remain in genomics analysis of the m^6A epitranscriptome. For example, little is known about how DNA sequence variations may affect the m^6A modification and the role of m^6A in common diseases. Besides, a computational method to analyze m^6A-seq data for differential methylation loci that is compatible with complex study design is lacking. In this thesis, we report two major endeavors to answer these questions. First, we mapped Quantitative Trait Loci (QTL) of m^6A peaks in 60 Yoruba lymphoblastoid cell lines. By analyzing these variants, we uncovered features associated with m^6A installation, including binding by specific RNA binding proteins (RBPs), RNA secondary structure, and transcriptional processes. Our joint analysis of QTL data of m^6A and related molecular traits suggests that the downstream effects of m^6A are heterogeneous and context-dependent. We identified new proteins that suppress translation of m^6A-modified transcripts. Integrated analysis with GWAS data shows that m^6A–QTLs are enriched with variants associated with a range of immune and blood related traits, and contribute significantly to the heritability of these traits. Second, we developed RADAR, a comprehensive analytical tool for detecting differentially methylated loci in MeRIP-seq data. RADAR enables accurate identification of altered methylation sites by accommodating variability of pre-immunoprecipitation expression level and post-immunoprecipitation count using different strategies. In addition, it is compatible with complex study design when covariates need to be incorporated in the analysis. Through simulation and real datasets analyses, we show that RADAR leads to more accurate and reproducible differential methylation analysis results than alternatives.

CHAPTER 1

INTRODUCTION

1.1 Overview of the message RNA (mRNA) N^6-methyladenosine (m^6A) modification

1.1.1 Regulatory proteins of mammalian m^6A pathway

Chemical modifications to RNA species have been discovered for 50 years and over 100 types of modifications have been identified in cellular RNAs [1]. Among them, m^6A is the most abundant internal modification on the polyadenylated mRNA and non-coding RNA with approximately 4 m^6A sites per mRNA (the m^6A/A ratio is estimated to be 2-4% on mRNA) in eukaryotes. Levels of m^6A are dynamically regulated by both writers (m^6A methyltransferase complexes) and erasers (m^6A demethylases) [2]. Functional effects of m^6A can be modulated by reader proteins as well as through structure switch mechanism [1].

The first characterized m^6A writer complex is composed of multiple subunits including the methyltransferase-like 3 (METTL3), methyltransferase-like 14 (METTL14), Wilms tumor 1-associating protein (WTAP), Vir like m^6A methyltransferase associated (VIRMA), Zinc finger CCCH-type containing 13 (ZC3H13) and RNA binding motif protein 15/15B (RBM15/15B). Among them, METTL3 is the most core component functioning as the catalytic subunit while METTL14 is a essential component that facilitates the binding of the methyltransferase complex to the RNA [3, 4]. WTAP binds to METTL3/14 and regulates cellular localization of the complex, in addition to its role in promoting substrate recruitment [5]. VIRMA plays a role in directing m^6A deposition in 3'UTR of transcripts [6]. ZC3H13 facilitates nuclear localization of the writer complex [7]. RBM15/15B has been shown to bind to U-riched regions and thus may be responsible for the deposition of m^6A on certain RNAs [8]. Recently, another m^6A writer, methyltransferase-like 16 (METTL16) was discovered with two validated substrates including U6 small nuclear RNA and a hairpin in the

3'UTR of human *MAT2A* mRNA [9].

Two m^6A erasers have been described to date – ALKBH5 [10] and FTO [11, 12]. Indeed, the recent gold rush in the epitranscriptomics field started from the discovery of the m^6A demethylases in the 2010 by our lab [11, 10], which established the concept that RNA m^6A modification is reversible and can be regulated dynamically. FTO is a versatile demethylase that can work on different species of RNA including mRNA, ncRNA and tRNA; and on different base modifications including m^6A, m^6A_m and m^1A [12]. In contrast, ALKBH5 is known to demethylate m^6A on mRNA.

The downstream functions of m^6A are mediated by m^6A readers that preferentially bind m^6A over unmodified A and regulate mRNA processing of the modified transcripts. The most intensively-studied class of reader proteins contain a YT521-B homology (YTH) domain (aka YTH family proteins). These proteins include cytoplasmic reader YTHDF1 that has been shown to promote translation efficiency of modified transcripts by recruitment of translation initiation factors [13]; cytoplasmic reader YTHDF2, which has been shown to accelerate decay of modified transcripts by interacting with the CCR4-NOT deadenylase complex and the endoribo-nuclease complex RNase P/MRP [14, 15]; cytoplasmic reader YTHDF3, which has been suggested to promote translation efficiency of its targets [16] and nuclear reader YTHDC1 that mediates the splicing and expedited nuclear export of its target transcripts by preferably recruiting SRSF3 [17, 18] as well as accelerated nuclear decay of certain transcripts by interacting with nuclear exosome-targeting complex [19]. In addition to YTH protein family, several proteins containing the KH domain were identified to bind m^6A and regulate metabolism of modified transcripts. These proteins include insulin-like growth factor 2 mRNA-binding proteins 1-3 (IGF2BP1/2/3), which enhance the stability of modified nuclear mRNA [20]; and fragile X mental retardation 1 (FMR1), which affects both stability and translation of its targets possibly through interacting with YTHDF1 and YTHDF2 [21]. There is another type of non-canonical readers that response to m^6A through a structure switch mechanism. For example, presence of m^6A can remodel local RNA struc-

ture, hence affecting binding of certian RNA binding proteins (RBPs) such as heterogeneous nuclear ribonucleoproteins (HNRNPs) to the vicinity of m^6A sites, which in turn can regulate alternative splicing or processing of the target pre-mRNAs [22, 23].

1.1.2 The role of m^6A pathway in biological processes

The m^6A mediated regulatory pathways affect many biological processes including development, stress response, immune, and neuronal functions. For example, loss of *Mettl3* in mouse embryonic stem cells impairs proper differentiation as turn over of some important developmental regulators during cell fate transitions are regulated by m^6A [24]. In zebrafish, over one-third of maternal mRNA are m^6A modified and clearance of these maternal mRNA during the maternal-to-zygotic transition is facilitated by YTHDF2-mediated mRNA decay [25].

Under hypoxia stress, hypoxia-inducible factors (HIF-1α and HIF-2α) activate *ALKBH5* (encode m^6A demethylase) expression and thereby stabilize *NANOG* mRNA by inhibition of m^6A mediated mRNA decay [26]. In another study, researchers identified the gene encoding a m^6A reader – YTHDF1 – under positive selection for high-altitude adaptation and contributes to pathogenesis of non-small cell lung cancer (NSCLC) [27]. Another example of m^6A functions in stress response is the installation of m^6A at the 5'UTR of response genes upon heat shock, which in turn mediates the cap-independent translation of these genes [28].

In anti-tumor immunity, YTHDF1 has been shown to promote translation of lysosomal cathepsins in dendritic cells and hence inhibits the cross-presentation of the tumor antigen and the cross-priming of $CD8^+$ T cell *in vivo* [29]. In the innate immunity, m^6A has been reported to regulate main cytokines that drive the type I interferon response [30].

In neuron, m^6A has been shown to promote translation of YTHDF1 targets in response to stimuli, thereby facilitating learning and memory [31]. Loss of *Mettl14* (encode a key component of m^6A writer complex) in mouse brain delays cortical neurogenesis and impairs decay of transcripts involved in lineage specification [32].

1.1.3 Transcriptome-wide mapping of m^6A

Methylated RNA immunoprecipitation sequencing (MeRIP-seq) [33, 34] is the most widely used technique in mRNA modification studies, which enables us to survey the m^6A epi-transcriptome using various study designs: (1) identifying the location of modification on the transcript by performing peak calling on samples of certain phenotype or experimental condition [33, 34, 35] and (2) identifying differentially methylated loci by comparing MeRIP-seq samples across different phenotypical or experimental groups [36, 37, 38]. MeRIP-seq works similarly to ChIP-seq, which enriches m^6A-containing RNA by immunoprecipitation (IP) reaction from fragmented RNA using m^6A-specific antibody. Library constructed from m^6A-IP-eluted RNA is then compared with the input library, which is constructed from the initial RNA fragments pool prior to the IP reaction, to determine the enrichment of m^6A-containing RNA fragments across the transcriptome.

Two enhanced versions of MeRIP-seq were later developed: photo-crosslinking-assisted m^6A sequencing (PA-m^6A-seq) [39] and m^6A individual-nucleotide-resolution cross-linking and immunoprecipitation (miCLIP) [40]. PA-m^6A-seq is combination of MeRIP-seq and photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation sequencing (PAR-CLIP-seq). Specifically, 4-thiouridine (4SU) is incorporated into RNA when 4SU is added to growth medium. After incubating m^6A-antibody with 4SU-containing RNA, 365 nm UV light is used to initiate 4SU-based photo-crosslinking, which is followed by RNase digestion to remove unprotected regions of RNA. This step results in RNA fragments of approximately 30 nucleotide (nt) long, which narrows the m^6A peaks called from PA-m^6A-seq data [39]. miCLIP, identifies putative m^6A sites by mutation and truncation signatures induced by RNA-antibody photo-crosslink. Unlike PA-m^6A-seq, miCLIP doesn't incorporate 4SU to enhance photo-crosslink and uses a 254 nm UV light to trigger the crosslink. This method further increases the resolution of putative m^6A sites to single-base [40]. These two enhanced versions of MeRIP-seq, although improved the resolution of the m^6A sites mapping, have their own limitations. PA-m^6A-seq requires 4SU uptake during cell culture, which

means this method is only applicable to cultured cells while tissue samples are frequently encountered in real studies. Besides, the complicated experimental procedures increase the cost of the experiment (4SU is expensive) and induce more source of technical variations. miCLIP relies on antibody-specific mutation/truncation signatures to determine the m^6A sites. In practice, we find this method not very robust, probably because the signatures observed in the reference [40] is specific to the batch of antibody they used. Conversely, MeRIP-seq is relatively simple to perform and yields robust result, making it the most prevalent method used in the epitranscriptomics field to date.

Identifying differentially methylated loci is the most commonly utility of the transcriptome-wide m^6A mapping. Early studies performing qualitative analysis compared peaks called in one experimental group versus peaks in another group and identified peaks unique to each experimental group as differentially methylated peaks. However, many differential peaks identified by this method are caused by boundary cases at the peak-detection threshold rather than true presence/absence of peaks as noted in recent studies [38, 36]. To enable cross-group comparisons, a few count-based methods have been developed [41, 42, 43, 44, 36]. These methods are based on different statistical models and are compatible with different study designs.

1.2 Overview of quantitative trait loci (QTL) mapping

Human individuals differ from each other by millions of DNA sequence variations (genetic variations). Some of these genetic variations in turn, contribute to phenotypical variations including physiological, morphological variations and predispositions to diseases as revealed by rapidly accumulating Genome-wide association studies (GWAS). However, GWAS results usually cannot tell the mechanism of how genetic variants affect phenotypical variations. QTL mapping emerged as an effective approach to study the mechanism by which genetic variations exert their effects on phenotypical variations [45, 46]. QTL mapping identifies sets of genetic variants that are correlated with a particular molecular trait that is involved

in gene regulation processes such as gene (mRNA) expression level, translation efficiency, protein level and alternative splicing events. Such regulatory variants can be leveraged to understand gene regulation processes [45, 47, 48, 49] as well as genetics architecture of complex traits [46, 50, 51, 52].

1.2.1 Biological systems for QTL study

Several systems have been deployed in QTL studies including cell lines and various types of human tissues. One of the most popular cell line system is the lymphoblastoid cell lines (LCLs) derived from the HapMap and the thousand genomes project [53], particularly the Yoruba cohort. One important benefit of using the LCLs system in QTL studies is the availability of genotype data. Yoruba cohort is African ancestry and harbors more genetic diversity compared to other thousand genomes project populations as African is a founder population, making Yoruba LCLs an ideal system for QTL studies. A series of studies using 60-100 Yoruba LCLs have mapped the *cis*-QTLs of a number of molecular traits along the gene expression regulation cascade from chromatin accessibility to protein level [49, 54, 55, 56, 57, 58, 59, 48, 60, 47]. These studies not only identified *cis*-regulatory variants of each regulatory mechanism, but also provided a chance to investigate the correlation between each regulatory mechanism. For example, many concerted regulatory mechanism changes across genotypes can be linked to sequence alteration in transcription factor (TF) binding sites, suggesting TF binding may underlie general properties of chromatin states [45]. Joint analysis of QTLs along the regulatory cascade revealed that effect sizes are highly correlated between consecutive steps (e.g. transcription rate to expression, expression to ribosome loading, etc.), reflecting a percolation of genetic effects from transcription through the regulatory cascade [47]. Meanwhile, the reduced effect in each step moving along the regulatory cascade (e.g. expression QTL (eQTL) tends to have significantly reduced effects on protein level compared to ribosome loading) suggesting that additional mechanisms exist to buffer the impact from one molecular phenotype to the next [47, 48]. In addition to

the Yoruba cohort, effort has been made to map eQTLs in 462 LCLs across five human population [61], which not only increased the power to identify more eQTLs, but also enabled the investigation of shared and population-specific eQTLs [62].

Recently, induced pluripotent stem cells (iPSCs) emerged as a promising system for QTL studies. Researchers established a panel of iPSCs from 58 intensively-studied Yoruba LCLs [63]. This panel of iPSCs have genotype data readily available as they are genetically identical to the LCLs from which they are derived. Moreover, they are capable of being induced into numerous cell types, enabling mapping of cell type specific QTLs as well as time-resolved regulatory QTLs during cellular differentiation [63, 64].

Another important type of biological system for QTL studies is primary human tissues. Blood is one of the most easy-to-access sample that can be obtained from living donors. This enabled QTL studies of relatively large sample size. For example, an eQTL study using 2,752 peripheral blood samples from twins identified 6,988 high-confidence *cis*-eQTLs and 165 *trans*-eQTLs, which were replicated in another set of unrelated subjects [65]. Another study used $CT4^+$ T cells derived from blood of healthy donors to study the regulatory QTL of activated T cell, which identified thousands of chromatin accessibility QTLs (ATAC QTLs) and co-accessibility QTLs as well as hundreds of eQTLs [66]. Analysis of these regulatory QTLs revealed insights into the mechanism of transcription regulation from TF binding to gene expression in a physiological relevant system under stimulated condition. In addition to blood-related tissue, many regulatory variants that are specific to certain solid tissues may contribute to the etiology of human diseases. However, unlike blood, samples from solid tissues can only be obtained from post-mortem individuals. A consortium effort has been put to coordinate sample collection and analysis of a big number of human solid tissues, known as Genotype-Tissue Expression (GTEx) project [67, 68]. This project has enabled us to explore how one set of DNA sequence could exert different effects on different tissues as well as the shared effects across multiple tissues [69, 70].

1.2.2 Leverage regulatory QTLs to understand genetic architecture of complex traits

Molecular QTLs, are enriched with human complex traits-associated variants, and can be leveraged to identify disease susceptibility variants and genes [50, 71, 72, 73, 74]. Many studies found regulatory QTLs, particularly eQTLs, are enriched with GWAS signals. This enrichment is often the most significant when the cell/tissue type from which the QTLs were mapped matches the complex trait-related tissue [75, 76]. For example, regulatory QTLs mapped in blood-derived immune cells are mostly enriched with GWAS signal of immune-related traits [47, 77, 66] while regulatory QTLs identified using human brain sample show enrichment with GWAS variants mostly in neuropsychiatric disease [35, 78].

GWAS results often identify a region that contains multiple genes and many genetic variants in high linkage disequilibrium (LD) to be associated with a complex trait while the ultimate goal is to uncover the causal gene and genetic variant. Regulatory QTLs can be leveraged to prioritize likely causal variants given the assumption that variants exert their effects on complex traits by altering the regulatory mechanism of pathological genes. Empirical based approaches, e.g. Regulatory Trait Concordance (RTC) [79], have been utilized to prioritize functional variants in GWAS regions as candidate causal variant of GWAS traits until several integrative Bayesian methods have been developed. These methods include eCAVIAR [80], Coloc [81] and Enloc [72] with the former two methods directly assuming a fixed value for the enrichment of molecular QTLs in GWAS variants while the later estimates this parameter from the data. All three methods run model-based colocalization tests to identify putative causal variants of both molecular QTLs and GWAS trait. Another popular approach for integrating regulatory QTLs and GWAS is gene-based association study such as PrediXcan [82] and transcriptome-wide assocaition study (TWAS) [83]. These methods build predictive models of molecular trait from regulatory QTLs, e.g. eQTLs, and test for association between expression level imputed from genotype data and phenotype. The benefits of this type of approach include reduced multiple test burden (increased power) and

easy interpretation for follow studies.

1.3 Dissertation overview

Despite rapid progress in the m^{6}A field, genomics analysis of m^{6}A epitranscriptome, particularly quantitative analysis, has notable gaps. As introduced in previous chapter that genetics of molecular traits across several steps of gene expression regulatory have been characterized. However, how DNA sequence variation could affect m^{6}A has not been explored. Besides, a flexible method to analyzed m^{6}A-seq (MeRIP-seq) data for differentially methylated loci that can compatible with complex study design is not available, despite tremendous amount of m^{6}A-seq has been generated in dozens of studies [84]. In this dissertation, I will present my works on genetics analysis of m^{6}A epitranscriptome using a QTL mapping approach as well as an R package I developed to perform differential methylation analysis on MeRIP-seq data with compatibility to complex study designs.

Chapter 2 presents genetics analyses of human mRNA m^{6}A modification and its contribution to the genetics of complex phenotypes. The work present in this chapter also appears in journal article Zijie Zhang$^{\#}$, Kaixuan Luo$^{\#}$, Zhongyu Zou, Maguanyun Qiu, Jiakun Tian, Laura Sieh, Hailing Shi, Yuxin Zou, Gao Wang, Allen C. Zhu, Min Qiao, Zhongshan Li, Matthew Stephens*, Xin He*, Chuan He*. Genetic Analyses Support the Contribution of mRNA N^6-methyladenosine (m^{6}A) Modification to Human Disease Heritability. *Nature Genetics*, in press.

Chapter 3 presents a new statistical method we developed to analyze MeRIP-seq data for differentially methylated loci. The work present in this chapter also appears in journal article Zijie Zhang, Qi Zhan, Mark Eckert, Allen Zhu, Agnieszka Chryplewicz, Dario F. De Jesus, Decheng Ren, Rohit N. Kulkarni, Ernst Lengyel, Chuan He* & Mengjie Chen*. RADAR: differential analysis of MeRIP-seq data with a random effect model. *Genome Biology* 20, 294 (2019).

In the appendix, I attached abstracts of three additional publications in which I am co-

first author. In these works, I applied epitranscriptomic analysis techniques on m^6A to study the role of mRNA m^6A modification in acquired drug resistance in leukemia cells, type 2 diabetic islets as well as human metapneumovirus (RNA virus) with collaborators from other institutes. Of note, the challenges of lacking a statistical method for differential methylation analysis compatible with complex study design I encountered during the epitranscriptomic analysis of type 2 diabetic islets was a key factor that motivated me to develop the R package RADAR presenting in Chapter 3.

CHAPTER 2

SYSTEMATIC GENETIC ANALYSIS OF mRNA m^6A METHYLATION

2.1 Introduction

Despite rapid progress in the m^6A field, our understanding of m^6A regulation and function has notable gaps. Among all adenosine sites on mRNA, only a small fraction is m^6A modified and we know little about what factors control this specificity. *De novo* motif analysis on current m^6A-seq data as well as CLIP-seq data of METTL3/14 revealed a consensus motif – RRACH. However, RRACH motif is wide-spread across the transcriptome and only a relative small percentage of RRACH motifs on transcripts are modified. This suggests that additional sequence features are involved in m^6A deposition specificity. The downstream functions of m^6A are believed to depend on m^6A reader proteins. However, we have limited understanding of how RNA sequence contexts may affect the recognition of m^6A by readers as well as downstream effects. Indeed, our list of m^6A reader proteins may well be incomplete [21]. At the phenotypic level, dysregulation of m^6A has been implicated in cancer progression [85, 86, 87, 88, 89]. However, we know very little about how m^6A variation may contribute to other common human diseases.

To fill these gaps, we took a genetic approach based on mapping variants associated with m^6A levels in mRNA transcripts, or m^6A quantitative trait loci (m^6A QTLs). QTL mapping of molecular traits has provided unique insights into gene regulation and facilitated identification of susceptibility variants and genes from GWAS data as introduced in above section. We mapped m^6A QTLs using a well-characterized cohort of Yoruba LCLs, for which QTL data of multiple molecular traits are available [55, 54, 59, 49, 56, 58, 57, 48]. We found that the m^6A consensus motif (RRACH), while highly enriched, explains only a small fraction of m^6A QTLs. We observed that m^6A QTLs are enriched in RBP target sites, RiboSNitches (variants affecting RNA secondary structure) and transcriptional features,

suggesting that these factors are important regulators of m^6A installation. By integrating with other molecular QTL data, we found that regulatory effects of m^6A on downstream traits such as translation likely vary across m^6A sites in a context-dependent manner.

We conducted joint analysis of m^6A QTLs and genome-wide association studies (GWAS) data. Current efforts to characterize GWAS variants have largely focused on transcriptional effects. However, recent studies, employing different approaches from colocalization to heritability analyses, estimate that eQTLs explain only 10-25% of GWAS signals [50, 90, 91]. To fill this gap, researches have suggested other mechanisms such as RNA splicing [47, 35]. In our analysis, we found that m^6A QTLs are enriched for risk variants of a range of complex traits, particularly autoimmune diseases and blood-cell-related traits. The contribution of m^6A QTLs to heritability of these traits is roughly half of eQTLs and comparable to splicing QTLs (sQTLs) mapped in the same cohort LCLs. Treating m^6A level as molecular traits, we performed TWAS [83] of these traits and identified a number of m^6A sites and genes. Taken together, our results demonstrate that m^6A variation is an important link between genetic and phenotypic variations.

2.2 Material and methods

2.2.1 Human lymphoblastoid cell line

Human lymphoblastoid cell line (LCL) of 60 Yoruba individuals were purchased from Coriell Institute (cells were at 3rd - 5th passages when received). These 60 individuals were chosen by the availability of other molecular QTLs data in previous studies that can be used for integrated analysis with m^6A QTLs. These regulatory QTLs data include transcription rate QTLs, eQTLs, decay QTLs, ribosome loading QTLs (ribo-seq QTLs) and protein QTLs [55, 59, 57, 48]. Upon receiving, cells were split into flasks as technical replicates and were processed independently thereafter. Cells were cultured and propagated in RPMI 1640 medium with 15% FBS at 37°C and 5% CO2 until harvest.

2.2.2 HeLa cell line

Human HeLa cell line used in this study was purchased from ATCC (CCL-2) and grown in DMEM (Gibco, 11995) media supplemented with 10% FBS and 1% 100 × Pen/Strep (Gibco) at 37°C and 5% CO2 until harvest. Transfection was achieved by using Lipofectamine RNAiMAX (Invitrogen) for siRNAs.

2.2.3 RNA extraction and m^6A-sequencing

Cells were harvested by 1000× g centrifuge. Total RNA was extracted from cell pellets using TRIzol (Invitrogen) and Direct-Zol RNA extraction kit (Zymo Research cat. R2072) according to the manufacturer's instruction. mRNA was further purified with Dynabeads mRNA DIRECT purification kit (Thermo Fisher, cat. 61011) by robot (KingFisher Duo Prime System, ThermoFisher Scientific) for 12 samples at a time. mRNA was adjusted to 15ng/μl in 100 μl and fragmented using Bioruptor ultrasonicator (Diagenode) with 30s on/off for 30 cycles.

Approximately 50 ng of fragmented mRNA was saved as input sample and 1,450 ng was subject to m^6A-immunoprecipitation (m^6A-IP) with EpiMark N^6-Methyladenosine enrichment kit (NEB cat. E1610S) according to manufacturer's protocol. The selection of this antibody is based on a benchmark of several available commercial m^6A antibodies (data not shown), in which we found NEB antibody giving the highest fold enrichment of m^6A peaks. To minimize the variation due to IP experiment, which is often a great source of technical noise in IP-based sequencing, m^6A-IP was performed by robot (KingFisher Duo Prime System, ThermoFisher Scientific) for 12 samples at a time. Though a monoclonal antibody was used, we further controlled for lot variation by pooling several tubes of antibody prior to aliquot to each of the 60 samples.

RNA eluted from m^6A-IP was cleaned using RNA Clean and Concentrator (Zymo Research, cat. R1013). Input and IP samples were then used to prepare library with KAPA mRNA Hyper Kit (Roche, Cat. KK8541). A total of 240 libraries (duplicates per individual,

each with an input and IP) were constructed in three batches. All libraries were sequenced by the HiSeq4000 platform at SE50 mode at the sequencing core facility at the University of Chicago. For each batch of library constructed, all libraries (with distinct index) were pooled and sequenced at a lane together for 3-5 repetitive lanes. This study design balanced the lane effect on each batch of library. In sum, approximately 30 million reads were obtained for each library and reads from technical replicates were pooled to result in 60 million reads for each input and IP sample per individual.

2.2.4 Sequence alignment and joint peak calling

For each dataset, the raw sequencing data were mapped to the hg19 reference genome by Hisat2 [92] with parameter –known-splicesite-infile ⟨splice-file extracted from Refseq hg.19 GTF file⟩ -k 1. We used WASP [93] to control for the alignment bias due to genetic variations. The BAM files obtained from alignment are used as an input file for reads quantification.

To call m^{6}A peaks jointly across samples, we first divided genes (concatenated exons) into 50 base pair (bp) consecutive bins where read counts of input and IP sample were quantified. Second, we applied a two tailed Fisher's exact test to call bins significantly enriched in IP vs. input. Specifically, we constructed a contingency table consisting of the read counts of a bin in the input (a) and in the IP (b), and the median read count of the bins in the gene containing that bin in the input (c) and in the IP (d). The odds ratio is represented as $\frac{b\times c}{a\times d}$. The FDR control procedure was performed on each gene and an FDR $< 5\%$ cutoff was used to call a bin peak for each sample. Third, to obtain a common set of peaks for all QTL analysis, we define joint-m^{6}A-peaks by requiring a bin to be called as peak in at least 5 individuals. Neighboring bins that satisfied this criterion were merged into a single peak. Then, a pair of read counts (the input and IP) were obtained for each of the joint m^{6}A peaks. Finally, we filtered out peaks with zero read in any of the samples.

To obtain consensus motif of m^{6}A, we used Homer2 [94] to search for *de novo* motifs in m^{6}A peak sequences with the parameter -len 5,6,7 -rna -S 5 -noknown. As a background

control, we extracted sequences from random peaks of 200 bp size that were sampled from mRNA transcripts.

To visualize the distribution of m^6A peaks on the transcript, we generated meta-gene plot using the R package Guitar [95] with default settings.

2.2.5 Genotype data and imputation

We downloaded the latest thousand genome project "combined variant calling data release" [53] where 50 samples of ours are covered in this dataset. For the rest 10 samples, there are 8 sample covered in the chip array genotyped data from the thousand genome. For the two individual that are not covered in the thousand genome genotype data, we obtained their genotype data from HapMap and liftovered the hg18 coordinate to match the hg19 coordinate of others. To fill the missing genotypes of these 10 individuals that are not covered in the thousand genome combined variant call dataset, we pre-phased and imputed missing genotypes using Impute2 [96, 97]. Overall, we obtained genotypes for 9,821,958 SNPs that have MAF $> 5\%$.

2.2.6 m^6A QTLs mapping

Unlike common omics data (e.g. ChIP-seq) that quantifies molecular trait by a single variable (e.g. read counts in a ChIP-seq peak), m^6A-seq experiments are characterized by a pair of input and immunoprecipitation (IP) measurements. For a given testing window (as defined by joint m^6A-peaks), the read counts of IP (immunoprecipitated) and input (regular RNA-seq) in individual i are denoted as $Y_i^{(1)}$ and $Y_i^{(0)}$, respectively. Let $T_i^{(1)}$ and $T_i^{(0)}$ be the library size of IP and input, respectively. We define log odds ratio (log-OR) as the m^6A quantitative phenotype:

$$y_i = \frac{Y_i^{(1)}/T_i^{(1)}}{Y_i^{(0)}/T_i^{(0)}} \tag{2.1}$$

Correction for possible confounders: Various factors (such as IP efficiency and GC

content) may distort this value. Our strategy is to adjust for this offset term to account for covariates of peaks in sample-specific fashion.

We define observed log-OR of a peak j in sample i as:

$$\tilde{y}_{ij} = \log_2 \frac{Y_{ij}^{(1)}/T_i^{(1)}}{Y_{ij}^{(0)}/T_i^{(0)}} \tag{2.2}$$

Leaning how $\tilde{y}_{ij}$ depends on IP efficiency and GC content in each sample would allow us to correct for deviation of $\tilde{y}_{ij}$ due to those factors in a sample-specific manner.

Adjusting for IP efficiency: Variation of overall IP efficiency across individuals can have impact on the expected fraction of reads in IP. We adjust for this variation by estimating the difference of IP-efficiency between each sample i and the average across samples, using a strategy similar to WASP [93]. Let $\bar{y}_j$ be the average log-OR of peak j across all samples. We can plot $\tilde{y}_{ij}$ vs. $\bar{y}_j$ for all peaks in a sample i (**Figure 2.1a**). For samples with low IP efficiency, the lines would fall below diagonal line, and for samples with high IP efficiency, above the diagonal line.

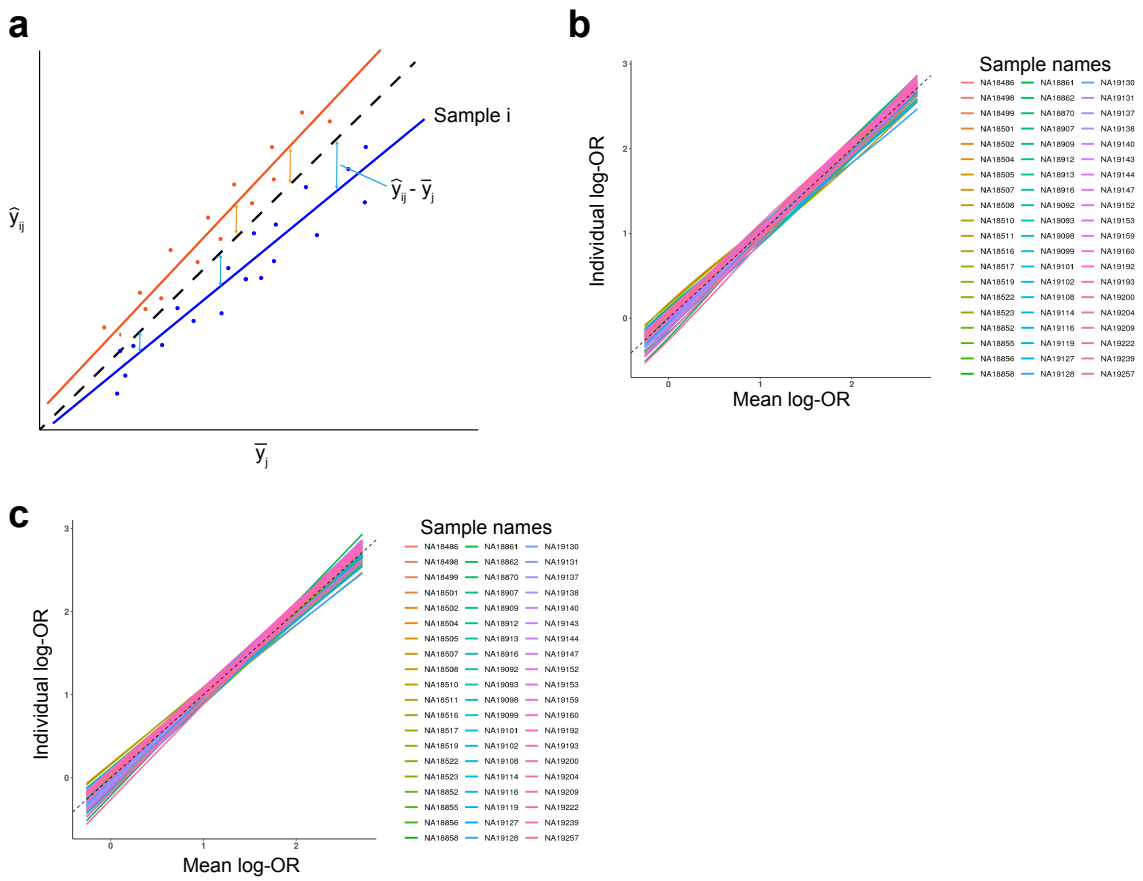


Figure 2.1: Fitting the overall IP efficiency correction term. The fraction of reads coming from the peaks varies between experiments. We adjust for this variation by estimating the difference of IP-efficiency between each sample i and the average across samples as illustrated in panel **a**. We first fitted a quadratic function of individual log odds ratio vs. the mean log odds ratio across all samples as shown in panel **b**. Since the fitted curves are approximately linear, we simplified to fit the individual log odds ratio as a linear function of the mean log odds ratio across all samples as shown in panel **c**.

In order to capture the variation of IP efficiency across samples, we first fitted a quadratic function of $\tilde{y}_{ij}$ against $\bar{y}_j$ for each sample, as shown in **Figure 2.1b**. Since the fitted curves are approximately linear, we then fitted a linear model for simplicity (**Figure 2.1c**).

Let $\hat{y}_{ij}$ be the expected log-OR for a given peak j in sample i (based on the fitted line), then our correction term for peak j in sample i is:

$$\Delta K_{ij} = \hat{y}_{ij} - \bar{y}_j \tag{2.3}$$

Adjusting for GC content: Another technical covariate that has been shown to influence read count representation is GC content [93, 98]. To adjust for the GC content bias, we group peaks of the same GC content into bins. Let b_{il} be the average log-OR of all peaks in bin l of sample i, and let $b_{.l}$ be the average log-OR of all peaks in bin l over all samples. The difference of the two should reflect the GC effect. However, a sample may have higher log-OR across all peaks because of IP efficiency or other technical factors, so we should adjust for that. Let $b_{i.}$ be the average log-OR of all peaks of sample i, and $b_{..}$ be the average log-OR of all peaks of all samples. We define the deviation of bin l in sample i as:

$$F_{il} = (b_{il} - b_{.l}) - (b_{i.} - b_{..}) \tag{2.4}$$

As shown in **Figure 2.2** , the effect of GC content varies from sample to sample. We fitted a quadratic function of F_{il} vs. its GC content to get the correction term ΔF_{il} for each peak given its GC content bin l in sample i.

Our log fold enrichment for sample i in peak j adjusting for IP efficiency and GC content is:

$$LogOR_{ij} = \tilde{y}_{ij} + \Delta K_{ij} + F_{il} \tag{2.5}$$

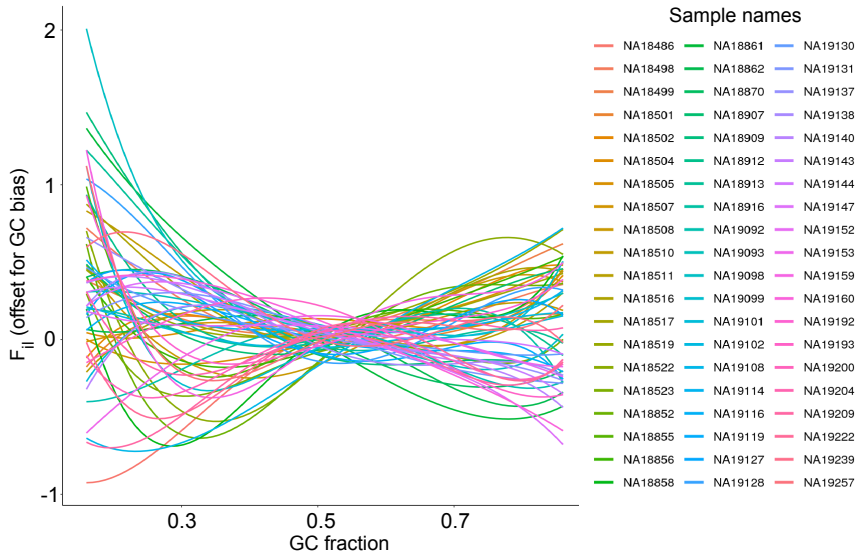


Figure 2.2: Fitting the GC bias correction term. The GC content of a region can affect the sequencing depth of that region. This influence varies from sample to samples. To capture and correct for this GC bias, we fitted a correction term $F_i l$ for each peak as a quadratic function of GC fraction of the peaks for each sample.

We next standardized the log-OR by subtracting out the mean and dividing by the standard deviation of each peak followed by quantile normalization. We applied a linear model implemented in FastQTL [99] to test the association between phenotypes and genotypes, adjusting for 15 principal components (PCs). The number of PCs was chosen to maximize power. We tested *cis*-associations between peaks and SNPs within 100 kb, as we found the signals for m^6A-SNP association are mainly enriched within the 100 kb window flanking the m^6A peak.

To account for multiple genetic variants tested for each peak, we performed 1,000 rounds of permutation and used the beta-approximation scheme in the FastQTL to obtain empirical P values for each peak. We then used Storey's qvalue method [100] to obtain the false discovery rate (FDR), accounting for multiple peaks tested. SNP level FDR was obtained by applying Storey's qvalue method to nominal P values for the association tests.

2.2.7 Spatial distribution and genomic annotation of the m^{6}A QTLs

To characterize the spatial distribution of the m^{6}A-associated SNPs with respect to the m^{6}A peaks, we calculated the distance from the SNP to the center of the corresponding peak. Since m^{6}A is an RNA modification, the distances were calculated with respect to the transcript strand where a negative distance indicates an upstream location of the transcript and vice versa. m^{6}A QTL SNPs were assigned to 5'UTR, CDS, 3'UTR, Intron and Intergenic annotation using the R package ChIPseeker [101]. For SNPs that could be assigned to different annotation due to multiple isoforms of a gene, the annotation priority was set to be UTR > CDS > Intron > Intergenic.

2.2.8 Comparisons between m^{6}A QTLs vs. eQTLs and sQTLs

To compare m^{6}A QTLs with eQTLs (both at FDR < 10%), we compared the overlap of ePeak-harboring genes and eGenes in the same cohort of Yoruba LCL samples [55]. Then, for those genes with both ePeak and eGene mapped, we computed the pairwise distances and LD between the lead ePeak SNPs and the eGene SNPs. To compare m^{6}A QTLs with sQTLs (both at FDR < 10%), we used the sQTL data from a slightly larger cohort of GEUVADIS Yoruba LCL samples (n = 87) [47]. We mapped intron clusters with at least one significant sQTL (denoted as eSplicing intron clusters) and compared the overlap of ePeak-harboring genes and genes containing eSplicing intron clusters. For genes with multiple m^{6}A ePeaks and/or multiple eSplicing intron clusters, we computed the pairwise distances and LD between all pairs of the lead ePeak SNPs and the eSplicing SNPs.

2.2.9 Annotation of the m^{6}A QTLs and enrichment analysis

Our functional annotations include m^{6}A consensus motif (RRACH) in m^{6}A peaks, RBP CLIP-seq peaks in K562 and HepG2 cells from ENCODE [102], TF and histone modifications ChIP-seq peaks in LCLs from ENCODE59, experimentally determined RiboSNitch [103] and

predicted microRNA binding site by TargetScan [104].

To annotate SNPs by the m^6A consensus motif (RRACH) in m^6A peaks, we used MotifBreakR [105] to find instance of m^6A motifs overlapping with SNPs (note this software matches motifs to either allele of the polymorphic site). We then intersected these motif matches with the joint peaks to obtain motifs in m^6A peaks. For RBP CLIP-seq peaks, we intersected the peaks of the two replicates and from the two cell lines to obtain a peak set that are consistent across replicates and cell lines. These peaks shared across cell lines are more likely to be functional in LCL than those in single cell line. To define ChIP-seq peaks for TFs and histone markers, we chose peaks that are "optimal IDR peaks" as defined by the ENCODE processing pipeline. We used the microRNA binding sites predicted by TargetScan [104], limit to the sites that are targeted by microRNAs expressed in the LCLs. MicroRNA expression data was obtained from the microRNA-seq data of LCLs samples from GEUVADIS [61]. We defined a microRNA being expressed in LCLs by requiring the median read count across individuals $\geqq 5$.

We used two methods to test the enrichment of each functional annotations in m^6A QTLs. In the first method, we took the independent SNPs from the fine-mapped m^6A QTLs (see Fine-mapping m^6A QTLs, eQTLs and sQTLs Section) by choosing SNPs with the maximum posterior inclusion probability (PIP) per credible set. We then compared the number of QTLs vs. the number of random control SNPs overlapping with a functional annotation using the two-tailed Fisher's exact test. To generate the control SNP set, we used a modified version of SNPsnap [106] to sample 100 sets of SNPs that match the allele frequencies, numbers of SNPs in LD, distances to the nearest genes, gene density as well as annotations about SNP locations relative to genes (5'UTR, CDS, 3'UTR, intron and intergenic regions).

In the second method, we used Torus [107] as an alternative to assess the enrichment of functional annotation in the m^6A QTLs. Torus fits a logistic regression model to estimate an enrichment parameter for each annotation, which enables joint analysis of multiple

annotations.

2.2.10 Fine mapping m^{6}A QTLs, eQTLs and sQTLs

Many significant m^{6}A QTLs are likely not causal variants but tagging the causal SNPs due to LD. For each ePeak, there could be multiple SNPs all showing significant association with m^{6}A level because genotypes of these SNPs are all highly correlated with each other (SNPs in proximity tend to travel together). Thus, using P value thresholds to select SNPs may include many non-causal SNPs that can significantly dilute signals in downstream analyses. A coarse way to select independent SNPs is to perform LD clumping, which selects the SNP with the lowest P value from each set of highly correlated SNPs in sliding windows. However, causal SNPs may not always show the strongest effect due to complex LD structure and when statistical power is not large [108].

To better identify independent associations and likely causal variants, we performed fine-mapping of m^{6}A QTLs using the state-of-art tool SuSiE [109]. SuSiE is a Bayesian variable selection regression method for selecting SNPs of non-zero effect from highly correlated SNPs, which can leverage functional annotations as priors for variable selection. We used the standard version of SuSiE that takes individual-level phenotype and genotype data. For SuSiE parameters, the maximal causal variants per region was set to 3 and estimate_prior_variance = TRUE. With this setting, we let SuSiE to estimate the percentage of variance explained from the data.

Molecular QTLs are often enriched for functional annotations as some of the functional annotations contribute to the mechanism of molecular QTLs. For example, a SNP close to a peak and located in an RBP binding site would have a higher prior probability of being a causal variant of m^{6}A. To make use of these information in fine-mapping, we derived informative priors from the functional annotation enrichment analysis using the program Torus with flag: -dump_prior to output priors for each SNP. To be conservative, we first fine-mapped m^{6}A QTLs with uniform prior inclusion probability and applied this version in

most of our analyses including enrichment analysis comparing m^6A QTLs with control SNPs by Fisher's exact test, m^6A QTL RBP-motif break analysis and partition of GWAS traits heritability analysis. Then, we performed another version of fine-mapping that leveraged RNA annotations including RiboSNitch, RBP binding sites and SNP-peak distances by using informative priors derived by Torus. We used the RNA-features-informed fine-mapping results in the S-LDSC analysis of enrichment of GWAS variants in m^6A QTNs.

Similarly, we fine-mapped eQTLs and sQTLs using SuSiE on individual-level expression and splicing data in GEUVADIS YRI LCL samples with uniform priors and the same parameter settings as fine-mapping m^6A QTLs.

SuSiE outputs posterior inclusion probability (PIP) and 95% confidence interval credible set. PIP of a SNP is the sum of posteriors over all models that include this SNP as causal, which represents the estimated probability of a SNP being causal variant. Credible set reports the minimal sets of SNPs containing all causal SNPs with 95% probability. Each credible set basically represents an independent effect with likely causal SNP/SNPs contained in this set. In current manuscript, we used credible set and ranked PIP to select for putative causal m^6A QTNs.

2.2.11 Evaluating the role of RBP binding in regulating m^6A level

To evaluate if RBP binding is causally linked to m^6A levels, we investigated directional consistency of SNP effects on RBP binding and on m^6A level. First, we learned motifs of each RBPs from the CLIP-seq data. For each RBP, we took the top 3,000 peaks as ranked by peak strength of each replicate and retain the ones that are consistent in both replicates. We then extend 5 bp at both sides for peaks that are shorter than 10 bp. The sequence of the resulted peaks served as target sequence for *de novo* motif search by Homer2. To generate matched background peaks for each RBP, we first generated a large set of random peaks of length 70 bp (the mean width of CLIP-seq peaks) on the transcribed region including both exons and introns. Then we annotated the genomic locations (5'UTR, CDS, Intron,

3'UTR etc.) of the top CLIP-seq peaks and drew random peaks with matched distribution of genomic annotation. At least one motifs of P value $< 1 \times 10^{-10}$ were obtained for 113 RBPs. For each RBP, the top 2 motifs by P value were used for motif correlation analysis of RBP binding. To visualize the motifs, we used the R package Loglas [110] to generate sequence logo plots that highlight both nucleotide conservations and depletions.

Next, we checked whether the impact of genetic variant on RBP binding is correlated with the effect on m^6A level, measured by m^6A QTL effect size, in a directionally consistent manner. To assess the effect of genetic variants on RBP binding affinity, we used MotifBreakR [105] to map SNPs that disrupts a RBP motif. A cutoff of P value $< 1 \times 10^{-3}$ was used to filter the motif match results in the parameter setting. To enhance signals, we used fine-mapped m^6A QTLs as described in the **Fine Mapping m^6A QTLs, eQTLs and sQTLs** sub-section. Some of the ePeaks don't have SNPs in fine-mapped credible set. This will likely result in insufficient number of SNPs for the motif break analysis. To include more SNPs, we used all SNPs with PIP > 0.5 and for ePeaks without SNP PIP > 0.5, we included the maximum PIP SNPs to select likely causal SNPs in this analysis. For each RBP, we chose the motif-breaking SNPs and peak pairs that are within 0.5 kb range with the assumption that m^6A is mainly affected by RBPs bound close to the m^6A sites. RBPs with less than 5 SNP-peak pairs were not included in the analysis. In total, we assessed the correlation between binding affinity change of 37 RBP motifs and m^6A QTL effect sizes. Linear model was used to assess the correlation and Storey's qvalue method was used for FDR control.

2.2.12 Estimating contribution of RNA features and TFs to the m^6A QTLs

To compare the relative contribution of RBP binding, secondary structure, RRACH motif, microRNA binding and TF binding to the installation of m^6A, we estimated the proportion of putative causal m^6A QTNs falling in each of these annotation categories. Specifically, we took all SNPs from the credible sets of SuSiE fine-mapping and summed the PIP of SNPs

in each annotation category to obtain an estimation about the expected "number" of causal SNPs related to each mechanism. To compute the proportion, we divided the summed PIP of each annotation category by the sum of PIP across all SNPs in credible sets.

2.2.13 Joint analysis of transcription-rate QTLs and m^6A QTLs

. We downloaded published transcription-rate-QTL data [47] where transcript rate was measured by 4SU-seq at two labeling time points (30min and 60 min) in the same cohort of YRI samples as in our study. 4SU-seq labels newly synthesized RNA using nucleotide analog 4-thiouridine (4SU), allowing for pulldown of the labeled newly synthesized RNA for sequencing. We note that 4SU-seq quantifies the overall transcription rate, but does not distinguish different events (e.g. PolII pausing vs. slow progression) leading to transcription rate change.

2.2.14 Validating m^6A methyltransferase interactions with TFs

To experimentally validate that some TFs interact with m^6A installing machinery, we performed co-immunoprecipitation (co-IP) experiment for several TFs following a modified protocol from an earlier study [111]. Briefly, 150 μl LCL cell pellet was washed with 10 volumes of PBS once, then washed with 10 volumes of hypotonic lysis buffer (10 mM Tris-HCl pH = 7.5, 10mM KCl, 2mM $MgCl_2$, 0.2 mM EDTA, 0.2 mM DTT, 10mM sodium butyrate, 1× protease and phosphatase inhibitor cocktail) once. The pellet was resuspended in 8 volumes of hypotonic lysis buffer for 5 minutes to swell cells. We then homogenized the swollen cells using the "loose" pestle of a 2 ml Dounce homogenizer for 200-300 strokes. Nuclei were pelleted at 800 g for 5 minutes followed by washing once with hypotonic lysis buffer. The nuclei pellet was then resuspended in 4 volumes of nuclear lysis buffer (20 mM Tris-HCl pH = 7.5, 50 mM KCl, 100 mM NaCl, 2 mM $MgCl_2$, 10% glycerol, 0.1% Tween20, 0.2 mM EDTA, 0.2 mM DTT, 10mM sodium butyrate, 1× protease and phosphatase inhibitor cocktail). The nuclei were homogenized by 150 strokes of the "tight" pestle of a 2 ml Dounce

homogenizer. After nuclear lysis, Turbo DNase (Invitrogen, cat. AM2238) was added at a 1:20 ratio and the mix was incubated at 16°C with rotation for 1 hr. The resulting lysate

Target protein	Purpose	Antibody	Note
METTL3	Western blot	Cell signaling #86132	1:1000 dilution
WTAP	Western blot	Cell signaling #41934	1:1000 dilution
VIRMA	Western blot	Proteintech 25712-1-AP	1:300 dilution
RBBP5	Western blot	Cell signaling #13171	1:1000 dilution
BACH1	Western blot	Bethyl A303-058A	1:2000 dilution
RBBP5	Immunoprecipitation	Bethyl A300-109A	5 ug/∼1mg lysate
BACH1	Immunoprecipitation	Bethyl A303-058A	5 ug/∼1mg lysate
IP detection	Western blot	Abcam VeriBlot ab131366	1:300 dilution

Table 2.1: Antibodies used in Co-IP experiment. The detailed information for the antibodies used to test the interaction between TFs of interest and m^6A methyltransferase complex.

was then cleared by centrifuge at 16,000 g, 4°C for 20 minutes. Immunoprecipitations were performed by incubating cleared lysate with antibody specific to TFs (see **Table 2.1** for details of antibodies used) and 20 μl of corresponding protein A/G beads for 4 hours at 4°C. We then applied 5 rounds of stringent washes using dialysis buffer (20 mM Tris-HCl pH 7.5, 50 mM KCl, 100 mM NaCl, 2 mM $MgCl_2$, 10% glycerol, 0.2 mM EDTA, 0.2 mM DTT, 10mM sodium butyrate, 1× protease and phosphatase inhibitor cocktail) followed by elution in 1× Laemmli SDS sample buffer.

2.2.15 Joint analysis of m^6A QTLs and other molecular traits QTLs

We collected QTL data of multiple molecular traits in YRI LCLs from earlier studies, including transcription rate, mRNA levels, mRNA decay, mRNA splicing, ribosome loading and protein levels. We used processed phenotype data and YRI genotypes complied in reference [47] (downloaded from http://eqtl.uchicago.edu/jointLCL/). To map *cis*-QTLs for these molecular phenotypes, we chose SNPs within 100 kb windows around genes using linear regression by FastQTL, and regressed out PCs (the number PCs was chosen to maximize the number of detected QTLs in each molecular phenotype).

To map alternative polyadenylation (APA) QTL, we followed a procedure used in an early study [47] to generate APA data from RNA-seq data. Specifically, using the RNA-seq data we generated (the input of the m^{6}A-seq), we predicted and quantified APA events based on sequencing coverage at the 3'UTR regions using a modified version of DaPars [112] as described in Li et al [47]. We find 7,617 putative APA sites. Using the ratio of distal to proximal polyA site usages as quantitative phenotypes, we tested their association with genotypes within 100 kb range by FastQTL. 7 PCs were included to maximize the QTL discovery. At SNP-level FDR $< 10\%$ (Storey's qvalue method), we obtained 3,586 APA-QTLs.

To estimate QTL sharing between m^{6}A and other molecular phenotypes, we followed the procedure of Li et al [47], we first ascertained m^{6}A QTLs at a given P value threshold. We then limited our analysis to the lead SNP per m^{6}A locus; thus, the SNPs we include are largely independent. We next assessed the proportion of non-null ($\pi 1$) from P values of the ascertained SNPs in another molecular phenotype, using Storey's method in the qvalue R package [100]. 80% bootstrap confidence intervals for the our $\pi 1$ estimates were computed by resampling P values with replacement 100 times. Control SNPs were randomly sampled across the genome.

We investigated the QTL effect size correlation between m^{6}A and 5 related molecular phenotypes studied in above $\pi 1$ sharing analysis. We ascertained lead m^{6}A QTLs and assessed correlation between their effects on m^{6}A and each of other 5 molecular traits using a linear model. To understand how presence of RBP in vicinity of m^{6}A site may influence the correlation between m^{6}A and other molecular traits, we repeated effect size correlation analysis stratifying RBP bindings. We ascertained lead m^{6}A QTLs, and grouped m^{6}A peaks bound by different RBPs, requiring the genomic intervals of RBP binding sites to be entirely within m^{6}A peaks (results are similar if we require only 1 bp overlap). 82 RBPs with at least 50 data points (peak-SNP pairs) were selected. In each group of m^{6}A peaks bound by an RBP, we assessed the correlations of effect sizes between m^{6}A QTLs and QTLs of

other molecular phenotypes. When computing effect sizes for molecular QTLs, we used the slope of the linear regression as a measure of effect size, and did not regress out PCs as that could modify effect size estimates [47]. Significant correlations of effect sizes between m^6A QTLs and QTLs of other molecular phenotypes were selected at FDR (Benjamini & Hochberg method) 10% threshold.

2.2.16 Re-analysis of ribosome-profiling data of METTL3 and YTHDF1 knockdown in HeLa cells

To validate our finding of heterogeneous effect of m^6A on downstream molecular traits, we used translation efficiency as an example and re-analyzed the ribosome profiling data of METTL3 (m^6A methyltransferase) and YTHDF1 (m^6A reader) depleted Hela cells from a published study [13]. The original paper focused on the target transcripts of YTHDF1 as defined by transcripts harboring YTHDF1-bound m^6A peaks. To systematically examine the effect of m^6A depletion on translation efficiency, we first stratified all transcripts with m^6A peaks and examined the distribution of the $\log_2$ fold change of translation efficiency. To understand the heterogeneous effect of m^6A on translation efficiency, we then stratified transcripts by RBP-bound m^6A peaks (we define RBP targets by overlaps of RBP eCLIP-seq peaks with m^6A peaks in HeLa cell line) and Welch's two-sample t-test was used to test the $\log_2$ fold change in translation efficiency in RBP targets vs. non-targets, upon METTL3 knockdown. We reported 32 RBPs that showed significant difference of translation efficiency in targets vs. non-targets with FDR $< 5\%$. In order to further understand the heterogeneous effect of m^6A on translation efficiency, we also re-analyzed the ribosome profiling data of YTHDF1 (m^6A reader) depleted Hela cells from reference [13]. We find the distribution of $\log_2$ fold change of translation efficiency in all transcripts are shifted towards down-regulation (**Figure 2.3a**), suggesting improper normalization in the original processed data (**Figure 2.3b**). We normalized the libraries using the median-of-ratio method implemented in DE-Seq2 [113] (**Figure 2.3c**). We examined the impact of YTHDF1 depletion on translation

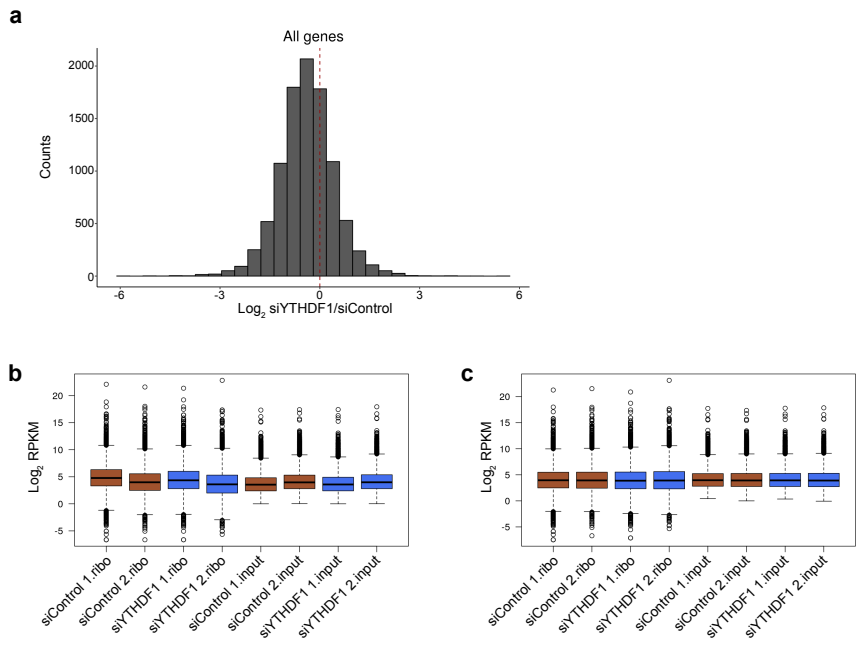


Figure 2.3: Re-normalize ribosome profiling libraries. We analyzed the ribosome profiling data from Wang et al. *Cell* 2015 and found the overall distribution of translation efficiency changes are shifted towards down-regulation as shown in panel a. We checked the distribution of read counts in individual samples of the analyzed data in the original paper as shown in b (n = 9,742 genes). The distribution of normalized read counts in each individual sample are shown in c (n = 9,742 genes). The lower and upper hinges correspond to the first and third quartiles. Horizontal line indicates median value, and whiskers correspond to the value no further than 1.5x inter-quartile range.

efficiency by overlay distributions (histograms) of YTHDF1 targets and non-targets. Surprisingly, we find YTHDF1 knockdown led to overall reduction of translation efficiency in transcripts harboring YTHDF1-bound m^6A sites, but there are 33% (proportion based on considering translation efficiency $\log_2$ fold change $<$ -0.5 or $>$ 0.5 as having effects) YTHDF1 targets showing opposite effects.

2.2.17 Validating YBX3 function in repressing translation efficiency

We depleted YBX3 using siRNA (Qiagen, Cat No. SI00355019) in Hela cells and performed polysome profiling as described previously [13] to assess the effect on translation efficiency. We quantified transcripts level of selected targets in three polysome-bound fractions: mono-

some, polysome1 and polysome2 (as indicated in **Supplementary Figure 2.5a**) and non-polysome-bound fraction for further gene-specific analysis. We selected 5 YBX3 target genes from the analysis of ribosome profiling data in m^6A depleted cells; all 5 YBX3-targets have m^6A peaks overlapping with YBX3 CLIP-seq peaks and showed increased TE upon m^6A depletion (*IGS15, HSP90B1, CDT1, MRPL4, HIST3H2A*). As negative controls, we selected 3 YTHDF1-targets, which showed decreased TE upon m^6A depletion (*THRAP3, PTPRM, UGGT1*). For each gene, we normalized the monosome, polysome1 and polysome2 fractions by the non-polysome-bound fraction to obtain a translation efficiency estimation.

2.2.18 Heritability and enrichment analysis of GWAS summary statistics using stratified LD score regression (S-LDSC)

We download summary statistics of 45 phenotypes from UK Biobank [114], the Price lab [115] and the GWAS Catalog (https://www.ebi.ac.uk/gwas/summary-statistics). The GWAS traits and corresponding references are listed in **Table 2.2**.

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Platelet count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Lymphocyte Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Rheumatoid Arthritis	Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376 (2013).

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Red Cell Distribution Width	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Reticulocyte Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Red Blood Cell Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Coronary Artery Disease	Data on coronary artery disease & myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.ORG
Hypothyroidism	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Hemoglobin Concentration	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Eosinophil Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Leukocyte Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Body Mass Index	Akiyama, M. et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nature Genetics 49, 1458 (2017).
Menarche Age	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Autism	Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet 381, 1371-1379 (2013)
Total Cholesterol	Willer, C., Schmidt, E., Sengupta, S. et al. Discovery and refinement of loci associated with lipid levels. Nat Genet 45, 1274–1283 (2013)
Granulocyte Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016)
Height	Wood, A.R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics 46, 1173 (2014)

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Neuroticism	Okbay, A., Baselmans, B., De Neve, J. et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat Genet48, 624–633 (2016)
Schizophrenia	Schizophrenia Working Group of the Psychiatric Genomics, C. et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421 (2014)
Menopause Age	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015)
Neutrophil Count	Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429.e19 (2016).
Low-Density Lipoprotein	Teslovich, T., Musunuru, K., Smith, A. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010)
Alzheimers	Lambert, J.-C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics 45, 1452 (2013).
High-Density Lipoprotein	Teslovich, T., Musunuru, K., Smith, A. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010)

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Allergy	Ferreira, M.A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nature Genetics 49, 1752 (2017).
Depressive Symptoms	Okbay, A., Baselmans, B., De Neve, J. et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat Genet48, 624–633 (2016).
Anorexia	Boraska, V., Franklin, C., Floyd, J. et al. A genome-wide association study of anorexia nervosa. Mol Psychiatry 19, 1085–1094 (2014)
Bipolar Disorder	Sklar, P., Ripke, S., Scott, L. et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet 43, 977–983 (2011)
Respiratory and Ear-nose-throat Disease	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Breast Carcinoma	Michailidou, K., Lindström, S., Dennis, J. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92–94 (2017)
Triglycerides	Teslovich, T., Musunuru, K., Smith, A. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010)

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Dermatology	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Celiac	Dubois, P., Trynka, G., Franke, L. et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42, 295–302 (2010)
Inflammatory Bowel Disease	Luo, Y. et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nature Genetics 49, 186 (2017)
Allergy Eczema Diagnosed	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Ulcerative Colitis	Jostins, L., Ripke, S., Weersma, R. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012)
Lupus	Bentham, J., Morris, D., Cunninghame Graham, D. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet 47, 1457–1464 (2015)

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Type 2 Diabetes	Scott, R.A. et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes, db161253 (2017).
Auto Immune Traits (Sure)	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Crohn's Disease	Jostins, L., Ripke, S., Weersma, R. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012)
Asthma	Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779 (2015).
Primary Biliary Cirrhosis	Cordell, H., Han, Y., Mells, G. et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat Commun 6, 8019 (2015)
Fasting Clucose	Manning, A., Hivert, M., Scott, R. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 44, 659–669 (2012)

Table 2.2: Summary of GWAS data.

GWAS trait	Reference
Ovarian Carcinoma	Phelan, C., Kuchenbaecker, K., Tyrer, J. et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat Genet 49, 680–691 (2017)
Type 1 Diabetes	Bradfield, J.P. et al. A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. PLOS Genetics 7, e1002293 (2011)

To test enrichment of GWAS signals in m^{6}A QTLs, we first extracted GWAS SNPs that also belong to m^{6}A QTLs (association P value $< 1 \times 10^{-4}$) and plotted the QQ-plot of the GWAS P values for those SNPs. Similarly, we plotted GWAS SNPs that are also eQTLs or sQTLs (association P value $< 1 \times 10^{-4}$) for comparison. Genome-wide GWAS P values were also plotted as a control.

We next performed heritability and enrichment analysis of GWAS summary statistics using stratified LD-score regression (S-LDSC) [115]. S-LDSC partitions the heritability of genomic annotations using GWAS summary statistics and estimates the enrichment as a ratio of the proportion of heritability explained by an annotation divided by the proportion of SNPs in that annotation. We partitioned the heritability of complex traits and estimated heritability enrichment of m^{6}A QTL, eQTL and sQTL. We then constructed a probabilistic (continuous-valued) annotation using PIP estimates from SuSiE fine-mapping with RNA-features-informed prior or uniform prior inclusion probability. We applied S-LDSC to our QTL-based annotations using separate models for each QTL annotation and joint model with all three types of QTL annotations together. In our S-LDSC analysis, we adjusted for various baseline annotations of SNPs using a baselineLD model [115], including gene annotations (coding, UTRs, intron, promoter), MAF bins and LD-related annotations. We did not include functional annotations such as enhancer markers in our baseline model, because

these annotations are likely correlated with the QTL features of interest (e.g. enhancers are enriched in eQTLs), and including them will bias our estimated enrichment. To estimate heritability explained by molecular QTLs, we constructed a binary annotation containing all SNPs at given SNP-level FDR cutoffs. We repeated the analysis on m^{6}A QTL, eQTL, sQTL at thresholds of 20%, 10% and 5% FDR. We find our conclusion is robust at varying thresholds.

2.2.19 TWAS and heritability analysis of m^{6}A peaks

TWAS was performed using the FUSION [83] software. We trained regression models (LASSO, Elastic Net and top SNPs) using our own m^{6}A data in LCLs, published RNA-seq data in YRI LCLs [55], and splicing data [47] using GEUVADIS YRI LCLs data [61], and the corresponding YRI genotype data. In m^{6}A-TWAS analysis, we computed weights for each m^{6}A peak using LASSO and Elastic Net models as well as regression model with the single most significantly associated SNPs (using the R function FUSION.compute_weights.R provided by FUSION with parameter –models lasso,enet,top1). The best performing model in cross-validation was selected for each peak to perform imputation. We used 100 kb *cis*-window, and restricted the genotypes to the set of markers in the LD reference panel (1000 Genomes European samples) provided on the FUSION website (http://gusevlab.org/projects/fusion/), as we used the LD reference data for the GWAS-m^{6}A association analysis. 19,130 m^{6}A peaks had estimated heritability. We obtained trained weights for 918 peaks with estimated heritability significantly greater than 0 (with default parameter hsq $P = 0.01$). We then performed imputation of genetically determined m^{6}A levels and estimated GWAS-m^{6}A associations. We selected genome-wide significant m^{6}A peaks/genes at Bonferroni corrected P value < 0.05. Similarly, we built prediction model of gene expression as well as splicing (introns with missing values were ignored) and estimated the GWAS-gene expression as well as GWAS-splicing associations using FUSION.

2.2.20 Colocalization analysis of m^6A QTLs and GWAS variants

Our colocalization analysis was performed using the Approximate Bayes Factor (ABF) test implemented in software Coloc [81], which has been incorporated in the TWAS/FUSION pipeline. Coloc computes five posterior probabilities (PP0, PP1, PP2, PP3 and PP4), each corresponding to a hypothesis – H_0: no association with either trait; H_1: association with trait 1, not with trait 2; H_2: association with trait 2, not with trait 1; H_3: association with trait 1 and trait 2, two independent SNPs; H_4: association with trait 1 and trait 2, one shared SNP. We ran coloc incorporated in the FUSION pipeline with default parameters for TWAS-significant associations (using the R function Fusion.assoc_test.R in FUSION software with - -coloc_P flag) and used PP4 to assess evidence of colocalization. We visualized the colocalization of m^6A QTL and GWAS associations using LocusCompareR package (https://github.com/boxiangliu/locuscomparer).

2.3 Results

2.3.1 Analysis of m^6A peaks in LCLs

We used m^6A-seq [33, 34] to profile m^6A levels across the transcriptome in LCLs derived from 60 Yoruba individuals. We obtained on average 60 million reads for unmodified (input) and immunoprecipitated (IP) mRNA libraries for each cell (**Figure 2.4a**). To identify m^6A sites that are present in most of the samples, we called peak jointly on all samples. Specifically, we called m^6A peak in consecutive bins for each sample (see Method Section) and defined a joint significant bin if this bin is significantly enriched for IP read counts in more than 5 samples. We then merged neighboring joint significant bins as a single peak. We then filtered out peaks that have zero read count in any of the sample. With this procedure, we identified 20,044 peaks located in transcripts of 8,448 genes, with an average peak length of 351-bp (**Supplementary Figure 2.1a**). Consistent with previous reports [33, 34],our joint m^6A peaks are enriched near start and stop codons (**Supplementary Figure 2.1b-c**) and

are mostly located in CDS and 3' UTR. Sequences within m^{6}A peaks are enriched with the canonical RRACH motif (**Supplementary Figure 2.1d**).

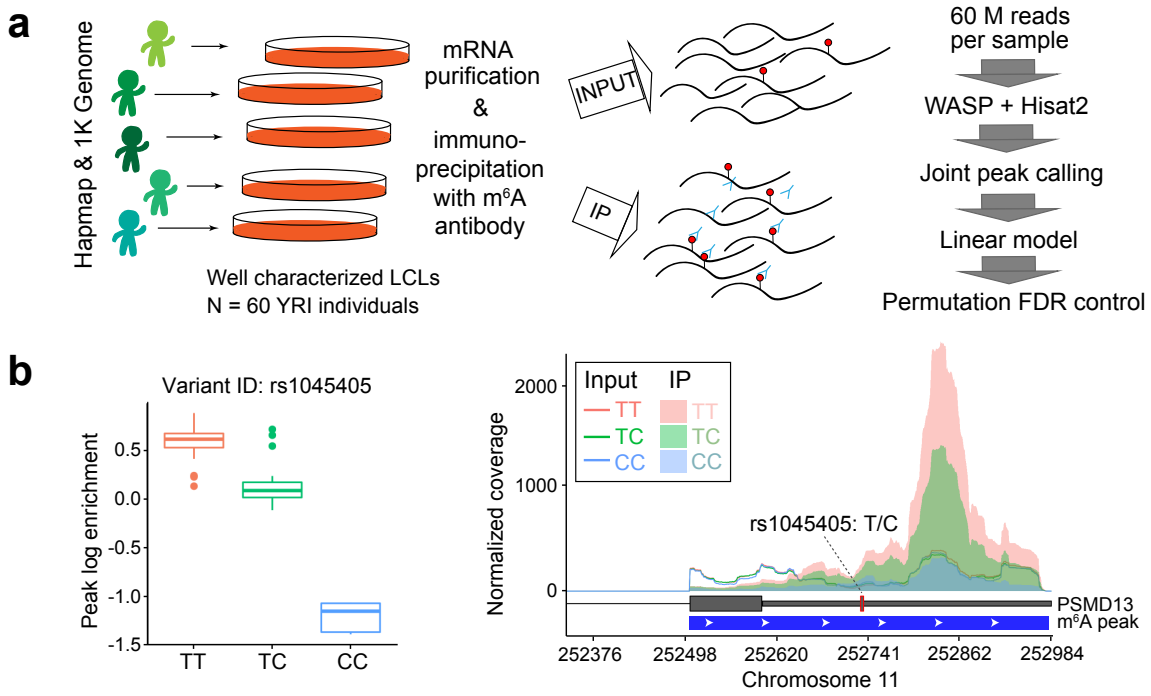


Figure 2.4: Overview of m^{6}A QTL mapping. **a**, Overall study design and workflow of m^{6}A QTL mapping. The linear regression model for association testing, adjusting for GC content and IP efficiency. **b**, An example of m^{6}A QTL. The left panel shows the box plot of m^{6}A levels grouped by the genotype of the example m^{6}A QTL (rs1045405). n = 60 biologically independent samples. The lower and upper hinges correspond to the first and third quartiles. Horizontal line indicates median value, and whiskers correspond to the value no further than 1.5 × the inter-quartile range. The right panel shows the mean coverage of each genotype at the m^{6}A peak. The m^{6}A peak is shown by the blue track and the gene model by the gray track. The coverages of input and IP libraries are shown in lines and shadows, respectively.

2.3.2 Mapping genetic variants associated with m^{6}A

We tested association between m^{6}A level of each peak and nearby genetic variants using a linear model, accounting for GC content and other covariates (**Figure 2.4b**, see Method Section). To determine a proper window size for *cis*-QTL mapping, we first tested all SNPs within 2 Mb of m^{6}A peaks (**Supplementary Figure 2.1e**). Most SNPs strongly associated

with m^{6}A are within 100 kb of m^{6}A peaks (**Figure 2.5a**). We therefore restrict our *cis*-tests to SNP-peak pairs within 100 kb. The resulting P values show a strong deviation from the null expectation, while P values from permutations are consistent with the null, indicating that the test is well-calibrated (**Figure 2.5b**). We used a permutation scheme implemented in FastQTL [99] to account for multiple genetic variants tested per peak and qvalue method to account for multiple peaks tested. This results in 822 peaks with at least one significant *cis*-m^{6}A QTL (denoted as ePeaks, following the literature of eQTLs), at 10% FDR (**Supplementary Figure 2.1f**). Most of these ePeaks (86%) have single causal effect (**Supplementary Figure 2.1g**), based on computational fine-mapping analysis using SuSiE [109].

We quantified the contribution of genetic variation to inter-individual variation of m^{6}A levels by estimating the *cis*-heritability of each peak. Most peaks have low heritability values, with 918 peaks having heritability > 0 (**Supplementary Figure 2.2a**). Heritability of ePeaks is higher with median 0.31 (**Supplementary Figure 2.2b**).

We next examined the distribution of m^{6}A QTLs relative to gene-based features, using the program Torus [107]. m^{6}A QTLs are most enriched in 3' UTR ($\log_2$-odds ratio 4.9, $\log_2$-OR hereafter) followed by 5' UTR ($\log_2$-OR 4.5) and CDS ($\log_2$-OR 3.8), but not in intergenic repressive regions as marked by H3K27me3 (**Supplementary Figure 2.2c**).

Comparing m^{6}A QTLs (derived from mRNA m^{6}A peaks) to eQTLs mapped in the same LCLs, we find that genes containing ePeaks and eQTLs are largely distinct (**Figure 2.5c**). In genes with both m^{6}A and expression QTLs, lead SNPs of two types of QTLs are mostly $>$ 10 kb apart and in low linkage disequilibrium (LD) (**Figure 2.5d**, **Supplementary Figure 2.2d**). Comparison of m^{6}A QTLs to sQTLs shows similar patterns (**Supplementary Figure 2.2e-g**). These results suggest that m^{6}A QTLs and eQTLs/sQTLs are distinct types of molecular QTLs.

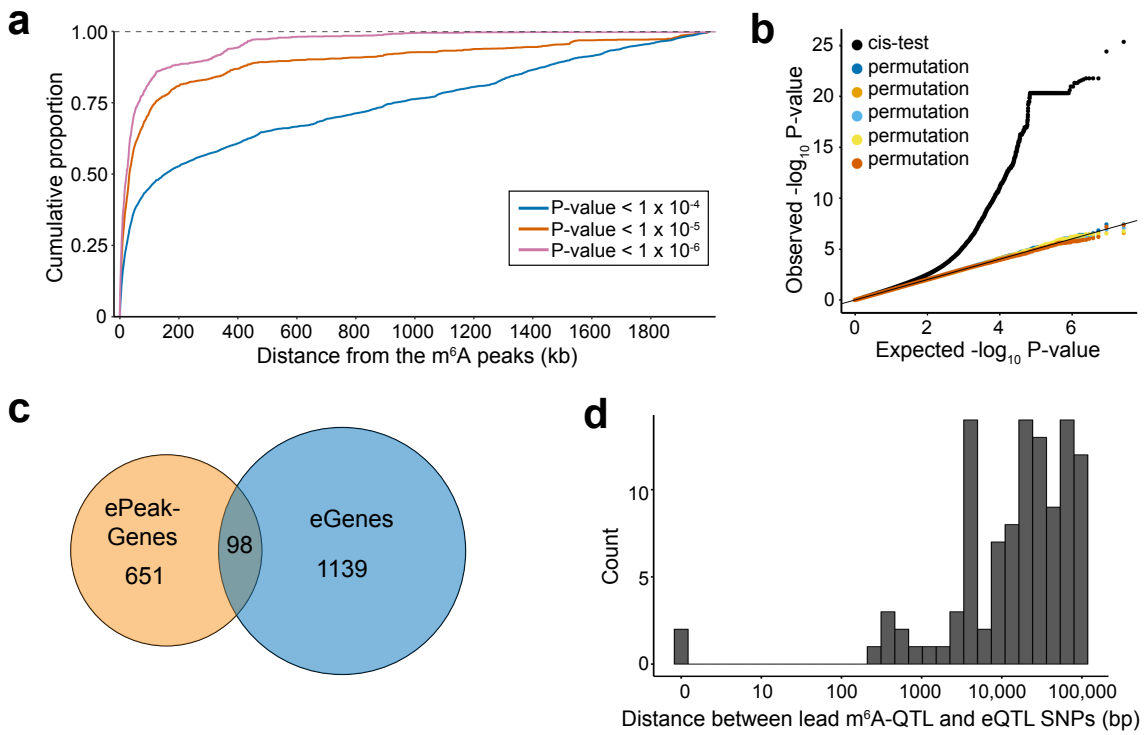


Figure 2.5: Mapping common genetic variants associated with m^{6}A. a, Spatial distribution of m^{6}A QTLs represented by cumulative fraction of SNPs with increasing distance from m^{6}A peaks at varying P value cutoffs of SNP-peak association. **b**, Quantile-quantile (QQ) plot of P values. *cis* tests (n = 60 individuals) results are plotted in black and results of five permutation tests are shown in different colors. **c**, Overlap between ePeak-harboring genes and eGenes (both at FDR < 10%) mapped in the same cohort of YRI LCL samples. **d**, Distribution of the distance between the lead ePeak SNP and the eGene SNP in genes that have both ePeak and eGene mapped.

2.3.3 m^{6}A QTLs are enriched in RNA-related features

To understand which factors may determine m^{6}A deposition, we analyzed features of m^{6}A QTLs and their surrounding sequences. We annotated SNPs using RNA-related features including m^{6}A consensus motif (RRACH) contained in m^{6}A peaks, binding sites of 121 RBPs (ENCODE eCLIP-seq peaks [102]), RiboSNitches [103] (genetic variants changing RNA secondary structure), and predicted microRNA binding sites [104]. We used two approaches to test enrichment. In our primary analysis, we used Fisher's Exact test, comparing possible causal variants of m^{6}A from fine-mapping using SuSiE and background control SNPs. To

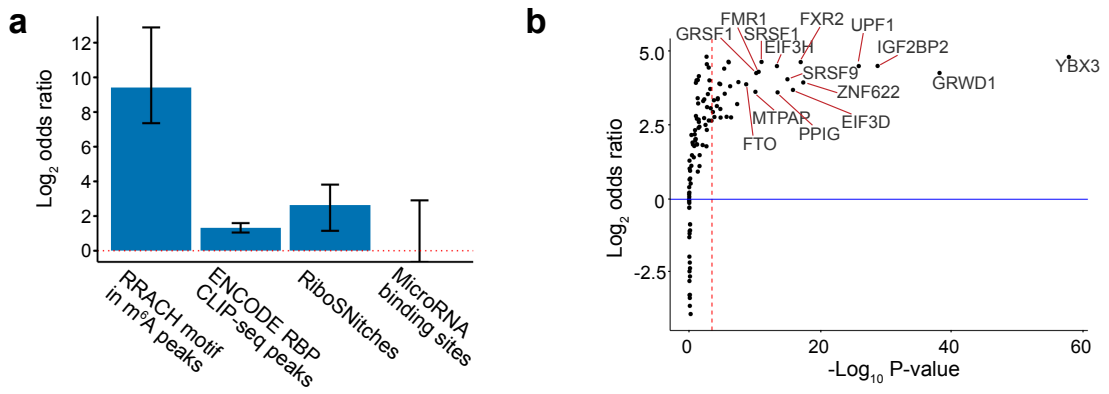


Figure 2.6: Functional features enriched in m^{6}A QTLs. **a**, $\log_2$ odds ratio enrichment of fine-mapped m^{6}A QTLs (SNP with the highest posterior inclusion probability, or PIP, in each credible set) vs. random control SNPs in RNA features by Fisher's exact test. Error bars represent 95% confidence intervals from two tailed tests. **b**, Enrichment of m^{6}A QTLs in RNA binding protein (RBP) binding sites of individual RBPs using eCLIP-seq data from ENCODE50 by Torus39,49. The red dashed line represents the Bonferroni corrected P value 0.05 threshold.

control possible confounding factors, we randomly sampled 100 sets of control SNPs that match key properties of m^{6}A QTLs including minor allele frequencies, numbers of SNPs in LD, distances to the nearest genes, gene density as well as annotations about SNP locations relative to genes (5'UTR, CDS, 3'UTR, intron and intergenic regions). We find m^{6}A consensus motif in m^{6}A peaks highly enriched in m^{6}A QTLs with odds ratio (OR) 685 ($P = 1.0 \times 10^{-33}$).

However, only 12% of m^{6}A QTLs contained in m^{6}A peaks (most QTLs are outside peaks) disrupt the consensus motif. Despite this, we find m^{6}A QTLs tend to be located in proximity to the consensus motif as additional 33% of m^{6}A QTLs contained in m^{6}A peaks are located within 50 bp of the motif and 46% within 100 bp, suggesting many m^{6}A QTLs may indirectly affect binding of the methyltransferase, demethylase or reader proteins to the consensus motif. We also find enrichment in RiboSNitches (OR 6.2, $P = 5.9 \times 10^{-4}$) and RBP binding sites (OR 2.5, $P = 8.3 \times 10^{-19}$) but not predicted microRNA targets (**Figure 2.6a**). As a secondary analysis, we tested enrichment using Torus, which accounts for uncertainty

of causal variants due to LD. This analysis reveals similar results (**Supplementary Figure 2.3a**).

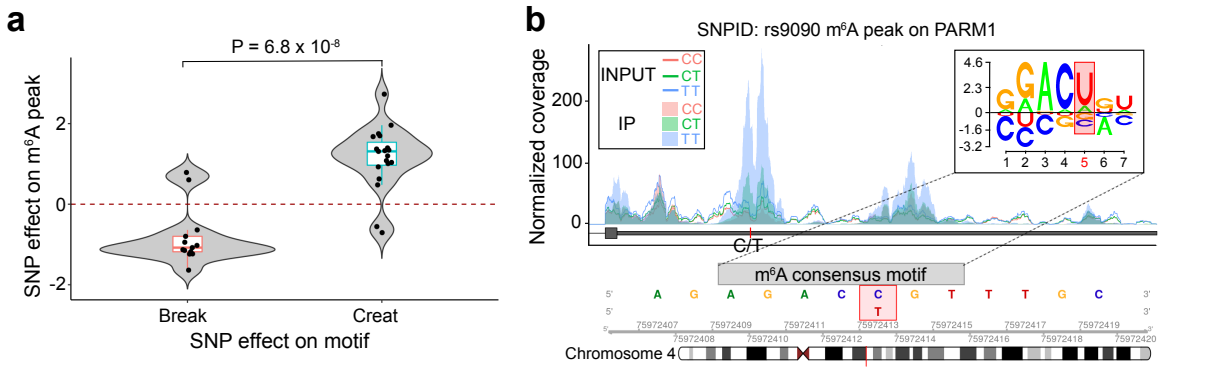


Figure 2.7: m^{6}A QTLs may affect m^{6}A level via disrupting RRACH motif. a, Distribution of m^{6}A QTL effect sizes between SNPs creating vs. breaking the m^{6}A consensus motif. *P* value was computed using Welch's test (n = 32 SNPs). The lower and upper hinges correspond to the first and third quartiles. Horizontal line indicates median value, and whiskers correspond to the value no further than 1.5× inter-quartile range. **b**, An m^{6}A QTL example illustrating how a genetic variant disrupting a RRACH motif could lead to m^{6}A variation.

We tested enrichment for each RBP (**Supplementary Table 2.1**) and microRNA (**Supplementary Table 2.2**) separately using Torus. Interestingly, several of the most enriched RBPs are known m^{6}A-interacting proteins including FTO, a m^{6}A demethylase [12, 11], and IGF2BP2, a m^{6}A reader protein that stabilizes nuclear RNA [20] (**Figure 2.6b**). Although we didn't observe enrichment of m^{6}A QTLs in combined microRNA binding sites, analyses of individual microRNAs show enrichment of m^{6}A QTLs in binding sites of hsa-miR-582-5p and hsa-miR-331-3p. This finding is in line with previous reports that microRNA could affect m^{6}A levels [116].

2.3.4 m^{6}A QTLs' effects on RBP motifs are correlated with their effects on m^{6}A level

The enrichment of binding sites of an RBP in m^{6}A QTLs could occur if binding sites of an RBP co-occur with *cis*-elements regulating m^{6}A, without the RBP playing a direct role in

m^{6}A deposition. We reason that if the RBP is causal, alterations in motif scores (disruption

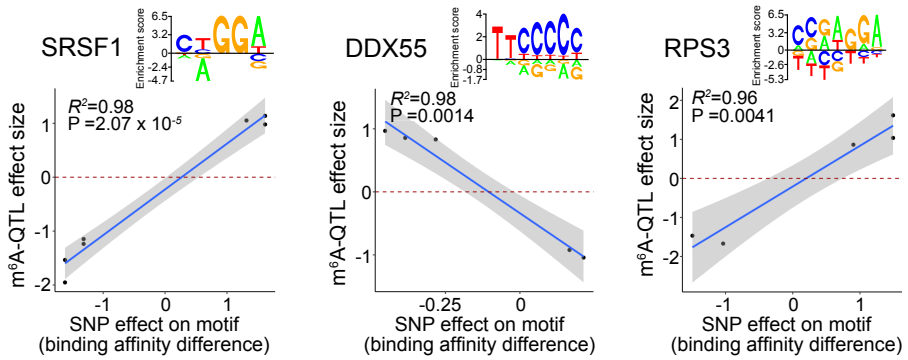


Figure 2.8: m^{6}A QTL effects on m^{6}A levels correlate with their effects on RBP motifs. RBPs for which changes in binding affinities are significantly correlated with fine-mapped m^{6}A QTL effect sizes (all SNPs with PIP > 0.5, and maximum PIP SNPs for ePeaks without SNP PIP > 0.5). Changes in binding affinity are represented by the alteration of motif match scores from the reference to alternative allele. Shaded region and line show the 95% confidence interval and fitted line from the linear model (n = 7, 5 and 5 SNPs for SRSF1, DDX55 and RPS3, respectively).

or creation) of SNPs should correlate with their effects on m^{6}A deposition. We limited this correlation analysis to fine-mapped m^{6}A QTLs that also have significant effects on motif score. As a proof of principle, DNA variants creating a consensus motif are much more likely to be positively associated with m^{6}A levels (**Figure 2.7a**, an example in **Figure 2.7b**). We tested 29 RBPs with > 5 data points, and identified three RBPs with significant correlations at FDR 10% (**Figure 2.8**). Interestingly, SRSF1 is a known splicing factor [117], suggesting a possible connection of splicing with m^{6}A deposition.

2.3.5 m^{6}A QTLs are enriched in transcription features

Recent studies suggest that the deposition of m^{6}A may occur co-transcriptionally and be influenced by transcription processes [111, 118, 119]. We used our m^{6}A QTLs and ENCODE ChIP-seq data from LCLs to examine the potential link between m^{6}A and transcription [120, 56]. We observed significant enrichment (Fisher's exact test) of fine-mapped m^{6}A

QTLs in RNAPII, phosphor-RNAPII and transcription factor binding sites (TFBSs) as well as in histone marks of promoters (H3K4me3), enhancers (H3K4me1, H3K27ac) and active transcription (H3K36me3) (**Figure 2.9a**). The enrichment of m^{6}A QTLs in H3K36me3, the most enriched feature, remains strong when conditioned on other histone modifications in enrichment test using Torus (**Supplementary Figure 2.3b**).H3K36me3 was shown by a recent study [121] to be recognized by the m^{6}A writer protein METTL14 to facilitate m^{6}A installation on mRNA, thus validating our finding.

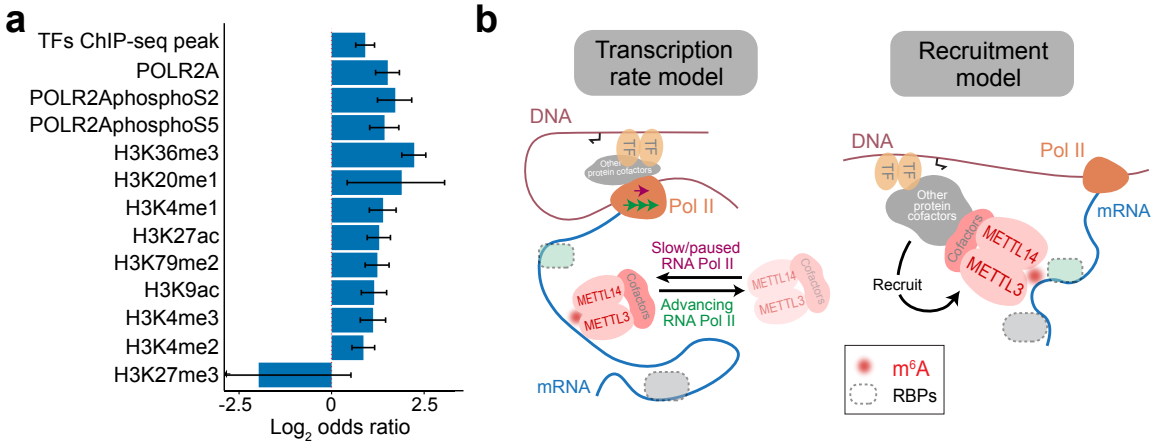


Figure 2.9: m^{6}A installation is coupled with transcriptional processes. a, Enrichment of fine-mapped m^{6}A QTLs (SNP with the highest posterior inclusion probability, or PIP, in each credible set) in chromatin features by the two-sided Fisher's exact test comparing m^{6}A QTLs to control SNPs. The error bars represent 95% confidence intervals. **b**, Two possible models of m^{6}A regulation through transcription.

We then compared contributions of RNA-related and transcriptional features (TFBSs) to m^{6}A QTLs. We used fine-mapping to quantify the probability of a SNP being a causal variant of m^{6}A, known as Posterior Inclusion Probabilities (PIPs). We estimated the proportion of causal variants attributed to a feature by summing the PIPs of all variants located within that feature (see Method Section). Using this approach, we find that TFBSs and RBP binding sites make about equal contribution to m^{6}A QTLs (17.8% and 15.8%, respectively) and RRACH motif contributes 1.95% (**Supplementary Figure 2.3c**).

These findings support a tight connection between transcriptional processes and m^{6}A

installation. Two models have been suggested to explain this connection (**Figure 2.9b**). In the first model ("transcription rate model"), m^6A installation is affected by the progression rate of RNAPII, with fast progression associated with lower m^6A methylation [118]. In the second model ("TF recruitment model"), the methyltransferase complex is recruited to mRNA by TFs, e.g. ZFP217 [119], CEBPZ [85] or adaptor proteins, e.g. SMAD2/3 [111].

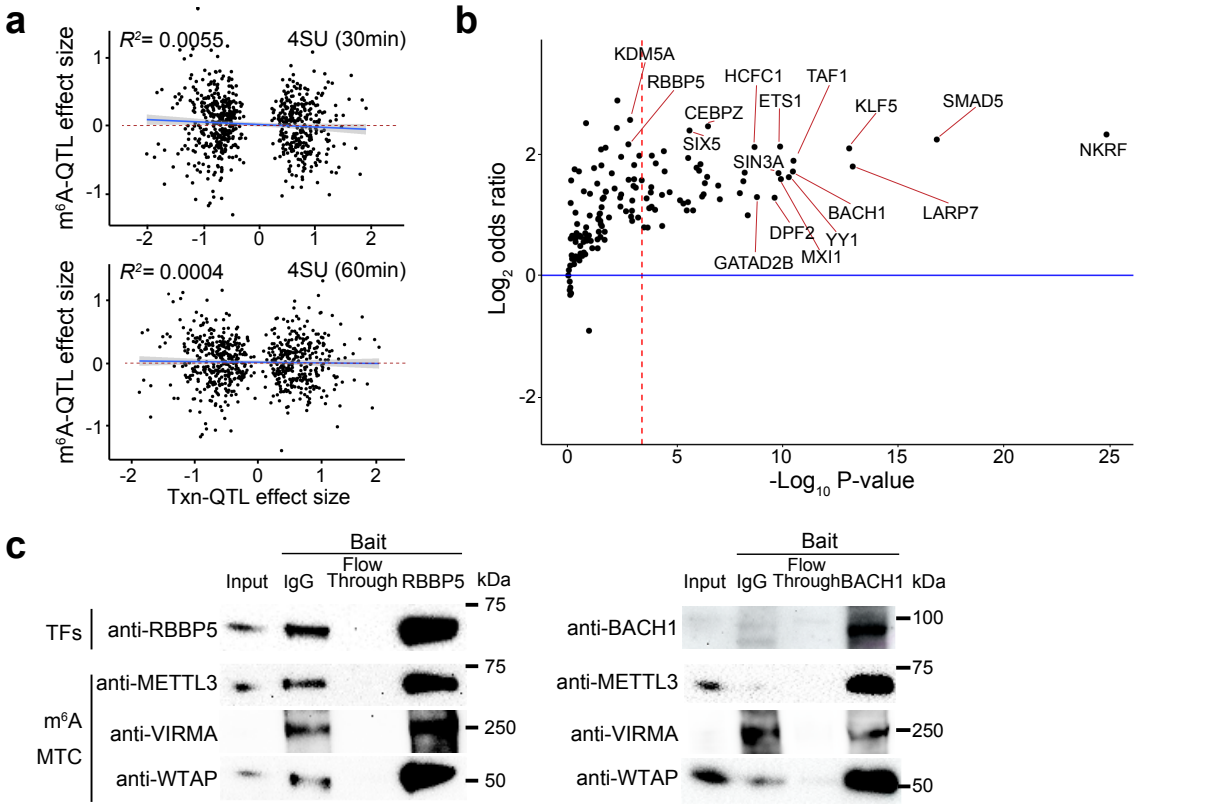


Figure 2.10: Test two models of transcriptional regulation of m^6A. a, Effect sizes of ascertained transcription rate QTLs (Txn-QTLs) vs. their effects on m^6A level. The transcription rate was measured by 4sU-seq in an earlier study [13]. 4sU-seq of 30 mins 4sU labeling (upper, n = 698 SNPs) and 60 mins 4sU labeling (lower, n = 688 SNPs) show similar results. Shaded region and line show the 95% confidence interval and fitted line from the linear model. **b**, Enrichment of m^6A QTL in transcription factor (TF) binding sites of individual TFs conditioned on H3K27ac peaks by Torus analysis. The red dashed line shows the Bonferroni corrected P value 0.05 cutoff. **c**, Western blot of transcription factor (TF) co-IP experiment. 10% of lysate was loaded as "input". The cropped blot of each TF of interest is shown, as well as three m^6A methyltransferase complex components—METTL3, WTAP and VIRMA. These experiments were repeated twice with similar results.

If the transcription rate model is correct, we expect correlation between variant effects on

transcription rate and variant effects on m^{6}A level in the matched transcript. To assess this, we ascertained the lead SNPs of transcription-rate-QTLs from the same LCL samples [47], but find little correlation between transcription rates and m^{6}A effect sizes (**Figure 2.10a**). As a positive control, we observed strong correlation of transcription-rate-QTL effects with eQTL effects (R^2 0.69 and 0.65) and with protein-QTL effects (R^2 0.37 and 0.42) in the matched transcripts (**Supplementary Figure 2.3d, e**). These data suggest that overall transcription rate may not determine m^{6}A deposition in LCLs. It remains possible that other mechanisms such as RNAPII pausing explain the observed correlation between m^{6}A and transcription rates in an earlier study [118].

To examine the "TF recruitment" model, we used Torus to assess enrichment of m^{6}A QTLs for binding sites of individual TF while accounting for enrichment of m^{6}A QTLs in H3K27ac, a general transcription marker. 50 TFs are significantly enriched at Bonferroni corrected P value < 0.05 (**Figure 2.10b**). We then selected a few of these based on literature review and performed co-immunoprecipitation (co-IP) experiments. Two TFs robustly pull down m^{6}A methyltransferase components in LCLs, including RBBP5, a component of COMPASS histone H3K4 methylase complex, and BACH1, a regulator of oxidative stress [122, 123] (**Figure 2.10c**), supporting the "TF recruitment model".

2.3.6 QTL sharing between m^{6}A and related molecular traits

It is generally believed that specific reader proteins recognize m^{6}A and mediate downstream effects [2, 1]. The best-studied readers are known to promote translation (e.g. YTHDF1), mRNA degradation (e.g. YTHDF2), or affect mRNA nuclear processing (e.g. YTHDC1) [13, 111, 18, 17]. We use m^{6}A QTLs as natural perturbations of m^{6}A to explore its effects on five possible downstream traits: mRNA expression, ribosome binding, protein level, mRNA decay rate and alternative polyadenylation (APA) [55, 48, 59].

We first ascertained lead m^{6}A QTLs at different P value thresholds, and then estimated the percentage of m^{6}A QTLs that are also QTLs of other traits using Storey's π1 method

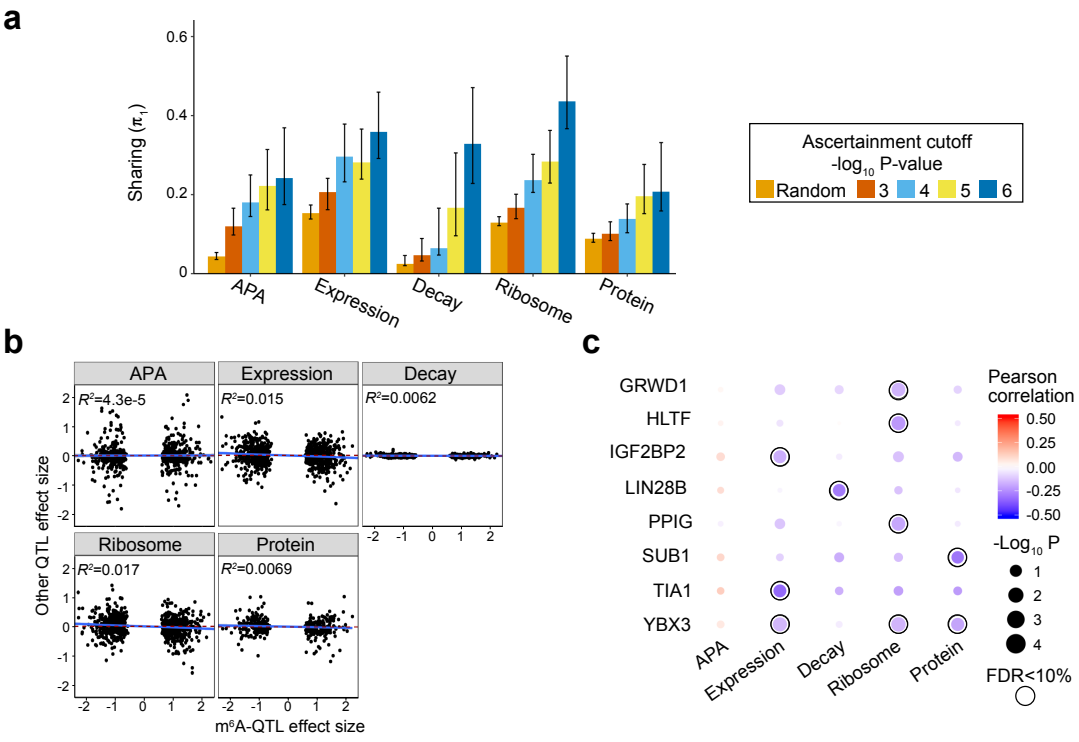


Figure 2.11: Joint analysis of m^{6}A QTLs and other molecular QTLs. **a**, The estimated fractions of m^{6}A QTLs shared with other molecular phenotypes, measured by $\pi 1$, the fraction of true positives. The five bars in each panel correspond to random SNPs and m^{6}A QTLs at different P value cutoffs. Error bars show 80% confidence intervals (n = 100 bootstraps). **b**, Low correlations of effect sizes between m^{6}A QTLs and related molecular QTLs, estimated from linear regression (n = 709, 884, 742, 884 and 393 SNPs-gene pairs for APA, Expression, Decay, Ribosome and Protein, respectively). **c**, Correlations of effect sizes between m^{6}A QTLs and related molecular traits QTLs stratified by the m^{6}A sites bound by different RBPs. Correlations are determined by linear regression. Shown are RBPs having at least one trait significantly correlated trait with m^{6}A at FDR $< 10\%$. Pearson correlations are shown by color code and P value by dot size.

following the procedure in reference [100, 47]. We find m^{6}A QTLs are more likely to be other QTLs than random SNP-gene pairs, with increased sharing at more stringent P value thresholds (**Figure 2.11a**), suggesting functional connections between m^{6}A and other molecular phenotypes, as expected from earlier studies [2, 1]. The $\pi 1$ values are generally lower than those between QTLs along the cascade from transcription to protein [47]. One potential problem is that sharing of m^{6}A QTLs and other molecular QTLs may be confounded by eQTLs, as transcription may influence both m^{6}A and other traits. However, the majority

of m^6A QTLs mapped in this study are not chromatin-located eQTLs, suggesting that in practice, this is not a large concern (**Supplementary Figure 2.4a**).

2.3.7 m^6A-QTLs effect sizes are correlated with related molecular trait QTLs effect size in a context-dependent manner

Based on our current understanding that m^6A function is mediated by reader proteins with certain downstream effects (e.g. increase of translation efficiency by YTHDF1), we hypothesize that m^6A QTLs and other molecular QTLs have consistent effect directions. To test this hypothesis, we first confirmed that molecular traits along the cascade from transcription to translation show high positive correlations in QTL effects (**Supplementary Figure 2.4b**). Surprisingly, the effect sizes of m^6A QTLs and other molecular traits are poorly correlated (**Figure 2.11b**). This lack of overall correlation suggests that effects of m^6A on downstream molecular phenotypes may be heterogeneous, with mechanisms varying across transcripts.

The context dependency of m^6A function may be partially mediated by RBPs bound near m^6A sites. For example, binding by different m^6A readers may lead to different downstream effects. To examine this hypothesis, we stratified our effect size correlation analysis by m^6A peaks bound by different RBPs (using eCLIP-seq data). In 8 RBPs, we observed significant correlations (FDR $< 10\%$) between effect sizes of m^6A QTLs and at least one related molecular trait (**Figure 2.11c**). This result suggests that m^6A function may vary according to binding of specific RBPs.

2.3.8 Re-analysis of ribosome-profiling data in METTL3 depleted HeLa cells

To further investigate context-dependent effects of m^6A, we made use of data from an earlier study [13] of m^6A effects on translation in HeLa cells. This study examined the impacts of m^6A depletion (by METTL3 knockdown) and YTHDF1 (m^6A reader) knockdown on translation efficiency (TE) of all transcripts, measured by ribosome profiling. Across all m^6A

modified transcripts, the effects of m^6A depletion on TE are heterogeneous, with similar numbers of up- and down-regulated genes (**Supplementary Figure 2.4c**). To assess the impact of RBP context, we compared the effects of m^6A depletion on TE of transcripts containing m^6A sites targeted by an RBP vs. transcripts not targeted. Among 121 tested RBPs, 33 showed statistically significant differences in RBP-targeted m^6A transcripts vs. non-target transcripts (FDR $< 5\%$) (**Figure 2.12a**). Again, the effects are quite heterogeneous with almost equal numbers of RBPs involved in up- and down-regulation of TE upon m^6A depletion. This list of RBPs includes all four RBPs: YBX3, GRWD1, HLTF and PPIG, we identified from m^6A vs. ribosome QTL effect size correlation analysis (**Figure 2.11c**). Furthermore, the effect directions are consistent between two analyses: m^6A depletion resulted in higher TE of the RBP's targets, in agreement with negative correlations of m^6A versus ribosome QTL effects (**Figure 2.11c and Figure 2.12a**). These results provide independent support to the hypothesis that the effects of m^6A on TE depend on contexts, in particular binding of specific RBPs. Interestingly, even in transcripts targeted by the classical m^6A reader, YTHDF1, the effect of m^6A may be more complex than previously thought. While depletion of YTHDF1 leads to an overall reduction of TE in transcripts harboring YTHDF1-bound m^6A peaks, approximately 33% YTHDF1 targets show opposite effects (**Supplementary Figure 2.4d**). This observation suggests the possibility that the action of reader proteins may be modulated by additional, yet-to-identify factors, diversifying m^6A effects.

2.3.9 Validation of YBX3 as a translation repressor for YBX3-bound m^6A-modified transcripts

We validated a newly identified m^6A effector protein, YBX3, as a translation repressor of m^6A-modified and YBX3-bound transcripts (**Figure 2.12b**). We knocked down YBX3 in HeLa cell and performed polysome profiling followed by RT-qPCR. We find more RNAs in polysome-bound fractions in YBX3-depleted cells compared with control (**Supplementary**

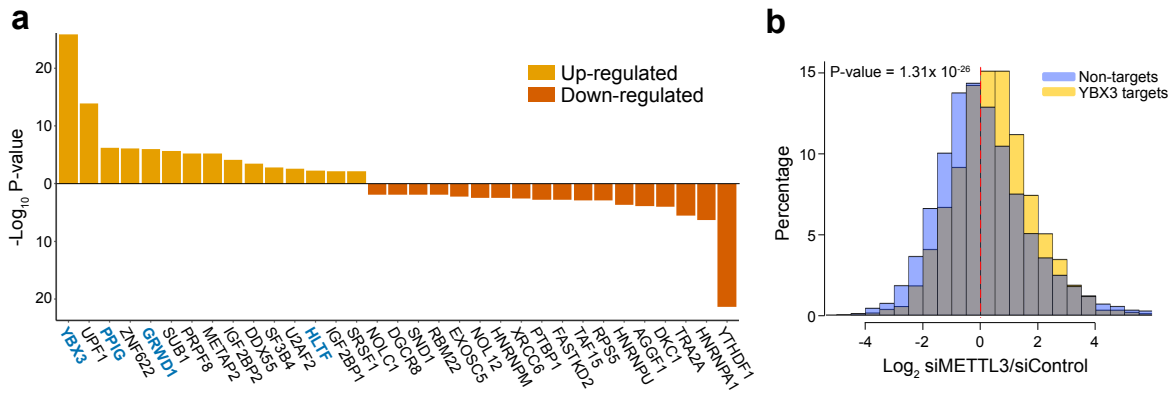


Figure 2.12: Re-analysis of ribosome profiling data in METTL3 knockdown HeLa cells. **a**, RNA binding proteins (RBPs) that modulate the impact of m^{6}A depletion on translation efficiency. For each RBP, Welch's two-sided t-test is used to test the $\log_2$ fold change in translation efficiency in RBP targets vs. non-targets, upon METTL3 knockdown. RBP targets are defined by transcripts harboring m^{6}A peaks that are bound by certain RBP. Shown are RBPs with FDR $< 5\%$ (Benjamini & Hochberg method). RBPs with significant correlation between m^{6}A QTL and ribosome-QTL effect sizes are highlighted in blue. **b**, Distribution of $\log_2$ fold change in translation efficiency of YBX3 targets, upon m^{6}A (METTL3) depletion, in comparison with non-targets. P value is computed by Welch's two-sided t-test.

Figure 2.5a), suggesting YBX3 as a translation repressor. To further validate YBX3 function, we selected five transcripts harboring m^{6}A peaks overlapping with YBX3-bound sites, all of which show elevated TE upon METTL3 knockdown (**Figure 2.12b**). We quantified these transcripts in three polysome-bound fractions using RT-qPCR upon YBX3 knockdown. TE of these target transcripts is elevated in YBX3 depleted cells compared with control in all but one case (**Supplementary Figure 2.5b**). As a negative control, three YTHDF1-targeted m^{6}A transcripts do not show TE elevation upon YBX3 depletion. These results suggest that YBX3 likely mediates m^{6}A effect by repressing translation of YBX3-bound m^{6}A transcripts. This effect is opposite from YTHDF1 effect (increasing translation), providing a partial explanation of why m^{6}A downstream effects appear heterogeneous (**Figure 2.11b and Figure 2.12a**).

2.3.10 m^{6}A-QTLs are enriched for GWAS signals

To study the role of m^{6}A QTLs in human phenotypic variations, we collected GWAS summary statistics of 45 complex traits with an emphasis on immune and blood-related traits. For comparison, we also included eQTLs and splicing QTLs (sQTLs) from LCLs. All three types of QTLs show excess of low P values in GWAS of these traits (**Supplementary Figure 2.6a**), e.g. lymphocyte counts (**Figure 2.13a**).

2.3.11 Partition of GWAS trait heritability among functional features by S-LDSC

We used Stratified LD score regression (S-LDSC) to formally test enrichment of GWAS variants in m^{6}A QTLs [115, 124, 125]. S-LDSC is a tool for assessing how the heritability of a complex trait is partitioned among functional features, while controlling for LD, allele frequency and other baseline features. Following a previously used strategy [50, 72], we fine-mapped m^{6}A QTLs (using SuSiE) and used the resulting PIPs as an annotation, representing likely causal m^{6}A variants (known as Quantitative Trait Nucleotides, or QTNs). We find 10- to 20-fold enrichment of heritability in m^{6}A QTNs in several selected traits (**Figure 2.13b**, **Supplementary Figure 2.6b**). The enrichment increases to 15- to 35-fold (**Supplementary Figure 2.7a**), when we used m^{6}A-related annotations (**Supplementary Figure 2.7b**), such as RBP binding, as priors to improve fine-mapping (see Methods). Including QTNs of expression and splicing in S-LDSC only modestly reduced the enrichment level (**Figure 2.13b**). We note, however, that m^{6}A may affect expression (e.g. by changing mRNA stability) and pre-mRNA splicing, therefore adjusting eQTLs and sQTLs likely leads to underestimation of m^{6}A QTL enrichment. Expanding S-LDSC analysis to all 45 traits, we find that m^{6}A-PIPs are enriched in most immune and blood traits (**Figure 2.13c**, **Supplementary Figure 2.7c**), and a small number of other traits such as Coronary Artery Disease (CAD) and age at menopause, in which immune systems may play a significant role

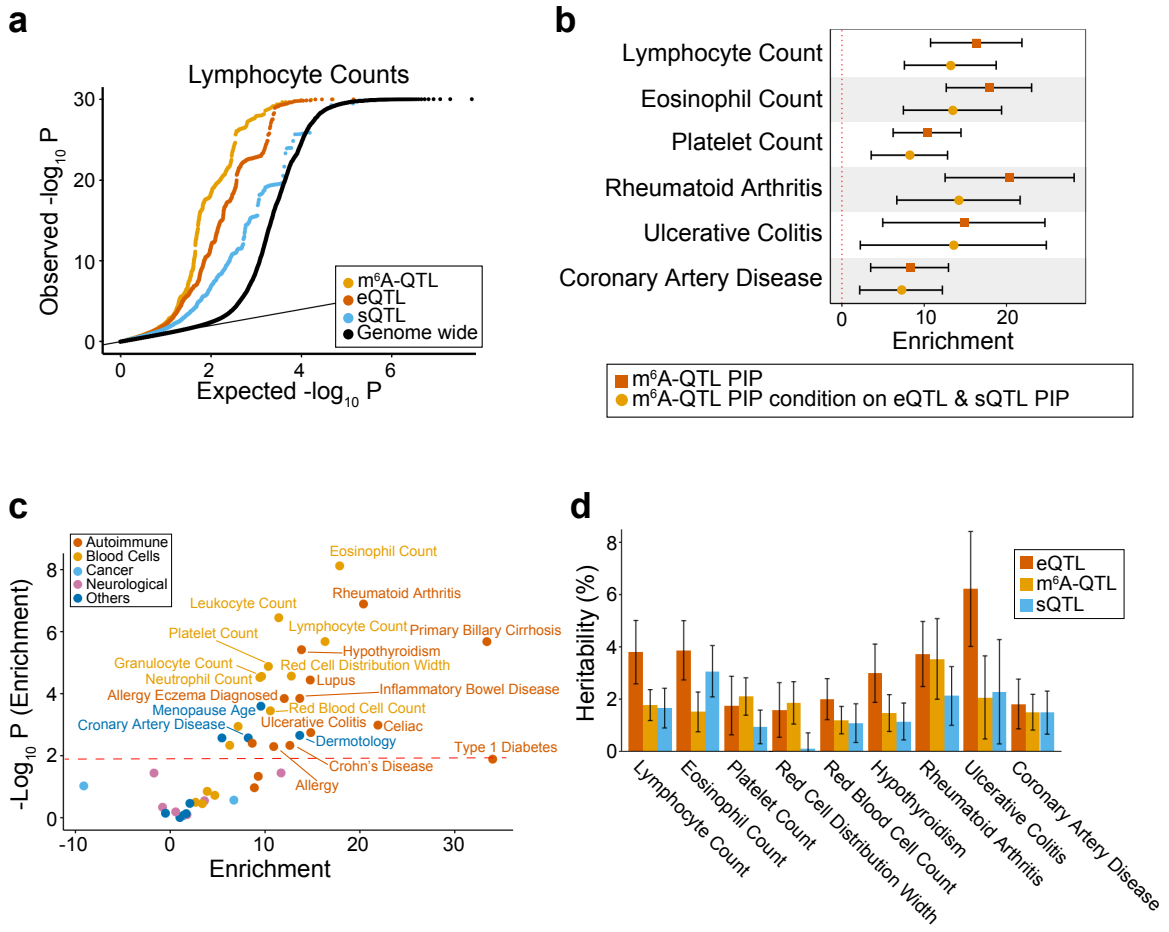


Figure 2.13: Integrated analysis of m^{6}A QTLs and GWAS data of human complex traits. **a**, Quantile-quantile (QQ) plot of lymphocyte count GWAS P values. m^{6}A QTLs, eQTLs and sQTLs are shown in comparison with genome wide SNPs. GWAS SNPs are binary annotated using m^{6}A QTLs, eQTLs and sQTLs with P value $< 1 \times 10^{-4}$. **b**, Enrichment of selected immune and blood GWAS trait heritability estimated by S-LDSC [115]. We used two QTL annotations: (1) m^{6}A QTL continuous annotation using PIP from fine-mapping (with uniform prior); (2) m^{6}A QTL PIP annotation conditional on eQTL and sQTL PIP annotations (with uniform prior). Error bars represent the 95% confidence intervals. **c**, Summary of GWAS traits heritability enrichment for all 45 traits using m^{6}A QTL PIP (with uniform priors) as annotation conditional on the baseline LD model. The dashed line shows the significance threshold at FDR 5%. **d**, Proportion of GWAS traits heritability explained by m^{6}A QTLs, eQTLs and sQTLs. Because it would be hard to estimate heritability contribution using PIP (continues annotation) from fine-mapping, we used binary annotations at SNP-level FDR 10% threshold in this analysis. Error bars represent standard errors.

[126, 127, 128]. These results thus confirm the specificity of our finding as we mapped m^{6}A QTL in a cell line of immune lineage, and are consistent with the known role of m^{6}A in the immune system [129, 130, 131, 29].

Using S-LDSC, we compared the relative contributions to trait heritability by m^{6}A QTLs, eQTLs and sQTLs (FDR $< 10\%$). For traits in which m^{6}A QTNs show enrichment (**Figure 2.13c**), m^{6}A QTLs explain about 2-4% of heritability, comparable to sQTLs and roughly 50-100% of the heritability explained by eQTLs (**Figure 2.13d**, **Supplementary Figure 2.8**). These estimates are likely conservative, as many QTLs below the FDR cutoff may contribute to trait heritability. Nevertheless, given the established roles of eQTLs and sQTLs in complex traits, this relative comparison suggests that m^{6}A QTLs can make a significant contribution to heritability of complex traits.

2.3.12 Transcriptome-wide association studies (TWAS) using m^{6}A QTLs as a molecular trait

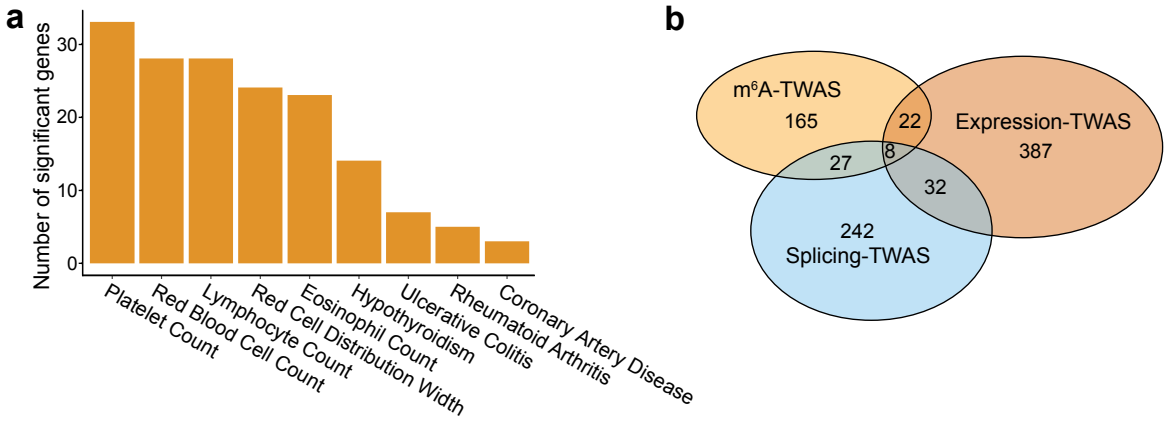


Figure 2.14: m^{6}A-TWAS result summary and comparison with expression-TWAS and splicing-TWAS. a, Number of significant m^{6}A-TWAS genes in selected immune and blood-related traits. **b**, Overlaps between significant genes discovered by TWAS analyses using m^{6}A, expression and splicing as molecular-level phenotypes.

To highlight the potential of using m^{6}A QTLs to identify specific risk genes, we performed TWAS [83, 132] using m^{6}A as a molecular-level trait. We built prediction models of m^{6}A

levels using genetic variants as explanatory variables, then assessed if genetically determined m^{6}A levels correlate with specific phenotypes using TWAS/FUSION [83]. We find a number of m^{6}A peaks passing Bonferroni threshold across a range of immune and blood-related traits (**Figure 2.14a**) as well as several other phenotypes (**Supplementary Figure 2.9a**). These results show limited overlap, at the level of genes, with TWAS results using eQTLs and sQTLs mapped in LCLs (**Figure 2.14b**). This suggests that m^{6}A variation represents a distinct path from genetic to phenotypic variation.

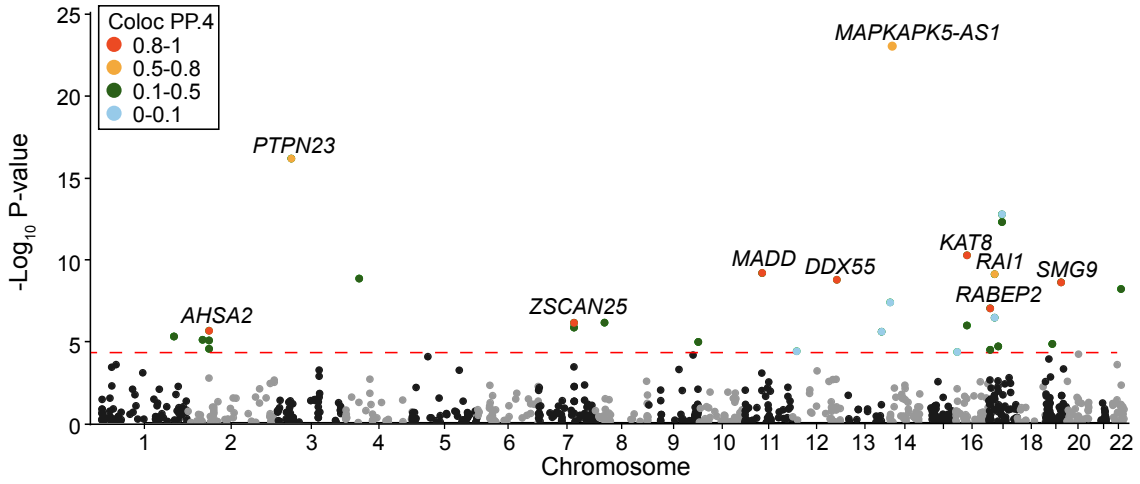


Figure 2.15: m^{6}A-TWAS identify m^{6}A peaks associated with lymphocyte count. Manhattan plot of m^{6}A-TWAS associations of lymphocyte count. The dashed line shows the Bonferroni corrected P value threshold of 0.05. Genes are colored by Coloc PP.4 (posterior probability of GWAS trait and m^{6}A QTL sharing common genetic causal variants). 10 genes with Coloc PP.4 > 0.5 are labeled.

We performed an in-depth analysis of lymphocyte count. m^{6}A-TWAS identified a total of 30 significant m^{6}A peaks in 28 genes (**Figure 2.15**). Since TWAS associations can result from LD and/or pleiotropic effects [132], we conducted colocalization analysis [81] to identify cases where a single causal variant drives both m^{6}A QTL and GWAS association. Among 30 peaks, 10 have high colocalization probabilities (PP4 from Coloc > 0.5) (**Supplementary Table 2.3**). As an example, a m^{6}A peak in the *DDX55* gene shows high colocalization probability (PP4 = 0.929). The SNP driving colocalization result, rs3204541, is the top

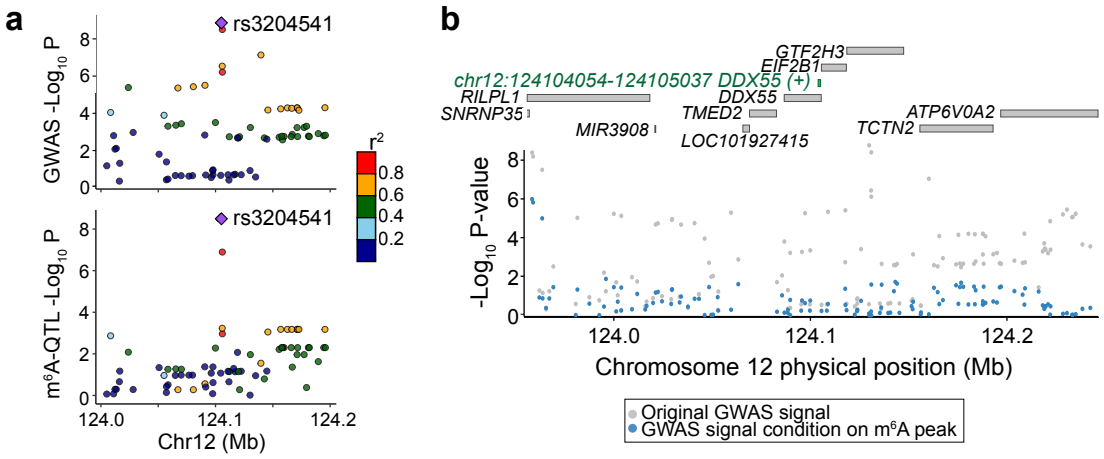


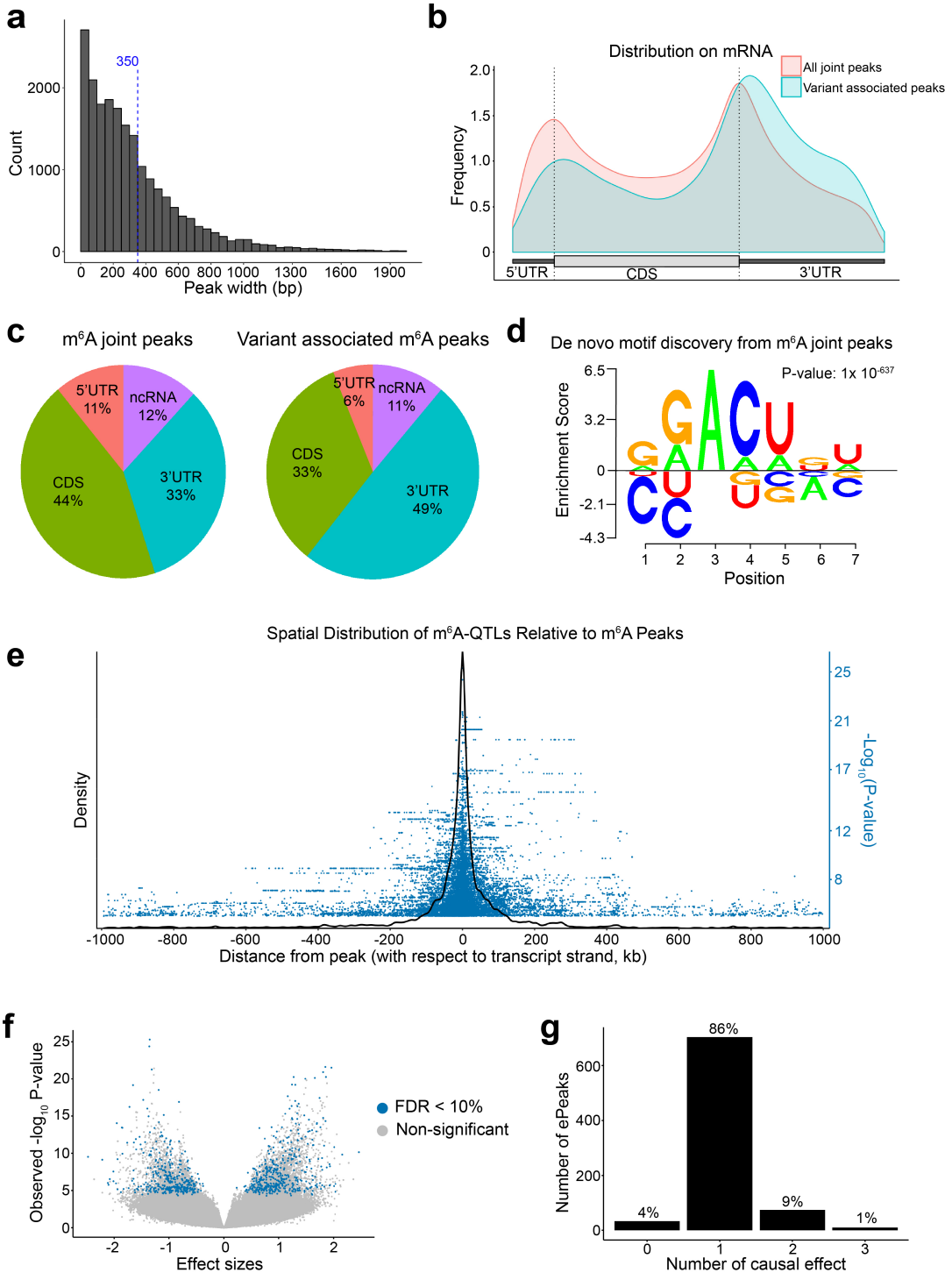
Figure 2.16: Colocalization of m^{6}A QTL and lymphocyte count variant at *DDX55* locus. a, Aligned Manhattan plots of GWAS and m^{6}A QTL at *DDX55* locus generated by LocusCompare. SNPs are colored by LD (r^2) with the lead m^{6}A QTL (rs3204541). **b**, Manhattan plot of GWAS association signal of lymphocyte count at *DDX55* locus before (gray dots) and after (blue dots) conditioning on the TWAS-predicted m^{6}A level. The top panel labels all genes within 200 kb and the significant m^{6}A peak (green).

SNP in both m^{6}A QTL and GWAS (**Figure 2.16a**). A conditional association test adjusting for m^{6}A shows that the TWAS association almost fully explains the GWAS signal in the region (**Figure 2.16b**). The same m^{6}A peak in *DDX55* is also found by m^{6}A-TWAS in leukocyte counts (**Supplementary Figure 2.9b and c**). *DDX55* is a DEAD-Box Helicase gene, and its paralog gene, *DDX10* is implicated in Myelodysplastic Syndrome, a disease with abnormal blood cell counts [133]. Importantly, *DDX55* is not found by expression or splicing-TWAS (P values $\geqq 0.1$ in both cases). Together, our TWAS results highlight the promise of using m^{6}A QTLs to reveal mechanisms in GWAS loci where genetic effects are not mediated by expression or splicing.

Gene	Peak	Best GWAS SNP	BEST GWAS Z	m^6A QTL SNP	m^6A QTL Z	m^6A QTL GWAS Z	N SNP	TWAS Z	TWAS P	TWAS Bonferroni P	Coloc PP.0	Coloc PP.1	Coloc PP.2	Coloc PP.3	Coloc PP.4
AHSA2	chr2:61405847 -61406244 _AHSA2_+	rs1177290	5.04	rs777585	-5.15	4.867	50	-4.7291	2.25E-06	2.02E-03	0.023	0.008	0.064	0.02	0.885
PTPN23	chr3:47452762 -47452960 _PTPN23_+	rs295433	12.238	rs13075233	-3.88	-8.361	34	8.3621	6.16E-17	5.53E-14	0.00	0.00	0.288	0.019	0.693
ZSCAN25	chr7:99217253 -99219031 _ZSCAN25_+	rs776746	5.06	rs7808022	5.78	4.923	42	4.9528	7.32E-07	6.57E-04	0.001	0.008	0.003	0.017	0.97
MADD	chr11:47306107 -47308026 _MADD_+	rs11828339	-6.56	rs1051006	5.53	-6.10	53	-6.1757	6.59E-10	5.92E-07	0.00	0.00	0.045	0.026	0.929
MAPKAPK5 -AS1	chr12:112277871 -112278218 _MAPKAPK5-AS1_-	rs4346023	-10.22	rs3177647	3.58	-10.081	46	-10.0584	8.44E-24	7.58E-21	0.00	0.00	0.423	0.022	0.554
DDX55	chr12:124104056 -124105037 _DDX55_+	rs3204541	6.02	rs3204541	5.28	6.024	74	6.0247	1.69E-09	1.52E-06	0.00	0.00	0.007	0.001	0.992
KAT8	chr16:31129034 -31138368 _KAT8_+	rs2359612	7.162	rs4527034	5.72	6.02	28	6.5616	5.32E-11	4.78E-08	0.00	0.00	0.004	0.004	0.992
RABEP1	chr17:5284729 -5286967 _RABEP1_+	rs1046368	5.55	rs1806261	4.90	5.10	64	5.3343	9.59E-08	8.61E-05	0.004	0.002	0.072	0.03	0.893
RAI1	chr17:17696391 -17696987 _RAI1_+	rs3818717	7.08	rs11649804	-5.85	5.982	66	-6.1491	7.79E-10	7.00E-07	0.00	0.00	0.114	0.234	0.652
SMG9	chr19:44235350 -44235742 _SMG9_-	rs346527	6.68	rs12669	4.11	6.462	92	5.9655	2.44E-09	2.19E-06	0.00	0.00	0.174	0.011	0.815

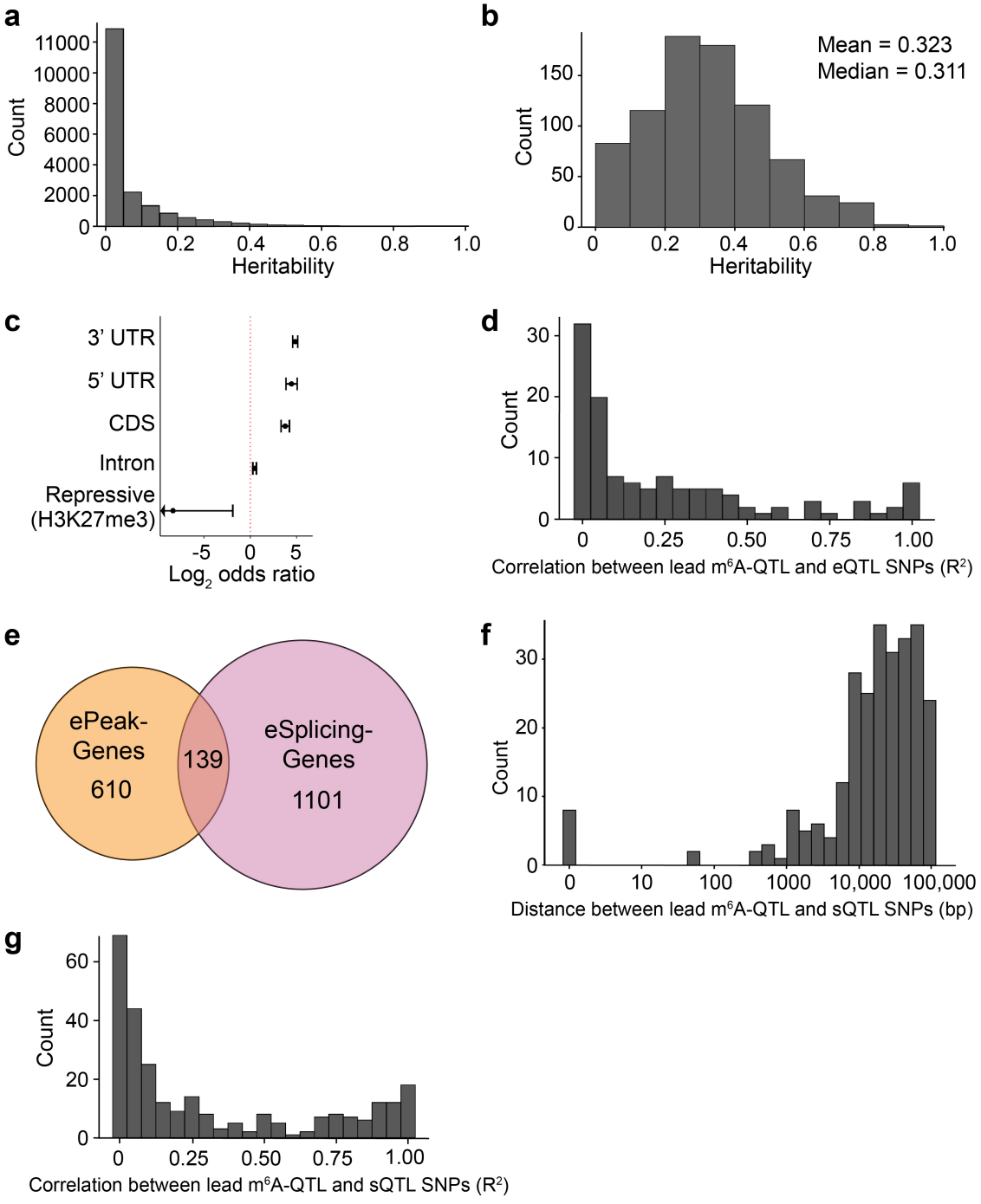
Table 2.3: Coloc result for m^6A-TWAS peaks in lymphocyte count. The colocalization test results for the m^6A-TWAS significant peaks in lymphocyte count that have Coloc PP4 > 0.5. In Coloc result, PP4 represent the posterior probability for the hypothesis that association with m^6A and lymphocyte count has one shared causal SNP.

2.4 Supplementary figures



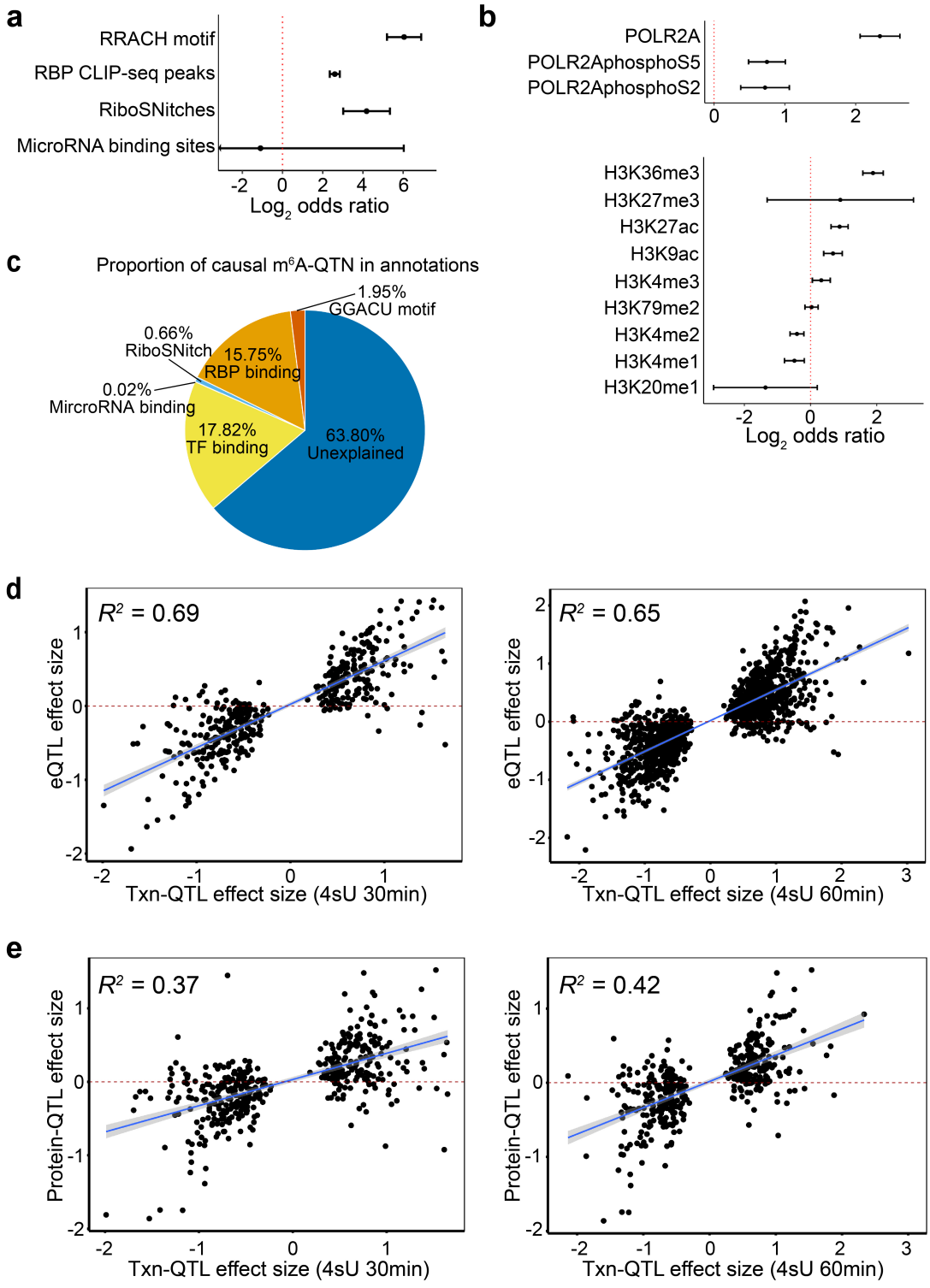
Supplementary Figure 2.1: Joint m^{6}A peak calling and QTL mapping.

Supplementary Figure 2.1: **a**, Distribution of merged m^6A peak length. Dash line marks the mean peak width. **b**, Distribution of all m^6A peaks vs. ePeaks on a meta-gene. **c**, Proportion of all m^6A peaks vs. ePeaks in each genomic annotation. **d**, m^6A motif learned by Homer2, and visualized using EDlogo package. **e**, Spatial distribution of m^6A QTLs illustrated by density plot of SNP to peak distances of m^6A QTL with nominal P value $< 1 \times 10^{-4}$ in a 2Mb window surrounding m^6A peaks. We also showed the significance by the -$\log_{10}$ P value of the association tests in the blue dots. **f**, Volcano plot of overall statistics of m^6A QTLs with peak-level FDR $< 10\%$ (ePeaks). **g**, Distribution of the number of causal effects of ePeaks (FDR $< 10\%$) by SuSiE fine-mapping with uniform prior. We set SuSiE parameters L = 3 (assuming at most three causal effects) and coverage = 0.95 (95% coverage for credible sets).



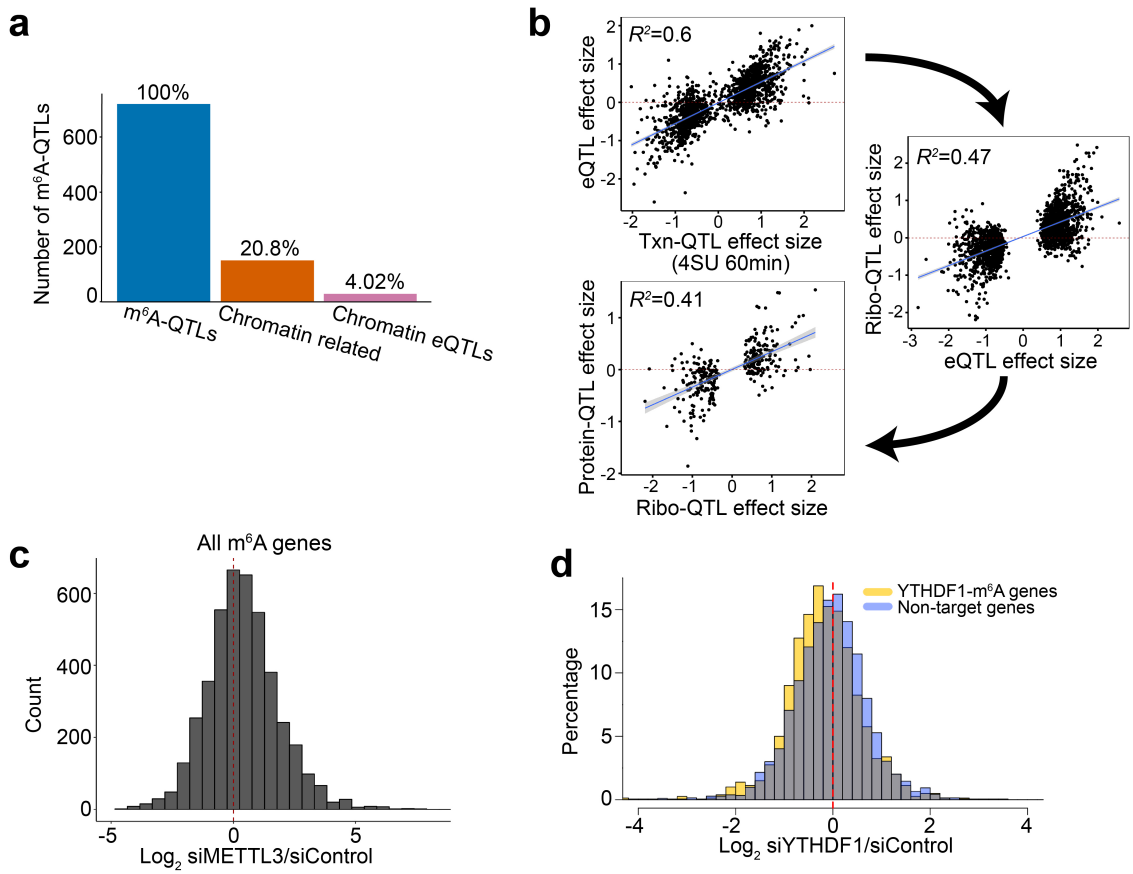
Supplementary Figure 2.2: Heritability of m^{6}A peaks and independence of m^{6}A QTLs, eQTL and sQTLs. a, Distribution of estimated heritability of the 19,130

Supplementary Figure 2.2: peaks included in the TWAS analysis, in which 918 peaks had estimated heritability significantly greater than 0 (minimum heritability P value of 0.01). **b**, Distribution of estimated heritability of ePeaks (n = 822 peaks). **c**, Enrichment ($\log_2$ odds ratio) of m^6A QTLs in gene annotations. **d**, Distribution of the LD between the lead ePeak SNP and the eGene SNP in genes that have both ePeak and eGene mapped. **e**, Overlap between ePeak-harboring genes and eSplicing-harboring (splicing event with at least one significant sQTL) gene (both at FDR $< 10\%$) mapped in YRI LCL samples. **f**, Distribution of the distance between the lead ePeak SNP and the eSplicing SNP in genes that have both ePeak and eSplicing mapped. **g**, Distribution of the LD between the lead ePeak SNP and eSplicing SNP in genes that have both ePeak and eSpicing mapped.

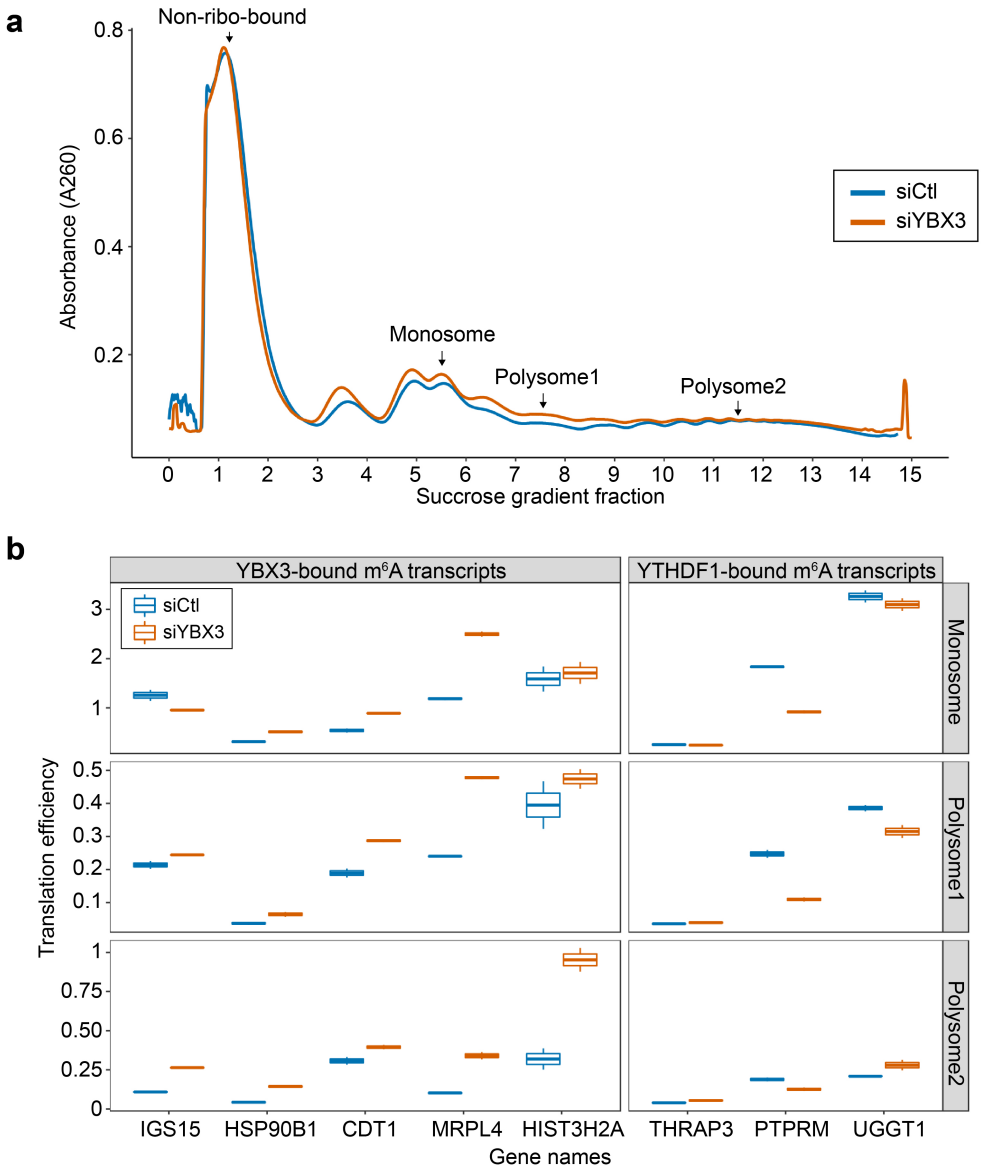


Supplementary Figure 2.3: Contribution of RNA features and transcriptional features to m^{6}A variation. a, Enrichment of m^{6}A QTLs in RNA related features by

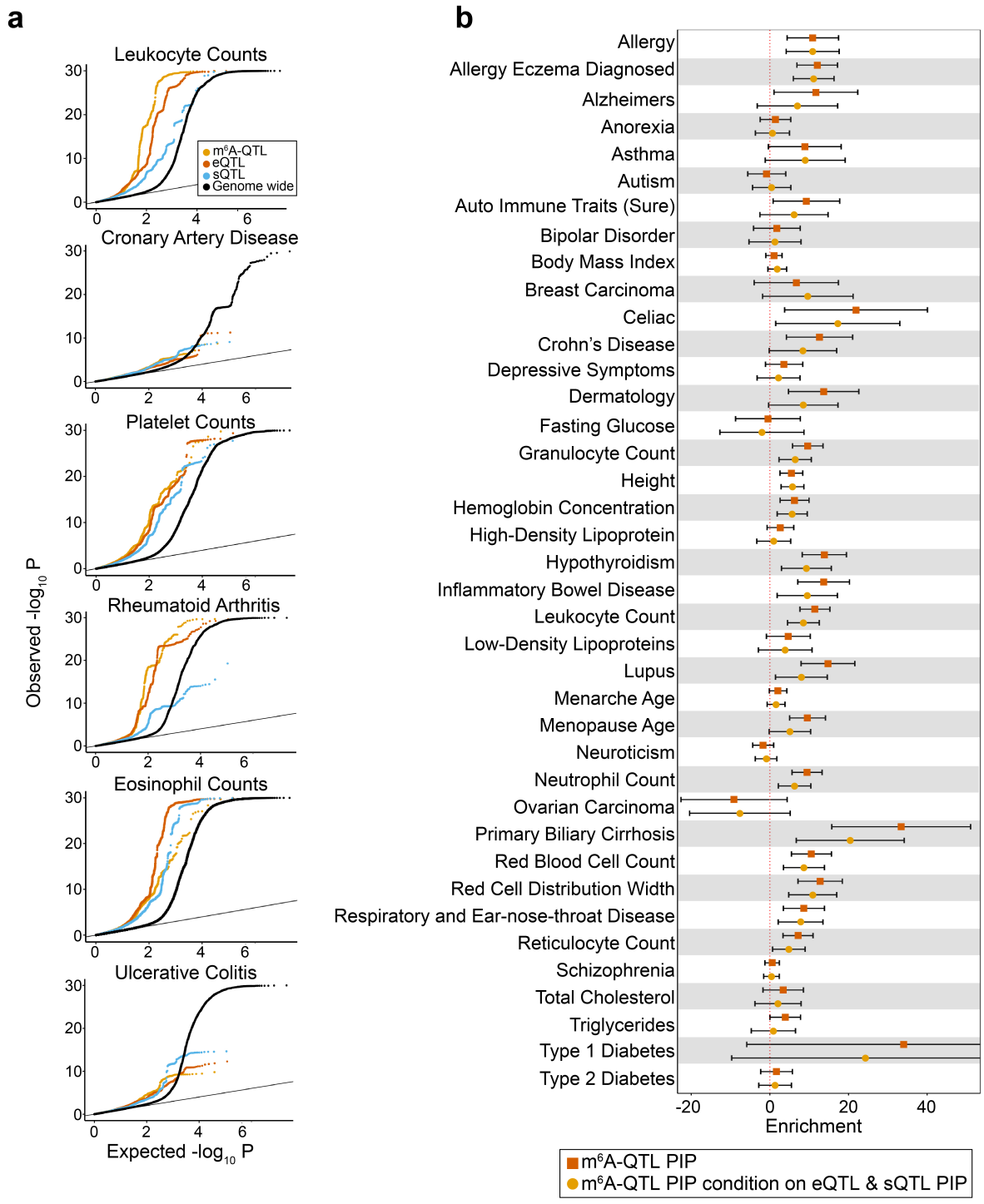
Supplementary Figure 2.3: Torus. Error bars represent the 95% confidence intervals. **b**, Enrichment of m^6A QTLs in the binding sites of RNA polymerase2 subunit A (POLR2A), and phosphorylated POLR2A at two residues (S2 and S5) by Torus joint analysis of all annotations (upper panel), and enrichment of m^6A QTLs in histone modifications from Torus joint analysis. Error bars indicate the 95% confidence intervals. **c**, Proportion of putative causal m^6A QTNs in RNA features and transcription factor binding site annotations (see Methods). **d-e**, To confirm that transcription rate affects mRNA and protein level, we ascertained transcription rate QTLs (Txn-QTLs) and assessed the correlation between transcription rate (Txn)-QTL effect sizes (30 min and 60 min 4sU labelling, respectively) and eQTL effect size (panel **d**, n = 425 and 1,387 SNP-gene pairs), and protein-QTL effect sizes (panel **e**, n = 425 and 408 SNP-gene pairs). Correlation is computed using linear regression. Fitted lines and 95% confidence intervals are shown in blue lines and shades.



Supplementary Figure 2.4: Downstream effects of m^{6}A are context dependent. **a**, The number and fraction of m^{6}A QTLs in chromatin related genomic regions (using the union of promoter and enhancer regions annotated by ChromHMM in GM12878 cell line), and in chromatin related eQTLs (eQTLs with nominal P value < 0.05 and also in promoter and enhancer regions). **b**, High correlations of effect sizes between molecular QTLs along the cascade from transcription to translation. Correlation is computed using linear regression, in which fitted lines and 95% confidence intervals are shown in blue lines and shades. **c**, Log_2 fold change of translation efficiency of m^{6}A methylated transcripts in METTL3 knockdown vs. controls. **d**, Log_2 fold changes of translation efficiency upon YTHDF1 (m^{6}A reader protein) knockdown. Transcripts harboring YTHDF1-bound m^{6}A peaks are labeled in yellow and non-targets in blue.

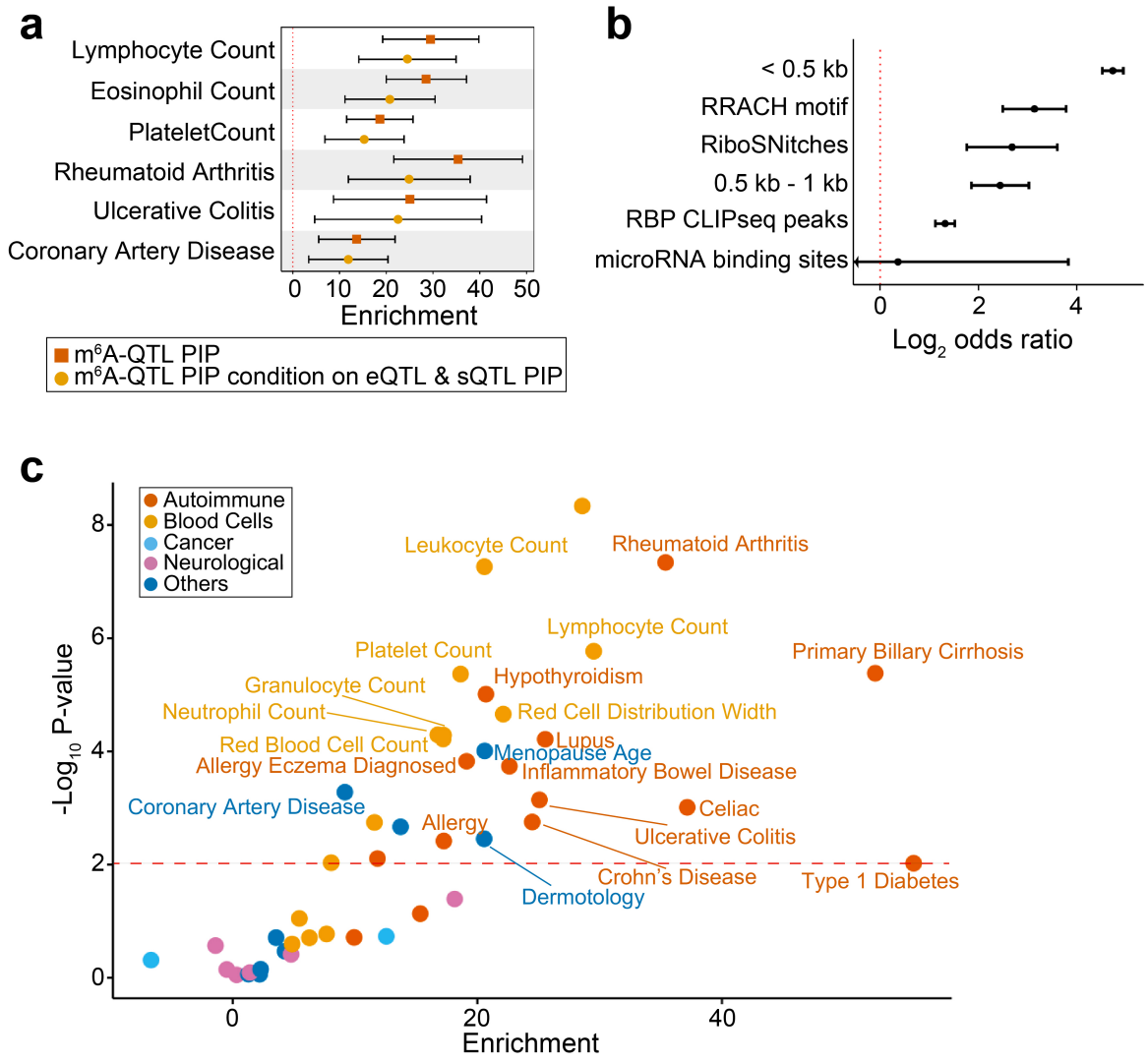


Supplementary Figure 2.5: YBX3 mediates translation efficiency of m^{6}A modified transcripts. **a**, Sucrose gradient A260 absorbance profile from YBX3 knockdown and control Hela cells. The arrows indicate the fraction sampled for subsequent qPCR analysis of YBX3 target transcripts. This experiment is repeated 2 times. **b**, Translation efficiency of YBX3 targets in comparison with YTHDF1 targets. We accounted for mRNA level variation by dividing polysome-bound fraction by the non-polysome-bound fraction. Transcript levels are quantified using RT-qPCR. Three polysome-bound fractions, as indicated in panel **a**, are sampled from sucrose gradient fractionation. 2 technical replicates were measured to obtain the data. The lower and upper hinges correspond to the first and third quartiles. Horizontal line indicates median value, and whiskers correspond to the value no further than 1.5x inter-quartile range.

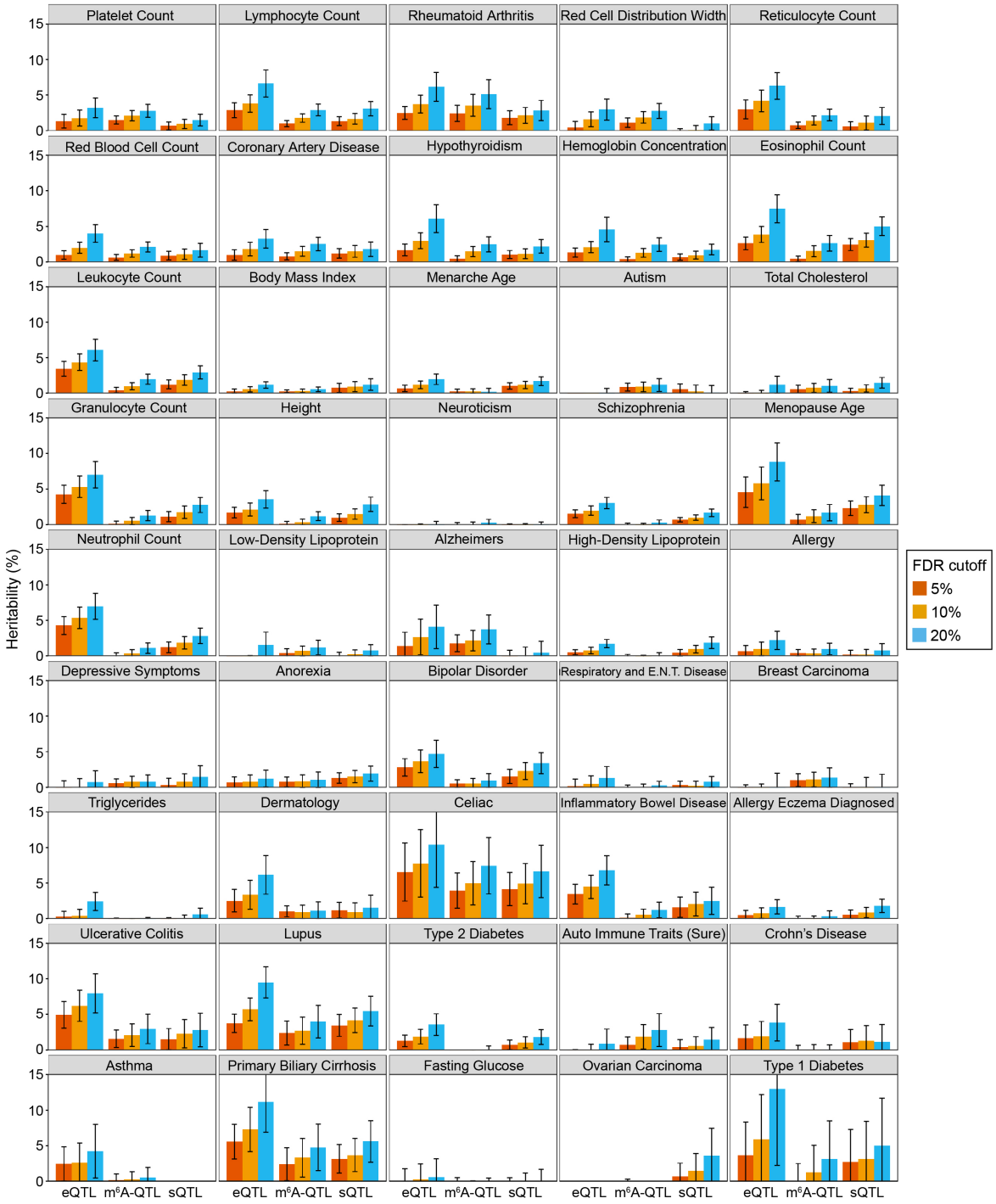


Supplementary Figure 2.6: Enrichment of GWAS signal in m^{6}A QTLs. **a**, Quantile-quantile (QQ) plots of P values from GWAS of selected traits. m^{6}A QTLs, eQTLs and sQTLs are shown in comparison with genome wide SNPs. GWAS SNPs are binary annotated using m^{6}A QTLs, eQTLs and sQTLs with P value $< 1 \times 10^{-4}$. **b**, Enrichment of GWAS trait heritability assessed by stratified LD-score regression (S-LDSC). Shown are the

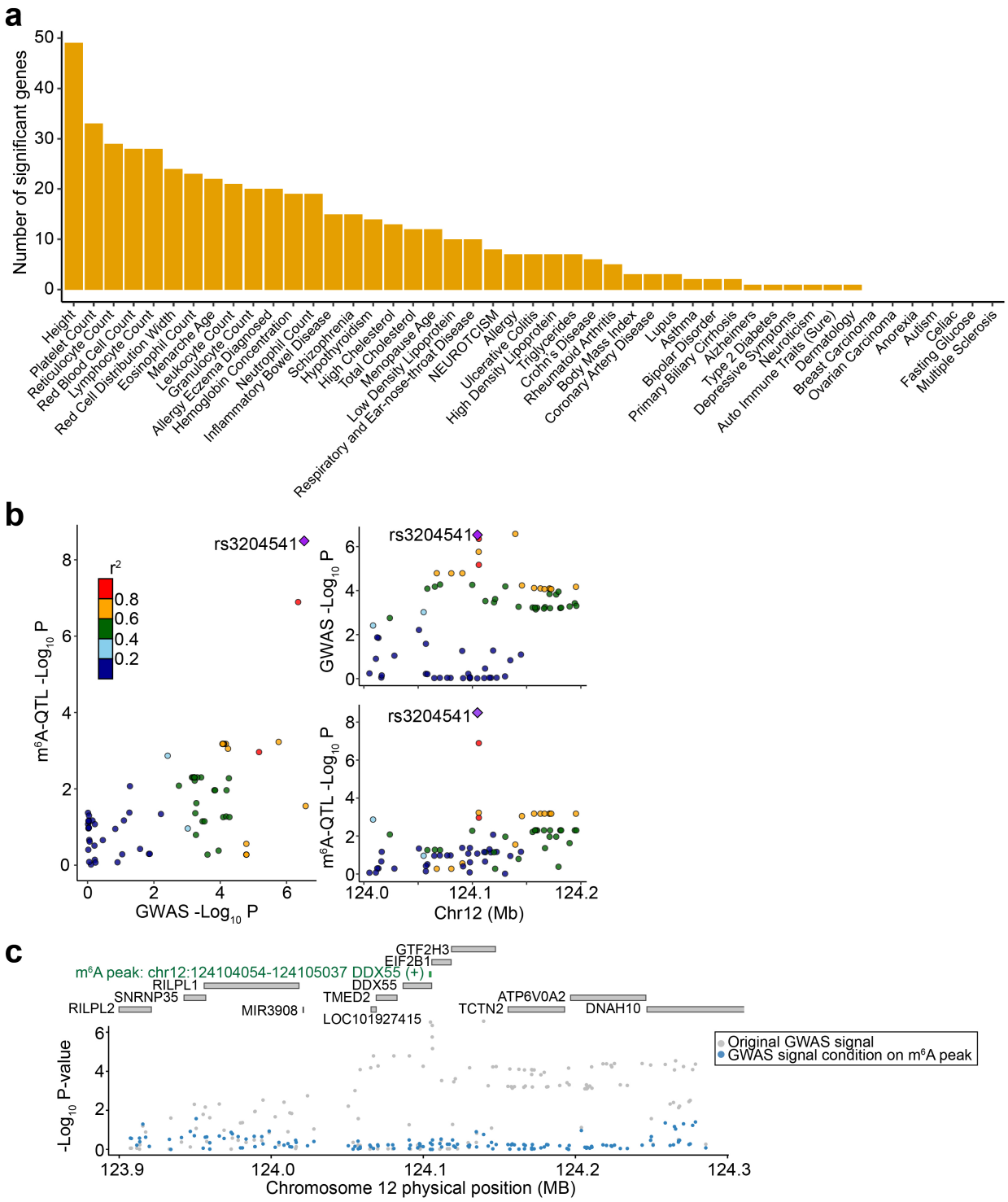
Supplementary Figure 2.6: results of GWAS traits not reported in **Figure 2.13b**. Posterior inclusion probability (PIPs) in this analysis are derived from SuSiE with default (uniform) priors. Error bars represent the 95% confidence intervals.



Supplementary Figure 2.7: Enrichment of complex trait heritability in m^{6}A QTNs using RNA-features-informed priors. **a**, Enrichment of selected immune and blood GWAS trait heritability assessed by stratified LD-score regression (S-LDSC). PIPs of m^{6}A QTLs are derived from SuSiE using RNA-features-informed priors. PIPs of eQTL and sQTL are based on uniform prior. Error bars represent 95% confidence intervals. **b**, Enrichment parameters of m^{6}A-related annotations that are used to derive priors (by Torus) for SuSiE fine-mapping. Error bars represent the 95% confidence intervals. **c**, Summary of GWAS traits heritability enrichment analysis using m^{6}A QTL PIP (using RNA-feature informed priors) as annotation. The -log$_{10}$ P value is plotted against the enrichment of heritability. The dots are colored by disease category. The red dashed line shows FDR 5% threshold.



Supplementary Figure 2.8: Partitioning complex trait heritability by m^{6}A QTLs, eQTLs and sQTLs. Heritability is assessed by S-LDSC in which QTLs are binary annotated with varying SNP-level FDR thresholds of 5%, 10%, and 20%. Error bars represent standard errors.



Supplementary Figure 2.9: m^{6}A-TWAS identifies putative risk genes in human complex traits. a, Number of significant m^{6}A-TWAS genes in all 45 GWAS traits. Significance is defined by the Bonferroni corrected P value 0.05. **b**, LocusCompare plot showing the scatter plot and aligned Manhattan plots of leukocyte count GWAS and m^{6}A QTL association signal at the *DDX55* locus. **c**, Manhattan plot of GWAS association signals

Supplementary Figure 2.9: before and after conditioning on the TWAS-predicted m^6A level (gray and blue dots, respectively) for the leukocyte count at the *DDX55* locus.

2.5 Supplementary tables

Supplementary Table 2.1: Summary of Torus result for RBP binding sites

Feature	Log Odds Ratio	CI05	CI95	*P* value
DDX3X	3.666	2.766	4.566	1.42E-15
NCBP2	3.598	2.366	4.830	1.04E-08
SRSF7	3.553	2.145	4.962	7.72E-07
SUB1	3.484	2.577	4.390	4.80E-14
DDX6	3.349	1.204	5.494	2.21E-03
YBX3	3.342	2.936	3.748	1.48E-58
LSM11	3.230	1.935	4.525	1.02E-06
SRSF1	3.227	2.299	4.155	9.38E-12
FXR2	3.223	2.486	3.960	1.02E-17
DDX55	3.218	1.934	4.503	9.18E-07
TRA2A	3.173	1.160	5.186	2.01E-03
IGF2BP2	3.132	2.588	3.677	1.98E-29
EIF3H	3.129	2.317	3.941	4.26E-14
UPF1	3.129	2.553	3.704	1.47E-26
RPS3	3.094	1.245	4.943	1.04E-03
LIN28B	3.069	1.757	4.381	4.54E-06
FMR1	2.994	2.114	3.874	2.59E-11
GRWD1	2.967	2.520	3.413	7.36E-39
GRSF1	2.965	2.077	3.853	5.97E-11
DGCR8	2.892	0.234	5.549	3.29E-02

Feature	**Log Odds Ratio**	**CI05**	**CI95**	***P* value**
SRSF9	2.819	2.131	3.508	1.06E-15
PUS1	2.803	0.014	5.592	4.89E-02
IGF2BP3	2.801	0.021	5.582	4.84E-02
DHX30	2.794	-0.275	5.864	7.45E-02
SND1	2.793	1.160	4.427	8.07E-04
UCHL5	2.753	1.779	3.728	3.13E-08
ZNF622	2.744	2.124	3.364	4.15E-18
NPM1	2.736	-0.456	5.929	9.31E-02
DDX42	2.717	1.455	3.980	2.48E-05
RBM27	2.704	1.467	3.941	1.83E-05
FTO	2.701	1.818	3.584	2.03E-09
TIA1	2.655	1.614	3.695	5.63E-07
IGF2BP1	2.591	1.110	4.072	6.06E-04
EIF3D	2.569	1.960	3.179	1.53E-16
MTPAP	2.524	1.763	3.285	7.99E-11
PPIG	2.511	1.862	3.160	3.37E-14
GEMIN5	2.479	1.428	3.530	3.78E-06
NOL12	2.471	0.956	3.986	1.39E-03
LARP4	2.374	0.002	4.746	4.98E-02
METAP2	2.373	1.235	3.510	4.30E-05
FXR1	2.350	1.222	3.478	4.44E-05
FUBP3	2.347	0.713	3.982	4.90E-03
U2AF2	2.326	1.115	3.537	1.67E-04
PUM2	2.308	0.627	3.989	7.12E-03
RBM15	2.251	0.081	4.420	4.19E-02

Feature	**Log Odds Ratio**	**CI05**	**CI95**	***P* value**
HLTF	2.236	1.434	3.039	4.82E-08
PRPF8	2.189	1.119	3.258	5.98E-05
BUD13	2.165	0.802	3.528	1.85E-03
PPIL4	2.139	0.932	3.346	5.14E-04
POLR2G	2.121	1.153	3.089	1.75E-05
CDC40	2.051	0.957	3.145	2.38E-04
XRN2	1.956	-0.319	4.231	9.20E-02
PCBP2	1.937	1.139	2.735	1.96E-06
BCCIP	1.933	0.937	2.928	1.40E-04
SAFB2	1.923	1.179	2.667	4.06E-07
CSTF2	1.919	1.047	2.791	1.61E-05
SF3A3	1.907	0.570	3.244	5.18E-03
SUPV3L1	1.892	-0.127	3.912	6.64E-02
AGGF1	1.872	0.131	3.614	3.52E-02
AKAP8L	1.850	0.748	2.952	1.00E-03
DROSHA	1.846	0.826	2.866	3.89E-04
HNRNPA1	1.816	0.145	3.486	3.31E-02
EXOSC5	1.747	0.586	2.908	3.19E-03
DDX24	1.703	0.200	3.207	2.65E-02
CPSF6	1.677	0.106	3.249	3.65E-02
EIF4G2	1.629	0.596	2.663	2.02E-03
ZRANB2	1.623	-0.452	3.698	1.25E-01
GTF2F1	1.524	-0.500	3.549	1.40E-01
GNL3	1.510	-2.358	5.379	4.44E-01
HNRNPK	1.415	-0.257	3.088	9.74E-02

Feature	**Log Odds Ratio**	**CI05**	**CI95**	***P* value**
TNRC6A	1.409	-0.326	3.145	1.12E-01
HNRNPU	1.381	-0.343	3.106	1.17E-01
U2AF1	1.332	-1.189	3.852	3.00E-01
XPO5	1.292	-0.067	2.651	6.24E-02
EFTUD2	1.286	-0.939	3.511	2.57E-01
DDX59	1.285	-0.061	2.631	6.13E-02
KHDRBS1	1.273	0.321	2.225	8.77E-03
FKBP4	1.251	-0.512	3.014	1.64E-01
NKRF	1.244	0.445	2.043	2.28E-03
SUGP2	1.036	0.132	1.941	2.49E-02
DKC1	1.034	-1.212	3.280	3.67E-01
CSTF2T	0.985	-0.497	2.467	1.93E-01
FUS	0.907	-10.244	12.058	8.73E-01
RBFOX2	0.786	-1.267	2.840	4.53E-01
KHSRP	0.779	0.093	1.465	2.60E-02
SF3B1	0.720	-5.988	7.427	8.33E-01
FAM120A	0.669	-2.820	4.158	7.07E-01
HNRNPC	0.649	0.008	1.291	4.75E-02
SF3B4	0.487	-1.866	2.841	6.85E-01
HNRNPUL1	0.464	-4.736	5.664	8.61E-01
TARDBP	0.369	-0.937	1.675	5.80E-01
NOLC1	0.283	-13.598	14.165	9.68E-01
NONO	0.156	-2.574	2.886	9.11E-01
SBDS	0.150	-8.527	8.827	9.73E-01
EIF4G1	0.112	-13.794	14.018	9.87E-01

Feature	**Log Odds Ratio**	**CI05**	**CI95**	***P* value**
RBM5	0.057	-1.726	1.841	9.50E-01
QKI	0.002	-3.716	3.719	9.99E-01
ILF3	-0.012	-0.717	0.694	9.73E-01
EWSR1	-0.036	-2.068	1.995	9.72E-01
YWHAG	-0.064	-3.274	3.145	9.69E-01
GPKOW	-0.086	-2.333	2.161	9.40E-01
TAF15	-0.194	-4.226	3.839	9.25E-01
SLBP	-0.201	-7.400	6.997	9.56E-01
HNRNPM	-0.587	-3.773	2.600	7.18E-01
SFPQ	-0.593	-3.693	2.508	7.08E-01
PTBP1	-0.732	-4.568	3.105	7.08E-01
RPS11	-0.748	-7.365	5.868	8.25E-01
SLTM	-0.805	-9.048	7.438	8.48E-01
XRCC6	-0.892	-19.791	18.007	9.26E-01
FUBP1	-1.349	-23.807	21.108	9.06E-01
TROVE2	-1.488	-12.375	9.400	7.89E-01
RBM22	-1.621	-10.901	7.658	7.32E-01
LARP7	-1.694	-22.118	18.731	8.71E-01
FASTKD2	-1.813	-12.543	8.917	7.41E-01
SERBP1	-1.921	-28.844	25.001	8.89E-01
NSUN2	-2.246	-16.475	11.982	7.57E-01
TBRG4	-2.251	-27.745	23.242	8.63E-01
AUH	-2.320	-18.914	14.274	7.84E-01
SMNDC1	-2.505	-19.787	14.777	7.76E-01
RPS5	-2.697	-12.453	7.059	5.88E-01

Feature	Log Odds Ratio	CI05	CI95	*P* value
TIAL1	-7.706	-95.713	80.301	8.64E-01

Supplementary Table 2.2: Summary of Torus result for microRNA binding sites

Feature	Log Odds Ratio	CI05	CI95	*P* value
hsa-miR-361-5p	6.581	-6.685	19.848	3.31E-01
hsa-miR-582-5p	6.128	1.102	11.154	1.69E-02
hsa-miR-423-5p	5.796	-2.137	13.729	1.52E-01
hsa-miR-331-3p	5.388	0.760	10.017	2.25E-02
hsa-miR-25-3p	4.545	-2.729	11.819	2.21E-01
hsa-miR-182-5p	4.004	-5.464	13.472	4.07E-01
hsa-miR-212-5p	2.682	-3.155	8.519	3.68E-01
hsa-miR-17-5p	2.310	-15.122	19.741	7.95E-01
hsa-miR-491-5p	1.699	-66.480	69.879	9.61E-01
hsa-miR-362-5p	1.666	-71.115	74.448	9.64E-01
hsa-miR-155-5p	0.002	-6.259	6.263	1.00E+00
hsa-miR-140-3p	-0.399	-20.306	19.508	9.69E-01
hsa-miR-141-3p	-0.629	-14.941	13.682	9.31E-01
hsa-miR-142-3p	-0.890	-50.892	49.111	9.72E-01
hsa-miR-328-3p	-1.048	-80.832	78.736	9.79E-01
hsa-miR-221-3p	-1.145	-48.109	45.819	9.62E-01
hsa-miR-22-3p	-1.184	-63.199	60.832	9.70E-01
hsa-miR-501-3p	-1.754	-237.077	233.569	9.88E-01
hsa-miR-21-5p	-1.922	-120.778	116.934	9.75E-01
hsa-miR-324-5p	-2.817	-49.904	44.271	9.07E-01
hsa-miR-339-5p	-2.884	-34.807	29.038	8.59E-01

Feature	Log Odds Ratio	CI05	CI95	*P* value
hsa-miR-340-5p	-3.265	-108.112	101.581	9.51E-01
hsa-miR-874-3p	-3.681	-55.803	48.442	8.90E-01
hsa-miR-142-5p	-4.230	-130.825	122.365	9.48E-01
hsa-miR-335-5p	-4.323	-67.139	58.493	8.93E-01
hsa-miR-150-5p	-4.598	-60.322	51.126	8.72E-01
hsa-miR-542-3p	-4.627	-155.664	146.410	9.52E-01
hsa-miR-28-5p	-4.739	-113.007	103.530	9.32E-01
hsa-miR-183-5p	-4.779	-88.582	79.024	9.11E-01
hsa-miR-532-5p	-4.828	-235.076	225.419	9.67E-01
hsa-miR-425-5p	-4.857	-139.573	129.858	9.44E-01
hsa-miR-532-3p	-5.169	-115.291	104.953	9.27E-01
hsa-miR-140-5p	-5.792	-322.868	311.284	9.71E-01
hsa-miR-186-5p	-5.824	-151.749	140.101	9.38E-01
hsa-miR-192-5p	-5.848	-131.021	119.324	9.27E-01
hsa-miR-296-5p	-6.181	-216.046	203.684	9.54E-01
hsa-miR-223-3p	-6.465	-155.177	142.246	9.32E-01
hsa-miR-330-3p	-6.614	-90.118	76.890	8.77E-01
hsa-miR-505-3p	-6.843	-166.110	152.423	9.33E-01
hsa-miR-423-3p	-7.215	-332.915	318.484	9.65E-01
hsa-miR-210-3p	-7.537	-202.749	187.675	9.40E-01
hsa-miR-486-5p	-7.610	-204.760	189.541	9.40E-01
hsa-miR-132-3p	-7.833	-94.687	79.021	8.60E-01
hsa-miR-877-5p	-8.036	-331.312	315.240	9.61E-01
hsa-miR-185-5p	-8.390	-253.073	236.294	9.46E-01
hsa-miR-342-3p	-8.461	-212.971	196.050	9.35E-01

Feature	**Log Odds Ratio**	**CI05**	**CI95**	***P* value**
hsa-miR-143-3p	-10.158	-181.952	161.635	9.08E-01

CHAPTER 3

R PACKAGE RADAR FOR DIFFERENTIAL ANALYSIS OF MERIP-SEQ DATA

3.1 Introduction

Methylated RNA immunoprecipitation sequencing (MeRIP-seq) [33, 34] is a key technique used in m^6A to survey the epitranscriptome. Early studies performing qualitative analysis compared peaks called in one experimental group versus peaks in another group and identified peaks unique to each experimental group as differentially methylated peaks. However, many differential peaks identified by this method are caused by boundary cases at the peak-detection threshold rather than true present/absent of peaks as noted in an recent study [38]. To enable cross-group comparisons, a few methods have been developed and applied to analyze differential methylation in MeRIP-seq data [41, 42, 43, 19]. ExomePeak uses Fisher's exact test for differential methylation identifications and its later version uses a likelihood ratio test based on the binomial distribution (termed "bltest") [41]. MeTPeak uses a beta-binomial model to infer differential peaks [42]. DRME and its improved version — QNB uses a model based on the negative binomial distribution [43, 44].

While existing methods have yielded promising results, they also have important drawbacks: (1) Current methods [41, 42, 43, 19]. designed for small sample size scenario ignore the existence of confounding factors and cannot accommodate complex study designs with covariates (such as age, gender, etc.) that are frequently encountered in patient or animal studies with larger sample sizes. (2) Most of the differential gene expression (DE) analysis tools such as edgeR, DESeq2 and Sleuth [113, 134, 135] are compatible with complex study designs. But they rely on models developed for RNA-seq experiment and cannot accommodate unique features of MeRIP-seq data. A standard MeRIP-seq experiment yields an Input and an Immunoprecipitation (IP) library for each sample. The Input library is the initial RNA fragments pool prior to the antibody pull-down — a measurement of RNA expression

level. The IP library represents the RNA fragments carrying modified bases captured by antibody pull-down — a measurement of methylated RNA abundance. RNA Differential Methylation (DM) is defined as the alteration of methylated RNA abundance conditioning on the RNA expression background. Thus, DM analysis requires assessment of RNA methylation change based on pre-IP and post-IP measurements in pairs. In contrast, DE analysis tools only compare a single read count measurement across samples. (3) Current MeRIP-seq specific tools [42, 44, 41] use peak (about 250 bp) read counts in the Input library as measurement of RNA expression for a gene (about 11 kb). However, the variability in a small genomic range across samples due to the sparsity of reads sampled can result in unwanted variation to the expression level estimation if using local read counts. QNB combines local read counts of both Input and IP as an estimator of the expression level to mitigate this problem. However, incorporating IP read counts can cofound pre-IP expression level with post-IP RNA abundance, leading to biased estimation of expression level. Inaccurate expression level measurement can lead to substantial false discoveries in the subsequent DM analysis. Thus, the utilization of Input library to account for pre-IP RNA expression level needs to be further optimized.

To combat these challenges and allow for accurate identification of differentially methylated loci, we present a novel approach to perform RNA methylAtion Differential Analysis in R (RADAR) for MeRIP-seq data. RADAR accounts for variation in pre-IP RNA and in post-IP read counts using different strategies. Specifically, RADAR uses gene-level read counts instead of peak-level read counts in the Input library as a robust measurement of the initial pre-IP RNA expression level. In addition, RADAR uses a flexible Poisson random effect model to accommodate over-dispersion in the post-IP read counts due to variability of biological replicates and noise introduced in the immunoprecipitation process. This generalized linear model framework enables incorporation of covariates in complex study designs.

We benchmarked the performance of RADAR with alternative methods on simulated data

by different data generating models. We showed RADAR achieved higher sensitivity and specificity compare to existing alternative methods. We also demonstrated the performance of RADAR on real MeRIP-seq data by applying it to four high quality m^6A-meRIP-seq (aka m^6A-seq) datasets generated by us and others, including an ovarian cancer dataset (GSE 119168) consisting of 7 normal fallopian tubes tissue from healthy individuals and 6 metastatic omental tumors, a Type 2 Diabetes (T2D, GSE 120024) dataset consisting of human islets from 8 type 2 diabetes patients and 7 healthy control patients with samples being processed in three batches due to different sample acquisition times, a mice liver (GSE 119490) dataset consisting of mouse liver from 4 wild type mice and 4 *Mettl14* knockout mice and a mice brain (GSE 113781) dataset consisting of 7 mouse cortex samples of stress exposed mice and 7 from control mice. We showed that our approach can accommodate distinct study designs and led to more sensitive and reproducible DM loci identification than possible alternatives.

3.2 Material and methods

3.2.1 Biological samples

Ovarian cancer samples. All human tissue samples were collected with informed consent under approved University of Chicago Institutional Review Board protocols and in accordance with the Declaration of Helsinki. All experiments were conducted in accordance with approved protocol guidelines and regulations. Six omental tumor tissues were collected from newly diagnosed patients with advanced, metastatic high-grade serous ovarian cancer during primary debulking surgery at the University of Chicago. Seven normal fallopian tube tissues were collected from patients with benign gynecological conditions at the time of surgery.

T2D samples. A minimum of 20,000 human islets equivalents (IEQs)/patient were obtained from the Integrated Islet Distribution Program (IIIDP) and Prodo Laboratories. Upon receipt, islets were cultured overnight in Miami Media #1A (Cellgro, USA) and then

handpicked, washed twice by self-sedimentation with ice-cold DPBS and pelleted for RNA isolation. All studies and protocols used were approved by the Joslin Diabetes Center's Committee on Human Studies (CHS#5-05). Samples from eight T2D patients and seven non-diabetic controls were collected for analyses in this study.

Mouse liver samples. Mouse liver tissues were collected from wild type Albumin-Cre;Mettl14+/+ and Albumin-Cre;Mettl14flox/flox;Cre liver specific conditional knockout mice [32]. Four wide-type and four Mettl14 cKO mice on 42% high fat diet for three months were sequenced and analyzed.

3.2.2 RNA extraction and m^6A-MeRIP-seq

Total RNA was extracted from tissues using TRIzol (Invitrogen) according to the manufacturer's instruction. For T2D and mouse liver samples, mRNA was further purified with Dynabeads mRNA DIRECT purification kit (Thermo Fisher, cat. 61011). mRNA was adjusted to 15 ng/μl in 100 μl and fragmented using Bioruptor ultrasonicator (Diagenode) with 30s on/off for 30 cycles. m^6A-immunoprecipitation (m^6A-IP) was performed using EpiMark N^6-Methyladenosine enrichment kit (NEB cat. E1610S). RNA eluted from m^6A-IP was cleaned using RNA Clean and Concentrator (Zymo Research, cat. R1013). Input and IP samples were then used to prepare library with KAPA mRNA Hyper Kit (Roche, Cat. KK8541). For fallopian tube and omental tumor tissues, total RNA was fragmented and directly subjected to m^6A-IP. Takara Pico-Input Strand-Specific Total RNA-seq for Illumina (Takara, Cat. 634413) was used to construct libraries from total RNA where ribosome-derived cDNA was removed before final library amplification. T2D and mouse liver libraries were sequenced by the HiSeq4000 platform at SE50 mode. The ovarian cancer libraries were sequenced by the NextSeq 500 platform at PE37 mode.

3.2.3 Data preparation

For each dataset, the raw sequencing data were mapped to the corresponding reference genome (hg38 for human, and mm10 for mouse) by Hisat2 [92] with parameter -x 1. The BAM files obtained from alignment are used as an input file for RADAR.

3.2.4 Read count pre-processing

RADAR takes a GTF file as an input for gene annotation and obtains a gene model using the R package GenomicFeatures [136]. Exons of a gene are concatenated to form the "longest isoform" transcript, which is then divided into bins of user defined size. The R package Rsamtools is used to extract and quantify aligned reads from BAM files in each bin. The gene-level counts of Input library are obtained by summing up bin-level read counts of each gene.

Normalization. Unlike previous methods [41, 42, 43, 19] that scale read counts to library sizes as a way of normalization, which can be strongly skewed by highly expressed genes [134]. RADAR considers Input and IP samples separately. The Input sample is essentially an RNA-seq library; therefore, we directly apply the median-of-ratios method implemented in DESeq2, which is robust to outliers, to estimate a sample-wise size factor for each sample from Input gene-level counts. In regard to the IP sample, the abundance of read counts $t_{i,j}$ depends on the abundance of RNA in the pre-IP RNA pool, the overall IP efficiency of that sample, the total sequencing depth of that IP library, and the methylation level of that locus. To normalize the variation due to sample-wise IP efficiency difference and sequencing depth variations of IP libraries, we estimate a sample-wise size factor from the fold enrichment $E_{i,j} = t_{i,j}/s_{gene_{m,j}}$ of the top 1% bins ranked by IP read counts where $t_{i,j}$ is IP read counts and $s_{gene_{m,j}}$ is the normalized geneSum (gene-level) read count at corresponding gene. The reason that only top bins are used to estimate the overall IP efficiency is to exclude the regions where IP read counts are mainly attributed to non-specific binding.

3.2.5 Adjust IP read count for pre-IP RNA expression level

To account for the pre-IP gene expression level variation in IP read counts, we compute a gene-wise size factor by centering normalized gene-level counts to 1. For each bin, we divide the normalized IP read counts by the gene-wise size factor of the corresponding gene. The resulting IP read counts now reflect the methylation level as other factors have all been accounted for. The adjusted IP read counts representing methylation levels are further used for DM tests.

3.2.6 Data filtering

We apply two filters to remove unwanted bins in the data: 1) We remove bins in which reads are depleted in IP libraries because read counts in these bins are likely attributed to non-specific binding during the immunoprecipitation; 2) We remove bins in which raw IP read counts are smaller than 15 (this cutoff can be defined by the user) because signals in regions without sufficient coverage will be too noisy and unreliable. Low IP read count also implies that the bin is likely a non-methylated region.

3.2.7 Model for DM test

For each bin, we model the processed IP read counts Y_i in the i-th sample as follows:

$$\begin{aligned} Y_i &\sim Poi(\lambda_i) \\ \log(\lambda_i) &= \mu + \mathbf{X}\boldsymbol{\beta} + e_i = \mu + X_0\beta_0 + \sum_{j=1}^{k} X_i\beta_i + e_i \end{aligned} \quad (3.1)$$

where λ_i is the mean of a Poisson distribution, μ is a bin specific intercept, $\mathbf{X}$ is the design matrix including the indicator of the groups of interest X_0 and covariates X_j $(j = 1, \ldots, k)$, $\boldsymbol{\beta}$ represent associated coefficients and e_i is a random effect following a log gamma distribution with a scale parameter ψ and mean equal to 1, i.e., $e_i \in \text{logGamma}(\psi, \psi)$,. Introducing

a new variable $w_i \in \text{Gamma}(\psi, \psi)$, we have $\lambda_i = e^{\mu_1 + X_i\beta} w_i$. The differential analysis is equivalent to test against the null hypothesis $\beta_0 = 0$.

After integrating out w_i, the likelihood of observing the data given all other parameters Θ is:

$$\begin{aligned} P(\mathbf{Y}|\Theta, -w_i) &= \int e^{-w_i e^{\mu+\mathbf{X}\boldsymbol{\beta}}} \frac{(w_i e^{\mu+\mathbf{X}\boldsymbol{\beta}})^{Y_i}}{Y_i!} \frac{\psi^\psi w_i^{\psi-1} e^{-\psi w_i}}{\Gamma(\psi)} dw_i \\ &= \frac{\psi^\psi}{Y_i!} \frac{\Gamma(Y_i+\psi)}{\Gamma(\psi)} \frac{(e^{\mu+\mathbf{X}\boldsymbol{\beta}})^{Y_i}}{(e^{\mu+\mathbf{X}\boldsymbol{\beta}}+\psi)^{Y_i+\psi}} \qquad (3.2) \end{aligned}$$

The marginal log likelihood of observing $\mathbf{Y}$ can be written as:

$$\begin{aligned} \log L(\mathbf{Y}) &= \sum_i^n [Y_i(\mu + \mathbf{X}\boldsymbol{\beta}) + \psi \log \psi + \log \Gamma(Y_i+\psi) - \log Y_i! \\ &\quad - \log \Gamma(\psi) - (Y_i+\psi)\log(e^{\mu+\mathbf{X}\boldsymbol{\beta}}+\psi)] \qquad (3.3) \end{aligned}$$

We use the gradient ascent algorithm to calculate maximum likelihood estimators of all the parameters, which involves the calculation of first derivatives.

$$\begin{aligned} \frac{\partial \log L(\mathbf{Y})}{\partial \beta} &= \sum_i^n \Big[Y_i - \big(Y_i+\psi\big)\frac{e^{\mu+\mathbf{X}\boldsymbol{\beta}}}{e^{\mu+\mathbf{X}\boldsymbol{\beta}}+\psi}\Big] X_i \qquad (3.4) \\ \frac{\partial \log L(\mathbf{Y})}{\partial \psi} &= \sum_i^n \Big[\log\psi + 1 - \frac{Y_i+\psi}{e^{\mu+\mathbf{X}\boldsymbol{\beta}}+\psi} - \log(e^{\mu+\mathbf{X}\boldsymbol{\beta}}+\psi) \\ &\quad + digamma(Y_i+\psi) - digamma(\psi)\Big] \qquad (3.5) \\ \frac{\partial \log L(\mathbf{Y})}{\partial \mu_1} &= \sum_i^n \Big[Y_i - \big(Y_i+\psi\big)\frac{e^{\mu+\mathbf{X}\boldsymbol{\beta}}}{e^{\mu+\mathbf{X}\boldsymbol{\beta}}+\psi}\Big] \qquad (3.6) \end{aligned}$$

In each iteration, the parameters are updated through $\Phi_{(t+1)} = \Phi_{(t)} + s_{(t+1)} \frac{\partial \log(\mathbf{Y})}{\partial \Phi}|_{\Phi=\Phi_{(t)}}$. The step size $s_{(t+1)}$ is determined by a line search algorithm. Finally, a Wald test is derived to test against $\beta_0 = 0$, i.e., the test statistics is $\widehat{\beta_0}/sd(\widehat{\beta_0})$ where $\widehat{\beta_0}$ is the MLE and $sd(\widehat{\beta_0})$ is estimated using observed Fisher information.

3.2.8 Post-processing

Since the DM tests are performed on consecutive bins on the mRNA, post-processing is needed to merge connected bins that contain reads derived from the same methylation site and report their genome coordinates instead of mRNA coordinates. Specifically, we filter all the bins under user-defined FDR cutoffs and merge adjacent significant bins to a single peak. To represent the mRNA peaks using the genome coordinate, we report the final result in BED12 format, which can specify exon blocks for an intron-spanning interval.

3.2.9 Simulation analysis

To assess the sensitivity and specificity of each methods on detecting true DM sites, we simulated dataset of 8 replicates with and without covariate. Since RADAR make inferential test on fold enrichment (pre-processed IP read counts adjusted for input RNA level variation), we first simulated this enrichment data (pre-processed IP read counts) using Model (1). We draw the distribution of gene-specific intercept parameter μ (equivalent to baseline sequencing depth of control samples), random effect parameter ψ from real data (T2D) to better reflect the property of real data. For each dataset, we simulated 26,324 sites where 20% of them were predefined as true DM sites with effect sizes of 0.5, 0.75 or 1. At pre-defined true DM sites, we simulated read counts with $\beta = 0.5$ (or 0.75 or 1) while $\beta = 0$ at null sites. Other alternative methods take Input and IP read counts as input data for DM tests. To convert our simulated enrichment data into paired Input and IP read counts, we used Input read counts from the real T2D dataset and generated corresponding IP read counts by rescaling simulated IP counts to match the IP/Input ratio in the real data.

We then applied RADAR, MeTPeak (version 1.1) [42], QNB (version 1.1.11) [44], Fisher's exact test, exomePeak (version 2.17.0) [41] to the simulated data. We used the Benjamini-Hochberg method to adjust for multiple comparisons. Using an FDR cutoff at 10% (or 1%, 5% ... in sliding threshold analysis), we obtained a set of predicted DM sites for each method. We then checked whether pre-defined true DM sites were predicted to be DM site.

To evaluate the performance, we computed sensitivity by dividing the number of overlaps between predicted DM sites and true sites by the number of true sites. We also computed the empirical FDR by dividing the number of predicted sites that is not in the true sites by the number of predicted sites. To evaluate effect of covariates on the performance, we repeated the above analysis using Model (1) with an additional binary categorical (such as gender) variable with an effect size of 2. There are 3 covariates included in the T2D data analysis. The covariate with the largest effect size ranges from 0 to 4. We chose an intermediate effect size of 2 to represent a moderately challenging scenario in real data.

Next, we repeated the simulation analysis with an alternative model as described in the QNB manuscript [44], termed as the "QNB model". Unlike Model (1) that directly simulates the enrichment as pre-processed IP-read counts, the QNB model simulates the paired Input and IP read counts separately, each following a negative binomial distribution. The original QNB model sets equal variance for both Input and IP data. However, we observe that in the MeRIP-seq, the IP read counts usually have higher variability than the Input read counts due to extra variation introduced during the IP process. Therefore, we modified the QNB model so that the variance parameter for IP is a magnitude higher than Input. Similarly, we generated data for the simple case as well as the difficult case with one confounding factor using the QNB model and applied each method to test for DM loci.

Coverage sub-sampling analysis

To demonstrate that the robust measurement of pre-IP RNA level implemented in RADAR improves the robustness to varying Input sequencing depth, we used T2D dataset as an example and performed sequencing depth sub-sampling analysis. We used the Sambamba [137] with parameters "view -h -t 20 -s 0.5 -f bam –subsampling-seed=1231" to sub-sample half of the reads from BAM files of Input samples. To count the overlap between results from sub-sampled and full data, we first obtained filtered bins that are shared in both datasets, then count the bin if it reached significant threshold in both datasets. To compare the log

fold change (logFC) estimates, we plotted the logFC estimated from the sub-sampled data against that estimated from the full data by each method.

De novo motif discovery and metagene analysis

To examine if the putative DM sites detected by RADAR are consistent with known characteristics of m^6A sites, performed de novo motif discovery analysis using the findMotifs function of Homer2 [94] with parameter "-len 5,6 -rna -p 20 -S 5 -noknown". A background sequence of randomly sampled peaks on transcriptome was used in the motif analysis.

Topological distribution of putative DM sites were plotted on a metagene using R package Guitar [95] with default settings.

Pathway enrichment analysis

A pathway enrichment analysis was performed on the DMGs identified from the ovarian cancer dataset using KEGG pathways [138] using the enrichKEGG function in R package ClusterProfiler [139].

Experimental validation by SELECT method

SELECT is an elongation and ligation-based method that can distinguish single m^6A site from A site [140]. Briefly, we design two oligos flanking the target m^6A/A site that leave a gap on the m^6A/A site. Then *Bst* DNA polymerase and SplintR ligase are used to fill the gap where m^6A hinders the elongation of the complementary oligo and thus prevent the gap to be filled. Finally, qPCR targeting the ligated oligo is used to quantify the abundance of the non-methylated RNA molecules. qPCR quantification targeting a nearby region on the target gene is used to normalize the gene expression variation. Since readout of SELECT method reflects relative abundance of non-methylated molecules, we expect the SELECT result to be inversely correlated to the m^6A levels.

Since SELECT method involves many steps for each site and is not feasible for high throughput analysis, we selected 6 sites including 2 DM sites that were only detected by RADAR and 4 DM sites that were implicated in T2D biology for experimental validation. We first matched RRACH motif in the putative DM peak and designed complementary oligos of 30 nt flanking the putative m^6A site. An additional 21 nt sequence at 5' of the up-probe and 20 nt sequence 3' of the down-probe were added to the oligos as universal primer sequence. For each DM site, we also designed a pair of primer targeting the gene harboring the DM site (see **Table 3.1** for oligo and primer sequences).

We applied the SELECT method to 4 control and T2D samples that have enough RNA material leftover from sequencing experiment. For each sample, 50 ng of total RNA was mixed with 0.8 μl up-probe and down-probe oligo (1uM) of each target m^6A site, 1 μl dTTP (100 μM), and 2 μl 10X CutSmart buffer (NEB) supplemented with H2O to 17 μl total volume. The mix was incubated at a temperature gradient: 90°C for 1min, 80°C for 1 min, 70°C for 1 min, 60°C for 1 min, 50°C for 1 min and then 40°C for 6 min. Subsequently, a 3 μl of enzyme mixture containing 0.5 μl Bst 2.0 DNA polymerase (0.02 U/μl) (NEB M0275S), 0.5 μl SplintR ligase (1U/μl) (NEB M0375S) and 2 μl ATP (5mM) was added in the former mixture to the final volume of 20 μl. The final reaction mixture was incubated at 40°C for 20 min then denatured at 80°C for 20 min. Then, qPCR reaction to quantify the "read through" oligos was assembled by 10 μl 2X qPCR master mix, 0.8 μl universal primer as designed in the oligo probes (10uM), 2 μl reaction from previous step and 7.2 μl H2O. To quantify the RNA expression level of each gene harboring the m^6A site, we first prepared the cDNA from 50 ng of total RNA using SuperScript VILO Master Mix (Thermo Fisher 11755050). Then 2 μl of cDNA were used for qPCR quantification of each gene. Finally, the gene expression level quantification was used to normalize the "read through" oligo probe quantification to obtain "relative read through" level for each site. Note the "relative read through" level reflect the non-methylated level, which is inversely correlated with m^6A site.

Name	Sequence
IGF1R_up	TAG CCA GTA CCG TAG TGC GTG CGC GAC GCA GTT CGC AAG ATC GCC CCG AAG
IGF1R_down	/5Phos/CC GGG TCA CAG GCG AGG CCG GCG AGG GGC CAG AGG CTG AGT CGC TGC AT
TRIB3_up	TAG CCA GTA CCG TAG TGC GTG AGC AAG ATG CAT AAG TAC CAT CCT TGG GAG
TRIB3_down	/5Phos/CT TAG AAA GCT CCC CAG GTT CGA GGC TGG GCA GAG GCT GAG TCG CTG CAT
CPEB2_up	TAG CCA GTA CCG TAG TGC GTG AGC GGC GGA GGC GGC GGC GGC GGC TTC GAG
CPEB2_down	/5Phos/CC GGA GGG TGG GGA AGG TGG GGA GGG CTG ACA GAG GCT GAG TCG CTG CAT
RNF213_up	TAG CCA GTA CCG TAG TGC GTG CCT TCT GAG GCA GAG GTG TAA GCG TTT CAG
RNF213_down	/5Phos/CC CAG ATC GGC TAC AGG GAG TGG CGC TCA GCA GAG GCT GAG TCG CTG CAT
MAFA_up	TAG CCA GTA CCG TAG TGC GTG GGC CTG GTG TCC ACG TCC TGT ACC GCG GAG
MAFA_down	/5Phos/CC GAG CCG AGG CCC CGA GAG GCC TGC GCG ACA GAG GCT GAG TCG CTG CAT
PDX1_up	TAG CCA GTA CCG TAG TGC GTG CTA ATT GAA TAC AAG GAG GCA AAT TCT AAG
PDX1_down	/5Phos/CT GAA CAG AAT ACA GAA AAT TCT GAC AGT CCA GAG GCT GAG TCG CTG CAT
IGF1R_qPCR_F	GCC GCT CAT TCA TTT TGA CT
IGF1R_qPCR_R	CTA GGC GAG GAA AAA CAA GC
TRIB3_qPCR_F	AAC CTT CAG TGC CTT CCA GA
TRIB3_qPCR_R	TGT TGT CAG CTC AAG GAT GC
CPEB2_qPCR_F	TTT CCA CCA AAA GGC TAT GC
CPEB2_qPCR_R	AGC CCT TAA TGG CCT AGG AA
RNF213_qPCR_F	ACA CCT CTG CCT CAG AAG GA
RNF213_qPCR_R	TGA AGG GGC ATT TTT AGC AC
MAFA_qPCR_F	GCG GAG AAC GGT GAT TTC TA
MAFA_qPCR_R	AAG GAA AGG GAG GCT GAG AA
PDX1_qPCR_F	AGC AGT GCA AGA GTC CCT GT
PDX1_qPCR_R	CAC AGC CTC TAC CTC GGA AC

Table 3.1: Oligo probes sequences and qPCR primer sequences. We designed an up and a down probe flanking the putative DM m^6A site leaving the m^6A nucleotide as a gap. For each pair of oligo probes, we designed an overhanging universal primer sequence at their 5' and 3' end, respectively. The table shows the sequence of oligo probes we used as well as the qPCR primers we used to quantify gene level variation.

3.2.10 Sample size analysis

To investigate the effect of sample size on the power of detecting DM sites, we simulated datasets of varying sample sizes from N = 2 to N = 8 using Model (1) without covariates. We assessed the performance by plotting the sensitivity of each method against its FDR using DM loci obtained at an FDR cutoff of 10%. Additionally, we also made the ROC curve for each sample size by varying the FDR cutoffs when selecting predicted sites.

3.3 Results

3.3.1 RADAR overcomes challenges in modeling MeRIP-seq data and accommodates complex study designs

Using BAM files as the input, RADAR first divides transcripts (concatenated exons) into 50-bp consecutive bins and quantifies pre-IP and post-IP read counts for each bin (**Figure 3.1a**). Unlike current differential methylation analysis methods [41, 42, 43, 19] that scale to library sizes as a way of normalization, which can be strongly skewed by highly expressed genes [134] (**Supplementary Figure 3.1**), RADAR uses the median-of-ratio method [113] implemented in DEseq2 to normalize the Input library for the sake of robustness. For the IP library, RADAR normalizes the fold enrichment computed from the IP counts divided by the Input counts, which takes both IP efficiency and IP library size variation into account.

After proper normalization across all samples, RADAR then calculates the methylation level for each bin conditioned on its pre-IP RNA expression level for each sample. In contrast to previous methods [41, 42, 43, 19] that use peak-level read counts in the Input library as its measurement of pre-IP RNA expression level, we use gene-level read counts as a more robust representation, which is defined as the total number of reads across all bins that span the same gene (**Figure 3.1a**). This choice is motivated by the observation that the median read coverage within each peak is very low —18 reads per peak (7 reads in a 50-bp bin) (**Supplementary Figure 3.2**) in a typical MeRIP-seq input sample of 20 million (map-

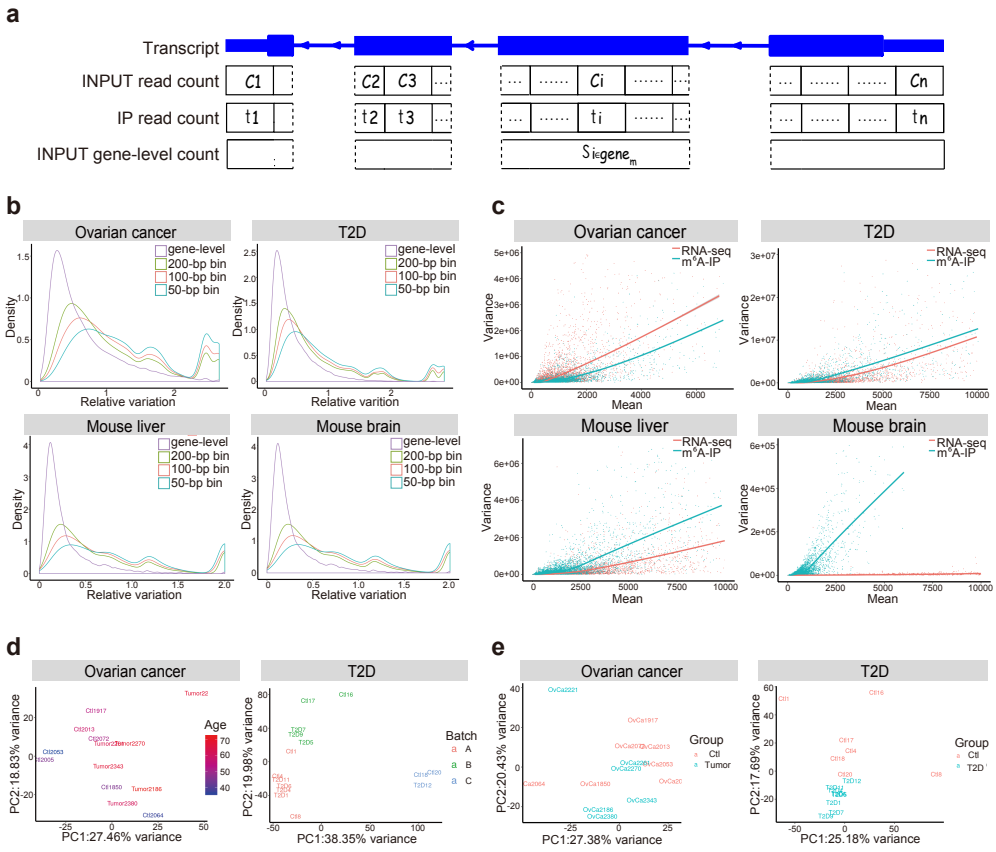


Figure 3.1: Unique features of m^{6}A-seq (MeRIP-seq) data. RADAR divides concatenated exons of a gene into consecutive bins and models the immunoprecipitation (IP)-enriched read counts in such bins. **a** depicts a pair of read counts in the Input and the IP library in the i-th bin as c_i and t_i. In the RADAR workflow, the gene-level read count of the Input library $s_{i \in gene_m}$ substitutes the bin-level read count c_i as the representation of the pre-IP RNA levels of the i-th bin. **b** compares the relative variation of gene-level and bin-level (local) read counts of different bin sizes in four m^{6}A-seq datasets, suggesting that unwanted variation can be reduced using gene-level counts as the estimates of pre-IP RNA levels. Panel **c** compares the cross-sample mean and variance of regular RNA-seq (pre-IP counts) and m^{6}A-seq (post-IP read counts adjusted for pre-IP RNA level variation) data in four m^{6}A-seq datasets. The fitted curvature of m^{6}A-seq can differ from that of RNA-seq, indicating that m^{6}A-seq may have a different mean-variance relationship from RNA-seq. Biological and experimental confounding factors are often encountered in patient samples. **d** shows the first two principal components (PCs) of m^{6}A enrichment in each dataset, where the samples are colored by covariates that need to be accounted for. m^{6}A enrichment was represented by IP sample read counts adjusted for pre-IP (Input) RNA level variation. **e** shows the first two PCs after regressing out known covariates — age in the ovarian cancer dataset and batch in the T2D dataset. After regressing out the covariate, samples are separated by disease conditions on the PCA plot.

pable) reads (**Supplementary Figure 3.3**). Over dispersion of low counts due to random sampling in the sequencing process can introduce substantial unwanted variation to the estimation of pre-IP RNA level. This can be further exacerbated by the uneven distribution of reads caused by local sequence characteristics such as GC content and mappability. Using gene-level counts as the estimate of pre-IP RNA expression level can mitigate the dispersion by increasing the number of reads (272 reads on average) and simultaneously diminishing the effects of sequence characteristics within a gene (**Figure 3.1a**). By comparing the variance of read counts across replicates at the gene-level with that at the bin-level, we show that the cross-sample variance is much less at the gene-level than at the bin-level in all three datasets. (**Figure 3.1b**).

RADAR models the read count distribution using a Poisson random effect model instead of a negative binomial distribution, which is commonly used in RNA-seq analysis [113, 134, 141] as well as in DRME and QNB for MeRIP-seq analysis [43, 44]. Negative binomial distribution-based models assume a quadratic relationship between mean read counts and their variance across all genes. We observe in real m^6A-seq datasets that the mean-variance relationship of post-IP counts across genes significantly differs from that of regular RNA-seq counts (i.e., pre-IP counts). The former does not always follow a similar quadratic curvature and can exhibit very different patterns of variability (**Figure 3.1c**, **Supplementary Figure 3.4**). To overcome these limitations, RADAR applies a more flexible generalized linear model framework (see Method) that captures variability through random effects.

Another important advancement of RADAR, compared to existing MeRIP-seq data analysis tools [41, 42, 43, 19], is the flexibility to incorporate covariates and permit more complex study design. Phenotypic covariates such as age and gender as well as experimental covariates such as batch information are often encountered in epitranscriptomic profiling studies with heterogeneous patient samples. Covariates such as litter and age are common in experimental animal studies. For example, in the ovarian cancer dataset, the age of the tissue donors is partially confounded with predictor variable – disease status. In the T2D islets

dataset, the variance of the first two principal components is confounded with the sequencing batch (**Figure 3.1d**). After regressing out the batch effect, the remaining variance can be better explained by disease status (**Figure 3.1e**). This indicates the importance of controlling for potential confounding factors when performing differential methylation tests. The generalized linear model framework in RADAR allows the inclusion of covariates and offers support for complex study designs.

3.3.2 Comparative benchmarks of different methods using simulated datasets

To evaluate the performance of RADAR in comparison to current methods, we applied RADAR and other methods for MeRIP-seq differential analysis including exomePeak, Fisher's exact test, MeTDiff and QNB on simulated datasets. We considered four scenarios: the proposed random effect model with/without covariates and the quad-negative binomial (QNB) model adopted from QNB [44] with/without covariates. For each scenario, we evaluated the sensitivity and false discovery rate (FDR) of different methods using ten simulated copies. We first simulated a dataset of 8 samples using the random effect model (Method Section Equation (3.1), denoted as the simple case). The Input library was directly drawn from the T2D dataset. We simulated IP read count adjusted for pre-IP expression level of each bin according to Equation (3.1) where μ is equal to mean log read count in the "control" group of T2D dataset. The final IP read counts were obtained by rescaling simulated data by the average IP/Input ratio observed in the T2D data. In total, we simulated three datasets of 26,324 sites in which 20% of sites are true positives with effect size of 0.5, 0.75 or 1, respectively.

For DM loci with an effect size of 0.5, RADAR achieved 29.1% sensitivity and 12.0% FDR at an FDR cutoff of 10%. At the same cutoff, exomePeak and Fisher's test achieved 72.8% sensitivity/52.5% FDR and 72.2% sensitivity/50.5% FDR, respectively. MeTDiff achieved 10.5% sensitivity and 16.2% FDR. QNB, on the contrary, did not own any power for the small effect size. When the effect size increased, RADAR achieved much higher sensitivity,

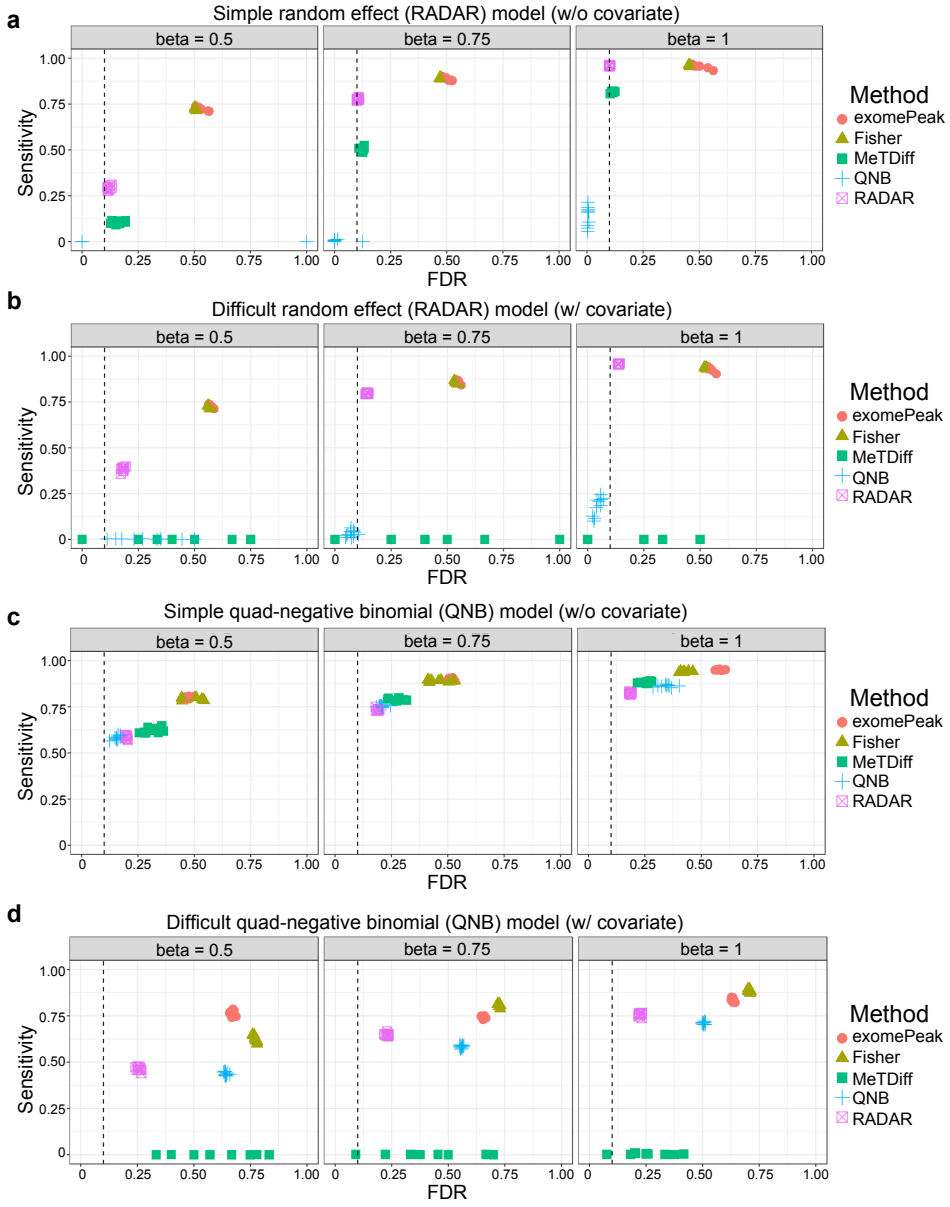


Figure 3.2: Benchmarking RADAR on two simulation models. We benchmarked RADAR and other alternative methods using two simulation models — a random effect (RADAR) model and a quad-negative-binomial (QNB) model. We simulated dataset of 8 replicates of varying true effect size (0.5, 0.75 and 1) with and without covariates. We compared the results at an FDR cutoff of 0.1 with simulated true sites. We show the sensitivity (fraction of true sites detected by the method at an FDR cutoff of 0.1) and false discovery rate (fraction of detected differential sites that are not true sites) of each method applied on data simulated by the random effect model without covariates in **a** and with covariates in **b**, the quad-negative-binomial model without covariates in **c** with covariates in **d**, respectively. The FDR cutoff used to select DM sites is labeled by a dashed line.

77.8% for an effect size of 0.75 and 95.7% for an effect size of 1, while FDR were well calibrated at 10.4% and 10.1%, respectively. exomePeak and Fisher's test both achieved 89% and 96% sensitivity for effect sizes of 0.75 and 1, respectively, but at the cost of unsatisfactory FDRs, which were greater than 46%. MeTPeak exhibited well-calibrated FDR (12.3% and 11.4%) and moderate sensitivity of 50.4% and 81.5% for effect sizes of 0.75 and 1, respectively. QNB only had low power for an effect size of 1 (beta = 1, 13.9% sensitivity and 0.5% FDR). Overall, for the simple case without covariates, RADAR achieved high sensitivity while maintained low FDR at varying true effect sizes (**Figure 3.2a**). We then applied the above analysis at varying FDR cutoff and found RADAR achieved the highest sensitivity at a fixed level of empirical FDR (**Supplementary Figure 3.5a**). We note that exomePeak and Fisher's test achieve high sensitivity at all effect sizes as combining read counts across replicates of the same group helped to gain power. As a tradeoff, they fail to account for within-group variability resulted in high FDR. On the contrary, RADAR and MeTDiff exhibit well-calibrated FDR while achieve high sensitivity at same levels as exomePeak for large effect sizes. QNB is overconservative and possessed little power.

We next applied the aforementioned methods to the proposed model with a covariate (effect size equal to 2, denoted as the difficult case) (**Figure 3.2b**). As a result, at an FDR cutoff of 10%, RADAR achieve 38.4%, 79.7% and 95.7% sensitivity with empirical FDRs slightly higher than those in the simple case (18.2%, 14.4% and 13.7% for effect sizes of 0.5, 0.75 and 1, respectively). MeTDiff, with similar performance as RADAR in the simple case, lose power in the difficult case due to incapability of accounting for confounding factors. exomePeak, Fisher's test and QNB behave similarly as in the simple case. The advantage of RADAR over other methods is robust to the choice of FDR cutoff as shown in **Supplementary Figure 3.5b**. In summary, RADAR outperformes existing alternatives in both cases.

Taking the covariate model with a DM effect size of 0.75 as an example, we also checked the distributions of effect size estimates and P values obtained from each method. In all

methods, effect sizes are overall correctly estimated with estimates for "true" sites centered at 0.75 (**Supplementary Figure 3.6a**) and that for null sites centered at zero (**Supplementary Figure 3.6b**). However, we note the distribution of beta estimates is narrower for RADAR, especially in the difficult case, suggesting a more confident estimation. P values of exomePeak and Fisher's test at null sites are enriched near zero, indicating over-detection of false positive signals (**Supplementary Figure 3.6c**). We also observed many large P values obtained by QNB for "true" sites in both cases and MeTDiff in the difficult case, which suggested a high false negative rate (**Supplementary Figure 3.6d**).

We then repeated simulation studies using the QNB model. Instead of setting the variances of Input and IP libraries equal as presented in the QNB paper, we let the variance of IP read count be larger than that of Input. This setting better reflects our observation in the real data as extra noise can be introduced during immunoprecipitation process for IP reads generation (**Supplementary Figure 3.4**). In the simple case without covariates, RADAR exhibits the lowest empirical FDR (18.9% and 18.5%) despite of slightly lower sensitivity comparing to other methods (73.5% and 82.3%) when the effect sizes are relatively large (for effect sizes of 0.75 and 1). QNB performs better when the effect size is small with 58.6% sensitivity and 15.6% FDR for an effect size of 0.5 (**Figure 3.2c**). The results are consistent when we evaluated their performance with different FDR cutoffs. Overall, QNB performs slightly better than RADAR with an effect size of 0.5. RADAR achieve similar sensitivity but better-calibrated FDR when effect sizes equal to 0.75 and 1 (**Supplementary Figure 3.5c**). In the model with covariates, RADAR exhibits the lowest empirical FDR, with 25.8%, 23.0% and 22.5% at effect sizes of 0.5, 0.75 and 1, respectively, while other methods either fail to detect the signal or have a higher empirical FDR. Specifically, MeTDiff has sensitivity below 0.5% at varying effect sizes and QNB reached FDRs of 64.1%, 55.8% and 50.5% for effect sizes of 0.5, 0.75 and 1, respectively, at an FDR cutoff of 10% (**Figure 3.2d**). The advantage of RADAR over alternative methods hold in the difficult case at varying cutoffs (**Supplementary Figure 3.5d**). In summary, RADAR outperforms other existing methods

in most scenarios, particularly when covariates are present.

3.3.3 Comparative benchmarks of different methods using four real m^{6}A-seq datasets

Next, we compared the performance of different methods using four real m^{6}A-seq datasets: ovarian cancer (GSE119168), T2D (GSE120024), mouse liver (GSE119490) and mouse brain (GSE113781). To evaluate the sensitivity of different methods, we first checked the distributions of P values obtained from corresponding DM tests (**Figure 3.3**). In the ovarian cancer, T2D and mouse liver data, Fisher's test and exomePeak detect the most signals as the P values are most dense near zero. In these three datasets, RADAR also returns a desirable shape for the P values histogram in which P values are enriched near zero while uniformly distributed elsewhere. MeTDiff returns a desired shape only in the ovarian cancer and mouse liver datasets. QNB is overconservative in the ovarian cancer and T2D dataset. All methods fail to return enriched P values near zero for the mouse brain dataset, suggesting there is no or little signal in this dataset. This is consistent with the original publication that very few differential peaks were detected in this study [38].

To ensure that well-performed methods achieve high sensitivity while maintain a low FDR, we further performed permutation analyses to obtain the null distribution of P values for each dataset. Specifically, we shuffled the phenotype labels of samples such that the new labels are not associated with the true ones or any other important confounding factors. We expected the P values from a permutation test to follow a uniform distribution and the enriched P values near zero would be considered as false discoveries. For each dataset, we combined test statistics from 15 permuted copies and compared their distribution with the original tests (**Figure 3.4**). P values from Fisher's test and exomePeak are strongly enriched near zero and only slightly lower than those from the original tests. This suggests the strong signals detected by these two methods are likely to be false discoveries, consistent with the conclusion from simulation analysis. On the contrary, the histograms of P values from

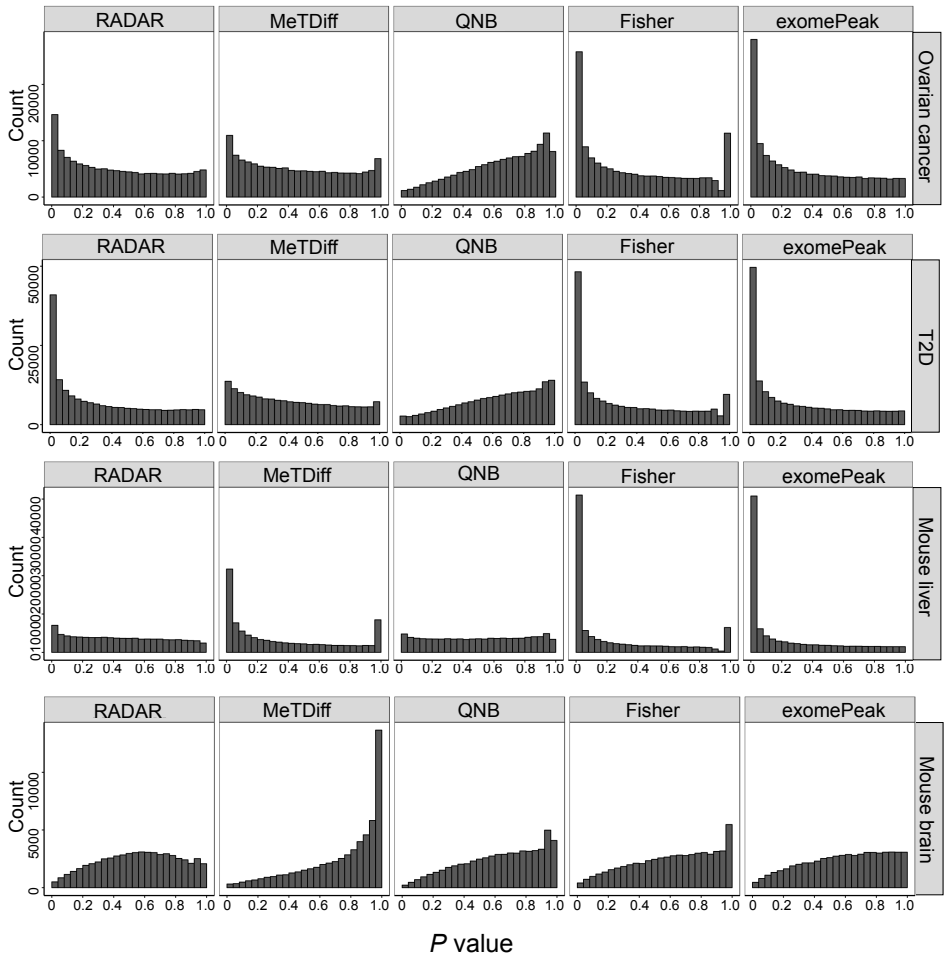


Figure 3.3: Sensitivity of benchmarked methods on real m^{6}A-seq data. We benchmarked RADAR and other alternative methods on four m^{6}A-seq data with different characteristics. Each panel shows the histogram of P values obtained from DM tests using RADAR, MeTDiff, QNB, Fisher's exact test and exomePeak on each dataset, respectively.

RADAR are close to flat in all datasets, indicating that strong signals detected by RADAR are more likely to be true. MeTDiff exhibits well-calibrated P values in the ovarian cancer and T2D data but enriches for small P values in the mouse liver data with an indicated high FDR. QNB test returns conservative P value estimates in all datasets. Taking together these analyses, we demonstrate that RADAR outperforms the alternatives by achieving high sensitivity and specificity simultaneously in real datasets.

To better demonstrate that RADAR detect DM sites with better sensitivity and speci-

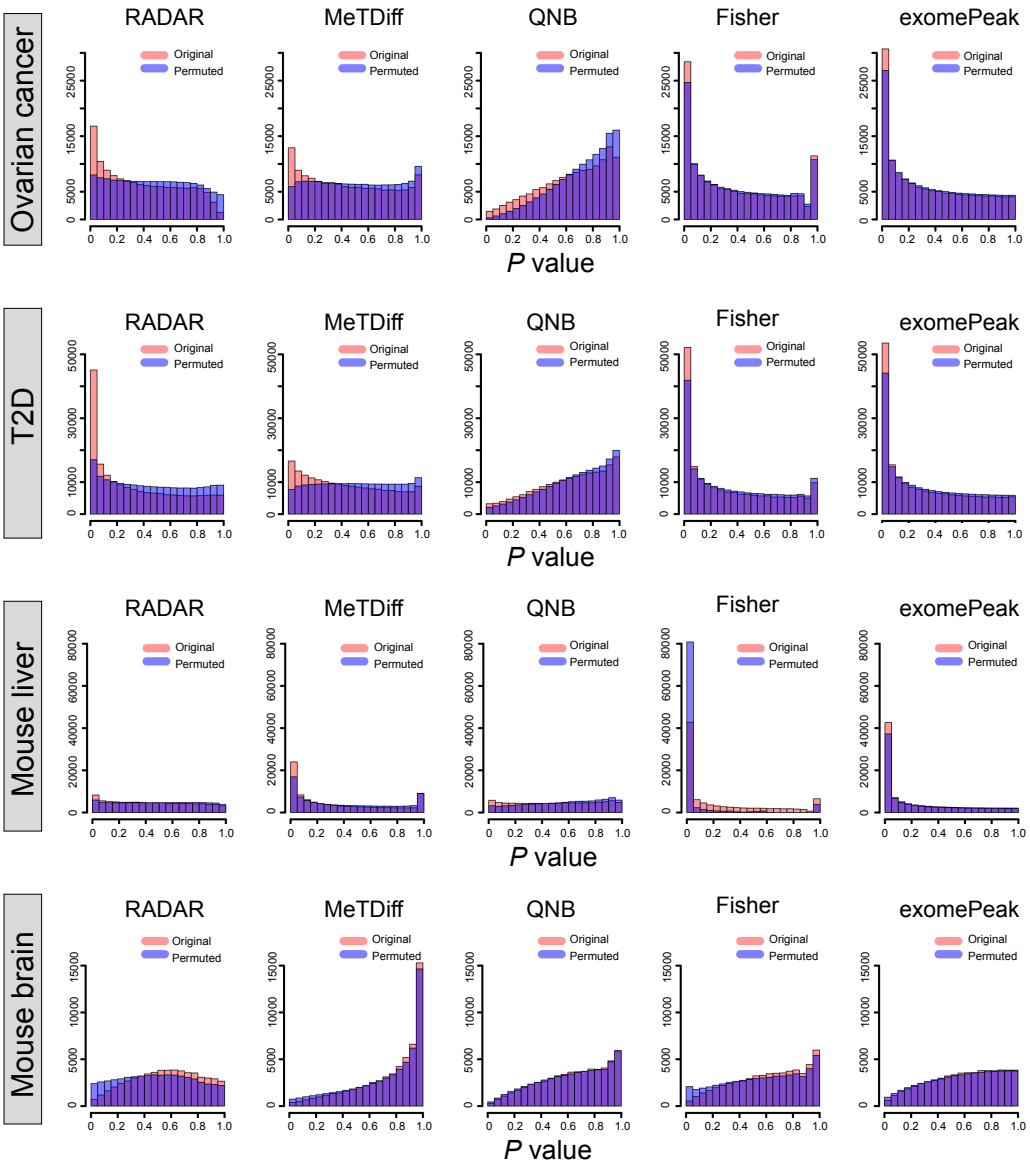


Figure 3.4: Benchmarking false positive signals using permutation analysis on real m^{6}A-seq data. To assess empirical FDR of the test, we permuted the phenotype labels of samples so that the new labels are not associated with true ones. Each panel shows the histograms of P values obtained from DM tests on 15 permuted copies (blue) and those from the tests on the original dataset (red).

ficity in real data, we show examples of DM site that is only detected by RADAR as well as likely false discovery sites identified by exomePeak and Fisher's test but not by RADAR in the T2D dataset. We plot sequence coverage of individual samples for the DM sites in

the RNF213 gene (**Supplementary Figure 3.7a**) and show despite of large variability in control samples, m^6A enrichment of T2D samples are consistently lower on this locus. Conversely, in the bogus DM sites detected by alternative methods (**Supplementary Figure 3.7b and c**), enrichment differences are mainly driven by one or two outlier samples in one group.

To further demonstrate the advantage of using gene-level read counts over local read counts to account for RNA expression level, we repeated the above analysis using post-IP counts adjusted by the local read counts of Input. We show that in the T2D dataset, gene-level adjustment not only enables stronger signal detection, but also lowers FDR as we observed that the permutation analysis using local counts adjustment results in undesired stronger signals around zero in the P value histogram (**Supplementary Figure 3.8**). In the ovarian cancer and the mouse liver datasets, local counts adjustment achieve higher signal detection but at the cost of a higher FDR. This analysis suggests that using gene-level read counts as the estimates of pre-IP RNA expression levels could effectively reduce FDR and lead to more accurate DM loci detection.

Attributed to the robust representation of pre-IP RNA expression level using gene-level read counts, RADAR's performance is more robust to the sequencing depth of Input samples. To demonstrate this, we applied RADAR on data created by sub-sampling the read counts of Input samples in the T2D dataset so that the sequencing depth is half of the full dataset (average 17.5 million reads). We compared the DM sites detected in the reduced dataset with the results obtained from the full dataset (**Supplementary Figure 3.9a**). Using a 10% FDR cutoff, RADAR-detected DM sites in the reduced dataset show the highest overlap with that in the full dataset. MeTDiff and QNB only have a few overlapping DM sites between the sub-sampled and full dataset. Fisher's test and exomePeak have slightly fewer overlaps comparing to RADAR but have more false discoveries. We further compared the log fold change (logFC) estimates from reduced and full datasets to check their consistency. As a result, we found reduced sequencing depth had the least impact on the logFC estimated by

RADAR while the estimates by others are much less reproducible with a shallower sequencing depth (**Supplementary Figure 3.9a**).

Unlike earlier pipelines that perform DM tests only on peaks identified from peak calling, RADAR directly tests on all filtered bins and reports DM sites. To check if the DM sites reported by RADAR are consistent with known characteristics of m^6A, we performed de-novo motif search on these sites and found DM sites detected in ovarian cancer, mouse liver and T2D datasets are enriched for known m^6A consensus motif (**Supplementary Figure 3.10a**) [33], suggesting DM sites reported by RADAR are mostly true. We also examined the topological distribution of these DM sites by metagene analysis (**Supplementary Figure 3.10b**). The distributions in ovarian cancer and mouse liver datasets are consistent with the topological distribution of common m^6A sites, indicating methylation changes occurred in these two datasets are not spatially biased. Interestingly, DM sites detected in T2D dataset are strongly enriched at 5'UTR, suggesting T2D related m^6A alteration are more likely to occur at 5'UTR.

3.3.4 RADAR analyses of m^6A-seq data connect phenotype with m^6A-modulated molecular mechanisms

Finally, we investigated whether DM test results obtained from RADAR would lead to better downstream interpretation. In the ovarian cancer dataset, we performed KEGG pathway enrichment analysis on the differential methylated genes (DMGs) detected by RADAR (**Figure 3.5a**). We found the detected DMGs are enriched with molecular markers related to ovarian cancer dissemination [142, 143]. For instance, we identified key regulators of the PI3K (enrichment P value 7.8×10^{-5}) and MAPK pathways (enrichment P value 1.1×10^{-4}), including hypo-methylated PTEN and hyper-methylated BCL2 (**Supplementary Figure 3.11**). Other notable DMGs include key markers of ovarian cancer such as MUC16 (CA-125) and PAX8, as well as genes that play key roles in ovarian cancer biology such as CCNE1 and MTHFR. Conversely, DMGs detected by MeTDiff are only enriched in three KEGG

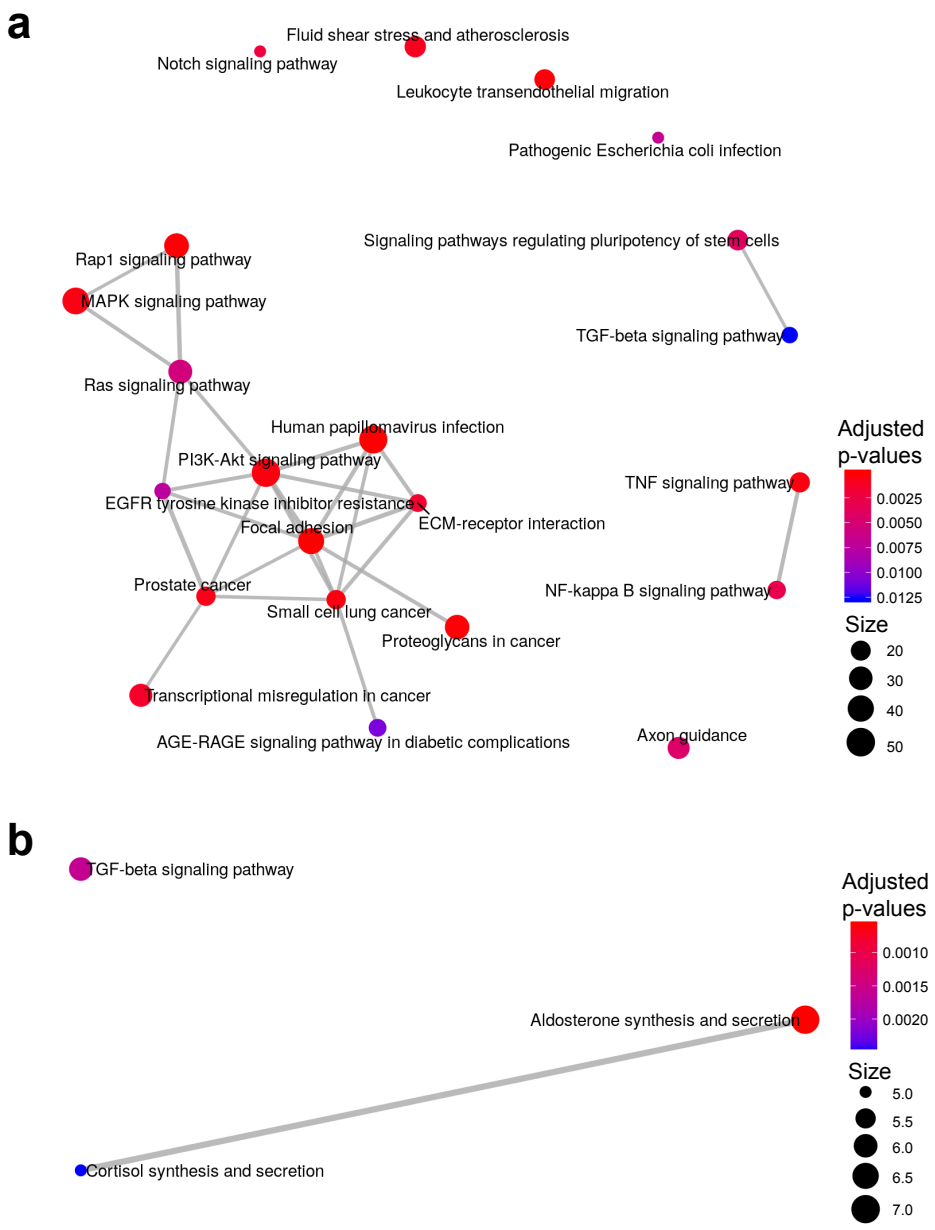


Figure 3.5: Pathways enriched in differential methylated genes identified in ovarian cancer and T2D datasets. We performed KEGG pathway enrichment analysis using ClusterProfiler on DMGs identified in the ovarian cancer dataset by RADAR **a** and MeTDiff **b**, respectively. The enrichment maps represent identified pathways as a network with edges weighted by the ratio of overlapping gene sets.

pathways (**Figure 3.5b**), most likely due to its inadequate power. We showed through permutation analysis that exomePeak and Fisher's test results included a significant portion of

false positives and could lead to biased downstream interpretations.

In the T2D dataset, DMGs identified by RADAR are enriched in related pathways including insulin signaling pathways, type 2 diabetes mellitus, mTOR pathways and AKT pathways, indicating a role that m^6A might play in T2D. We further analyzed these DMGs in related pathways and found the methylome of insulin/IGF1-AKT-PDX1 signaling pathway being mostly hypomethylated in T2D islets (**Supplementary Figure 3.12**). Impairment of this pathway resulting in down-regulation of PDX1 has been recognized as a mechanism associated with T2D where PDX1 is a critical gene regulating β-cells identity, cell cycle, and promote insulin secretion [144, 145, 146, 147]. Indeed, follow-up experiment on a cell line model validated the role of m^6A in tuning cell cycle and insulin secretion in β-cells and animal model lacking methyltransferase gene *Mettl14* in β-cells recapitulated key T2D phenotypes (results presented in a separate manuscript [148], see Appendix). To summarize, RADAR-identified DMGs enable us to pursue an in-depth analysis of the role that m^6A methylation plays in T2D. On the contrary, due to the incapability to take sample acquisition batches as covariates, the alternative methods are underpowered to detect DM sites in T2D dataset and could not lead to any in-depth discovery of m^6A biology in T2D islets. These examples suggest that MeRIP-seq followed by RADAR analysis could further advance functional studies of RNA modifications.

3.3.5 Validation of RADAR-detected DM sites by the SELECT method

Recently, Xiao et al. developed an elongation and ligation-based qPCR amplification method (termed SELECT) for single nucleotide-specific detection of m^6A [140]. This method relies on mechanism different from antibody pulldown-based MeRIP-seq to detect m^6A, making it a suitable method for validating DM sites discovered by RADAR analysis. We selected 6 DM sites (**Table 3.2**) including 2 sites only detected by RADAR and 4 sites in genes important in β-cell for experimental validation using the SELECT method. Among 6 validated sites, the β-cells regulator PDX1 and RADAR-specific DM sites show significant m^6A level alteration

with P value 0.009 and 0.017, respectively (**Figure 3.6**). Three other sites: IGF1R in

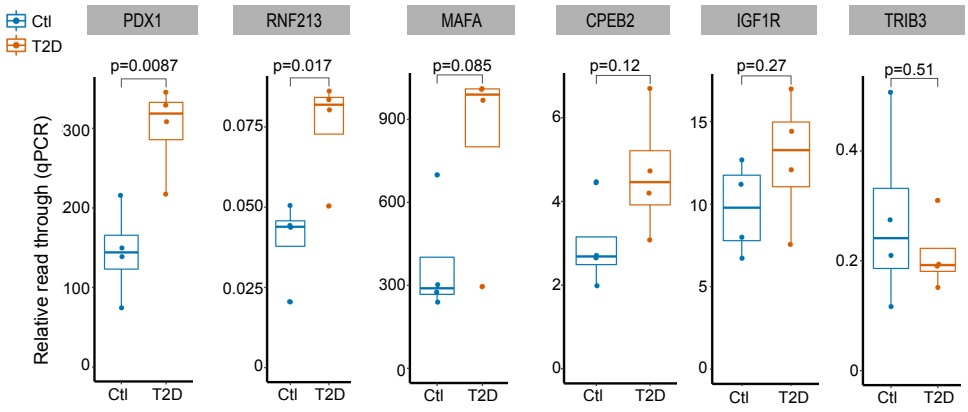


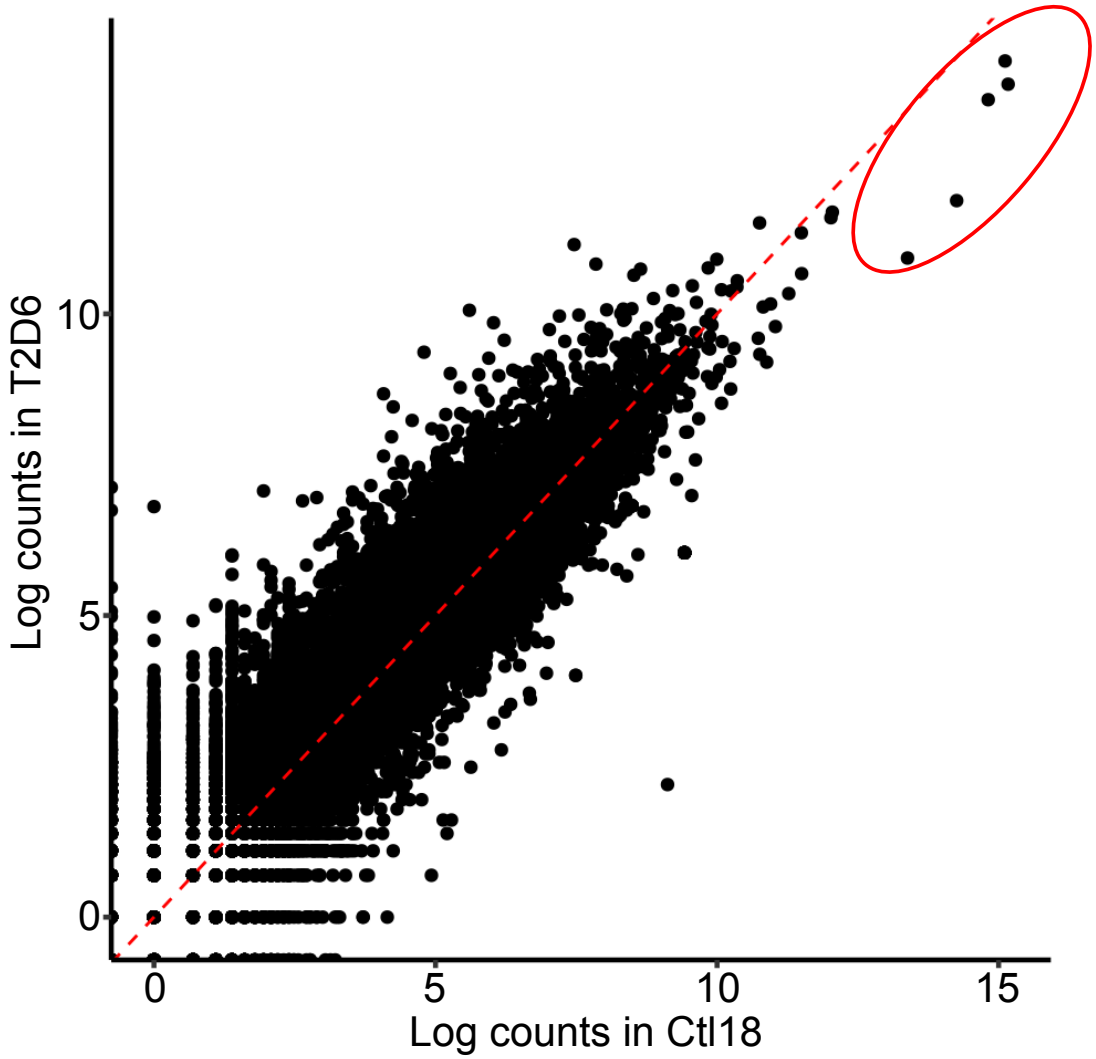
Figure 3.6: Experimental validation of RADAR-detected DM sites using SELECT method. We applied antibody independent method SELECT on T2D samples (N = 4). Shown are SELECT results of 6 putative DM sites for validation. SELECT measures the relative abundance of non-methylated RNA molecules of target locus as represented by the elongation and ligation "read through" of olilgo probes. Thus, SELECT results—"relative read through"—are inversely correlated with m^6A level.

the insulin/IGF1-AKT-PDX1 signaling pathway, MAFA—another important regulator of β-cells function and RADAR-specific DM site in CPEB2 show m^6A changes consistent with RADAR result despite of not reaching statistical significance. The sites in the TRIB3 gene are similarly methylated in control and T2D samples as measured by SELECT. Overall, five out of six experimentally validated sites are supported by orthogonal evidence by SELECT, confirming the reliability of RADAR-detected differential methylation sites.

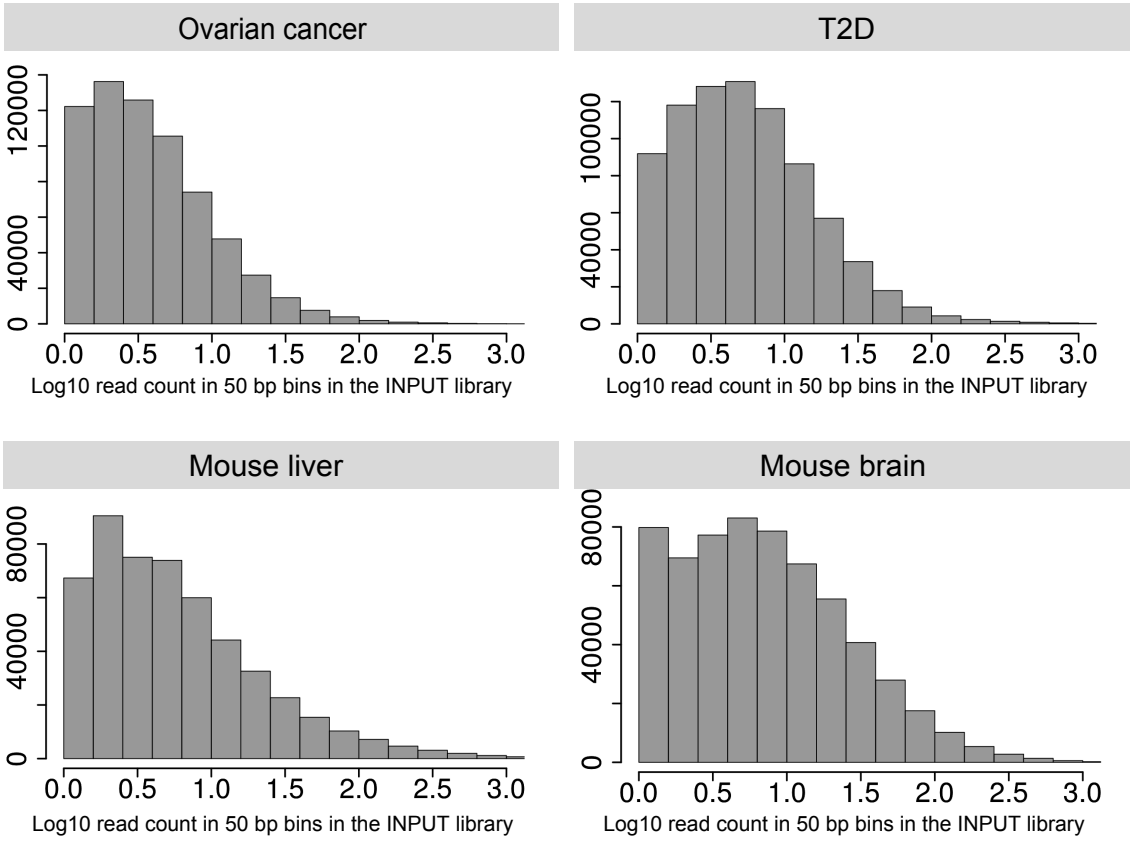
chr	start	end	name	score	strand	thickStart	thickEnd	itemRgb	blockCount	blockSizes	blockStarts	logFC	*P* value
chr15	98648988	98649137	IGF1R	0	+	98649038	98649087	0	1	149	0	-0.663268535758347	2.84108370870014e-07
chr4	15003223	15003372	CPEB2	0	+	15003273	15003322	0	1	149	0	-1.46982006559669	1.05070727955586e-06
chr17	80395488	80395637	RNF213	0	+	80395538	80395587	0	1	149	0	-1.2727126729966	2.01338246309221e-06
chr8	143429196	143429395	MAFA	0	-	143429246	143429345	0	1	199	0	-0.838372746377749	5.26955131288313e-08
chr13	27924434	27924583	PDX1	0	+	27924484	27924533	0	1	149	0	-0.696256238421844	7.3903106081821e-05

Table 3.2: Selected DM sites for experimental validation. The table shows the peak information of selected putative DM site from RADAR analysis. The genome coordinate is based on hg38. The shown peak table was extended 50 bp towards both upstream and downstream to search for RRACH motif match because the RNA molecules for m^6A-seq was fragmented to 150 nt but our sequence reads were only 50 bp. Consequently, the estimated peak locations could have position shift from the real peak for up to 100 bp. The extension is intended to take this uncertainty into account.

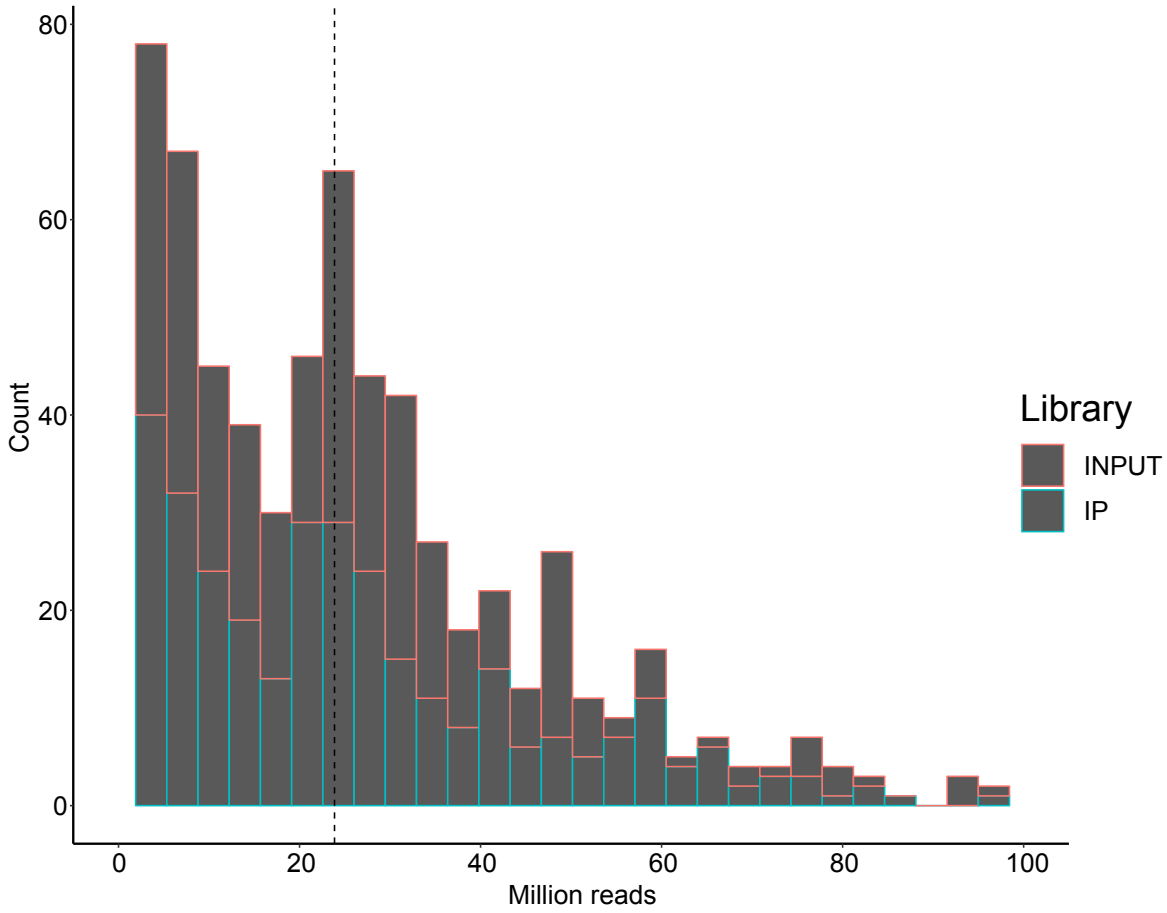
3.4 Supplementary figures



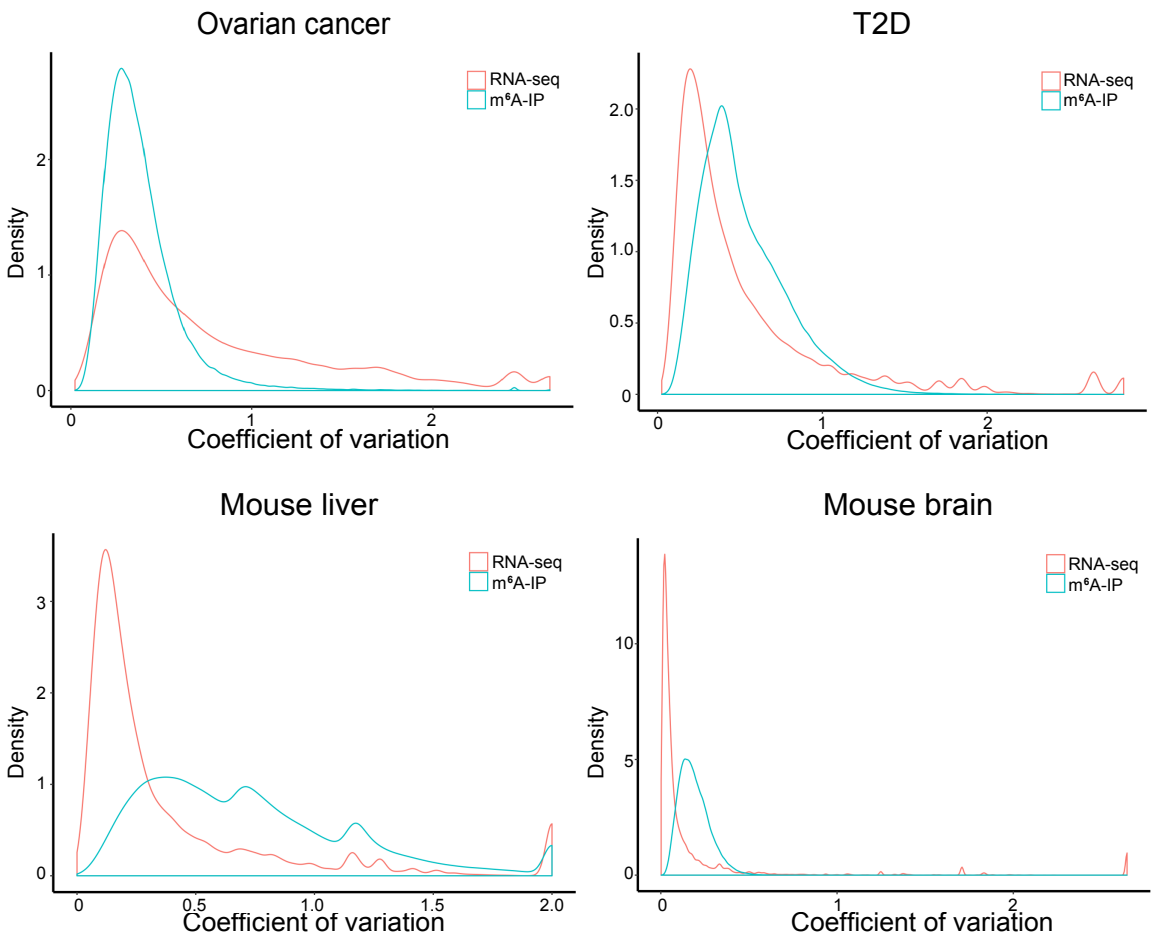
Supplementary Figure 3.1: Scatter plot of read count. The input library log read counts of a sample from one experimental group are plotted against a sample from another experimental group. Shown is an example scenario where highly expressed genes can result in underestimation of other genes when normalizing by total coverage. The highly expressed genes that can strongly influence the scaling factor estimation are highlighted by red circles.



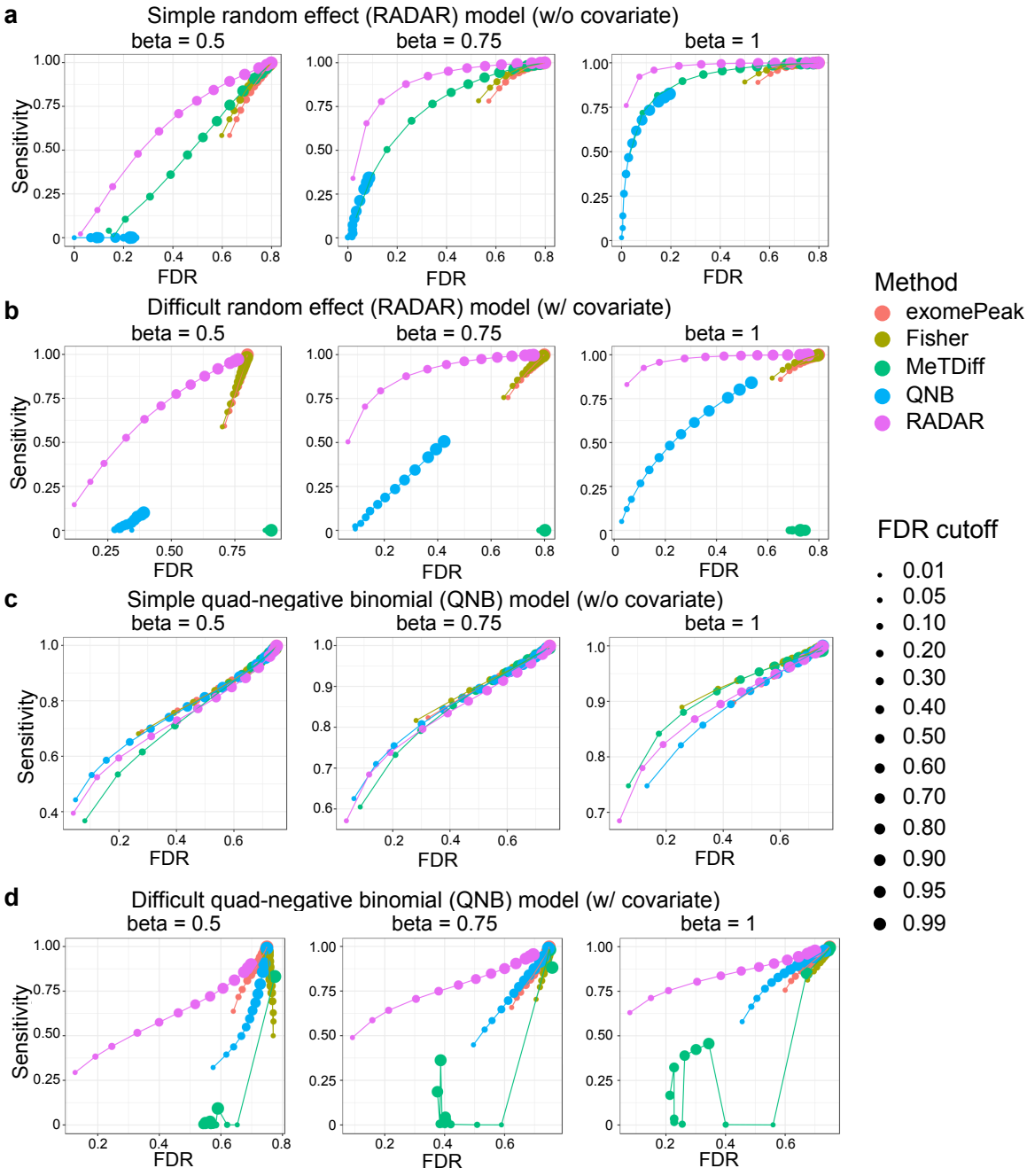
Supplementary Figure 3.2: Read count distribution of Input data. Distribution of read count c_i (**Figure 3.1a**) in a 50 bp bin of input library from real m^{6}A-seq datasets. The read count is shown in log10 scale.



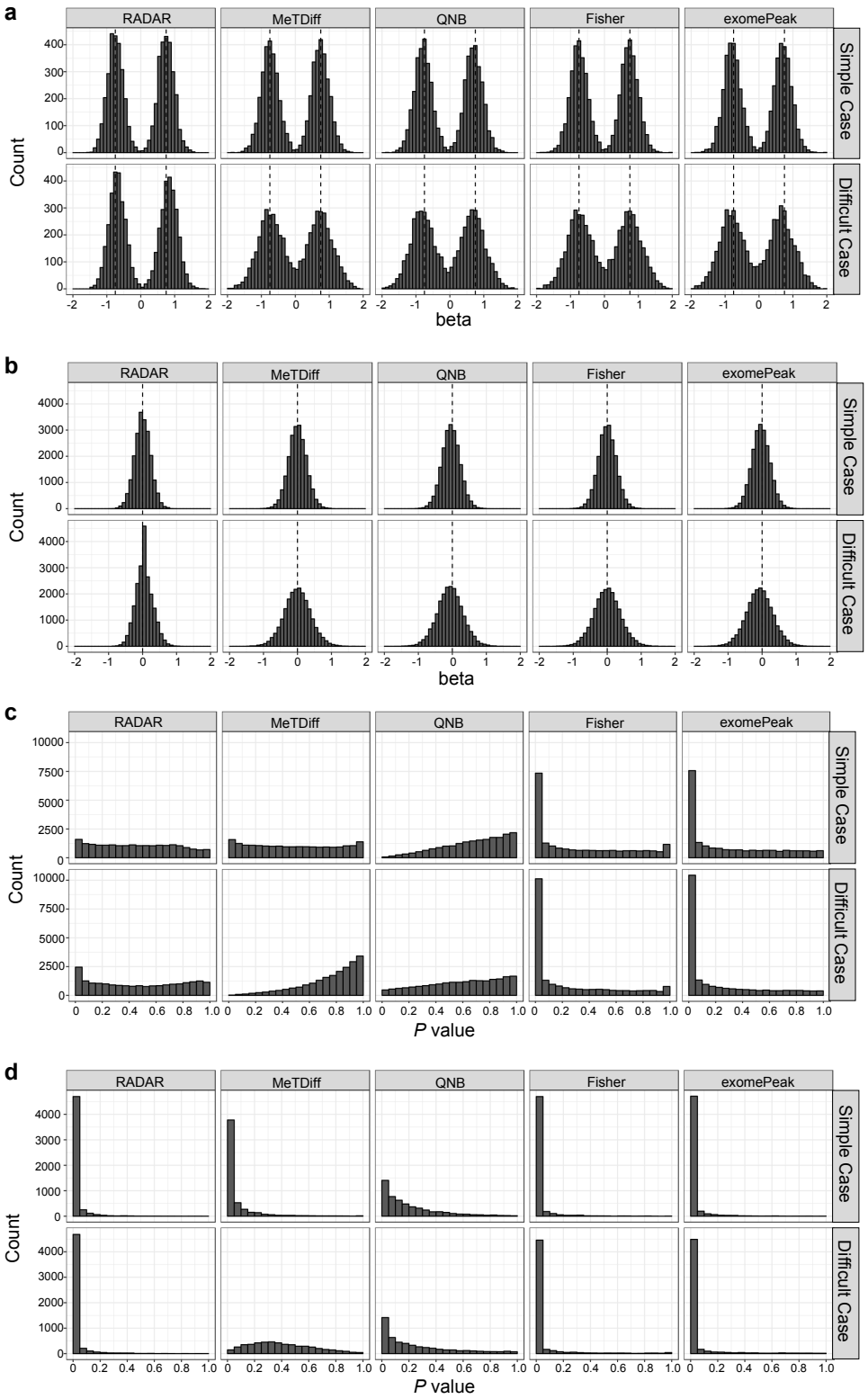
Supplementary Figure 3.3: Sequencing depth distribution of m^{6}A-seq in published literatures. Distribution of sequencing depth (million reads) drawn form a m^{6}A-seq database [84] is shown by histogram. The database included 339 datasets from published literatures.



Supplementary Figure 3.4: Variability distribution comparing RNA-seq and m^{6}A-seq (MeRIP-seq). Density plot comparing variabilities of m^{6}A-seq data with RNA-seq data. Variability is represented by coefficient of variation.

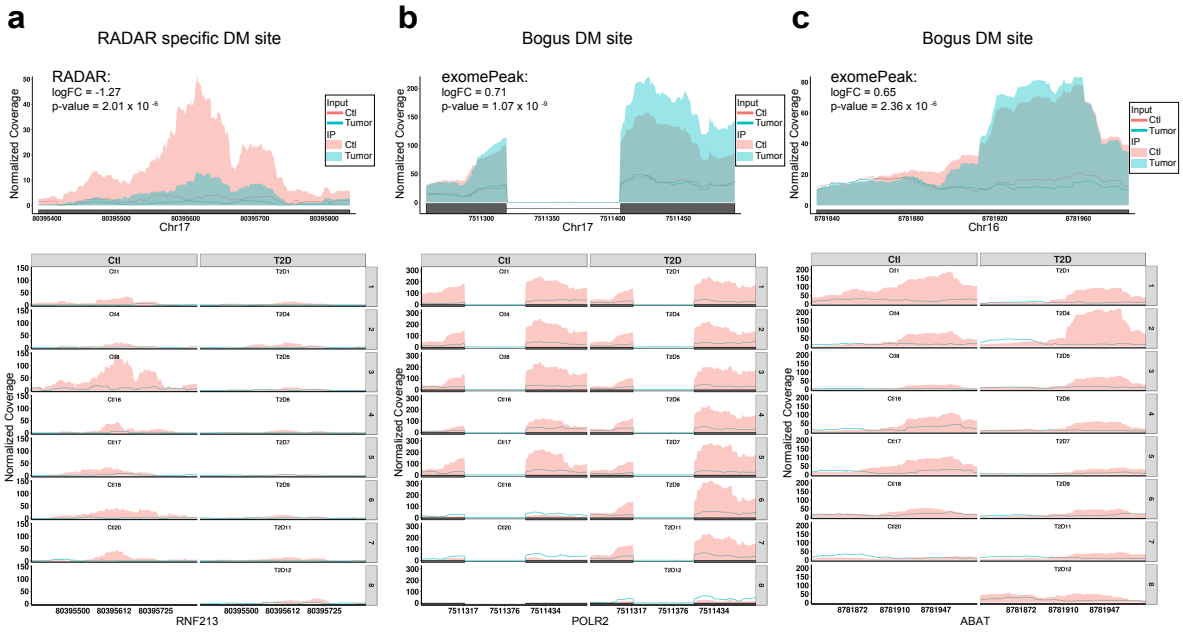


Supplementary Figure 3.5: Evaluating performances of benchmarked methods on simulated data using sliding thresholds. We evaluated the performance of RADAR and other methods by comparing the sensitivity and empirical FDR obtained by varying FDR threshold for selecting DM sites. The threshold of selecting DM sites are labeled by the size of data points.

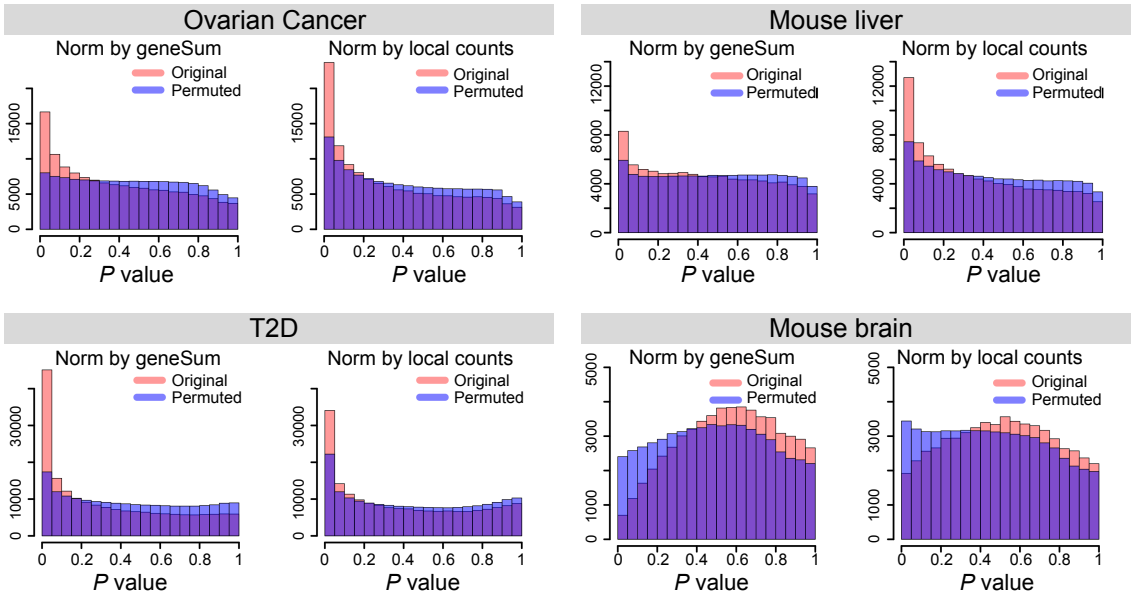


Supplementary Figure 3.6: *P* value and effect size estimates

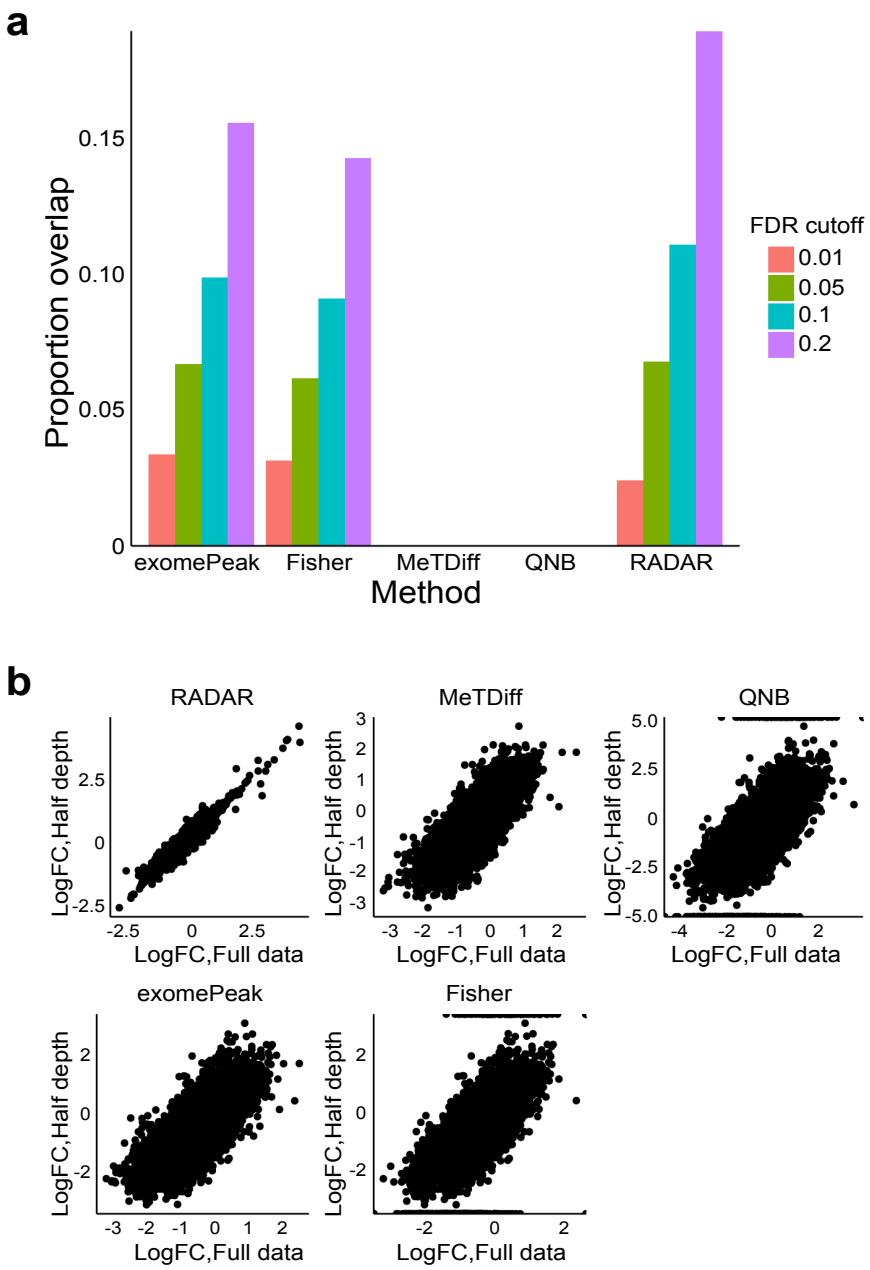
Supplementary Figure 3.6: on simulated data. Using data simulated by random effect model of effect size = 0.75, we evaluated the precision of effect size and P value estimates. (a) shows the distribution of effect size estimates in true differential sites and (b) shows the distribution of effect size estimates in Null sites where the true effect size is labeled by dashed line. (c) shows P values distribution for Null sites where P values are expected to be uniformly distributed. (d) shows P values distribution for true differential sites where P values are expected to be distributed near zero. In all panels, "simple case" refers to simulated data without covariates while "difficult case" refers to simulated data with a covariate.



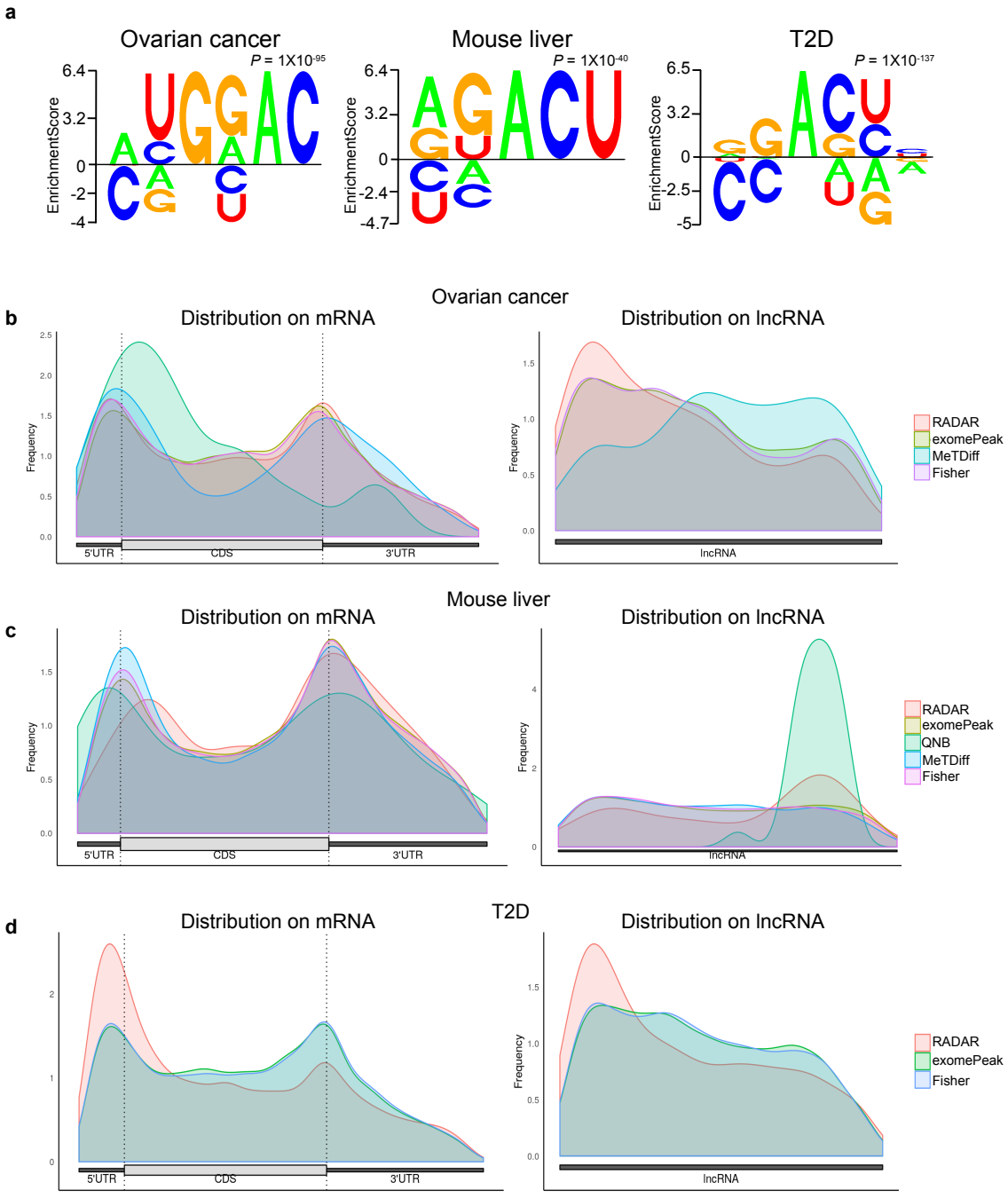
Supplementary Figure 3.7: Coverage plot of individual samples for example DM sites and bogus sites in the T2D dataset. We visualize raw data by showing coverage plot for three examples m^6A sites. **a** shows a putative DM site that was only detected by RADAR but missed by other methods. **b** shows a bogus DM site where difference between two groups was mainly driven by two strongly hypomethylated samples in the control group instead of consistent change among replicates. **c** shows another bogus DM site that were mainly driven by an outlier hypermethylated sample in the T2D samples.



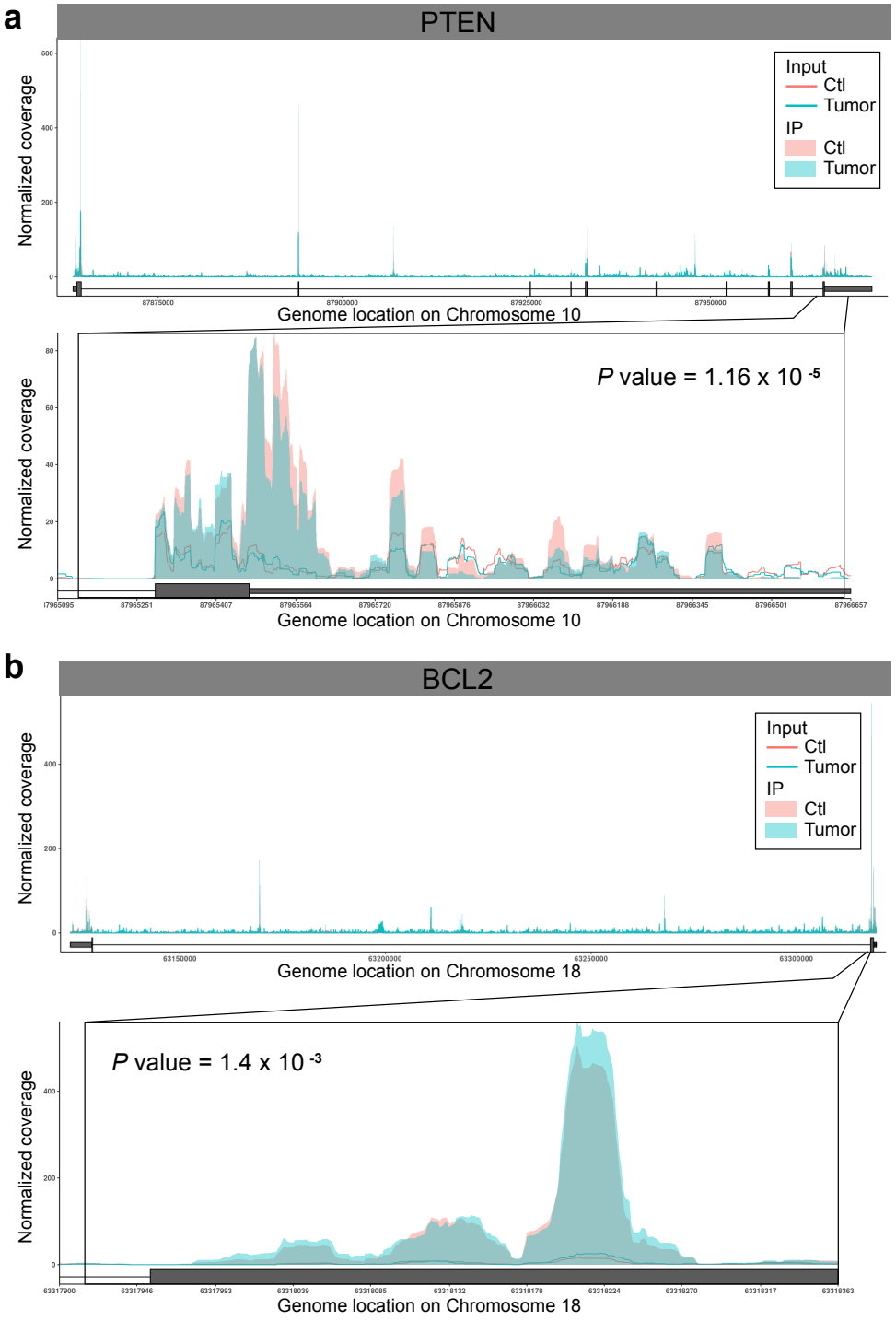
Supplementary Figure 3.8: Compare the methods to adjust for gene expression level. Local peak/bin read counts or gene level read counts of Input library can be used to account for pre-IP gene expression level variation. We compared the performance of two strategies to adjust for gene expression variation and showed the histogram of P values from original tests and permutation tests.



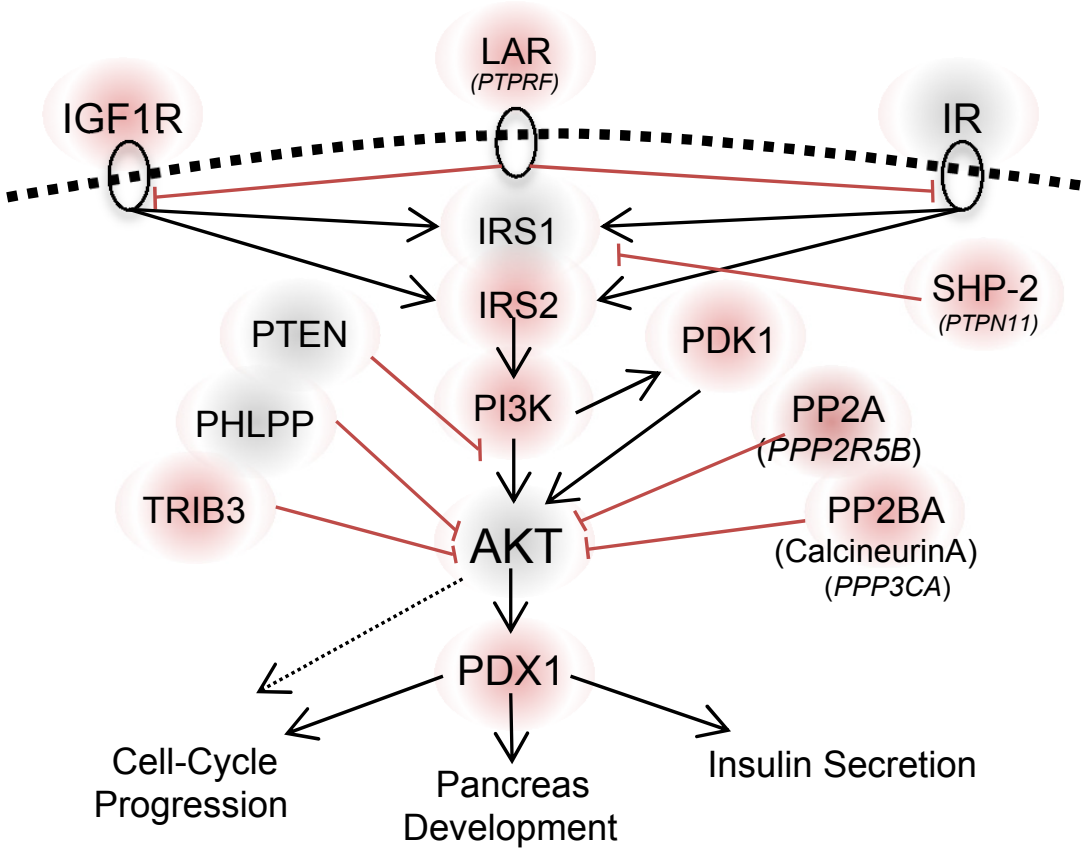
Supplementary Figure 3.9: Compare results obtained from shallower sequence depth with that from original depth. We sub-sampled sequence reads from Input libraries of the T2D dataset to obtain a dataset of shallower sequence depth (half of the original data). We applied the benchmarked methods to the sub-sampled data and compared the result with the result obtained from the original data. We show the proportion of sites positively identified in both datasets in **a** and plotted the estimated logFC against each other in **b**.



Supplementary Figure 3.10: Motif analysis and topological distribution of putative DM sites. We performed de-novo motif search analysis using Homer2 on the putative DM sites detected by RADAR on ovarian cancer, mouse liver and T2D datasets. **a** shows RADAR-detected DM sites were enriched for known m^6A consensus motif—RRACU. **b** shows metagene plots of putative DM sites detected by different methods (method that detected too few DM sites in given dataset were not shown).



Supplementary Figure 3.11: Coverage plot to visualize differential m^{6}A peaks in ovarian cancer. Average coverage of each group is plotted for **a** PTEN and **b** BCL2. The coverages of both Input and IP are normalized by the expression level of target gene so that the coverages of IP samples are directly comparable.



Supplementary Figure 3.12: Representation of Insulin/IGF1-AKT-PDX1 pathway. The diagram shows the Insulin/IGF1-AKT-PDX1 signaling pathway based on KEEG and Wikipathway annotations and depicts several m^6A hypomethylated genes (red shade) and unchanged genes (grey shade) in T2D as compared to Controls.

CHAPTER 4

CONCLUSION

4.1 Summary and significance

In **Chapter 2**, we report for the first time a systematic genetic analysis of the most abundant mRNA modification—N^6-methyladenosine. Our analysis reveals new insights into m^6A regulation, highlighting the importance of both RNA-features (e.g. RBPs, secondary structure) and transcriptional regulation (e.g. TF binding). We find that the functional effects of m^6A on downstream processes, in particular translation, can be highly heterogeneous and depend on binding of specific RBPs. Our integrated analysis of m^6A QTLs with GWAS supports the role of m^6A as an important link from genetic to phenotypic variations.

Using an analysis that correlates SNP effects on RBP motifs and m^6A levels, we identified specific RBPs such as SRSF1, that may be m^6A regulators. This analysis, however, has some limitations. It may not be able to distinguish RBPs from the same families that share similar motifs. Due to small sample size of our study, it may also be underpowered to detect many more RBPs regulating m^6A. The enrichment of m^6A QTLs in transcription-related features supports an emerging connection between mRNA modification and transcriptional control [119, 111, 118]. As a support of the "recruitment model" (**Figure 2.9b**), TFs binding sites are enriched in m^6A QTLs and several TFs interact with m^6A methyltransferase complex in LCLs. Interestingly, we found that RBBP5 and BACH1, which show robust and strong interaction with m^6A writers in LCLs, show only weak interaction with m^6A writers in HepG2 and A549 cells (data not shown). Given additional TF-methyltransferase interactions reported previously in stem cell [119, 111] and AML [85], we think TF-methyltransferase interactions may broadly exist and participate in cell-type specific m^6A regulation.

Previous studies found that m^6A promotes translation efficiency and mRNA decay via interactions with reader proteins [1]. Our results add additional insight into this model, suggesting that m^6A effects on downstream processes, e.g. translation, are much more

heterogeneous across transcripts than previously appreciated. We identified RBPs that may influence the effects of m^6A, including some with reported functions in RNA processing (**Figure 2.11c, Figure 2.12a**), including YBX3 [149], and HNRNPA1 [150]. The RBPs uncovered here provide a resource for future studies.

We hypothesize two possible mechanisms that explain context-dependent m^6A effects. First, there may be more m^6A reader proteins, with potentially different effects, than are currently known; some could be readers that respond to m^6A through structure-switch mechanism [22].Alternatively, the functions of RNA regulators may depend on m^6A, even if they do not directly bind and recognize m^6A nor respond through structure switch (and hence not readers). These proteins may bind the motif that harbors m^6A or a motif nearby m^6A sites, and compete with bona fide reader proteins on the modified transcripts. Future studies are

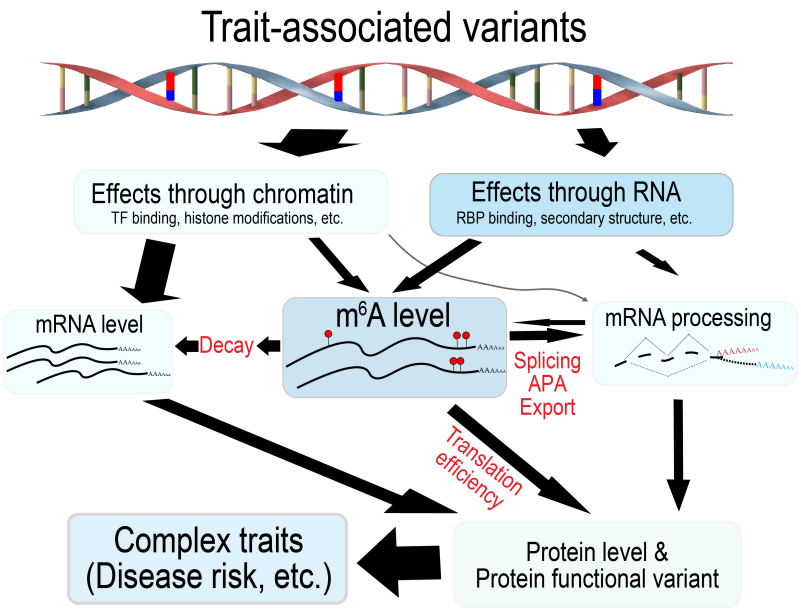


Figure 4.1: m^6A modification mediates the impact of genetic variation on human complex traits. Genetic variation exerts its impact on complex traits through varies mechanisms. As one of these mechanisms, we propose that variation of m^6A modification may lead to variation of mRNA processing, including mRNA decay, splicing, APA, export and translation efficiency. These variations in turn may change protein levels and functions, and lead to phenotypic variations.

needed to assess these RBPs and their interactions as well as competition in RNA binding.

Our integrated analysis of m^6A QTLs and GWAS highlights the importance of m^6A to the etiology of complex traits and adds to the growing evidence that post-transcriptional regulation (PTR) plays a key role in common diseases. Genetic variants affecting RNA-processing are almost as common as, and are largely independent from, those affecting transcription [151]. These variants have been implicated in a number of diseases including Cystic Fibrosis, Type 2 Diabetes, Crohn's disease and lung cancer [151]. Identifying variants with PTR effects, however, is more challenging than transcriptional effects. Mapping m^6A QTLs may be an effective strategy to address this challenge, given the central role of m^6A modification in almost every step of RNA processing (**Figure 4.1**).

In **Chapter 3**, we developed and described an R package RADAR for differential methylation analysis using a random effect model. Using simulation and real m6A-seq datasets, we demonstrated that RADAR can achieve higher sensitivity with lower FDR than existing methods in the DM analysis. Taking advantage of newly developed SELECT method for experimental validation, we verified that RADAR analysis can uncover true differentially methylated sites. RADAR is a general framework that can be applicable to comparative profiling by MeRIP-seq of various types of RNA modifications including but not limited to N^6-methyladenosine, N^1-methyladenosine and 5-methylcytosine. It also offers great flexibility to adopt to a wide range of mean-variance relationships in the data and accommodate different study designs. We believe RADAR will greatly advance our knowledge of the functions of post-transcriptional modifications.

4.2 Limitations and future directions

We mapped the m^6A QTLs using the most-intensively studied cohort of Yoruba LCL samples to facilitate integrated analysis of multiple molecular QTLs. However, we mapped only 822 ePeaks with current cohort of samples, which limited the statistical power to discover new factors that may regulate m^6A. For example, in the analysis to search for RBPs causally

linked to m^6A, we had very limited number of data points (fine-mapped m^6A QTNs) to test correlation between SNP effects on RBP motif and m^6A levels. Therefore, a larger scale study with expanded sample size is needed to map a more comprehensive set of m^6A QTLs and reveal new mechanisms of m^6A regulation. In this study, we analyzed the contribution of RNA features and TFs to the m^6A variation, which advanced our understanding of factors contributing to m^6A variation. However, there are 63% unexplained causal variants (**Supplementary Figure 2.3c**), possibly due to incomplete annotation of functional genomics regions. With the rapidly growing data, the analytical framework we built in this work will continue to provide insight into the mechanisms regulating m^6A variation.

One potential problem of our m^6A QTLs and GWAS joint analysis is the mismatch between population ancestries of QTL (African) and GWAS (mostly European) data. The impact of this mismatch, however, is likely limited. Studies have suggested that associations with complex traits, especially causal variants, are broadly shared across populations [152]. A systematic study with multiple complex traits estimated that more than 80% of causal variants are shared between Europeans and Asians [153]. In another study, TWAS on asthma using eQTL models trained on data from Europeans and Africans gave broadly similar results [154]. Given these findings, we think most m^6A QTNs (causal variants) in Yoruba LCLs are likely shared in Europeans. Therefore population mismatch likely has a small impact in our S-LDSC analysis, which used PIPs as SNP annotations; and in our TWAS, where results are often driven by single shared variant between molecular QTL and GWAS [155]. Finally, we note that population mismatch will generally reduce the signal, i.e. sharing of QTL and GWAS effects, leading to underestimation of enrichment in S-LDSC and false negatives in TWAS, but not false positive findings.

In terms of epitranscriptome analysis of m^6A in general, sample size is an important parameter of the study design that directly affects the power and reproducibility of an inferential test [156]. DM analyses based on less than 3 biological replicates are still common practice in the RNA epigenetics field and have been shown to exhibit poor reproducibility

[157]. Using the simulation model we described in **Chapter 3**, we explored the influence

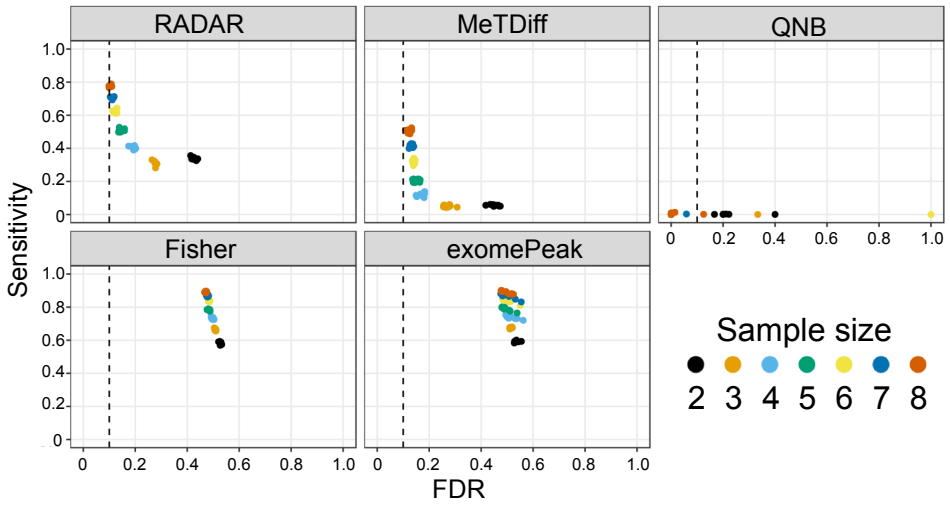


Figure 4.2: The influence of sample size on the statistical power of differential methylation analysis. Sensitivity vs. empirical FDR for each method on simulated data with different number of replicates (2 to 8) at 10% FDR. Each data point represents the results on one of ten simulated copies. Sample sizes are labeled by colors.

of sample size on the power and reproducibility of DM detection. we ran tests on 10 copies of simulated data with effect size equal to 0.75 (roughly two-fold enrichment difference) from two replicates (commonly used in the literature) to eight replicates (up-to-date highest in the literature). We show that at an FDR cutoff of 10% to select DM loci, empirical FDR increases rapidly as the sample size gets smaller and less than 5 (**Figure 4.2**). When the sample size is greater than 6, improvement of empirical FDR is slow while sensitivity climbs rapidly. Our results show the number of replicates greatly influences sensitivity and reproducibility of DM detection as each additional replicate can bring significant gain of area under ROC curve (**Figure 4.3**). Our simulation thus support the observation that very few m^6A peak changes reported in literature were found reproducible [157] given the inadequate sample size. To ensure adequate power and reliable DM analysis, we strongly suggest using no less than 5 biological replicates when surveying the alterations in the epitranscriptome of human samples with commonly sequenced library sizes ($\geqq$ 20 million mappable reads).

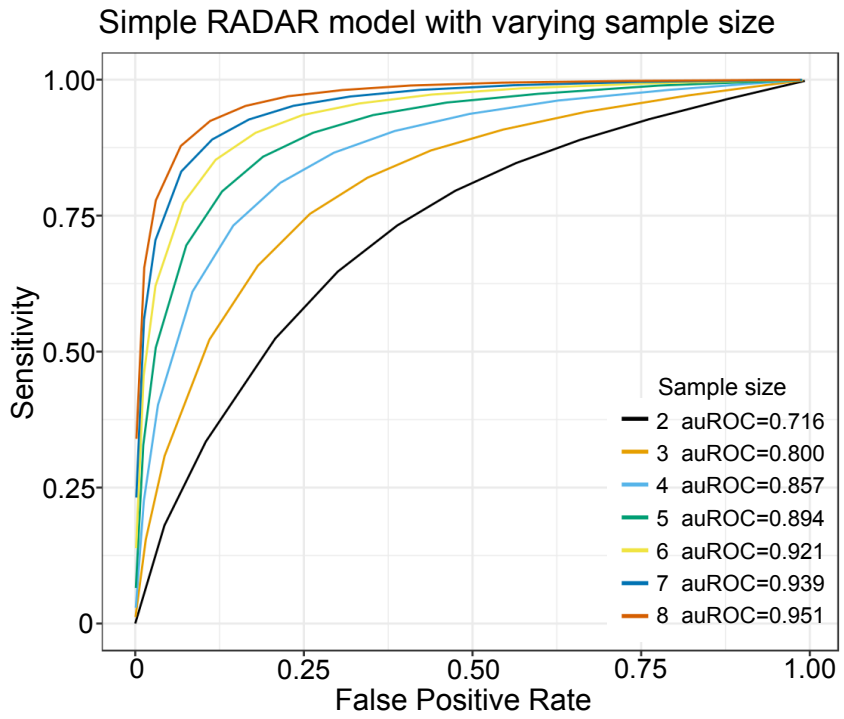


Figure 4.3: Analysis of statistical power and the number of replicates. We plotted the sensitivity against the empirical FDR by varying the FDR cutoff for selecting predicted differential sites. The larger the area under the curve, the larger the power of the test.

The works presented in this dissertation implicated m^6A variation as an important mechanism contributing to human diseases. With the analytic framework and software we developed, we believe epitranscriptomic analysis (with adequate sample size) will continue to uncover new knowledge of how m^6A affects gene regulation and complex phenotypes. Moving forward, we think there are three main challenges and opportunities to leverage m^6A QTLs to study disease genetics. First, more work needs to be done to characterize the possible mechanisms of how m^6A QTLs influence phenotypes. Second, eQTLs or sQTLs are often cell type- and condition-specific [158, 159]. For m^6A, recent studies suggest that its effects on decay or translation are probably strongest in cells undergoing differentiation [25, 24] or stimulation [28, 31]. Thus, a major future direction is to map m^6A QTLs under various disease-related cellular and physiological contexts. Third, recent work has shown that chromosome-associated regulatory RNA (carRNA) m^6A methylation regulates transcription

[19]. QTL studies of carRNA m^6A may reveal new insights into genetics of mammalian transcriptional regulation and human complex traits.

Appendices

APPENDIX A

ADDITIONAL m^{6}A-RELATED WORK DURING MY THESIS RESEARCH

In this appendix, I list and present the abstracts of three m^{6}A-related research projects I participated and co-first authored during my thesis research. It's worth noting that the challenges of lacking a statistical method compatible with complex study design in the epi-transcriptome analysis of type 2 diabetic islets (De Jesus et al. 2019) listed below were the original motivation for me to work with Prof. Mengjie Chen to develop the R package RADAR as introduced in **Chapter 3**.

A Dynamic mRNA N^6-Methyladenosine Methylome Regulates Acquired Resistance to Tyrosine Kinase Inhibitors

Fei Yan#, Aref Al-Kali#, Zijie Zhang#, Jun Liu, Jiuxia Pang, Na Zhao, Chuan He*, Mark R. Litzow*, Shujun Liu*.

Cell Research **28**, 1062–1076 (2018)

Abstract

N^6-methyladenosine (m^{6}A) on mRNAs is critical for various biological processes, yet whether m^{6}A regulates drug resistance remains unknown. Here we show that developing resistant phenotypes during tyrosine kinase inhibitor (TKI) therapy depends on m^{6}A reduction resulting from FTO overexpression in leukemia cells. This deregulated FTO-m^{6}A axis pre-exists in naïve cell populations that are genetically homogeneous and is inducible/reversible in response to TKI treatment. Cells with mRNA m^{6}A hypomethylation and FTO upregulation demonstrate more TKI tolerance and higher growth rates in mice. Either genetic or pharmacological restoration of m^{6}A methylation through FTO deactivation renders resistant

cells sensitive to TKIs. Mechanistically, the FTO-dependent m^{6}A demethylation enhances mRNA stability of proliferation/survival transcripts bearing m^{6}A and subsequently leads to increased protein synthesis. Our findings identify a novel function for the m^{6}A methylation in regulating cell fate decision and demonstrate that dynamic m^{6}A methylome is an additional epigenetic driver of reversible TKI-tolerance state, providing a mechanistic paradigm for drug resistance in cancer.

m^{6}A mRNA Methylation Regulates Human β-cell Biology in Physiological States and in Type 2 Diabetes

Dario F. De Jesus[#], Zijie Zhang[#], Sevim Kahraman[#], Natalie K. Brown, Mengjie Chen, Jiang Hu, Manoj K. Gupta, Chuan He[*] & Rohit N. Kulkarni[*]

Nature Metabolism **1**, 765–774 (2019)

Abstract

The regulation of islet cell biology is critical for glucose homeostasis. N^6-methyladenosine (m^{6}A) is the most abundant internal messenger RNA (mRNA) modification in mammals. Here, we report that the m^{6}A landscape segregates human type 2 diabetes (T2D) islets from controls significantly better than the transcriptome and that m^{6}A is vital for β-cell biology. m^{6}A sequencing in human T2D islets reveals several hypomethylated transcripts that are involved in cell-cycle progression, insulin secretion, and the insulin/IGF1–AKT–PDX1 pathway. Depletion of m^{6}A levels in EndoC-βH1 cells induces cell-cycle arrest and impairs insulin secretion by decreasing AKT phosphorylation and PDX1 protein levels. β-cell-specific *Mettl14* knockout mice, which display reduced m^{6}A levels, mimic the islet phenotype in human T2D with early diabetes onset and mortality owing to decreased β-cell proliferation and insulin degranulation. Our data underscore the significance of RNA methylation in regulating human β-cell biology, and provide a rationale for potential therapeutic targeting of m^{6}A

modulators to preserve β-cell survival and function in diabetes.

N^6-Methyladenosine Modification Enables Viral RNA to Escape Recognition by RNA Sensor RIG-I

Mijia Lu#, Zijie Zhang#, Miaoge Xue, Boxuan Simen Zhao, Olivia Harder, Anzhong Li, Xueya Liang, Thomas Z. Gao, Yunsheng Xu, Jiyong Zhou, Zongdi Feng, Stefan Niewiesk, Mark E. Peeples, Chuan He & Jianrong Li

Nature Microbiology **5** ,584–598 (2020)

Abstract

Internal N^6-methyladenosine (m^6A) modification is one of the most common and abundant modifications of RNA. However, the biological roles of viral RNA m^6A remain elusive. Here, using human metapneumovirus (HMPV) as a model, we demonstrate that m^6A serves as a molecular marker for innate immune discrimination of self from non-self RNAs. We show that HMPV RNAs are m^6A methylated and that viral m^6A methylation promotes HMPV replication and gene expression. Inactivating m^6A addition sites with synonymous mutations or demethylase resulted in m^6A-deficient recombinant HMPVs and virion RNAs that induced increased expression of type I interferon, which was dependent on the cytoplasmic RNA sensor RIG-I, and not on melanoma differentiation-associated protein 5 (MDA5). Mechanistically, m^6A-deficient virion RNA induces higher expression of RIG-I, binds more efficiently to RIG-I and facilitates the conformational change of RIG-I, leading to enhanced interferon expression. Furthermore, m^6A-deficient recombinant HMPVs triggered increased interferon in vivo and were attenuated in cotton rats but retained high immunogenicity. Collectively, our results highlight that (1) viruses acquire m^6A in their RNA as a means of mimicking cellular RNA to avoid detection by innate immunity and (2) viral RNA m^6A can serve as a target to attenuate HMPV for vaccine purposes.

REFERENCES

[1] I. A. Roundtree, M. E. Evans, T. Pan, and C. He. Dynamic rna modifications in gene expression regulation. *Cell*, 169(7):1187–1200, 2017.

[2] Y. Fu, D. Dominissini, G. Rechavi, and C. He. Gene expression regulation mediated through reversible m^6a rna methylation. *Nat Rev Genet*, 15(5):293–306, 2014.

[3] Jianzhao Liu, Yanan Yue, Dali Han, Xiao Wang, Ye Fu, Liang Zhang, Guifang Jia, Miao Yu, Zhike Lu, Xin Deng, Qing Dai, Weizhong Chen, and Chuan He. A mettl3–mettl14 complex mediates mammalian nuclear rna N^6-adenosine methylation. *Nature Chemical Biology*, 10:93, 2013.

[4] Ping Wang, Katelyn A Doxtader, and Yunsun Nam. Structural basis for cooperative function of mettl3 and mettl14 methyltransferases. *Molecular Cell*, 63(2):306–317, 2016.

[5] Xiao-Li Ping, Bao-Fa Sun, Lu Wang, Wen Xiao, Xin Yang, Wen-Jia Wang, Samir Adhikari, Yue Shi, Ying Lv, Yu-Sheng Chen, Xu Zhao, Ang Li, Ying Yang, Ujwal Dahal, Xiao-Min Lou, Xi Liu, Jun Huang, Wei-Ping Yuan, Xiao-Fan Zhu, Tao Cheng, Yong-Liang Zhao, Xinquan Wang, Jannie M. Rendtlew Danielsen, Feng Liu, and Yun-Gui Yang. Mammalian wtap is a regulatory subunit of the rna N^6-methyladenosine methyltransferase. *Cell Research*, 24:177, 2014.

[6] Yanan Yue, Jun Liu, Xiaolong Cui, Jie Cao, Guanzheng Luo, Zezhou Zhang, Tao Cheng, Minsong Gao, Xiao Shu, Honghui Ma, Fengqin Wang, Xinxia Wang, Bin Shen, Yizhen Wang, Xinhua Feng, Chuan He, and Jianzhao Liu. Virma mediates preferential m^6A mrna methylation in 3'utr and near stop codon and associates with alternative polyadenylation. *Cell Discovery*, 4(1):10, 2018.

[7] Jing Wen, Ruitu Lv, Honghui Ma, Hongjie Shen, Chenxi He, Jiahua Wang, Fangfang Jiao, Hang Liu, Pengyuan Yang, Li Tan, Fei Lan, Yujiang Geno Shi, Chuan He, Yang Shi, and Jianbo Diao. Zc3h13 regulates nuclear rna m^6A methylation and mouse embryonic stem cell self-renewal. *Molecular Cell*, 69(6):1028–1038.e6, 2018.

[8] D. P. Patil, C. K. Chen, B. F. Pickering, A. Chow, C. Jackson, M. Guttman, and S. R. Jaffrey. m^6A rna methylation promotes xist-mediated transcriptional repression. *Nature*, 537(7620):369–373, 2016.

[9] Tomohiko Aoyama, Seisuke Yamashita, and Kozo Tomita. Mechanistic insights into m^6A modification of U6 snRNA by human METTL16. *Nucleic Acids Research*, 48(9):5157–5168, 04 2020.

[10] Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min Huang, Charles J. Li, Cathrine B. Vågbø, Yue Shi, Wen-Ling Wang, Shu-Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan, Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne Klungland, Yun-Gui Yang, and Chuan He. Alkbh5

is a mammalian rna demethylase that impacts rna metabolism and mouse fertility. *Molecular Cell*, 49(1):18–29, 2013.

[11] Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang, Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, and Chuan He. N^6-methyladenosine in nuclear rna is a major substrate of the obesity-associated fto. *Nature chemical biology*, 7(12):885–887, 2011.

[12] Jiangbo Wei, Fange Liu, Zhike Lu, Qili Fei, Yuxi Ai, P. Cody He, Hailing Shi, Xiaolong Cui, Rui Su, Arne Klungland, Guifang Jia, Jianjun Chen, and Chuan He. Differential m^6A, m^6A_m, and m^1a demethylation mediated by fto in the cell nucleus and cytoplasm. *Molecular Cell*, 71(6):973–985.e5, 2018.

[13] Xiao Wang, Boxuan Simen Zhao, Ian A. Roundtree, Zhike Lu, Dali Han, Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He. N^6-methyladenosine modulates messenger rna translation efficiency. *Cell*, 161(6):1388–1399, 2015.

[14] Xiao Wang, Zhike Lu, Adrian Gomez, Gary C. Hon, Yanan Yue, Dali Han, Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, Bing Ren, Tao Pan, and Chuan He. N^6-methyladenosine-dependent regulation of messenger rna stability. *Nature*, 505:117, 2013.

[15] Hao Du, Ya Zhao, Jinqiu He, Yao Zhang, Hairui Xi, Mofang Liu, Jinbiao Ma, and Ligang Wu. Ythdf2 destabilizes m^6A-containing rna through direct recruitment of the ccr4–not deadenylase complex. *Nature Communications*, 7(1):12626, 2016.

[16] Hailing Shi, Xiao Wang, Zhike Lu, Boxuan S. Zhao, Honghui Ma, Phillip J. Hsu, Chang Liu, and Chuan He. Ythdf3 facilitates translation and decay of N^6-methyladenosine-modified rna. *Cell Research*, 27:315, 2017.

[17] Wen Xiao, Samir Adhikari, Ujwal Dahal, Yu-Sheng Chen, Ya-Juan Hao, Bao-Fa Sun, Hui-Ying Sun, Ang Li, Xiao-Li Ping, Wei-Yi Lai, Xing Wang, Hai-Li Ma, Chun-Min Huang, Ying Yang, Niu Huang, Gui-Bin Jiang, Hai-Lin Wang, Qi Zhou, Xiu-Jie Wang, Yong-Liang Zhao, and Yun-Gui Yang. Nuclear m^6A reader ythdc1 regulates mrna splicing. *Molecular Cell*, 61(4):507–519, 2016.

[18] Ian A. Roundtree, Guan-Zheng Luo, Zijie Zhang, Xiao Wang, Tao Zhou, Yiquang Cui, Jiahao Sha, Xingxu Huang, Laura Guerrero, Phil Xie, Emily He, Bin Shen, and Chuan He. Ythdc1 mediates nuclear export of N^6-methyladenosine methylated mrnas. *eLife*, 6:e31311, 2017.

[19] Jun Liu, Xiaoyang Dou, Chuanyuan Chen, Chuan Chen, Chang Liu, Meng Michelle Xu, Siqi Zhao, Bin Shen, Yawei Gao, Dali Han, and Chuan He. N^6-methyladenosine of chromosome-associated regulatory rna regulates chromatin state and transcription. *Science*, 367(6477):580–586, 2020.

[20] H. Huang, H. Weng, W. Sun, X. Qin, H. Shi, H. Wu, B. S. Zhao, A. Mesquita, C. Liu, C. L. Yuan, Y. C. Hu, S. Huttelmaier, J. R. Skibbe, R. Su, X. Deng, L. Dong,

M. Sun, C. Li, S. Nachtergaele, Y. Wang, C. Hu, K. Ferchen, K. D. Greis, X. Jiang, M. Wei, L. Qu, J. L. Guan, C. He, J. Yang, and J. Chen. Recognition of rna N^6-methyladenosine by igf2bp proteins enhances mrna stability and translation. *Nat Cell Biol*, 20(3):285–295, 2018.

[21] Raghu R. Edupuganti, Simon Geiger, Rik G. H. Lindeboom, Hailing Shi, Phillip J. Hsu, Zhike Lu, Shuang-Yin Wang, Marijke P. A. Baltissen, Pascal W. T. C. Jansen, Martin Rossa, Markus Müller, Hendrik G. Stunnenberg, Chuan He, Thomas Carell, and Michiel Vermeulen. N^6-methyladenosine (m^6A) recruits and repels proteins to regulate mrna homeostasis. *Nature Structural & Molecular Biology*, 24:870, 2017.

[22] Nian Liu, Qing Dai, Guanqun Zheng, Chuan He, Marc Parisien, and Tao Pan. N^6-methyladenosine-dependent rna structural switches regulate rna–protein interactions. *Nature*, 518:560, 2015.

[23] Claudio R Alarcón, Hani Goodarzi, Hyeseung Lee, Xuhang Liu, Saeed Tavazoie, and Sohail F Tavazoie. Hnrnpa2b1 is a mediator of m^6A-dependent nuclear rna processing events. *Cell*, 162(6):1299–1308, 2015.

[24] Michaela Frye, Bryan T. Harada, Mikaela Behm, and Chuan He. Rna modifications modulate gene expression during development. *Science*, 361(6409):1346, 2018.

[25] Boxuan Simen Zhao, Xiao Wang, Alana V. Beadell, Zhike Lu, Hailing Shi, Adam Kuuspalu, Robert K. Ho, and Chuan He. m^6a-dependent maternal mrna clearance facilitates zebrafish maternal-to-zygotic transition. *Nature*, 542(7642):475–478, 2017.

[26] Chuanzhao Zhang, Debangshu Samanta, Haiquan Lu, John W. Bullen, Huimin Zhang, Ivan Chen, Xiaoshun He, and Gregg L. Semenza. Hypoxia induces the breast cancer stem cell phenotype by hif-dependent and alkbh5-mediated m^6A-demethylation of nanog mrna. *Proceedings of the National Academy of Sciences*, 113(14):E2047–E2056, 2016.

[27] Yulin Shi, Songqing Fan, Mengge Wu, Zhixiang Zuo, Xingyang Li, Liping Jiang, Qiushuo Shen, Peifang Xu, Lin Zeng, Yongchun Zhou, Yunchao Huang, Zuozhang Yang, Jumin Zhou, Jing Gao, Hu Zhou, Shuhua Xu, Hongbin Ji, Peng Shi, Dong-Dong Wu, Cuiping Yang, and Yongbin Chen. Ythdf1 links hypoxia adaptation and non-small cell lung cancer progression. *Nature Communications*, 10(1):4892, 2019.

[28] Jun Zhou, Ji Wan, Xiangwei Gao, Xingqian Zhang, Samie R. Jaffrey, and Shu-Bing Qian. Dynamic m^6A mrna methylation directs translational control of heat shock response. *Nature*, 526:591, 2015.

[29] Dali Han, Jun Liu, Chuanyuan Chen, Lihui Dong, Yi Liu, Renbao Chang, Xiaona Huang, Yuanyuan Liu, Jianying Wang, Urszula Dougherty, Marc B. Bissonnette, Bin Shen, Ralph R. Weichselbaum, Meng Michelle Xu, and Chuan He. Anti-tumour immunity controlled through mrna m^6A methylation and ythdf1 in dendritic cells. *Nature*, 566(7743):270–274, 2019.

[30] Roni Winkler, Ella Gillis, Lior Lasman, Modi Safra, Shay Geula, Clara Soyris, Aharon Nachshon, Julie Tai-Schmiedel, Nehemya Friedman, Vu Thuy Khanh Le-Trilling, Mirko Trilling, Michal Mandelboim, Jacob H. Hanna, Schraga Schwartz, and Noam Stern-Ginossar. m^6A modification controls the innate immune response to infection by targeting type i interferons. *Nature Immunology*, 20(2):173–182, 2019.

[31] Hailing Shi, Xuliang Zhang, Yi-Lan Weng, Zongyang Lu, Yajing Liu, Zhike Lu, Jianan Li, Piliang Hao, Yu Zhang, Feng Zhang, You Wu, Jary Y. Delgado, Yijing Su, Meera J. Patel, Xiaohua Cao, Bin Shen, Xingxu Huang, Guo-li Ming, Xiaoxi Zhuang, Hongjun Song, Chuan He, and Tao Zhou. m^6A facilitates hippocampus-dependent learning and memory through ythdf1. *Nature*, 563(7730):249–253, 2018.

[32] Ki-Jun Yoon, Francisca Rojas Ringeling, Caroline Vissers, Fadi Jacob, Michael Pokrass, Dennisse Jimenez-Cyrus, Yijing Su, Nam-Shik Kim, Yunhua Zhu, Lily Zheng, Sunghan Kim, Xinyuan Wang, Louis C. Doré, Peng Jin, Sergi Regot, Xiaoxi Zhuang, Stefan Canzar, Chuan He, Guo-li Ming, and Hongjun Song. Temporal control of mammalian cortical neurogenesis by m^6A methylation. *Cell*, 171(4):877–889.e17, 2017.

[33] Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and Gideon Rechavi. Topology of the human and mouse m^6A rna methylomes revealed by m^6A-seq. *Nature*, 485:201, 2012.

[34] Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christopher E. Mason, and Samie R. Jaffrey. Comprehensive analysis of mrna methylation reveals enrichment in 3' utrs and near stop codons. *Cell*, 149(7):1635–1646, 2012.

[35] A. Takata, N. Matsumoto, and T. Kato. Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun*, 8:14519, 2017.

[36] Zijie Zhang, Qi Zhan, Mark Eckert, Allen Zhu, Agnieszka Chryplewicz, Dario F. De Jesus, Decheng Ren, Rohit N. Kulkarni, Ernst Lengyel, Chuan He, and Mengjie Chen. Radar: differential analysis of merip-seq data with a random effect model. *Genome Biology*, 20(1):294, 2019.

[37] Dario F. De Jesus, Zijie Zhang, Sevim Kahraman, Natalie K. Brown, Mengjie Chen, Jiang Hu, Manoj K. Gupta, Chuan He, and Rohit N. Kulkarni. m^6A mrna methylation regulates human β-cell biology in physiological states and in type 2 diabetes. *Nature Metabolism*, 1(8):765–774, 2019.

[38] Mareen Engel, Carola Eggert, Paul M. Kaplick, Matthias Eder, Simone Röh, Lisa Tietze, Christian Namendorf, Janine Arloth, Peter Weber, Monika Rex-Haffner, Shay Geula, Mira Jakovcevski, Jacob H. Hanna, Dena Leshkowitz, Manfred Uhr, Carsten T. Wotjak, Mathias V. Schmidt, Jan M. Deussing, Elisabeth B. Binder, and Alon Chen. The role of m^6a/m-rna methylation in stress response regulation. *Neuron*, 99(2):389–403.e9, 2018.

[39] Kai Chen, Zhike Lu, Xiao Wang, Ye Fu, Guan-Zheng Luo, Nian Liu, Dali Han, Dan Dominissini, Qing Dai, Tao Pan, and Chuan He. High-resolution N^6-methyladenosine (m^6a) map using photo-crosslinking-assisted m^6a sequencing. *Angewandte Chemie (International ed. in English)*, 54(5):1587–1590, 2015.

[40] Bastian Linder, Anya V. Grozhik, Anthony O. Olarerin-George, Cem Meydan, Christopher E. Mason, and Samie R. Jaffrey. Single-nucleotide-resolution mapping of m^6A and m^6Am throughout the transcriptome. *Nature Methods*, 12(8):767–772, 2015.

[41] J. Meng, Z. Lu, H. Liu, L. Zhang, S. Zhang, Y. Chen, M. K. Rao, and Y. Huang. A protocol for rna methylation differential analysis with merip-seq data and exomepeak r/bioconductor package. *Methods*, 69(3):274–81, 2014.

[42] X. Cui, L. Zhang, J. Meng, M. K. Rao, Y. Chen, and Y. Huang. Metdiff: A novel differential rna methylation analysis for merip-seq data. *IEEE/ACM Trans Comput Biol Bioinform*, 15(2):526–534, 2018.

[43] L. Liu, S. W. Zhang, F. Gao, Y. Zhang, Y. Huang, R. Chen, and J. Meng. Drme: Count-based differential rna methylation analysis at small sample size scenario. *Anal Biochem*, 499:15–23, 2016.

[44] L. Liu, S. W. Zhang, Y. Huang, and J. Meng. Qnb: differential rna methylation analysis for count-based small-sample sequencing data with a quad-negative binomial model. *BMC Bioinformatics*, 18(1):387, 2017.

[45] Athma A. Pai, Jonathan K. Pritchard, and Yoav Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLOS Genetics*, 11(1):e1004857, 2015.

[46] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[47] Yang I. Li, Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

[48] Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation from rna to protein. *Science*, 347(6222):664–667, 2015.

[49] Jacob F. Degner, Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E. Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482:390, 2012.

[50] Farhad Hormozdiari, Steven Gazal, Bryce van de Geijn, Hilary K. Finucane, Chelsea J. T. Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke O'Connor, Alexander Gusev, Eleazar Eskin, and Alkes L. Price. Leveraging molecular quantitative

trait loci to understand the genetic architecture of diseases and complex traits. *Nature Genetics*, 50(7):1041–1047, 2018.

[51] Junho Choe, Shuibin Lin, Wencai Zhang, Qi Liu, Longfei Wang, Julia Ramirez-Moya, Peng Du, Wantae Kim, Shaojun Tang, Piotr Sliz, Pilar Santisteban, Rani E. George, William G. Richards, Kwok-Kin Wong, Nicolas Locker, Frank J. Slack, and Richard I. Gregory. mrna circularization by mettl3–eif3h enhances translation and promotes oncogenesis. *Nature*, 561(7724):556–560, 2018.

[52] Bogdan Pasaniuc and Alkes L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18:117, 2016.

[53] Consortium The Genomes Project, Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, et al. A global reference for human genetic variation. *Nature*, 526:68, 2015.

[54] Nicholas E. Banovich, Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F. Degner, John D. Blischak, Julien Roux, Jonathan K. Pritchard, and Yoav Gilad. Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS genetics*, 10(9):e1004663–e1004663, 2014.

[55] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464:768, 2010.

[56] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksan-

dra Pekowska, Nastaran Heidari, Ghia Euskirchen, Wolfgang Huber, Jonathan K Pritchard, Carlos D Bustamante, Lars M Steinmetz, Anshul Kundaje, and Michael Snyder. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5):1051–1065, 2015.

[57] Graham McVicker, Bryce van de Geijn, Jacob F. Degner, Carolyn E. Cain, Nicholas E. Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K. Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science (New York, N.Y.)*, 342(6159):747–749, 2013.

[58] Helena Kilpinen, Sebastian M. Waszak, Andreas R. Gschwind, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, Eugenia Migliavacca, Michaël Wiederkehr, Maria Gutierrez-Arcelus, Nikolaos I. Panousis, Alisa Yurovsky, Tuuli Lappalainen, Luciana Romano-Palumbo, Alexandra Planchon, Deborah Bielser, Julien Bryois, Ismael Padioleau, Gilles Udin, Sarah Thurnheer, David Hacker, Leighton J. Core, John T. Lis, Nouria Hernandez, Alexandre Reymond, Bart Deplancke, and Emmanouil T. Dermitzakis. Coordinated effects of sequence variation on dna binding, chromatin structure, and transcription. *Science*, 342(6159):744, 2013.

[59] Athma A. Pai, Carolyn E. Cain, Orna Mizrahi-Man, Sherryl De Leon, Noah Lewellen, Jean-Baptiste Veyrieras, Jacob F. Degner, Daniel J. Gaffney, Joseph K. Pickrell, Matthew Stephens, Jonathan K. Pritchard, and Yoav Gilad. The contribution of rna decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLOS Genetics*, 8(10):e1003000, 2012.

[60] Can Cenik, Elif Sarinay Cenik, Gun W. Byeon, Fabian Grubert, Sophie I. Candille, Damek Spacek, Bilal Alsallakh, Hagen Tilgner, Carlos L. Araya, Hua Tang, Emiliano Ricci, and Michael P. Snyder. Integrative analysis of rna, translation, and protein levels reveals distinct regulatory variation across humans. *Genome research*, 25(11):1610–1621, 2015.

[61] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, Consortium The Geuvadis, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501:506, 2013.

[62] Xiaoquan Wen, Francesca Luca, and Roger Pique-Regi. Cross-population joint analysis of eqtls: Fine mapping and functional annotation. *PLOS Genetics*, 11(4):1–29, 04 2015.

[63] Nicholas E. Banovich, Yang I. Li, Anil Raj, Michelle C. Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E. Burnett, Marsha Myrthil, Samantha M. Thomas, Courtney K. Burrows, Irene Gallego Romero, Bryan J. Pavlovic, Anshul Kundaje, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation across human ipscs and differentiated cells. *Genome Research*, 2017.

[64] B. J. Strober, R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019.

[65] Fred A. Wright, Patrick F. Sullivan, Andrew I. Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, Abdel Abdellaoui, Sandra Batista, Casey Butler, Guanhua Chen, Ting-Huei Chen, David D'Ambrosio, Paul Gallins, Min Jin Ha, Jouke Jan Hottenga, Shunping Huang, Mathijs Kattenberg, Jaspreet Kochar, Christel M. Middeldorp, Ani Qu, Andrey Shabalin, Jay Tischfield, Laura Todd, Jung-Ying Tzeng, Gerard van Grootheest, Jacqueline M. Vink, Qi Wang, Wei Wang, Weibo Wang, Gonneke Willemsen, Johannes H. Smit, Eco J. de Geus, Zhaoyu Yin, Brenda W. J. H. Penninx, and Dorret I. Boomsma. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–437, 2014.

[66] R. E. Gate, C. S. Cheng, A. P. Aiden, A. Siba, M. Tabaka, D. Lituiev, I. Machol, M. G. Gordon, M. Subramaniam, M. Shamim, K. L. Hougen, I. Wortman, S. C. Huang, N. C. Durand, T. Feng, P. L. De Jager, H. Y. Chang, E. L. Aiden, C. Benoist, M. A. Beer, C. J. Ye, and A. Regev. Genetic determinants of co-accessible chromatin regions in activated t cells across humans. *Nat Genet*, 50(8):1140–1150, 2018.

[67] The GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[68] Project eGTEx, Barbara E. Stranger, Lori E. Brigham, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Gene Kopen, William F. Leinweber, John T. Lonsdale, Alisa McDonald, Bernadette Mestichelli, Kevin Myer, Brian Roe, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Jason Bridge, Barbara A. Foster, Bryan M. Gillard, Ellen Karasik, Rachna Kumar, Mark Miklos, Michael T. Moser, Scott D. Jewell, Robert G. Montroy, Daniel C. Rohrer, Dana R. Valley, David A. Davis, Deborah C. Mash, Sarah E. Gould, Ping Guan, Susan Koester, A. Roger Little, Casey Martin, Helen M. Moore, Abhi Rao, Jeffery P. Struewing, Simona Volpi, Kasper D. Hansen, Peter F. Hickey, Lindsay F. Rizzardi, Lei Hou, Yaping Liu, Benoit Molinie, Yongjin Park, Nicola Rinaldi, Li Wang, Nicholas Van Wittenberghe, Melina Claussnitzer, Ellen T. Gelfand, Qin Li, Sandra Linder, Rui Zhang, Kevin S. Smith, Emily K. Tsang, Lin S. Chen, Kathryn Demanelis, Jennifer A. Doherty, Farzana Jasmine, Muhammad G. Kibriya, Lihua Jiang, Shin Lin, Meng Wang, Ruiqi Jian, Xiao Li, Joanne Chan, Daniel Bates, Morgan Diegel, Jessica Halow, Eric Haugen, Audra Johnson, Rajinder Kaul, Kristen Lee, Matthew T.

Maurano, Jemma Nelson, Fidencio J. Neri, Richard Sandstrom, Marian S. Fernando, Caroline Linke, Meritxell Oliva, Andrew Skol, Fan Wu, Joshua M. Akey, Andrew P. Feinberg, Jin Billy Li, Brandon L. Pierce, John A. Stamatoyannopoulos, Hua Tang, Kristin G. Ardlie, Manolis Kellis, Michael P. Snyder, and Stephen B. Montgomery. Enhancing gtex by bridging the gaps between genotype, gene expression, and disease. *Nature Genetics*, 49:1664, 2017.

[69] Michelle C. Ward and Yoav Gilad. Cracking the regulatory code. *Nature*, 550(7675):190–191, 2017.

[70] Sarah M. Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51(1):187–195, 2019.

[71] Mark N. Lee, Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, Portia I. Chipendo, F. Ann Ran, Kamil Slowikowski, Lucas D. Ward, Khadir Raddassi, Cristin McCabe, Michelle H. Lee, Irene Y. Frohlich, David A. Hafler, Manolis Kellis, Soumya Raychaudhuri, Feng Zhang, Barbara E. Stranger, Christophe O. Benoist, Philip L. De Jager, Aviv Regev, and Nir Hacohen. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175), 2014.

[72] Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genetics*, 13(3):e1006646, 2017.

[73] Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas. *PLOS Genetics*, 6(4):e1000888, 2010.

[74] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K. Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Steven McCarroll, Benjamin M. Neale, Roel A. Ophoff, Michael C. O'Donovan, Gregory E. Crawford, Daniel H. Geschwind, Nicholas Katsanis, Patrick F. Sullivan, Bogdan Pasaniuc, Alkes L. Price, and Consortium Schizophrenia Working Group of the Psychiatric Genomics. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics*, 50(4):538–548, 2018.

[75] Halit Ongen, Andrew A. Brown, Olivier Delaneau, Nikolaos I. Panousis, Alexandra C. Nica, Emmanouil T. Dermitzakis, and G. TEx Consortium. Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 49(12):1676–1683, 2017.

[76] Jason M Torres, Eric R Gamazon, Esteban J Parra, Jennifer E Below, Adan Valladares-Salgado, Niels Wacher, Miguel Cruz, Craig L Hanis, and Nancy J Cox. Cross-tissue and tissue-specific eqtls: Partitioning the heritability of a complex trait. *The American Journal of Human Genetics*, 95(5):521–534, 2014.

[77] Sarah Kim-Hellmuth, Matthias Bechheim, Benno Pütz, Pejman Mohammadi, Yohann Nédélec, Nicholas Giangreco, Jessica Becker, Vera Kaiser, Nadine Fricker, Esther Beier, Peter Boor, Stephane E. Castel, Markus M. Nöthen, Luis B. Barreiro, Joseph K. Pickrell, Bertram Müller-Myhsok, Tuuli Lappalainen, Johannes Schumacher, and Veit Hornung. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nature communications*, 8(1):266–266, 2017.

[78] Heath E. O'Brien, Eilis Hannon, Matthew J. Hill, Carolina C. Toste, Matthew J. Robertson, Joanne E. Morgan, Gemma McLaughlin, Cathryn M. Lewis, Leonard C. Schalkwyk, Lynsey S. Hall, Antonio F. Pardiñas, Michael J. Owen, Michael C. O'Donovan, Jonathan Mill, and Nicholas J. Bray. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biology*, 19(1):194, 2018.

[79] Alexandra C. Nica, Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês Barroso, and Emmanouil T. Dermitzakis. Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLOS Genetics*, 6(4):1–11, 04 2010.

[80] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segrè, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[81] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genetics*, 10(5):e1004383, 2014.

[82] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, Dan L. Nicolae, Nancy J. Cox, Hae Kyung Im, and G. TEx Consortium. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[83] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H. Penninx, Rick Jansen, Eco J. C. de Geus, Dorret I. Boomsma, Fred A. Wright, Patrick F. Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J. Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T. Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L. Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.

[84] Shun Liu, Allen Zhu, Chuan He, and Mengjie Chen. Repic: a database for exploring the N^6-methyladenosine methylome. *Genome Biology*, 21(1):100, 2020.

[85] Isaia Barbieri, Konstantinos Tzelepis, Luca Pandolfini, Junwei Shi, Gonzalo Millán-Zambrano, Samuel C. Robson, Demetrios Aspris, Valentina Migliori, Andrew J. Bannister, Namshik Han, Etienne De Braekeleer, Hannes Ponstingl, Alan Hendrick, Christopher R. Vakoc, George S. Vassiliou, and Tony Kouzarides. Promoter-bound mettl3 maintains myeloid leukaemia by m^6a-dependent translation control. *Nature*, 552:126, 2017.

[86] Xiaolan Deng, Rui Su, Hengyou Weng, Huilin Huang, Zejuan Li, and Jianjun Chen. Rna N^6-methyladenosine modification in cancers: current status and perspectives. *Cell Research*, 28(5):507–517, 2018.

[87] Zejuan Li, Hengyou Weng, Rui Su, Xiaocheng Weng, Zhixiang Zuo, Chenying Li, Huilin Huang, Sigrid Nachtergaele, Lei Dong, Chao Hu, Xi Qin, Lichun Tang, Yungui Wang, Gia-Ming Hong, Hao Huang, Xiao Wang, Ping Chen, Sandeep Gurbuxani, Stephen Arnovitz, Yuanyuan Li, Shenglai Li, Jennifer Strong, Mary Beth Neilly, Richard A. Larson, Xi Jiang, Pumin Zhang, Jie Jin, Chuan He, and Jianjun Chen. Fto plays an oncogenic role in acute myeloid leukemia as a N^6-methyladenosine rna demethylase. *Cancer Cell*, 31(1):127–141, 2017.

[88] Jun Liu, Mark A. Eckert, Bryan T. Harada, Song-Mei Liu, Zhike Lu, Kangkang Yu, Samantha M. Tienda, Agnieszka Chryplewicz, Allen C. Zhu, Ying Yang, Jing-Tao Huang, Shao-Min Chen, Zhi-Gao Xu, Xiao-Hua Leng, Xue-Chen Yu, Jie Cao, Zezhou Zhang, Jianzhao Liu, Ernst Lengyel, and Chuan He. m^6A mrna methylation regulates akt activity to promote the proliferation and tumorigenicity of endometrial cancer. *Nature Cell Biology*, 20(9):1074–1083, 2018.

[89] Rui Su, Lei Dong, Chenying Li, Sigrid Nachtergaele, Mark Wunderlich, Ying Qing, Xiaolan Deng, Yungui Wang, Xiaocheng Weng, Chao Hu, Mengxia Yu, Jennifer Skibbe, Qing Dai, Dongling Zou, Tong Wu, Kangkang Yu, Hengyou Weng, Huilin Huang, Kyle Ferchen, Xi Qin, Bin Zhang, Jun Qi, Atsuo T. Sasaki, David R. Plas, James E. Bradner, Minjie Wei, Guido Marcucci, Xi Jiang, James C. Mulloy, Jie Jin, Chuan He, and Jianjun Chen. R-2hg exhibits anti-tumor activity by targeting fto/m^6a/myc/cebpa signaling. *Cell*, 172(1):90–105.e23, 2018.

[90] Sung Chun, Alexandra Casparino, Nikolaos A. Patsopoulos, Damien C. Croteau-Chonka, Benjamin A. Raby, Philip L. De Jager, Shamil R. Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics*, 49(4):600–605, 2017.

[91] Douglas W. Yao, Luke J. O'Connor, Alkes L. Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, 2020.

[92] Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature Biotechnology*, 37(8):907–915, 2019.

[93] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K. Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, 2015.

[94] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589, 2010.

[95] Xiaodong Cui, Zhen Wei, Lin Zhang, Hui Liu, Lei Sun, Shao-Wu Zhang, Yufei Huang, and Jia Meng. Guitar: An r/bioconductor package for gene annotation guided transcriptomic analysis of rna-related genomic features. *BioMed research international*, 2016:8367534–8367534, 2016.

[96] Bryan Howie, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *G3: Genes—Genomes—Genetics*, 1(6):457, 2011.

[97] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44:955, 2012.

[98] N. Kumasaka, A. J. Knights, and D. J. Gaffney. Fine-mapping cellular qtls with rasqual and atac-seq. *Nat Genet*, 48(2):206–13, 2016.

[99] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2016.

[100] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[101] Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383, 2015.

[102] Eric L. Van Nostrand, Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, Thai B. Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W. Yeo. Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nature Methods*, 13:508, 2016.

[103] Yue Wan, Kun Qu, Qiangfeng Cliff Zhang, Ryan A. Flynn, Ohad Manor, Zhengqing Ouyang, Jiajing Zhang, Robert C. Spitale, Michael P. Snyder, Eran Segal, and Howard Y. Chang. Landscape and variation of rna secondary structure across the human transcriptome. *Nature*, 505(7485):706–709, 2014.

[104] Vikram Agarwal, George W. Bell, Jin-Wu Nam, and David P. Bartel. Predicting effective microrna target sites in mammalian mrnas. *eLife*, 4:e05005, 2015.

[105] Simon G. Coetzee, Gerhard A. Coetzee, and Dennis J. Hazelett. motifbreakr: an r/bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics*, 31(23):3847–3849, 2015.

[106] Tune H. Pers, Pascal Timshel, and Joel N. Hirschhorn. Snpsnap: a web-based tool for identification and annotation of matched snps. *Bioinformatics*, 31(3):418–420, 2015.

[107] Xiaoquan Wen. Molecular qtl discovery incorporating genomic annotations using bayesian false discovery rate control. *Ann. Appl. Stat.*, 10(3):1619–1638, 2016.

[108] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

[109] Gao Wang, Abhishek K. Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, page 501114, 2018.

[110] Kushal K. Dey, Dongyue Xie, and Matthew Stephens. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinformatics*, 19(1):473, 2018.

[111] A. Bertero, S. Brown, P. Madrigal, A. Osnato, D. Ortmann, L. Yiangou, J. Kadiwala, N. C. Hubner, I. R. de Los Mozos, C. Sadee, A. S. Lenaerts, S. Nakanoh, R. Grandy, E. Farnell, J. Ule, H. G. Stunnenberg, S. Mendjan, and L. Vallier. The smad2/3 interactome reveals that tgfbeta controls m^6a mrna methylation in pluripotency. *Nature*, 555(7695):256–259, 2018.

[112] Zheng Xia, Lawrence A. Donehower, Thomas A. Cooper, Joel R. Neilson, David A. Wheeler, Eric J. Wagner, and Wei Li. Dynamic analyses of alternative polyadenylation from rna-seq reveal a 3'-utr landscape across seven tumour types. *Nature Communications*, 5:5274, 2014.

[113] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[114] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015.

[115] Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R. Day, Consortium ReproGen, Consortium Schizophrenia Working Group of the Psychiatric Genomics, Raci Consortium The, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R. B. Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J. Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Partitioning heritability

by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47:1228, 2015.

[116] Tong Chen, Ya-Juan Hao, Ying Zhang, Miao-Miao Li, Meng Wang, Weifang Han, Yongsheng Wu, Ying Lv, Jie Hao, Libin Wang, Ang Li, Ying Yang, Kang-Xuan Jin, Xu Zhao, Yuhuan Li, Xiao-Li Ping, Wei-Yi Lai, Li-Gang Wu, Guibin Jiang, Hai-Lin Wang, Lisi Sang, Xiu-Jie Wang, Yun-Gui Yang, and Qi Zhou. m^6A rna methylation is regulated by micrornas and promotes reprogramming to pluripotency. *Cell Stem Cell*, 16(3):289–301, 2015.

[117] Shipra Das and Adrian R. Krainer. Emerging functions of srsf1, splicing factor and oncoprotein, in rna metabolism and cancer. *Molecular cancer research : MCR*, 12(9):1195–1204, 2014.

[118] B. Slobodin, R. Han, V. Calderone, Jafo Vrielink, F. Loayza-Puch, R. Elkon, and R. Agami. Transcription impacts the efficiency of mrna translation via co-transcriptional N^6-adenosine methylation. *Cell*, 169(2):326–337 e12, 2017.

[119] Francesca Aguilo, Fan Zhang, Ana Sancho, Miguel Fidalgo, Serena Di Cecilia, Ajay Vashisht, Dung-Fang Lee, Chih-Hung Chen, Madhumitha Rengasamy, Blanca Andino, Farid Jahouh, Angel Roman, Sheryl R. Krig, Rong Wang, Weijia Zhang, James A. Wohlschlegel, Jianlong Wang, and Martin J. Walsh. Coordination of m^6A mrna methylation and gene transcription by zfp217 regulates pluripotency and reprogramming. *Cell Stem Cell*, 17(6):689–704, 2015.

[120] Encode Project Consortium The, Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57, 2012.

[121] Huilin Huang, Hengyou Weng, Kere Zhou, Tong Wu, Boxuan Simen Zhao, Mingli Sun, Zhenhua Chen, Xiaolan Deng, Gang Xiao, Franziska Auer, Lars Klemm, Huizhe Wu, Zhixiang Zuo, Xi Qin, Yunzhu Dong, Yile Zhou, Hanjun Qin, Shu Tao, Juan Du, Jun Liu, Zhike Lu, Hang Ying, Ana Mesquita, Celvie L. Yuan, Yueh-Chiang Hu, Wenju Sun, Rui Su, Lei Dong, Chao Shen, Chenying Li, Ying Qing, Xi Jiang, Xiwei Wu, Jun-Lin Guan, Lianghu Qu, Minjie Wei, Markus Muschen, Gang Huang, Chuan He, Jianhua Yang, and Jianjun Chen. Histone h3 trimethylation at lysine 36 guides m^6A rna modification co-transcriptionally. *Under revision*.

[122] Jiyoung Lee, Ali E. Yesilkanal, Joseph P. Wynne, Casey Frankenberger, Juan Liu, Jielin Yan, Mohamad Elbaz, Daniel C. Rabe, Felicia D. Rustandy, Payal Tiwari, Elizabeth A. Grossman, Peter C. Hart, Christie Kang, Sydney M. Sanderson, Jorge Andrade, Daniel K. Nomura, Marcelo G. Bonini, Jason W. Locasale, and Marsha Rich Rosner. Effective breast cancer combination therapy targeting bach1 and mitochondrial metabolism. *Nature*, 568(7751):254–258, 2019.

[123] Clotilde Wiel, Kristell Le Gal, Mohamed X. Ibrahim, Chowdhury Arif Jahangir, Muhammad Kashif, Haidong Yao, Dorian V. Ziegler, Xiufeng Xu, Tanushree Ghosh, Tanmoy Mondal, Chandrasekhar Kanduri, Per Lindahl, Volkan I. Sayin, and Martin O. Bergo. Bach1 stabilization by antioxidants stimulates lung cancer metastasis. *Cell*, 178(2):330–345.e22, 2019.

[124] Steven Gazal, Hilary K. Finucane, Nicholas A. Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M. Neale, Alexander Gusev, and Alkes L. Price. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, 49(10):1421–1427, 2017.

[125] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Consortium Schizophrenia Working Group of the Psychiatric Genomics, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47:291, 2015.

[126] Göran K. Hansson. Inflammation, atherosclerosis, and coronary artery disease. *The New England journal of medicine*, 352(16):1685–1695, 2005.

[127] Artika P. Nath, Scott C. Ritchie, Nastasiya F. Grinberg, Howard Ho-Fung Tang, Qin Qin Huang, Shu Mei Teo, Ari V. Ahola-Olli, Peter Würtz, Aki S. Havulinna, Kristiina Santalahti, Niina Pitkänen, Terho Lehtimäki, Mika Kähönen, Leo-Pekka Lyytikäinen, Emma Raitoharju, Ilkka Seppälä, Antti-Pekka Sarin, Samuli Ripatti, Aarno Palotie, Markus Perola, Jorma S. Viikari, Sirpa Jalkanen, Mikael Maksimow, Marko Salmi, Chris Wallace, Olli T. Raitakari, Veikko Salomaa, Gad Abraham, Johannes Kettunen, and Michael Inouye. Multivariate genome-wide association analysis of a cytokine network reveals variants with widespread immune, haematological, and cardiometabolic pleiotropy. *American journal of human genetics*, 105(6):1076–1090, 2019.

[128] Lisette Stolk, John R. B. Perry, Daniel I. Chasman, Chunyan He, Massimo Mangino, Patrick Sulem, Maja Barbalic, Linda Broer, Enda M. Byrne, Florian Ernst, Tõnu Esko, Nora Franceschini, Daniel F. Gudbjartsson, Jouke-Jan Hottenga, Peter Kraft, Patrick F. McArdle, Eleonora Porcu, So-Youn Shin, Albert V. Smith, Sophie van Wingerden, Guangju Zhai, Wei V. Zhuang, Eva Albrecht, Behrooz Z. Alizadeh, Thor Aspelund, Stefania Bandinelli, Lovorka Barac Lauc, Jacques S. Beckmann, Mladen Boban, Eric Boerwinkle, Frank J. Broekmans, Andrea Burri, Harry Campbell, Stephen J. Chanock, Constance Chen, Marilyn C. Cornelis, Tanguy Corre, Andrea D. Coviello, Pio d'Adamo, Gail Davies, Ulf de Faire, Eco J. C. de Geus, Ian J. Deary, George V. Z. Dedoussis, Panagiotis Deloukas, Shah Ebrahim, Gudny Eiriksdottir, Valur Emilsson, Johan G. Eriksson, Bart C. J. M. Fauser, Liana Ferreli, Luigi Ferrucci, Krista Fischer, Aaron R. Folsom, Melissa E. Garcia, Paolo Gasparini, Christian Gieger, Nicole Glazer, Diederick E. Grobbee, Per Hall, Toomas Haller, Susan E. Hankinson, Merli Hass, Caroline Hayward, Andrew C. Heath, Albert Hofman, Erik Ingelsson, A. Cecile J. W. Janssens, Andrew D. Johnson, David Karasik, Sharon L. R. Kardia, Jules Keyzer, Douglas P. Kiel, Ivana Kolcic, Zoltán Kutalik, Jari Lahti, Sandra Lai, Triin Laisk, Joop S. E. Laven, Debbie A. Lawlor, Jianjun Liu, Lorna M. Lopez, Yvonne V. Louwers, Patrik K. E. Magnusson, Mara Marongiu, Nicholas G. Martin, Irena Martinovic Klaric, Corrado Masciullo, Barbara McKnight, Sarah E. Medland, David Melzer, Vincent Mooser, Pau Navarro, Anne B. Newman, Dale R. Nyholt, N. Charlotte Onland-Moret, Aarno Palotie, Guillaume Paré, Alex N. Parker, Nancy L. Pedersen, et al. Meta-analyses identify 13 loci associated with age at menopause and highlight dna repair and immune pathways. *Nature Genetics*, 44(3):260–268, 2012.

[129] Hua-Bing Li, Jiyu Tong, Shu Zhu, Pedro J. Batista, Erin E. Duffy, Jun Zhao, Will Bailis, Guangchao Cao, Lina Kroehling, Yuanyuan Chen, Geng Wang, James P. Broughton, Y. Grace Chen, Yuval Kluger, Matthew D. Simon, Howard Y. Chang, Zhinan Yin, and Richard A. Flavell. m^6A mrna methylation controls t cell homeostasis by targeting the il-7/stat5/socs pathways. *Nature*, 548:338, 2017.

[130] Qingliang Zheng, Jin Hou, Ye Zhou, Zhenyang Li, and Xuetao Cao. The rna helicase ddx46 inhibits innate immunity by entrapping m^6a-demethylated antiviral transcripts in the nucleus. *Nature Immunology*, 18:1094, 2017.

[131] Gianluigi Lichinchi, Shang Gao, Yogesh Saletore, Gwendolyn Michelle Gonzalez, Vikas Bansal, Yinsheng Wang, Christopher E. Mason, and Tariq M. Rana. Dynamics of the human and viral m^6A rna methylomes during hiv-1 infection of t cells. *Nature Microbiology*, 1:16011, 2016.

[132] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan L. M. Björkegren, Hae Kyung Im, Bogdan Pasaniuc, Manuel A. Rivas, and Anshul Kundaje. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4):592–599, 2019.

[133] Katsuyuki Nakao, Maki Nishino, Kyoko Takeuchi, Masayasu Iwata, Akira Kawano, Yasuhito Arai, and Misao Ohki. Fusion of the nucleoporin gene, ¡i¿nup98¡/i¿, and

the putative rna helicase gene, ¡i¿dzxx10¡/i¿, by inversion 11 (p15q22) chromosome translocation in a patient with etoposide-related myelodysplastic syndrome. *Internal Medicine*, 39(5):412–415, 2000.

[134] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010.

[135] Harold Pimentel, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687–690, 2017.

[136] M. Lawrence, W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8):e1003118, 2013.

[137] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

[138] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(Database issue):D353–D361, 2017.

[139] G. Yu, L. G. Wang, Y. Han, and Q. Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS*, 16(5):284–7, 2012.

[140] Y. Xiao, Y. Wang, Q. Tang, L. Wei, X. Zhang, and G. Jia. An elongation- and ligation-based qpcr amplification method for the radiolabeling-free detection of locus-specific N^6-methyladenosine modification. *Angew Chem Int Ed Engl*, 57(49):15995–16000, 2018.

[141] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[142] Ji Luo, Brendan D. Manning, and Lewis C. Cantley. Targeting the pi3k-akt pathway in human cancer: Rationale and promise. *Cancer Cell*, 4(4):257–262, 2003.

[143] Y. Zhang, P. Kwok-Shing Ng, M. Kucherlapati, F. Chen, Y. Liu, Y. H. Tsang, G. de Velasco, K. J. Jeong, R. Akbani, A. Hadjipanayis, A. Pantazi, C. A. Bristow, E. Lee, H. S. Mahadeshwar, J. Tang, J. Zhang, L. Yang, S. Seth, S. Lee, X. Ren, X. Song, H. Sun, J. Seidman, L. J. Luquette, R. Xi, L. Chin, A. Protopopov, T. F. Westbrook, C. S. Shelley, T. K. Choueiri, M. Ittmann, C. Van Waes, J. N. Weinstein, H. Liang, E. P. Henske, A. K. Godwin, P. J. Park, R. Kucherlapati, K. L. Scott, G. B. Mills, D. J. Kwiatkowski, and C. J. Creighton. A pan-cancer proteogenomic atlas of pi3k/akt/mtor pathway alterations. *Cancer Cell*, 31(6):820–832 e3, 2017.

[144] Rohan K. Humphrey, Shu-Mei Yu, Luis E. Flores, and Ulupi S. Jhala. Glucose regulates steady-state levels of pdx1 via the reciprocal actions of gsk3 and akt kinases. *Journal of Biological Chemistry*, 285(5):3406–3416, 2010.

[145] Doris A. Stoffers, Noah T. Zinkin, Violeta Stanojevic, William L. Clarke, and Joel F. Habener. Pancreatic agenesis attributable to a single nucleotide deletion in the human ipf1 gene coding sequence. *Nature Genetics*, 15(1):106–110, 1997.

[146] Shuangli Guo, Chunhua Dai, Min Guo, Brandon Taylor, Jamie S. Harmon, Maike Sander, R. Paul Robertson, Alvin C. Powers, and Roland Stein. Inactivation of specific β cell transcription factors in type 2 diabetes. *The Journal of Clinical Investigation*, 123(8):3305–3316, 2013.

[147] Jennifer M. Oliver-Krasinski, Margaret T. Kasner, Juxiang Yang, Michael F. Crutchlow, Anil K. Rustgi, Klaus H. Kaestner, and Doris A. Stoffers. The diabetes gene pdx1 regulates the transcriptional network of pancreatic endocrine progenitor cells in mice. *The Journal of Clinical Investigation*, 119(7):1888–1898, 2009.

[148] Dario F. De Jesus, Zijie Zhang, Sevim Kahraman, Natalie K. Brown, Mengjie Chen, Jiang Hu, Manoj K. Gupta, Chuan He, and Rohit N. Kulkarni. m^6a mrna methylation regulates human β-cell biology in physiological states and in type 2 diabetes. *Nature Metabolism*, 2019.

[149] Elizabeth Snyder, Ramani Soundararajan, Manju Sharma, Andrea Dearth, Benjamin Smith, and Robert E. Braun. Compound heterozygosity for y box proteins causes sterility due to loss of translational repression. *PLOS Genetics*, 11(12):e1005690, 2015.

[150] Rajat Roy, Danielle Durie, Hui Li, Bing-Qian Liu, John Mark Skehel, Francesco Mauri, Lucia Veronica Cuorvo, Mattia Barbareschi, Lin Guo, Martin Holcik, Michael J. Seckl, and Olivier E. Pardo. hnrnpa1 couples nuclear export and translation of specific mrnas downstream of fgf-2/s6k2 signalling. *Nucleic Acids Research*, 42(20):12483–12497, 2014.

[151] Kassie S. Manning and Thomas A. Cooper. The roles of rna processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology*, 18(2):102–114, 2017.

[152] Deepti Gurdasani, Inês Barroso, Eleftheria Zeggini, and Manjinder S. Sandhu. Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, 20(9):520–535, 2019.

[153] Huwenbo Shi, Kathryn S. Burch, Ruth Johnson, Malika K. Freund, Gleb Kichaev, Nicholas Mancuso, Astrid M. Manuel, Natalie Dong, and Bogdan Pasaniuc. Localizing components of shared transethnic genetic architecture of complex traits from gwas summary data. *American Journal of Human Genetics*, 106(6):805–817, 2020.

[154] Lauren S. Mogil, Angela Andaleon, Alexa Badalamenti, Scott P. Dickinson, Xiuqing Guo, Jerome I. Rotter, W. Craig Johnson, Hae Kyung Im, Yongmei Liu, and Heather E. Wheeler. Genetic architecture of gene expression traits across diverse populations. *PLOS Genetics*, 14(8):e1007586, 2018.

[155] Anne Ndungu, Anthony Payne, Jason M. Torres, Martijn van de Bunt, and Mark I. McCarthy. A multi-tissue transcriptome analysis of human metabolites guides interpretability of associations based on multi-snp models for gene expression. *The American Journal of Human Genetics*, 106(2):188–201, 2020.

[156] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biol*, 17:13, 2016.

[157] Alexa B. R. McIntyre, Nandan S. Gokhale, Leandro Cerchietti, Samie R. Jaffrey, Stacy M. Horner, and Christopher E. Mason. Limits in the detection of m6a changes using merip/m6a-seq. *Sci Rep*, 10:6590, 2020.

[158] Benjamin J. Schmiedel, Divya Singh, Ariel Madrigal, Alan G. Valdovino-Gonzalez, Brandie M. White, Jose Zapardiel-Gonzalo, Brendan Ha, Gokmen Altay, Jason A. Greenbaum, Graham McVicker, Grégory Seumois, Anjana Rao, Mitchell Kronenberg, Bjoern Peters, and Pandurangan Vijayanand. Impact of genetic polymorphisms on human immune cell gene expression. *Cell*, 175(6):1701–1715.e16, 2018.

[159] Diego Calderon, Michelle L. T. Nguyen, Anja Mezger, Arwa Kathiria, Fabian Müller, Vinh Nguyen, Ninnia Lescano, Beijing Wu, John Trombetta, Jessica V. Ribado, David A. Knowles, Ziyue Gao, Franziska Blaeschke, Audrey V. Parent, Trevor D. Burt, Mark S. Anderson, Lindsey A. Criswell, William J. Greenleaf, Alexander Marson, and Jonathan K. Pritchard. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics*, 51(10):1494–1505, 2019.