Biological signatures of infectious disease resistance: Deciphering immunological mechanisms of resistance to tuberculosis (TB) by combining host genetics and immuno-transcriptomics

Cristian Ovadiuc

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2023

Committee:

Thomas Hawn

Kim Dill-McFarland

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

Abstract

Biological signatures of infectious disease resistance: Deciphering immunological mechanisms of resistance to tuberculosis (TB) by combining host genetics and immuno-transcriptomics

Cristian Ovadiuc

Chair of the Supervisory Committee:

Thomas Hawn

Department of Allergy and Infectious Diseases

Understanding the biological signatures for resistance to Mycobacterium tuberculosis (Mtb) in a unique group termed resisters (RSTR) is crucial. Despite extensive exposure to the pathogen, these individuals remain uninfected, highlighting elusive yet critical natural resistance pathways. In efforts to elucidate the molecular mechanisms behind the unique resistance to Mtb, we hypothesized that unique gene expression patterns in alveolar macrophages might explain this resistance phenomenon. Utilizing immuno-transcriptomic profiling and single nucleotide polymorphism (SNP) array genotyping, we analyzed alveolar macrophage samples from an observational cohort in Uganda, comparing RSTR and latent TB infection (LTBI) (N=49 participants). Differential gene expression analysis revealed elevated BMAL1 expression in RSTR relative to LTBI (false discovery rate [FDR], <0.25). Gene set enrichment analyses (GSEA) identified multiple gene sets differentiating RSTR and LTBI alveolar macrophages. The only significantly enriched gene set in RSTR compared to LTBI was the E2F biological pathway encoding cell cycle-related targets (false discovery rate [FDR], <0.2). Intriguingly, these findings allude to a potential convergence of these pathways with autophagy, a crucial defense mechanism against Mtb infection. This novel intersection between autophagy and circadian rhythm processes active in RSTRs point to a new avenue for host directed therapeutic interventions against tuberculosis (TB).

I. INTRODUCTION

Introduction to Tuberculosis (TB)

Mycobacterium tuberculosis (Mtb) has evolved alongside humans to become a highly successful airborne pathogen, infecting an estimated 23-32% of the global population. Upon exposure, three primary outcomes are possible: active tuberculosis (TB), latent Mtb infection, or resistance to infection. Of those exposed to an active index TB case, about 10% will develop active disease and the rest developing latent infection with the exception of rare cases of resistance (Simmons et al., 2018; Stein, 2023).

The only approved vaccine for TB in the past century, the Bacille Calmette-Guérin (BCG) vaccine, has limitations as it primarily protects against severe forms of TB in young children and infants but not in adults. Treatment of TB is further complicated by months-long antibiotic regimens, with the emerging challenge of antibiotic resistance (Orenstein et al., 2023). Consequently, TB has become a leading infectious disease killer surpassing all other pathogens, causing an average of 1.65 million deaths annually from 2010 to 2019, with most recent data indicating approximately 10 million new cases and 1.4 million deaths in 2019 alone (Dodd et al., 2023; Orenstein et al., 2023).

The challenge of combating Mtb is not due to a lack of effort but rather its complexity as an intracellular pathogen, having adapted to human hosts over thousands of years with a vast arsenal of countermeasures against host attempts to prevent and control Mtb infection (McHenry et al., 2020). Geographically, Mtb overlaps with HIV with an estimated 14 million TB cases present in HIV+ individuals, further complicating treatment by requiring combined therapeutics which often interact leading to complications and placing additional burden that Mtb vaccines and therapeutics can be safe and effective in HIV+ populations (Bruchfeld et al., 2015).

The infection of a third of the global population over thousands of years has allowed natural selection to give rise to a rare subset of individuals with genetic mutations that confer resistance to Mtb. By studying the individuals exhibiting these exceptional and rare traits, we hope to unravel the differences in their biology. This way, we exploit Mtb's unique strength given coevolution and host adaptation as a potential weakness or Achilles' heel, where natural selection and evolution gave rise to rare groups of individuals resistant to Mtb infection (Kwok et al., 2021). In support of this strategy, research shows that drug targets and therapeutics informed by human genetics data are much more likely to achieve clinical success while historically, medical countermeasures were discovered for other pathogens studying resistance mechanisms (Ochoa et al., 2022). Such was the case for HIV resistance identifying rare variation in the

CCR5 gene, and more recently a novel locus involving the CHD1L gene variant associated with controlling HIV infection by reducing viral load (McLaren et al., 2023; Simmons et al., 2018).

TB epidemiology and diagnostics

There is compelling epidemiological evidence suggesting a subset of individuals resist Mtb infection. However, variations in these findings might arise from inconsistencies in defining resistance criteria, such as the intensity and duration of exposure to the disease. With epidemiological data confirm high-intensity and long-term exposure to active TB without a positive tuberculin skin test (TST) result, it's plausible to conclude these individuals lack latent infections and are indeed resistant, known as RSTRs in the literature (Lu et al., 2019; Stein, 2023).

In a Ugandan cohort, 872 index cases with culture-confirmed TB and 2,585 household contacts were studied. Approximately 9% tested negative for latent TB infection (LTBI) using TST. A follow-up 8-10 years later, using both TST and more recent interferon-gamma release assay (IGRA) found 7% of the cohort likely to be resistant to Mtb infections (Stein et al., 2019). The follow-up duration determined the size of the RSTR population indicating why we see heterogeneity in the estimated number of RSTRs across independent studies. In contrast, an Indonesia study followed participants for 14 weeks to determine resistance which is really low given other research showing over 25% of individuals testing positive between 3 and 24 months post enrollment (Stein et al., 2020). In a South African miner cohort, epidemiological data revealed high transmission and environmental risks from close-quarter mining exposure to active Mtb cases. Some individuals, however tested negative for latent Mtb infection, suggesting around 13% might be RSTRs. Looking at historical epidemiological data, during a 1966 outbreak of TB on the USS Richard E. Byrd destroyer, 308 at-risk crew members were tested. Follow-up found that about 10% of those who were believed to be exposed never tested positive, indicating they are likely RSTRs (Simmons et al., 2018).

Defining the RSTR phenotype and genetic epidemiology of resistance to Mtb is complex (Stein, 2023). The current method relies on indirect measurements of Mtb infection, primarily looking at T cell interferon-gamma. The TST test measures hypersensitivity to mycobacterial antigens injected under the skin by quantifying the induration dimensions at the site of inoculation. While widely available, its main limitation is potential false positives, arising from factors like unrelated environmental mycobacterial exposures, prior BCG vaccinations or reduced immunity from HIV coinfections. To address these flaws, IGRA assays were developed, which measure whole blood interferon-gamma from Mtb specific antigen

CD4 T cells, providing more accurate results unaffected by prior BCG vaccinations or environmental mycobacterial interactions.

Human genetics determine Mtb resistance and susceptibility

A severe form of TB occurs when Mtb leaves the lungs and travels to the brain, leading to TB meningitis. This only develops in an estimated 1% of total infections and results from a breakdown in the innate immune response. Genetic variations in toll-like receptors (TLRs), which act as sensors activated by foreign pathogens, are involved in this process. Specifically, individuals with TLR2 variants are unable to contain Mtb allowing it to escape and spread to the brain (Thuong et al., 2007).

Immunogenomic discoveries in the study of Mtb can reveal essential host checkpoints that the pathogen must bypass to establish active disease, establish latency, or proceed from latency to reactivation and primary active disease (Arentz and Hawn, 2007). Although it might not prevent the infection itself, immunomodulation of TLR2 may inhibit Mtb from disseminating to the brain, and host-directed therapies could improve outcomes for this subset of cases showing potential of human genetic findings (Hawn et al., 2015; Thuong et al., 2007; Tobin et al., 2012). Rather than targeting many of these checkpoints and deficiencies at various stages, a more effective strategy would be to identify the factors responsible for preventing initial infection.

The resistance phenotype to Mtb infection is likely polygenic, influenced by a cascade of host defense responses. As one defense mechanism fails, another takes over. In areas where Mtb is hyperendemic, evidence suggests that some populations develop resistance very early in the infection process. Furthermore, both susceptibility and resistance appear to be heritable. However, a significant portion of this heritability remains unexplained. More research is needed to identify the genetic variations responsible given we see evidence of host genetics at play in twin studies, family linkage, and candidate gene studies (McHenry et al., 2021; Simmons et al., 2018).

Some genetic studies indicate that host genetics influence susceptibility to TB, with distinct genetic regions linked to resistance and protection from the disease. For example, polymorphisms in TOLLIP and ULK1 were associated with LTBI susceptibility, while the TST1 locus on chromosome 11p14 was linked to TNF secretion and TST negativity, and the TST2 locus on chromosome 5p15 was connected to the intensity of TST skin test reactivity (Simmons et al., 2018). Highlighting the importance of doing genetic studies in diverse population, a genome-wide association studies (GWAS) identified two risk loci for Mtb infection in Han Chinese population, rs12437118 and rs6114027 shown to decrease TGM6

expression in peripheral blood monocytes (PBMCs). Performing genetic studies across diverse population given different genetic backgrounds could partially explain why not all loci associated with Mtb resistance or susceptibly replicate across different populations (Zheng et al., 2018).

Shendure et al. explain five main challenges, detailing why we have yet to identify human genetic loci that explain heritability and large effect sizes in determining a clear gene or pathway of emphasis to target for common diseases (Shendure et al., 2019).

First, the authors note that disease risk is often influenced by many variants with small effects and limited individual predictive power. Second, for common diseases, GWAS variants often account for only a small minority of their heritability. As an example, they mention that the discovery of rare variants might account for a significant portion of the unexplained heritability, underscoring the need to transition from single nucleotide polymorphism (SNP) chip-based methods to whole genome sequencing (WGS). Third, there's a limited resolution of significantly associated loci. For example, a causal SNP in a region such as the 2-14 Mb range for the TST1 locus, can be challenging to pinpoint. Relying on linkage disequilibrium (LD), which reduced the cost of large-scale genetic studies, has been a drawback since it lacks the high resolution needed to identify a single causal SNP. Fourth, most loci are found in non-coding regions and are also present in cell-type-specific regulatory regions, making it challenging to decipher their potential biological mechanisms requiring additional cell-type context. Fifth, the omnigenic model suggests that due to the complexity and interconnectedness of genetic regulatory networks, GWAS might identify loci that, even if significant, are not necessarily involved and won't aid in elucidating the biological mechanisms related to the disease under study (Shendure et al., 2019). For these reasons and others related to study designs examining the genetic epidemiology of resistance to Mtb, such as using stringent definitions for the RSTR phenotype, there has yet to be a strong consistent genetic signal validated across multiple cohorts (Kwok et al., 2021; Stein, 2023).

It's important to emphasize that both GWAS and transcriptomics are complementary approaches. Biological signals undetected in the genetic architecture of the study population can be identified downstream by looking at RNA expression, with the caveat that it might be cell type-specific, so sampling different single cell types becomes crucial. The opposite is also true, where identifying a genetic loci, especially in gene regulatory regions, makes it difficult to predict downstream RNA gene expression and relevant biological pathways (Stein, 2023; Thuong et al., 2008).

Innate and Adaptive Resisters and Alveolar Macrophages

Evidence reveals a range of RSTR phenotypes, spanning both adaptive and innate resistance mechanisms. Innate resisters likely rely on alveolar macrophage pathways for infection prevention, whereas adaptive resisters rely on T cell and B cell responses for clearance (Simmons et al., 2018). Recent GWAS on the CHD1L gene highlighted its protective role against HIV infection in macrophages, but not T cells (McLaren et al., 2023). Considering human alveolar macrophages are Mtb's primary contact in the lung after inhalation and that blood samples, typically containing PBMCs, are commonly drawn in epidemiological studies due to the invasiveness of lung sampling for alveolar macrophages, potential resistance signals might be missed without examining multiple cell types (Simmons et al., 2022). This emphasizes the need to analyze biological signatures across distinct cell types and contrast their characteristics. Additionally, alveolar macrophages as antigen-presenting cells linking innate and adaptive immune responses in addition to playing a role in maintaining lung function, play a central role in host response to both infection and maintaining homeostasis (Ahmad et al., 2022; Joshi et al., 2018).

II. METHODS

Uganda Study Cohort

The Uganda based Kawempe Community Health Study identified and tracked a total of 872 index cases with culture confirmed pulmonary TB and 2,585 index case household contacts of pulmonary Mtb cases from 2002 to 2012, implementing serial TST testing over a span of two years (months 3,6,9,12,24). The study initiated a long-term follow-up a decade later to retrace the original cohort, focusing on HIV-negative contacts. The re-contacted subjects underwent a repeat TST and three IGRAs. Phenotype classifications were made using TST/IGRA test results from Uganda. Subjects resistant to Mtb infection demonstrated consistently negative results from all available TST/IGRA tests ruling out latent and asymptotic infection. Subjects with latent tuberculosis infection (LTBI) showed consistently positive TST/IGRA testing results on all available tests (Ma et al., 2014; Stein et al., 2019, 2018).

Biological Samples

Bronchoalveolar lavage (BAL) samples were collected from a specific subsegment within the right middle lobe of participant lungs. During this process, the technician carried out a lavage with six individual 30 cc aliquots of pre-heated normal saline using a wedged bronchoscope. The lavage fluid was then aspirated using a 50ml syringe and prepared into tubes containing 10^6 cells for Trizol preparation. The cells from the acquired samples were lysed in Trizol (Invitrogen) and subsequently frozen. Upon thawing, the aqueous phase was extracted after chloroform interaction, followed by the introduction of

100% ethanol. The processed lysates were then applied to miRNeasy mini columns in accordance with Qiagen manufacturer's instructions to isolate RNA for further analysis. Next, mRNA was extracted from the total RNA samples using Poly(A)+ RNA capture, and the isolated RNA was processed using a SMARTseq kit (Takara) for cDNA generation. Library preparation was conducted using the NexteraXT kit (Illumina), and sequencing was performed on a NextSeq 2000. The process generated paired-end reads of 50 base pairs, with each sample sequenced at a depth of 10 million reads per sample.

Raw RNA FASTQ reads to gene counts

The bioinformatics pipeline SEAsnake was used to analyze 23 RSTR samples and 26 LTBI samples for a total of 49 samples for the main study (Dill-McFarland et al., 2022). The general steps followed for the pipeline were: (1) sequence quality assessment with FASTQC (Andrews, 2010), (2) adapter removal and quality filtering with ADAPTERREMOVAL (Schubert et al., 2016), (3) alignment to the GRCh38 human reference genome using STAR (Dobin et al., 2013), (4) alignment quality assessment and filtering with samtools (Li et al., 2009), and (5) alignment quality assessment using Picard ("Picard toolkit," 2019). Upon completing these steps, Subread featureCounts was used to generate the raw gene count table by counting reads in gene exons (Liao et al., 2014).

Gene counts to voom – data cleaning, QC and normalization with edgeR and limma

Raw RNA-seq gene counts underwent pre-processing using the edgeR package (Robinson et al., 2010). Alignment metrics were calculated using Picard, and the extracted libID was merged with metrics, sample, and patient data into a single metadata R object. Sample quality was assessed using metrics such as total sequences passing QC, percent genome alignment, and the median coefficient of variation of coverage. Cut-offs applied were sequences over 1,000,000, CV coverage under 1, and alignment over 75%. Samples not passing these cutoffs were excluded (Figure S1). The biomaRt package was used for gene annotation to respective biotypes (e.g. protein coding, lcRNA, miRNA etc.) and the dataset was filtered to retain only protein-coding genes (Durinck et al., 2009). Principal component analysis (PCA) was utilized to identify sample outliers, characterized as deviations greater than 3 standard deviations on PC1 or PC2 which were subsequently removed from the dataset (Figure S1). Gene count and metadata were merged into a single DGElist object. Low-abundance and rare or zero count genes were excluded, retaining those with at least 0.5 Counts per Million (CPM) in three or more samples. The voom method was used to analyze the mean-variance relationship. Filtered data were visualized and rare genes filtered out were saved into a separate file. Gene expression was normalized using the Trimmed Mean of M-values (TMM) method, adjusting for library size and RNA composition variations. The voom transformation method converted counts to $\log_2$ CPM within each sample, adjusting for sample depth

differences and count data skewness. Gene-level weights were also computed to improve downstream linear modeling.

Host genetics and kinship analysis

Study participants were genotyped at over two million genetic markers with Illumina Infinium Expanded Multi-Ethnic Genotyping Array ($MEGA_{ex}$). Data was subjected to rigorous quality filtering using PLINK software, specifically filtering for Minor Allele Frequency (--maf < 0.05), Hardy-Weinberg Equilibrium P-value (--hwe < 1x10^-6), missing call rate for variants (--geno > 0.05), and missing call rate for samples (--mind > 0.03) (Purcell et al., 2007). Following quality control, a LD linkage disequilibrium filter was applied to select subset of independent SNPs using the parameters --indep-pairwise 50 5 0.1, which detects pairwise r2 greater than 0.1 in sliding windows of 50 base pairs with a 5 base pair slide. This resulted in 385,410 genetic variants for downstream analysis.

The analysis of the Uganda cohort, characterized by structured population and familial relatedness, used the GENESIS package in R, designed to account for these complexities (Gogarten et al., 2019; Gurdasani et al., 2019). The package's key algorithms, PC-AiR and PC-Relate, were used to estimate kinship, factoring in the cohort's unique genetic makeup. PC-AiR utilized kinship and ancestry divergence inputs to calculate the top ancestry principal components (PCs) necessary to adjust for population structure and ancestry. This adjustment was required to run PC-Relate for generating an accurate Genetic Relationship Matrix (GRM). To create the GRM, relatedness was determined via a KING-estimated kinship matrix. Next, the population structure from the unrelated subset of the sample was calculated by a KING-estimated kinship threshold of 0.0221. Using PC-Relate, kinship was recalculated by adjusting for these PCs, producing an estimate of kinship with the population structure removed (McHenry et al., 2021).

Linear statistical modeling

Gene expression was modeled as previously described (Dill-McFarland et al., 2023). Briefly the kimma R package allows for flexible linear modeling using kmFit function which can fit linear or mixed models with gene-level weights, covariates and random effects. Fitted linear models and p-values were estimated using restricted maximum likelihood where simple linear regression models were done in baseR using stats:lm and mixed effects using lme4:lmer. Gene set enrichment analysis (GSEA) was performed using the Molecular Signatures Database (MSigDB) looking at Hallmark, KEGG and Gene Ontology (GO) collections. Custom gene sets from weighted co-expression network analysis (WGCNA) modules looking at activated macrophage states were used as a custom gene set database running GSEA in addition to

previously discussed gene sets. Biological networks were visualized using StringDB (Szklarczyk et al., 2015).

III. RESULTS

Differential gene expression between RSTR and LTBI

This thesis aimed to elucidate the potential underlying biological mechanisms that confer Mtb infection resistance examining immuno-transcriptomic data from alveolar macrophages. Specifically, our aim was to identify genes and biological pathways showing differential expression between RSTR and LTBI cohorts.

Our sample collection comprised of 23 RSTR and 26 LTBI samples. An analysis of the epidemiological and clinical characteristics within the study cohort revealed a significant mean age difference between RSTR (mean 19.5) and LTBI (mean 24.9) groups (p=0.0064). Other measured variables such as sex, BMI, and risk score did not present a statistically significant difference (Table 1).

In the attempt to account for potential confounders that could distort gene expression levels, and misidentify true Mtb resistance genes, we built multiple statistical models accounting for potential confounders like age, sex, kinship, and exposure risk score. In total we built and evaluated four statistical models (1) ~RSTR, (2) ~RSTR + Age + Sex, (3) ~RSTR + Age + Sex + Kin and (4) ~RSTR + Risk Score. At a high level we used Akaike information criterion (AIC) metrics to calculate how well the linear model explains the data. Model goodness of fit was done by calculating residuals resulting from the predicted and actual gene expression values where smaller residuals for a given gene interpreted as the model is more accurate. A low AIC (-2, 2) is evidence there is minimal change in model fit when comparing two linear models. AIC values between 3-7 is moderate evidence of improved fit and AIC >10 is strong evidence of improved model fit.

Using age and sex as covariates to adjust for confounding improved model fit compared to the base model fitting 4051 genes better with >2 AIC out of total of 13808 genes. Model with age, sex and kinship did not improve fit, rather we see 13667 genes with AIC >10 modeled better without kinship. Given the cohort was a household contact study, there are related individuals in the study, and we considered kinship could confound gene expression data and adjusting for it could improve model fit as shown in other studies (Dill-McFarland et al., 2023). Lastly, modeling risk score we saw no model improvement over using age and sex with 8804 genes showing minimal change with AIC ~2 and 5004 genes better fit

using age and sex with AIC > 2. Based on AIC metrics we determined ~RSTR + Age + Sex as the best statistical model given the data (Figure 1).

Our analysis identified the transcription factor BMAL1 as being more significantly differentially expressed within RSTR individuals, after adjusting for age, sex, and kinship (FDR < 0.25; Figure 2). Even though ~RSTR + Age + Sex linear model is the best overall model, for BMAL1 specifically it is also the case that looking at AIC gene specific values we see ~RSTR + Age + Sex with an AIC of 56.6 compared to ~RSTR + Age, Sex, Kin AIC of 72. Lower AIC indicates a better fit and given the difference is >10 (i.e.,72 – 56.6 = 15.4) we can conclude we have high confidence ~RSTR + Age, Sex is a better fit for BMAL1. Furthermore, the statistical analysis indicates BMAL1 is not a high confidence differentially expressed gene (DEG) since the model with the best (i.e., age, sex with no kinship) indicates BMAL1 is not a significant DEG.

Differential biological pathways between RSTR and LTBI

Using gene set enrichment analysis (GSEA) considering fold change expression values across all genes (i.e., not only differentially expressed genes) and the Broad Molecular Signatures Database (MSigDB) we identified multiple Hallmark gene sets differentiating RSTR and LTBI alveolar macrophages. Notably, we found E2F targets (FDR < 0.2) which are genes encoding cell cycle related targets of E2F transcription factors as the only biological pathway enriched in the RSTR group. Leading edge gene analysis identified 39 genes driving enrichment subset from all 200 genes in Hallmark E2F biological targets shown in StringDB gene networks analysis (Figure 3; Figure 4).

In addition to already annotated gene sets in MSigDB, we ran GSEA on a set of WGCNA modules containing macrophages stimulated under different experimental conditions. We found Module 7 and Module 8 were enriched in the LTBI group with interferon-gamma stimulation consistent with previous work, and though there were no additional insights into RSTR group, this analysis adds a control layer confirming an expected finding validating the analysis.

How do alveolar macrophages compare to peripheral blood monocyte responses?

Recent work in the Hawn Lab looked at whether RSTR monocyte genes and pathways regulate resistance to Mtb infection. Though Mtb initially encounters alveolar macrophages in the lung it can infect other cells like monocytes to sustain infection and importantly it was shown there were differential genes and pathways comparing RSTR and LTBI monocytes from whole blood samples. The new data from this thesis project allows for comparing alveolar macrophages to monocytes. Mainly we compared significant

DEGs and biological pathways in alveolar macrophages to (1) Uganda and South Africa RSTR baseline or media only, (2) Uganda and South Africa RSTR Mtb+ and (3) Uganda RSTR HIV+ Mtb+.

(1) Uganda and South Africa RSTR baseline

Looking at unstimulated CD14+ monocyte immuno-transcriptomics, there were no statistically significant DEGs at an FDR < 0.2 for both Uganda and South Africa cohorts (Simmons et al., 2021). Additionally, looking at biological pathways it was found that metabolic biological signatures distinguish RSTR and LTBI monocytes. Notably, oxidative phosphorylation was significantly enriched in the RSTR group in both epidemiological cohorts, which was absent in alveolar macrophages and conversely BMAL1 and E2F were not differentially significant in monocytes.

(2) Uganda and South Africa RSTR Mtb+

Comparison of alveolar macrophages with monocytes stimulated by virulent Mtb strains ex vivo, we saw distinct transcriptional signatures separating RSTR and LTBI (Simmons et al., 2022). Notably, 260 DEGs were identified in the Uganda cohort, whereas only 5 DEGs were found in the South African cohort. Focusing in on Uganda, genes displaying the highest fold changes were associated with proinflammatory signaling. Network analysis of all 260 DEGs revealed involvement in cytokine production, TNF, and interferon-gamma signaling pathways. GSEA analysis found RSTR monocytes had lower TNFα signaling via NF-kB and inflammatory response hallmark gene sets when in media, which increased with Mtb stimulation. The oxidative phosphorylation pathway was enriched in RSTR compared to LTBI though at reduced levels. This might be explained by the host metabolic shift from oxidative phosphorylation to aerobic glycolysis upon Mtb infection, with RSTR monocytes showing reduced metabolic flexibility (Simmons et al., 2022). Across this comparison, neither BMAL1 nor E2F accounted for differences between RSTR and LTBI monocytes, as they were absent in both DEG and GSEA leading edge gene analyses.

(3) Uganda RSTR HIV+, Mtb+

Another strategy to isolate a RSTR signal is to focus on more extreme phenotypes such as HIV+ populations, where we expect HIV infections to increase susceptibility to TB and therefore resistance to Mtb infection in this subset of individuals can increase power of detecting DEGs or biological pathways (Panarella and Burkett, 2019). GSEA analyses of monocytes from HIV+ individuals only from Uganda found three hits with Hallmark myogenesis up in RSTR and Hallmark interferon-gamma response and allograft rejection up in LTBI. Consistent with previous findings comparing to alveolar macrophages signatures, there was no E2F differential expression detected in monocytes.

IV. DISCUSSION

Immuno-transcriptomic profiling revealed distinct biological signatures between RSTR and LTBI alveolar macrophages. Specifically, BMAL1 and E2F targets are elevated in RSTRs but not significantly associated with RSTR monocytes. While statistical evidence for BMAL1 differential expression is weak, it is worth exploring the potential biological significance in Mtb infection resistance, given previous research showing BMAL1 involved in controlling outcomes of other infectious diseases. Furthermore, there is possible overlap in function with E2F suggesting both might be active in the same autophagic biological mechanism leading to resistance.

BMAL1 is a transcription factor regulating circadian rhythms and biological clocks. It heterodimerizes with CLOCK to form a macromolecular protein complex that binds to DNA regulatory regions, controlling the expression of hundreds of clock-controlled genes (CCGs) (Diallo et al., 2020). Multiple splice variants of BMAL1 exist, including the Bmal1a isoform (hBmal1a). This variant, active in the cytoplasm and lacking the nuclear localization signal of the canonical form, is proposed to offer refined control of the molecular circadian clock (Lee et al., 2018). Hallmark E2F targets, regulated by nine related E2F transcription factors target around 200 genes. These genes primarily direct cell cycle, proliferation, division, apoptosis, and autophagy. Broadly we see an overlap in BMAL1 and E2F functions related to cell cycle progression (Thurlings and de Bruin, 2016) (Jiang et al., 2021).

Specific to antimicrobial function and cellular response to infection, we see BMAL1 and E2F functions overlap in controlling autophagy. Autophagy is mainly responsible for maintaining cellular homeostasis breaking down and recycling cellular contents (Deretic and Lazarou, 2022). This is particularly crucial given that Mtb is an intracellular bacterium that must evade host degradation mechanisms to survive and it has been shown autophagy is very important to defending against Mtb infection which is why we also see Mtb has evolved genes to counter host autophagy responses (Golovkine et al., 2023). Overexpressing BMAL1 has been observed to promote autophagy (Li et al., 2022). E2F transcription factors such as E2F4 and E2F1 are also known to enhance autophagy with an emerging role in DNA damage induced autophagy (Haim et al., 2015; Polager et al., 2008; Xiao et al., 2022). Interestingly, Mtb infection of macrophages is also known to induce DNA damage (Castro-Garza et al., 2018). Differential regulation of autophagy can therefore offer a plausible explanation for the RSTR phenotype.

Recent findings confirmed that autophagy offers protection against Mtb during the initial stages of infection, though emphasizing the complete mechanism remains undefined. Since its identification in

2004 as part of the innate immune response to Mtb, the exact role of autophagy has been ambiguous, due in part to its complex nature involving numerous genes in both canonical and noncanonical pathways (Aylan et al., 2023; Deretic and Wang, 2023; Golovkine et al., 2023; Typas, 2023). Finding an ideal single gene or isolated subnetwork to target for effective Mtb control continues to be challenging. However, the roles of BMAL1 and E2F in enhancing autophagy, coupled with their differential expression in alveolar macrophages from RSTRs, suggest they might be important in regulating autophagy for early clearance of Mtb infection.

Bmal1 is associated with other infection outcomes, underscoring its potential role in infectious disease progression, increasing the likelihood it plays a role in Mtb infection as well. Circadian rhythms are shown to have a large influence on the severity and progression of various infectious diseases, such as Streptococcus pneumoniae and Helicobacter pylori. In the context of Mtb infection for example, patients showed increased cough frequency during the daytime, linked to peak sputum bacillary loads. (Diallo et al., 2020).

As BMAL1 is involved in regulating biological clocks, this suggests that elements of host chronobiology may be important to understanding the RSTR phenotype. Rapamycin, known to enhance macrophage response via autophagy by inhibiting Mtb growth, also inhibits mTOR which is a pathway suppressed by BMAL1 overexpression showing up in RSTRs (Gutierrez et al., 2004; Khapre et al., 2014; Li et al., 2023; Singla et al., 2022). This inhibition by rapamycin has been linked to extended lifespan in various species, possibly due to the deceleration of cellular biological time (Blagosklonny, 2018; Harrison et al., 2009). Individual cells might have their own autonomous clocks (Mofatteh et al., 2021). Given BMAL1's expression patterns, we could speculate that RSTR alveolar macrophages might have desynchronized autonomous clocks, potentially decelerating cellular activities in relation to their surrounding environment.

After contact with Mtb, alveolar macrophages coordinate and transfer information to the adaptive immune system by presenting processed Mtb antigens via MHC-II and secrete IL-12, which is required to activate T cells (Harding and Boom, 2010; Matucci et al., 2014). Interestingly, autophagy has been shown to coordinate the delivery of microbial antigens to MHC class II for CD4+ T cell presentation (Crotzer and Blum, 2009). Activation of CD4+ T cells initiates a feedback loop when T cells release their own cytokines, further activating alveolar macrophages, promoting T cell and B cell proliferation, NK cell activity, and other signals like interferon-gamma initiating broad cellular responses to coordinate countermeasures against Mtb (O'Garra et al., 2013). Recently, it has been shown in the Uganda cohort that even though testing negative for TST and IGRA, RSTRs display evidence of non-canonical T cell

mediated cellular immunity to Mtb, suggesting there is some cryptic coordination with the adaptive immune system (Lu et al., 2019). Considering this interplay, the differential regulation of autophagy within RSTR alveolar macrophages could offer a unique window for these cells to process and respond to Mtb relative to LTBI.

Understanding the resistance mechanism could quickly lead to implementing therapeutics. For example, previous work identified NRAMP1 as a resistance gene in cows protecting from mycobacterium bovis (Mb) infection (Gao et al., 2017). Using clustered regularly interspaced short palindromic repeats (CRISPR), the authors inserted an additional copy of the NRAMP1gene in the cow germline genome. Transgenic cows were confirmed to be more resistant to Mb. Would we see a similar increase in Mtb infection resistance if we edited in additional BMAL1 or E2F genes in the human genome under the control of a cell-type-specific promoter, limiting the expression of the additional gene only to alveolar macrophages? Our analysis allows for asking new and interesting questions to test novel theories of Mtb resistance and guide the proper experimental setup.

Limitations

Main strength of the study comes from a very well-defined and characterized longitudinal cohort to minimize misclassification. It is important to consider that TST and IGRA relies on detecting Mtb specific T cells which could be residing in other non-accessible tissues leading to misclassification. Generally, it does not follow best scientific practices to rerun analysis multiple times, changing sample size or other parameters and keep results showing most statistical significance (Motulsky, 2014). For example, using more stringent filtering changing minimum CPM to 1 from 0.5 and min samples from 3 to 5 it resulted in an additional few hundred genes filtered out. More importantly, downstream analysis of this data set found an additional gene OAS1 up in LTBI (FDR <0.3) as well as GSEA picked up an additional Hallmark Hedgehog signaling pathway up in RSTR and Hallmark Apoptosis up in LTBI (FDR <0.3). This shows that having misclassification could potentially alter the results, however if a signal is robust enough it will still be present even with small changes in the data. Though more stringent filtering picked up some new genes and pathways, E2F and BMAL1 signals proved to be robust showing up enriched in the RSTR group as expected proving further validation the signal is significant and would very likely be robust to minor misclassification.

We used bulk RNA sequencing to provide an average gene expression profile across all alveolar macrophages in individual samples. While this method captures average gene expression, it overlooks the diverse subtypes of alveolar macrophages, including novel and rare cells that may be responsible for

coordinating resistance to Mtb (Kukurba and Montgomery, 2015; Li and Wang, 2021). On the other hand, newer and costlier techniques like single-cell RNA sequencing (scRNA-seq) makes it possible to separate and characterize alveolar macrophage subtypes. This allows for comparing gene expression across various cell subtypes and detecting rare and potentially important alveolar macrophage cells.

Co-evolving alongside humans over thousands of years, Mtb has evolved into multiple lineages with global circulation. Lineages vary in their virulence and other characteristics such as drug resistance (Allué-Guardia et al., 2021; Phelan et al., 2023; Saelens et al., 2019). Importantly, we do not have data on which specific Mtb lineages infected the index cases upon recruitment or good coverage of circulating lineages in Uganda over time. This is limiting in our analyses as the RSTR phenotype might only be resistant to lineages in Uganda and susceptible to lineages outside of Uganda whereas ideally, we want to identify global resistance to Mtb infection across all Mtb lineages. Specific to Uganda, there is some evidence of a wide diversity of Mtb lineages in circulation over time (Bazira et al., 2010; Micheni et al., 2021; Saifodine et al., 2014).

Conclusion

Our research findings highlight genes and biological pathways unique to RSTR alveolar macrophages, indicating biological signatures of TB disease resistance point to autophagic differences involving BMAL1 and E2F. Further studies, such as infecting RSTR alveolar macrophages with virulent Mtb strains and scRNA-seq analysis, can provide additional missing puzzle pieces in support or rejection of the relevance of autophagic signatures in explaining resistance to Mtb infection. Autophagic and circadian clock resistance mechanisms suggested by this work warrant further study.

References

Ahmad F, Rani A, Alam A, Zarin S, Pandey S, Singh H, Hasnain SE, Ehtesham NZ. 2022. Macrophage: A Cell With Many Faces and Functions in Tuberculosis. *Front Immunol* **13**:747799. doi:10.3389/fimmu.2022.747799

Allué-Guardia A, García JI, Torrelles JB. 2021. Evolution of Drug-Resistant Mycobacterium tuberculosis Strains and Their Adaptation to the Human Lung Environment. *Front Microbiol* **12**.

Andrews S. 2010. FASTQC. A quality control tool for high throughput sequence data.

Arentz M, Hawn TR. 2007. Tuberculosis Infection: Insight from Immunogenomics. *Drug Discov Today Dis Mech* **4**:231–236. doi:10.1016/j.ddmec.2007.11.003

Aylan B, Bernard EM, Pellegrino E, Botella L, Fearns A, Athanasiadi N, Bussi C, Santucci P, Gutierrez MG. 2023. ATG7 and ATG14 restrict cytosolic and phagosomal Mycobacterium tuberculosis replication in human macrophages. *Nat Microbiol* **8**:803–818. doi:10.1038/s41564-023-01335-9

Bazira J, Matte M, Asiimwe B, Joloba L. 2010. Genetic diversity of Mycobacterium tuberculosis in Mbarara, South Western Uganda. *Afr Health Sci* **10**:306–311.

Blagosklonny MV. 2018. Does rapamycin slow down time? *Oncotarget* **9**:30210–30212. doi:10.18632/oncotarget.25788

Bruchfeld J, Correia-Neves M, Källenius G. 2015. Tuberculosis and HIV Coinfection. *Cold Spring Harb Perspect Med* **5**:a017871. doi:10.1101/cshperspect.a017871

Castro-Garza J, Luévano-Martínez ML, Villarreal-Treviño L, Gosálvez J, Fernández JL, Dávila-Rodríguez MI, García-Vielma C, González-Hernández S, Cortés-Gutiérrez EI. 2018. Mycobacterium tuberculosis promotes genomic instability in macrophages. *Mem Inst Oswaldo Cruz* **113**:161–166. doi:10.1590/0074-02760170281

Crotzer VL, Blum JS. 2009. Autophagy and its role in MHC-mediated antigen presentation. *J Immunol Baltim Md 1950* **182**:3335–3341. doi:10.4049/jimmunol.0803458

Deretic V, Lazarou M. 2022. A guide to membrane atg8ylation and autophagy with reflections on immunity. *J Cell Biol* **221**:e202203083. doi:10.1083/jcb.202203083

Deretic V, Wang F. 2023. Autophagy is part of the answer to tuberculosis. *Nat Microbiol* **8**:762–763. doi:10.1038/s41564-023-01373-3

Diallo AB, Coiffard B, Leone M, Mezouar S, Mege J-L. 2020. For Whom the Clock Ticks: Clinical Chronobiology for Infectious Diseases. *Front Immunol* **11**:1457. doi:10.3389/fimmu.2020.01457

Dill-McFarland K, Benson B, rmsegnitz. 2022. BIGslu/SEAsnake: v1.0. doi:10.5281/zenodo.6471546

Dill-McFarland KA, Mitchell K, Batchu S, Segnitz RM, Benson B, Janczyk T, Cox MS, Mayanja-Kizza H, Boom WH, Benchek P, Stein CM, Hawn TR, Altman MC. 2023. Kimma: flexible linear mixed effects modeling with kinship covariance for RNA-seq data. *Bioinformatics* **39**:btad279. doi:10.1093/bioinformatics/btad279

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. doi:10.1093/bioinformatics/bts635

Dodd PJ, Shaweno D, Ku C-C, Glaziou P, Pretorius C, Hayes RJ, MacPherson P, Cohen T, Ayles H. 2023. Transmission modeling to infer tuberculosis incidence prevalence and mortality in settings with generalized HIV epidemics. *Nat Commun* **14**:1639. doi:10.1038/s41467-023-37314-1

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**:1184–1191. doi:10.1038/nprot.2009.97

Gao Y, Wu H, Wang Y, Liu Xin, Chen L, Li Q, Cui C, Liu Xu, Zhang J, Zhang Y. 2017. Single Cas9 nickase induced generation of NRAMP1 knockin cattle with reduced off-target effects. *Genome Biol* **18**:13. doi:10.1186/s13059-016-1144-4

Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, Rice KM, Conomos MP. 2019. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**:5346–5348. doi:10.1093/bioinformatics/btz567

Golovkine GR, Roberts AW, Morrison HM, Rivera-Lugo R, McCall RM, Nilsson H, Garelis NE, Repasy T, Cronce M, Budzik J, Van Dis E, Popov LM, Mitchell G, Zalpuri R, Jorgens D, Cox JS. 2023. Autophagy restricts Mycobacterium tuberculosis during acute infection in mice. *Nat Microbiol* **8**:819–832. doi:10.1038/s41564-023-01354-6

Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, Bouman H, Abascal F, Haber M, Tachmazidou I, Mathieson I, Ekoru K, DeGorter MK, Nsubuga RN, Finan C, Wheeler E, Chen L, Cooper DN, Schiffels S, Chen Y, Ritchie GRS, Pollard MO, Fortune MD, Mentzer AJ, Garrison E, Bergström A, Hatzikotoulas K, Adeyemo A, Doumatey A, Elding H, Wain LV, Ehret G, Auer PL, Kooperberg CL, Reiner AP, Franceschini N, Maher D, Montgomery SB, Kadie C, Widmer C, Xue Y, Seeley J, Asiki G, Kamali A, Young EH, Pomilla C, Soranzo N, Zeggini E, Pirie F, Morris AP, Heckerman D, Tyler-Smith C, Motala AA, Rotimi C, Kaleebu P, Barroso I, Sandhu MS. 2019. Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**:984-1002.e36. doi:10.1016/j.cell.2019.10.004

Gutierrez MG, Master SS, Singh SB, Taylor GA, Colombo MI, Deretic V. 2004. Autophagy is a defense mechanism inhibiting BCG and Mycobacterium tuberculosis survival in infected macrophages. *Cell* **119**:753–766. doi:10.1016/j.cell.2004.11.038

Haim Y, Blüher M, Slutsky N, Goldstein N, Klöting N, Harman-Boehm I, Kirshtein B, Ginsberg D, Gericke M, Guiu Jurado E, Kovsan J, Tarnovscki T, Kachko L, Bashan N, Gepner Y, Shai I, Rudich A. 2015. Elevated autophagy gene expression in adipose tissue of obese humans: A potential non-cell-cycle-dependent function of E2F1. *Autophagy* **11**:2074–2088. doi:10.1080/15548627.2015.1094597

Harding CV, Boom WH. 2010. Regulation of antigen presentation by Mycobacterium tuberculosis: a role for Toll-like receptors. *Nat Rev Microbiol* **8**:296–307. doi:10.1038/nrmicro2321

Harrison DE, Strong R, Sharp ZD, Nelson JF, Astle CM, Flurkey K, Nadon NL, Wilkinson JE, Frenkel K, Carter CS, Pahor M, Javors MA, Fernandez E, Miller RA. 2009. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**:392–395. doi:10.1038/nature08221

Hawn TR, Shah JA, Kalman D. 2015. New tricks for old dogs: countering antibiotic resistance in tuberculosis with host-directed therapeutics. *Immunol Rev* **264**:344–362. doi:10.1111/imr.12255

Jiang H, Garcia V, Yanum JA, Lee J, Dai G. 2021. Circadian clock core component Bmal1 dictates cell cycle rhythm of proliferating hepatocytes during liver regeneration. *Am J Physiol-Gastrointest Liver Physiol* **321**:G389–G399. doi:10.1152/ajpgi.00204.2021

Joshi N, Walter JM, Misharin AV. 2018. Alveolar Macrophages. *Cell Immunol* **330**:86–90. doi:10.1016/j.cellimm.2018.01.005

Khapre RV, Kondratova AA, Patel S, Dubrovsky Y, Wrobel M, Antoch MP, Kondratov RV. 2014. BMAL1-dependent regulation of the mTOR signaling pathway delays aging. *Aging* **6**:48–57. doi:10.18632/aging.100633

Kukurba KR, Montgomery SB. 2015. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* **2015**:951–969. doi:10.1101/pdb.top084970

Kwok AJ, Mentzer A, Knight JC. 2021. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nat Rev Genet* **22**:137–153. doi:10.1038/s41576-020-00297-6

Lee J, Park E, Kim GH, Kwon I, Kim K. 2018. A splice variant of human Bmal1 acts as a negative regulator of the molecular circadian clock. *Exp Mol Med* **50**:159. doi:10.1038/s12276-018-0187-x

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. doi:10.1093/bioinformatics/btp352

Li H, Li M, Chen K, Li Y, Yang Z, Zhou Z. 2022. The circadian clock gene ARNTL overexpression suppresses oral cancer progression by inducing apoptosis via activating autophagy. *Med Oncol* **39**:244. doi:10.1007/s12032-022-01832-7

Li H, Meng H, Xu M, Gao X, Sun X, Jin X, Sun H. 2023. BMAL1 regulates osteoblast differentiation through mTOR/GSK3β/β-catenin pathway. *J Mol Endocrinol* **70**:e220181. doi:10.1530/JME-22-0181

Li X, Wang C-Y. 2021. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci* **13**:1–6. doi:10.1038/s41368-021-00146-0

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**:923–930. doi:10.1093/bioinformatics/btt656

Lu LL, Smith MT, Yu KKQ, Luedemann C, Suscovich TJ, Grace PS, Cain A, Yu WH, McKitrick TR, Lauffenburger D, Cummings RD, Mayanja-Kizza H, Hawn TR, Boom WH, Stein CM, Fortune SM, Seshadri C, Alter G. 2019. IFN-γ-independent immune markers of Mycobacterium tuberculosis exposure. *Nat Med* **25**:977–987. doi:10.1038/s41591-019-0441-3

Ma N, Zalwango S, Malone LL, Nsereko M, Wampande EM, Thiel BA, Okware B, Igo RP, Joloba ML, Mupere E, Mayanja-Kizza H, Boom WH, Stein CM, for the Tuberculosis Research Unit (TBRU). 2014. Clinical and epidemiological characteristics of individuals resistant to M. tuberculosis infection in a longitudinal TB household contact study in Kampala, Uganda. *BMC Infect Dis* **14**:352. doi:10.1186/1471-2334-14-352

Matucci A, Maggi E, Vultaggio A. 2014. Cellular and Humoral Immune Responses During Tuberculosis Infection: Useful Knowledge in the Era of Biological Agents. *J Rheumatol Suppl* **91**:17–23. doi:10.3899/jrheum.140098

McHenry ML, Benchek P, Malone L, Nsereko M, Mayanja-Kizza H, Boom WH, Williams SM, Hawn TR, Stein CM. 2021. Resistance to TST/IGRA conversion in Uganda: Heritability and Genome-Wide Association Study. *eBioMedicine* **74**:103727. doi:10.1016/j.ebiom.2021.103727

McHenry ML, Williams SM, Stein CM. 2020. Genetics and evolution of tuberculosis pathogenesis: New perspectives and approaches. *Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis* **81**:104204. doi:10.1016/j.meegid.2020.104204

McLaren PJ, Porreca I, Iaconis G, Mok HP, Mukhopadhyay S, Karakoc E, Cristinelli S, Pomilla C, Bartha I, Thorball CW, Tough RH, Angelino P, Kiar CS, Carstensen T, Fatumo S, Porter T, Jarvis I, Skarnes WC, Bassett A, DeGorter MK, Sathya Moorthy MP, Tuff JF, Kim E-Y, Walter M, Simons LM, Bashirova A, Buchbinder S, Carrington M, Cossarizza A, De Luca A, Goedert JJ, Goldstein DB, Haas DW, Herbeck JT, Johnson EO, Kaleebu P, Kilembe W, Kirk GD, Kootstra NA, Kral AH, Lambotte O, Luo M, Mallal S, Martinez-Picado J, Meyer L, Miro JM, Moodley P, Motala AA, Mullins JI, Nam K, Obel N, Pirie F, Plummer FA, Poli G, Price MA, Rauch A, Theodorou I, Trkola A, Walker BD, Winkler CA, Zagury J-F, Montgomery SB, Ciuffi A, Hultquist JF, Wolinsky SM, Dougan G, Lever AML, Gurdasani D, Groom H, Sandhu MS, Fellay J. 2023. Africa-specific human genetic variation near CHD1L associates with HIV-1 load. *Nature*. doi:10.1038/s41586-023-06370-4

Micheni LN, Kassaza K, Kinyi H, Ntulume I, Bazira J. 2021. Diversity of Mycobacterium tuberculosis Complex Lineages Associated with Pulmonary Tuberculosis in Southwestern, Uganda. *Tuberc Res Treat* **2021**:5588339. doi:10.1155/2021/5588339

Mofatteh M, Echegaray-Iturra F, Alamban A, Dalla Ricca F, Bakshi A, Aydogan MG. 2021. Autonomous clocks that regulate organelle biogenesis, cytoskeletal organization, and intracellular dynamics. *eLife* **10**:e72104. doi:10.7554/eLife.72104

Motulsky HJ. 2014. Common Misconceptions about Data Analysis and Statistics. *J Pharmacol Exp Ther* **351**:200–205. doi:10.1124/jpet.114.219170

Ochoa D, Karim M, Ghoussaini Maya, Hulcoop DG, McDonagh EM, Dunham I. 2022. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nat Rev Drug Discov* **21**:551–551. doi:10.1038/d41573-022-00120-3

O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. 2013. The Immune Response in Tuberculosis. *Annu Rev Immunol* **31**:475–527. doi:10.1146/annurev-immunol-032712-095939

Orenstein WA, Offit PA, Edwards KM, Plotkin SA. 2023. Plotkin's Vaccines. Elsevier.

Panarella M, Burkett KM. 2019. A Cautionary Note on the Effects of Population Stratification Under an Extreme Phenotype Sampling Design. *Front Genet* **10**.

Phelan J, Gomez-Gonzalez PJ, Andreu N, Omae Y, Toyo-Oka L, Yanai H, Miyahara R, Nedsuwan S, de Sessions PF, Campino S, Sallah N, Parkhill J, Smittipat N, Palittapongarnpim P, Mushiroda T, Kubo M, Tokunaga K, Mahasirimongkol S, Hibberd ML, Clark TG. 2023. Genome-wide host-pathogen analyses reveal genetic interaction points in tuberculosis disease. *Nat Commun* **14**:549. doi:10.1038/s41467-023-36282-w

Picard toolkit. 2019. . *Broad Inst GitHub Repos*.

Polager S, Ofir M, Ginsberg D. 2008. E2F1 regulates autophagy and the transcription of autophagy genes. *Oncogene* **27**:4860–4864. doi:10.1038/onc.2008.117

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**:559–575.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140. doi:10.1093/bioinformatics/btp616

Saelens JW, Viswanathan G, Tobin DM. 2019. Mycobacterial Evolution Intersects With Host Tolerance. *Front Immunol* **10**:528. doi:10.3389/fimmu.2019.00528

Saifodine A, Fyfe J, Sievers A, Coelho E, Azam K, Black J. 2014. Genetic diversity of Mycobacterium tuberculosis isolates obtained from patients with pulmonary tuberculosis in Beira city, Mozambique. *Int J Mycobacteriology* **3**:94–100. doi:10.1016/j.ijmyco.2014.03.004

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* **9**:88. doi:10.1186/s13104-016-1900-2

Shendure J, Findlay GM, Snyder MW. 2019. Genomic Medicine–Progress, Pitfalls, and Promise. *Cell* **177**:45–57. doi:10.1016/j.cell.2019.02.003

Simmons JD, Dill-McFarland KA, Stein CM, Van PT, Chihota V, Ntshiqa T, Maenetje P, Peterson GJ, Benchek P, Nsereko M, Velen K, Fielding KL, Grant AD, Gottardo R, Mayanja-Kizza H, Wallis RS, Churchyard G, Boom WH, Hawn TR. 2022. Monocyte Transcriptional Responses to Mycobacterium tuberculosis Associate with Resistance to Tuberculin Skin Test and Interferon Gamma Release Assay Conversion. *mSphere* **7**:e00159-22. doi:10.1128/msphere.00159-22

Simmons JD, Stein CM, Seshadri C, Campo M, Alter G, Fortune S, Schurr E, Wallis RS, Churchyard G, Mayanja-Kizza H, Boom WH, Hawn TR. 2018. Immunological mechanisms of human resistance to persistent Mycobacterium tuberculosis infection. *Nat Rev Immunol* **18**:575–589. doi:10.1038/s41577-018-0025-3

Simmons JD, Van PT, Stein CM, Chihota V, Ntshiqa T, Maenetje P, Peterson GJ, Reynolds A, Benchek P, Velen K, Fielding KL, Grant AD, Graustein AD, Nguyen FK, Seshadri C, Gottardo R, Mayanja-Kizza H, Wallis RS, Churchyard G, Boom WH, Hawn TR. 2021. Monocyte metabolic transcriptional programs associate with resistance to tuberculin skin test/interferon-γ release assay conversion. *J Clin Invest* **131**:e140073. doi:10.1172/JCI140073

Singla R, Mishra A, Lin H, Lorsung E, Le N, Tin S, Jin VX, Cao R. 2022. Haploinsufficiency of a Circadian Clock Gene Bmal1 (Arntl or Mop3) Causes Brain-Wide mTOR Hyperactivation and Autism-like Behavioral Phenotypes in Mice. *Int J Mol Sci* **23**:6317. doi:10.3390/ijms23116317

Stein CM. 2023. Genetic epidemiology of resistance to M. tuberculosis Infection: importance of study design and recent findings. *Genes Immun* **24**:117–123. doi:10.1038/s41435-023-00204-z

Stein CM, Mayanja-Kizza H, Hawn TR, Boom WH. 2020. Importance of Study Design and Phenotype Definition in Ongoing Studies of Resistance to Latent Mycobacterium tuberculosis Infection. *J Infect Dis* **221**:1025–1026. doi:10.1093/infdis/jiz539

Stein CM, Nsereko M, Malone LL, Okware B, Kisingo H, Nalukwago S, Chervenak K, Mayanja-Kizza H, Hawn TR, Boom WH. 2019. Long-term Stability of Resistance to Latent Mycobacterium tuberculosis Infection in Highly Exposed Tuberculosis Household Contacts in Kampala, Uganda. *Clin Infect Dis* **68**:1705–1712. doi:10.1093/cid/ciy751

Stein CM, Zalwango S, Malone LL, Thiel B, Mupere E, Nsereko M, Okware B, Kisingo H, Lancioni CL, Bark CM, Whalen CC, Joloba ML, Boom WH, Mayanja-Kizza H. 2018. Resistance and Susceptibility to Mycobacterium tuberculosis Infection and Disease in Tuberculosis Households in Kampala, Uganda. *Am J Epidemiol* **187**:1477–1489. doi:10.1093/aje/kwx380

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**:D447-452. doi:10.1093/nar/gku1003

Thuong NTT, Dunstan SJ, Chau TTH, Thorsson V, Simmons CP, Quyen NTH, Thwaites GE, Thi Ngoc Lan N, Hibberd M, Teo YY, Seielstad M, Aderem A, Farrar JJ, Hawn TR. 2008. Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. *PLoS Pathog* **4**:e1000229. doi:10.1371/journal.ppat.1000229

Thuong NTT, Hawn TR, Thwaites GE, Chau TTH, Lan NTN, Quy HT, Hieu NT, Aderem A, Hien TT, Farrar JJ, Dunstan SJ. 2007. A polymorphism in human TLR2 is associated with increased susceptibility to tuberculous meningitis. *Genes Immun* **8**:422–428. doi:10.1038/sj.gene.6364405

Thurlings I, de Bruin A. 2016. E2F Transcription Factors Control the Roller Coaster Ride of Cell Cycle Gene Expression. *Methods Mol Biol Clifton NJ* **1342**:71–88. doi:10.1007/978-1-4939-2957-3_4

Tobin DM, Roca FJ, Oh SF, McFarland R, Vickery TW, Ray JP, Ko DC, Zou Y, Bang ND, Chau TTH, Vary JC, Hawn TR, Dunstan SJ, Farrar JJ, Thwaites GE, King M-C, Serhan CN, Ramakrishnan L. 2012. Host genotype-specific therapies can optimize the inflammatory response to mycobacterial infections. *Cell* **148**:434–446. doi:10.1016/j.cell.2011.12.023

Typas D. 2023. Autophagy counteracts Mycobacterium tuberculosis infection at early stages. *Nat Struct Mol Biol* **30**:720–720. doi:10.1038/s41594-023-01024-5

Xiao W, Wang J, Wang X, Cai S, Guo Y, Ye L, Li D, Hu A, Jin S, Yuan B, Zhou Y, Li Q, Tong Q, Zheng L. 2022. Therapeutic targeting of the USP2-E2F4 axis inhibits autophagic machinery essential for zinc homeostasis in cancer progression. *Autophagy* **18**:2615–2635. doi:10.1080/15548627.2022.2044651

Zheng R, Li Z, He F, Liu H, Chen Jianhua, Chen Jiayu, Xie X, Zhou J, Chen H, Wu X, Wu J, Chen B, Liu Yahui, Cui H, Fan L, Sha W, Liu Yin, Wang Jiqiang, Huang X, Zhang L, Xu F, Wang Jie, Feng Y, Qin L, Yang H, Liu Z, Cui Z, Liu F, Chen X, Gao S, Sun S, Shi Y, Ge B. 2018. Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nat Commun* **9**:4072. doi:10.1038/s41467-018-06539-w

	RSTR	LTBI	p-value[A]
No. subjects included	23	26	
Age at enrollment, mean (IQR)	19.5 (3)	24.9 (6)	0.0064
Sex, %male (n/N)	59 (13/22)	50 (13/26)	0.735
BMI (mean, IQR)	21.3 (4.63)	22.1 (2.49)	0.56
Exposure risk score, mean (SD)	5.91(1.33)	6.38(1.33)	0.15

[A] Statistical comparison made using Pearson Chi-Square for categorial variables or two-sample Wilcoxon rank-sum (Mann-Whitney) tests for continues variables.

Table 1. Epidemiologic and demographic description of Uganda alveolar macrophage donors.

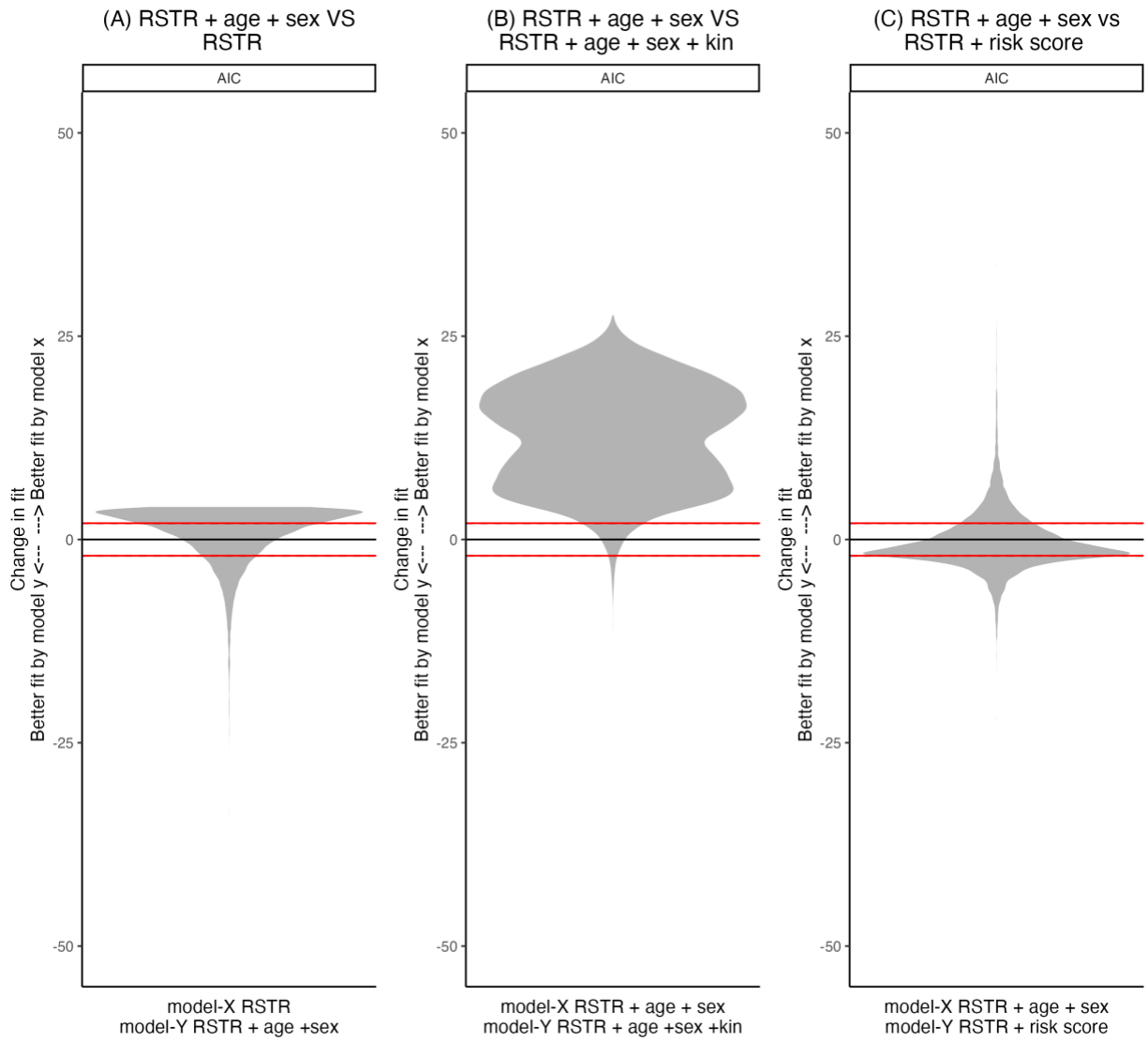


Figure 1. Statistical linear model fit was assessed by AIC. Red line indicates minimal change in AIC (-2,2).

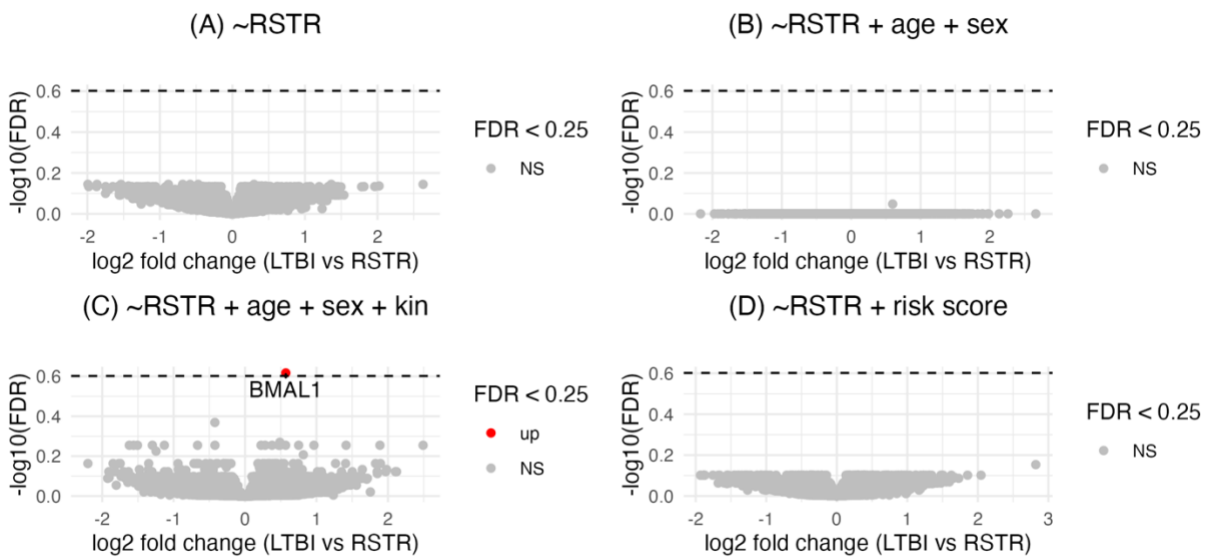


Figure 2. Volcano plots showing DEGs across four different statistical models in panels A – D. Only one DEG was statistically significant shown in panel C (FDR <0.25).

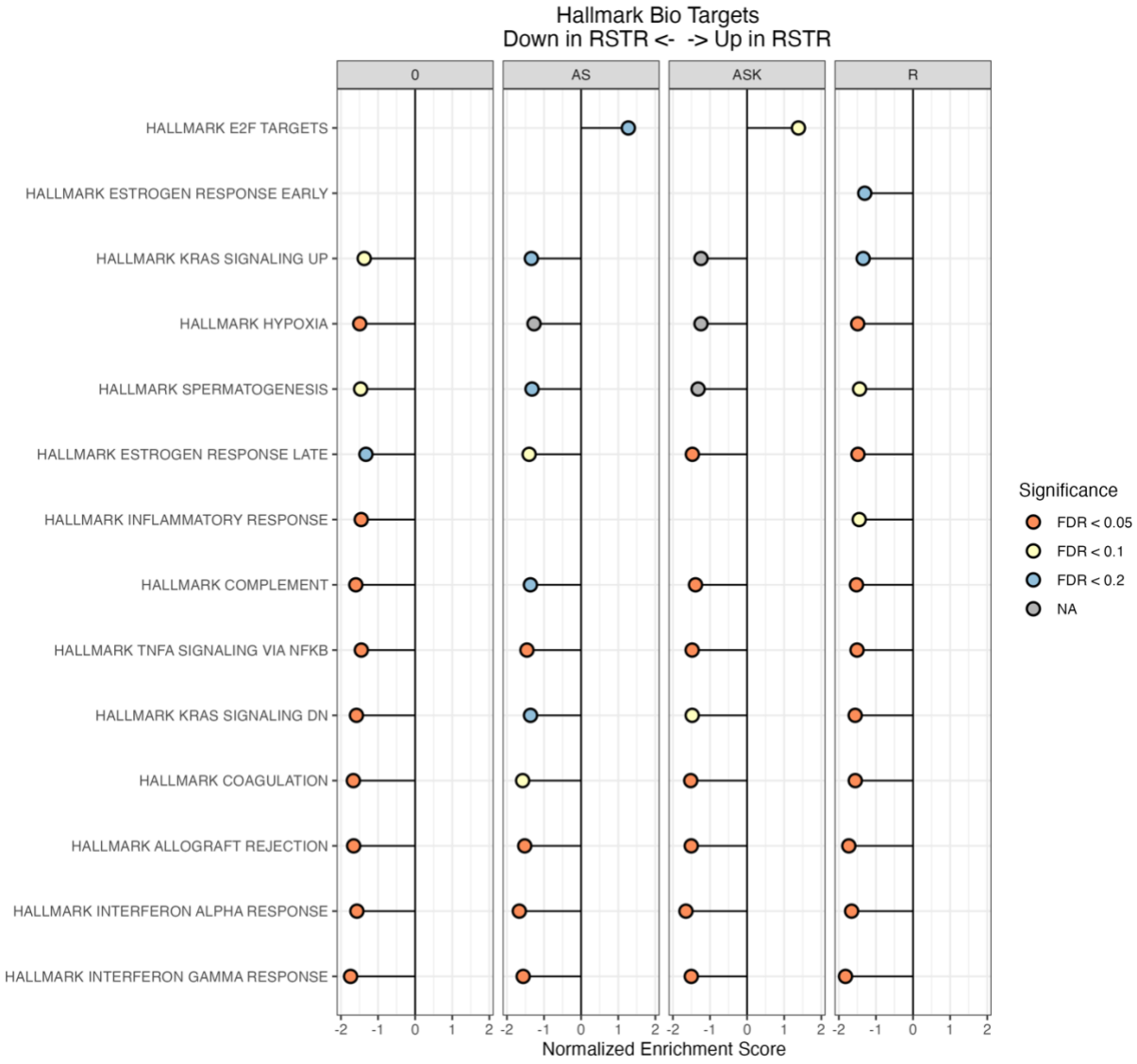


Figure 3. Gene set enrichment analysis (GSEA) identified pathways unique to RSTRs using fold changes across all genes. For each pathway, there is a normalized enrichment score (NES) plotted showing direction of enrichment and node color indicates significance. Each panel 0, AS, ASK, R indicate the four statistical models where (1) 0: ~RSTR, (2) AS: ~RSTR + Age + Sex, (3) ASK: ~RSTR + Age + Sex + Kin, and (4) R: ~RSTR + Risk Score.

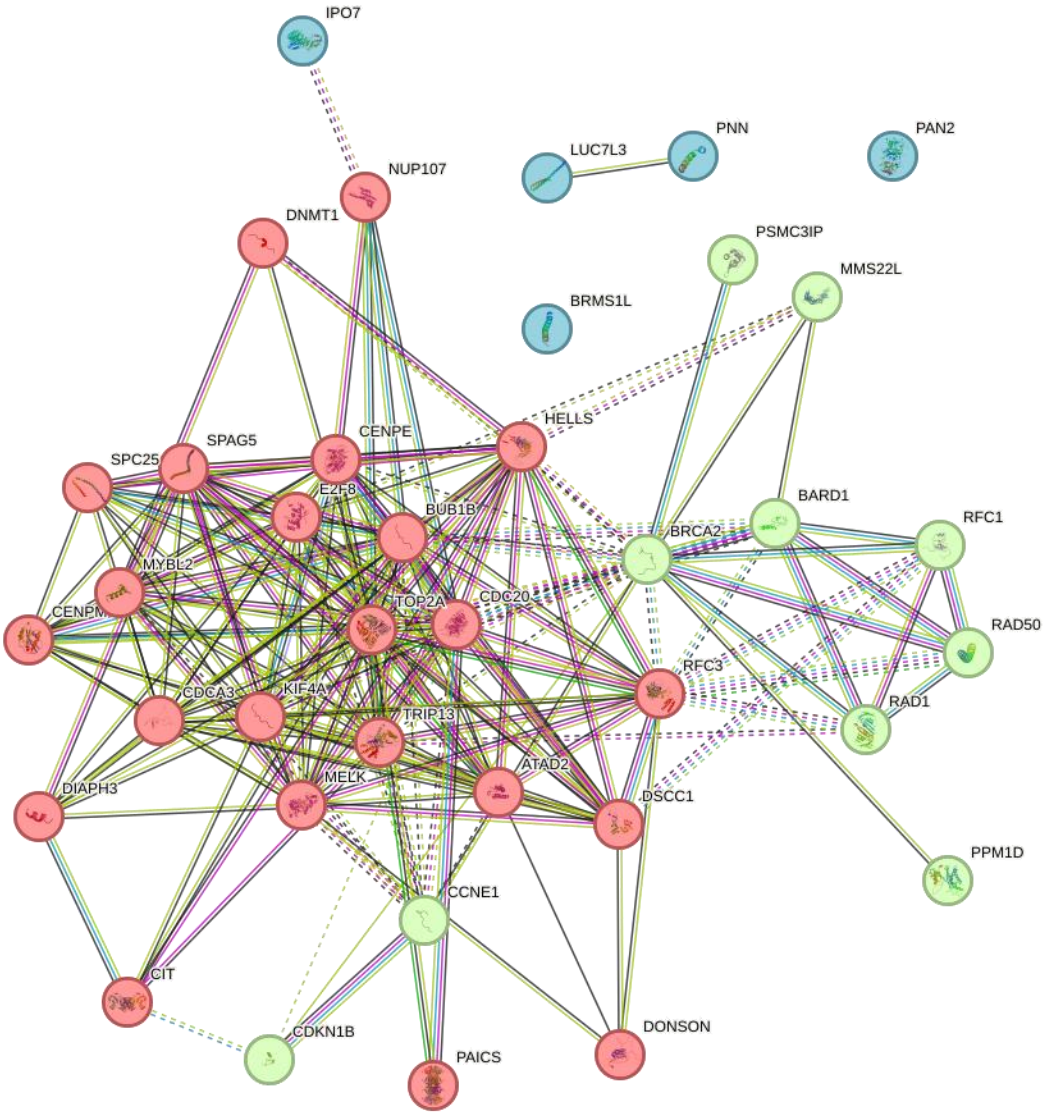
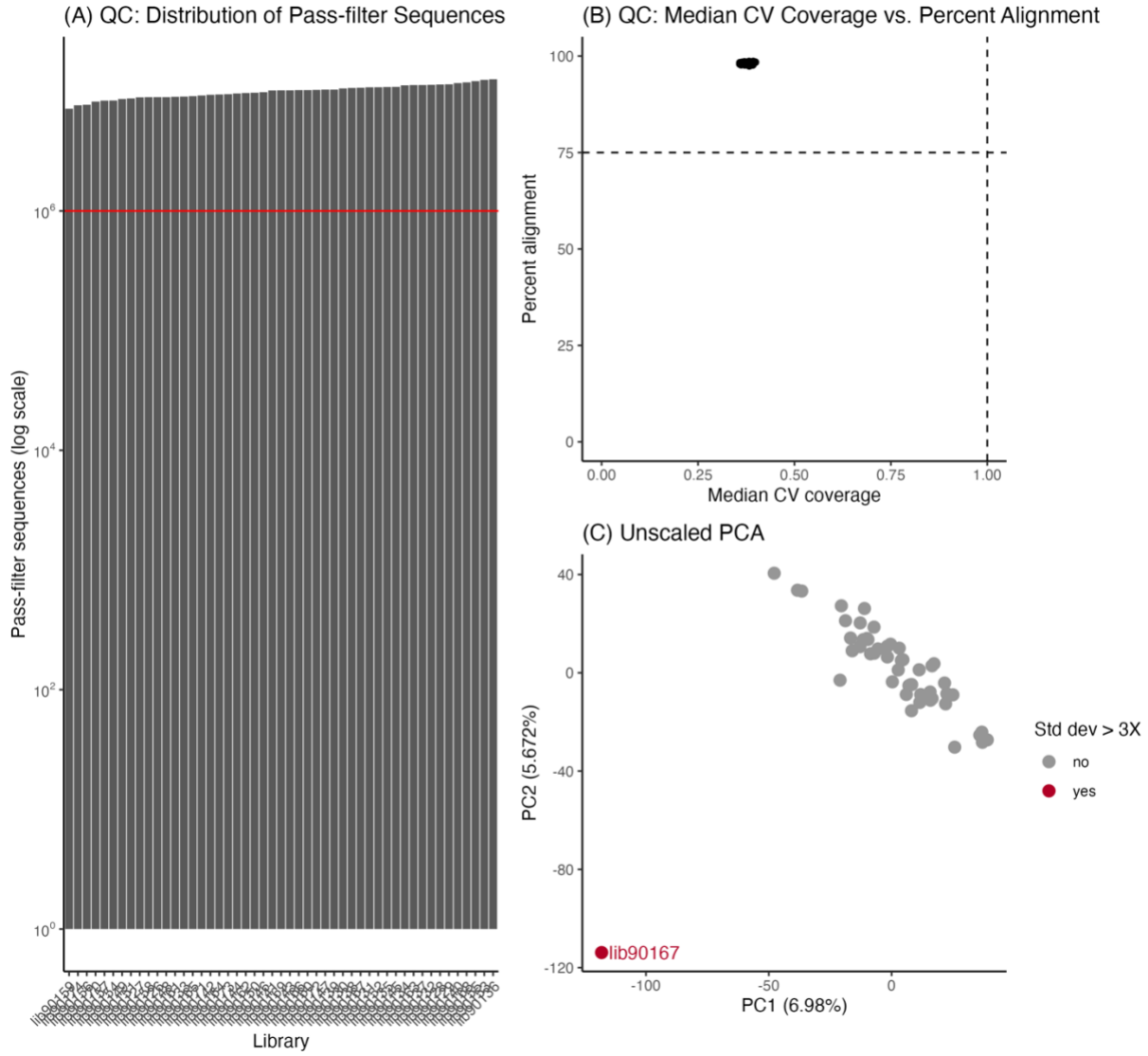


Figure 4. E2F leading edge gene network shows top 38 genes driving the E2F enrichment in RSTRs. K-means clustering shows three separate subclusters color coded and grouped by predicted functional similarity.

Supporting Information

Figure S1. Combined QC metrics indicate high quality measurements. (A) All libraries pass high total sequences. (B) All libraries pass high percent alignment and low CV coverage. (C) PCA identifies one RSTR library as an outlier removed from downstream analyses.



S2. E2F biological targets consist of about 200 genes. Of the 200, leading edge gene analysis identified 39 genes mainly responsible for driving E2F signal with highest fold changes in RSTRs annotated using StringDB.

GENE	Function Summary
ATAD2	ATPase family AAA domain-containing protein 2; May be a transcriptional coactivator of the nuclear receptor ESR1 required to induce the expression of a subset of estradiol target genes, such as CCND1, MYC and E2F1. May play a role in the recruitment or occupancy of CREBBP at some ESR1 target gene promoters. May be required for histone hyperacetylation. Involved in the estrogen-induced cell proliferation and cell cycle progression of breast cancer cells.
BARD1	BRCA1-associated RING domain protein 1; E3 ubiquitin-protein ligase. The BRCA1-BARD1 heterodimer specifically mediates the formation of 'Lys-6'-linked polyubiquitin chains and coordinates a diverse range of cellular pathways such as DNA damage repair, ubiquitination and transcriptional regulation to maintain genomic stability. Plays a central role in the control of the cell cycle in response to DNA damage. Acts by mediating ubiquitin E3 ligase activity that is required for its tumor suppressor function. Also forms a heterodimer with CSTF1/CSTF-50 to modulate mRNA processing and RNAP II [...]
BRCA2	Breast cancer type 2 susceptibility protein; Involved in double-strand break repair and/or homologous recombination. Binds RAD51 and potentiates recombinational DNA repair by promoting assembly of RAD51 onto single-stranded DNA (ssDNA). Acts by targeting RAD51 to ssDNA over double-stranded DNA, enabling RAD51 to displace replication protein-A (RPA) from ssDNA and stabilizing RAD51- ssDNA filaments by blocking ATP hydrolysis. Part of a PALB2-scaffolded HR complex containing RAD51C and which is thought to play a role in DNA repair by HR. May participate in S phase checkpoint activation. [...]
BRMS1L	Breast cancer metastasis-suppressor 1-like protein; Involved in the histone deacetylase (HDAC1)-dependent transcriptional repression activity. When overexpressed in lung cancer cell line that lacks p53/TP53 expression, inhibits cell growth.
BUB1B	Mitotic checkpoint serine/threonine-protein kinase BUB1 beta; Essential component of the mitotic checkpoint. Required for normal mitosis progression. The mitotic checkpoint delays anaphase until all chromosomes are properly attached to the mitotic spindle. One of its checkpoint functions is to inhibit the activity of the anaphase- promoting complex/cyclosome (APC/C) by blocking the binding of CDC20 to APC/C, independently of its kinase activity. The other is to monitor kinetochore activities that depend on the kinetochore motor CENPE. Required for kinetochore localization of CENPE. Neg [...]
CCND1	G1/S-specific cyclin-D1; Regulatory component of the cyclin D1-CDK4 (DC) complex that phosphorylates and inhibits members of the retinoblastoma (RB) protein family including RB1 and regulates the cell-cycle during G(1)/S transition. Phosphorylation of RB1 allows dissociation of the transcription factor E2F from the RB/E2F complex and the subsequent transcription of E2F target genes which are responsible for the progression through the G(1) phase. Hypophosphorylates RB1 in early G(1) phase. Cyclin D-CDK4 complexes are major integrators of various mitogenenic and antimitogenic signals. A [...]
CCNE1	G1/S-specific cyclin-E1; Essential for the control of the cell cycle at the G1/S (start) transition.
CDC20	Cell division cycle protein 20 homolog; Required for full ubiquitin ligase activity of the anaphase promoting complex/cyclosome (APC/C) and may confer substrate specificity upon the complex. Is regulated by MAD2L1: in metaphase the MAD2L1-CDC20-APC/C ternary complex is inactive and in anaphase the CDC20-APC/C binary complex is active in degrading substrates. The CDC20-APC/C complex positively regulates the formation of synaptic vesicle clustering at active zone to the presynaptic membrane in postmitotic neurons. CDC20-APC/C-induced degradation of NEUROD2 induces presynaptic differentia [...]
CDCA3	Cell division cycle-associated protein 3; F-box-like protein which is required for entry into mitosis. Acts by participating in E3 ligase complexes that mediate the ubiquitination and degradation of WEE1 kinase at G2/M phase (By similarity).
CDKN1B	Cyclin-dependent kinase inhibitor 1B; Important regulator of cell cycle progression. Inhibits the kinase activity of CDK2 bound to cyclin A, but has little inhibitory activity on CDK2 bound to SPDYA. Involved in G1 arrest. Potent inhibitor of cyclin E- and cyclin A-CDK2 complexes. Forms a complex with cyclin type D-CDK4 complexes and is involved in the assembly, stability, and modulation of CCND1-CDK4 complex activation. Acts either as an inhibitor or an activator of cyclin type D-CDK4 complexes depending on its phosphorylation state and/or stoichometry. Belongs to the CDI family.
CENPE	Centromere-associated protein E; Microtubule plus-end-directed kinetochore motor which plays an important role in chromosome congression, microtubule-kinetochore conjugation and spindle assembly checkpoint activation. Drives chromosome congression (alignment of chromosomes at the spindle equator resulting in the formation of the metaphase plate) by mediating the lateral sliding of polar chromosomes along spindle microtubules towards the spindle equator and by aiding the establishment and maintenance of connections between kinetochores and spindle microtubules. The transport of pole-pro [...]

CENPM	Centromere protein M; Component of the CENPA-NAC (nucleosome-associated) complex, a complex that plays a central role in assembly of kinetochore proteins, mitotic progression and chromosome segregation. The CENPA-NAC complex recruits the CENPA-CAD (nucleosome distal) complex and may be involved in incorporation of newly synthesized CENPA into centromeres.
CIT	Citron Rho-interacting kinase; Plays a role in cytokinesis. Required for KIF14 localization to the central spindle and midbody. Putative RHO/RAC effector that binds to the GTP-bound forms of RHO and RAC1. It probably binds p21 with a tighter specificity in vivo. Displays serine/threonine protein kinase activity. Plays an important role in the regulation of cytokinesis and the development of the central nervous system. Phosphorylates MYL9/MLC2.
DIAPH3	Protein diaphanous homolog 3; Actin nucleation and elongation factor required for the assembly of F-actin structures, such as actin cables and stress fibers. Required for cytokinesis, stress fiber formation and transcriptional activation of the serum response factor. Binds to GTP-bound form of Rho and to profilin: acts in a Rho-dependent manner to recruit profilin to the membrane, where it promotes actin polymerization. DFR proteins couple Rho and Src tyrosine kinase during signaling and the regulation of actin dynamics. Also acts as an actin nucleation and elongation factor in the nuc [...]
DNMT1	DNA (cytosine-5)-methyltransferase 1; Methylates CpG residues. Preferentially methylates hemimethylated DNA. Associates with DNA replication sites in S phase maintaining the methylation pattern in the newly synthesized strand, that is essential for epigenetic inheritance. Associates with chromatin during G2 and M phases to maintain DNA methylation independently of replication. It is responsible for maintaining methylation patterns established in development. DNA methylation is coordinated with methylation of histones. Mediates transcriptional repression by direct binding to HDAC2. In a [...]
DONSON	Protein downstream neighbor of Son; Replisome component that maintains genome stability by protecting stalled or damaged replication forks. After the induction of replication stress, required for the stabilization of stalled replication forks, the efficient activation of the intra-S-phase and G/2M cell-cycle checkpoints and the maintenance of genome stability.
DSCC1	Sister chromatid cohesion protein DCC1; Loads PCNA onto primed templates regulating velocity, spacing and restart activity of replication forks. May couple DNA replication to sister chromatid cohesion through regulation of the acetylation of the cohesin subunit SMC3.
E2F8	Transcription factor E2F8; Atypical E2F transcription factor that participates in various processes such as angiogenesis and polyploidization of specialized cells. Mainly acts as a transcription repressor that binds DNA independently of DP proteins and specifically recognizes the E2 recognition site 5'-TTTC[CG]CGC-3'. Directly represses transcription of classical E2F transcription factors such as E2F1: component of a feedback loop in S phase by repressing the expression of E2F1, thereby preventing p53/TP53-dependent apoptosis. Plays a key role in polyploidization of cells in placenta a [...]
FBXO5	F-box only protein 5; Regulator of APC activity during mitotic and meiotic cell cycle. During mitotic cell cycle plays a role as both substrate and inhibitor of APC-FZR1 complex. During G1 phase, plays a role as substrate of APC-FZR1 complex E3 ligase. Then switches as an inhibitor of APC-FZR1 complex during S and G2 leading to cell-cycle commitment. As APC inhibitor, prevents the degradation of APC substrates at multiple levels: by interacting with APC and blocking access of APC substrates to the D-box coreceptor, formed by FZR1 and ANAPC10; by suppressing ubiquitin ligation and chain [...]
HELLS	Lymphoid-specific helicase; Plays an essential role in normal development and survival. Involved in regulation of the expansion or survival of lymphoid cells. Required for de novo or maintenance DNA methylation. May control silencing of the imprinted CDKN1C gene through DNA methylation. May play a role in formation and organization of heterochromatin, implying a functional role in the regulation of transcription and mitosis (By similarity).
IPO7	Importin-7; Functions in nuclear protein import, either by acting as autonomous nuclear transport receptor or as an adapter-like protein in association with the importin-beta subunit KPNB1. Acting autonomously, is thought to serve itself as receptor for nuclear localization signals (NLS) and to promote translocation of import substrates through the nuclear pore complex (NPC) by an energy requiring, Ran-dependent mechanism. At the nucleoplasmic side of the NPC, Ran binds to importin, the importin/substrate complex dissociates and importin is re-exported from the nucleus to the cytoplasm [...]
KIF4A	Chromosome-associated kinesin KIF4A; Motor protein that translocates PRC1 to the plus ends of interdigitating spindle microtubules during the metaphase to anaphase transition, an essential step for the formation of an organized central spindle midzone and midbody and for successful cytokinesis. May play a role in mitotic chromosomal positioning and bipolar spindle stabilization.
LUC7L3	Luc7-like protein 3; Binds cAMP regulatory element DNA sequence. May play a role in RNA splicing.
MELK	Maternal embryonic leucine zipper kinase; Serine/threonine-protein kinase involved in various processes such as cell cycle regulation, self-renewal of stem cells, apoptosis and splicing regulation. Has a broad substrate specificity; phosphorylates BCL2L14, CDC25B, MAP3K5/ASK1 and ZNF622. Acts as an activator of apoptosis by phosphorylating and activating MAP3K5/ASK1. Acts as a regulator of cell cycle, notably by mediating phosphorylation of CDC25B, promoting

	localization of CDC25B to the centrosome and the spindle poles during mitosis. Plays a key role in cell proliferation and carcino [...]
MMS22L	Protein MMS22-like; Component of the MMS22L-TONSL complex, a complex that stimulates the recombination-dependent repair of stalled or collapsed replication forks. The MMS22L-TONSL complex is required to maintain genome integrity during DNA replication by promoting homologous recombination-mediated repair of replication fork-associated double- strand breaks. It may act by mediating the assembly of RAD51 filaments on ssDNA.
MND1	Meiotic nuclear division protein 1 homolog; Required for proper homologous chromosome pairing and efficient cross-over and intragenic recombination during meiosis (By similarity). Stimulates both DMC1- and RAD51-mediated homologous strand assimilation, which is required for the resolution of meiotic double- strand breaks.
MYBL2	Myb-related protein B; Transcription factor involved in the regulation of cell survival, proliferation, and differentiation. Transactivates the expression of the CLU gene.
NUP107	Nuclear pore complex protein Nup107; Plays a role in the nuclear pore complex (NPC) assembly and/or maintenance. Required for the assembly of peripheral proteins into the NPC. May anchor NUP62 to the NPC. Involved in nephrogenesis.
NUP155	Nuclear pore complex protein Nup155; Essential component of nuclear pore complex. Could be essessential for embryogenesis. Nucleoporins may be involved both in binding and translocating proteins during nucleocytoplasmic transport.
PAICS	Phosphoribosylaminoimidazole carboxylase and phosphoribosylaminoimidazolesuccinocarboxamide synthase; In the C-terminal section; belongs to the AIR carboxylase family. Class II subfamily.
PAN2	PAN2-PAN3 deadenylation complex catalytic subunit PAN2; Catalytic subunit of the poly(A)-nuclease (PAN) deadenylation complex, one of two cytoplasmic mRNA deadenylases involved in general and miRNA-mediated mRNA turnover. PAN specifically shortens poly(A) tails of RNA and the activity is stimulated by poly(A)-binding protein (PABP). PAN deadenylation is followed by rapid degradation of the shortened mRNA tails by the CCR4-NOT complex. Deadenylated mRNAs are then degraded by two alternative mechanisms, namely exosome-mediated 3'-5' exonucleolytic degradation, or deadenlyation-dependent [...]
PNN	Pinin; Transcriptional activator binding to the E-box 1 core sequence of the E-cadherin promoter gene; the core-binding sequence is 5'CAGGTG-3'. Capable of reversing CTBP1-mediated transcription repression. Auxiliary component of the splicing-dependent multiprotein exon junction complex (EJC) deposited at splice junction on mRNAs. The EJC is a dynamic structure consisting of core proteins and several peripheral nuclear and cytoplasmic associated factors that join the complex only transiently either during EJC assembly or during subsequent mRNA metabolism. Participates in the regulation [...]
PPM1D	Protein phosphatase 1D; Involved in the negative regulation of p53 expression. Required for the relief of p53-dependent checkpoint mediated cell cycle arrest. Binds to and dephosphorylates 'Ser-15' of TP53 and 'Ser-345' of CHEK1 which contributes to the functional inactivation of these proteins. Mediates MAPK14 dephosphorylation and inactivation. Is also an important regulator of global heterochromatin silencing and critical in maintaining genome integrity (By similarity).
PSMC3IP	Homologous-pairing protein 2 homolog; Plays an important role in meiotic recombination. Stimulates DMC1-mediated strand exchange required for pairing homologous chromosomes during meiosis. The complex PSMC3IP/MND1 binds DNA, stimulates the recombinase activity of DMC1 as well as DMC1 D-loop formation from double-strand DNA. This complex stabilizes presynaptic RAD51 and DMC1 filaments formed on single strand DNA to capture double- strand DNA. This complex stimulates both synaptic and presynaptic critical steps in RAD51 and DMC1-promoted homologous pairing. May inhibit HIV-1 viral protei [...]
RAD1	Cell cycle checkpoint protein RAD1; Component of the 9-1-1 cell-cycle checkpoint response complex that plays a major role in DNA repair. The 9-1-1 complex is recruited to DNA lesion upon damage by the RAD17-replication factor C (RFC) clamp loader complex. Acts then as a sliding clamp platform on DNA for several proteins involved in long-patch base excision repair (LP-BER). The 9-1-1 complex stimulates DNA polymerase beta (POLB) activity by increasing its affinity for the 3'-OH end of the primer-template and stabilizes POLB to those sites where LP-BER proceeds; endonuclease FEN1 cleavag [...]
RAD50	DNA repair protein RAD50; Component of the MRN complex, which plays a central role in double-strand break (DSB) repair, DNA recombination, maintenance of telomere integrity and meiosis. The complex possesses single-strand endonuclease activity and double-strand-specific 3'-5' exonuclease activity, which are provided by MRE11. RAD50 may be required to bind DNA ends and hold them in close proximity. This could facilitate searches for short or long regions of sequence homology in the recombining DNA templates, and may also stimulate the activity of DNA ligases and/or restrict the nuclease [...]
RFC1	Replication factor C subunit 1; The elongation of primed DNA templates by DNA polymerase delta and epsilon requires the action of the accessory proteins PCNA and activator 1. This subunit binds to the primer-template junction. Binds the PO-B transcription element as well as other GA rich DNA sequences. Could play a role in DNA transcription regulation as well as DNA replication and/or repair. Can bind single- or double-stranded DNA.

RFC2	Replication factor C subunit 2; The elongation of primed DNA templates by DNA polymerase delta and epsilon requires the action of the accessory proteins proliferating cell nuclear antigen (PCNA) and activator 1. This subunit binds ATP (By similarity).
RFC3	Replication factor C subunit 3; The elongation of primed DNA templates by DNA polymerase delta and epsilon requires the action of the accessory proteins proliferating cell nuclear antigen (PCNA) and activator 1.
SPAG5	Sperm-associated antigen 5; Essential component of the mitotic spindle required for normal chromosome segregation and progression into anaphase. Required for chromosome alignment, normal timing of sister chromatid segregation, and maintenance of spindle pole architecture. In complex with SKAP, promotes stable microtubule- kinetochore attachments. May contribute to the regulation of separase activity. May regulate AURKA localization to mitotic spindle, but not to centrosomes and CCNB1 localization to both mitotic spindle and centrosomes. Involved in centriole duplication. Required for C [...]
SPC25	Kinetochore protein Spc25; Acts as a component of the essential kinetochore-associated NDC80 complex, which is required for chromosome segregation and spindle checkpoint activity. Required for kinetochore integrity and the organization of stable microtubule binding sites in the outer plate of the kinetochore. The NDC80 complex synergistically enhances the affinity of the SKA1 complex for microtubules and may allow the NDC80 complex to track depolymerizing microtubules. Belongs to the SPC25 family.
TOP2A	DNA topoisomerase 2-alpha; Control of topological states of DNA by transient breakage and subsequent rejoining of DNA strands. Topoisomerase II makes double- strand breaks. Essential during mitosis and meiosis for proper segregation of daughter chromosomes. May play a role in regulating the period length of ARNTL/BMAL1 transcriptional oscillation (By similarity).
TRIP13	Pachytene checkpoint protein 2 homolog; Plays a key role in chromosome recombination and chromosome structure development during meiosis. Required at early steps in meiotic recombination that leads to non-crossovers pathways. Also needed for efficient completion of homologous synapsis by influencing crossover distribution along the chromosomes affecting both crossovers and non-crossovers pathways. Also required for development of higher- order chromosome structures and is needed for synaptonemal-complex formation. In males, required for efficient synapsis of the sex chromosomes and for [...]

Uganda RSTR Alveolar Macrophages RNAseq

Part 1: Data cleaning, QC and normalization with edgeR and limma

Contents

Overview

0. Setup

Packages

```
#Data manip
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
#RNAseq cleaning
library(RNAetc)
library(limma)
library(edgeR)
#Plots
library(BIGpicture)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(ggrepel)
library(patchwork)
# Misc
library(tinytex)
#Note we do not load biomaRt because it has conflicts with the tidyverse.
```

```r
# We instead call its functions with biomaRt::

set.seed(321)
```

Data description

1. Format data

1.1 Aligment metrics

Metrics obtained from from Picard RnaSeqMetrics.

```r
metric <- read_csv("data_raw/230124-P512-1qcMetricsSummary.csv") %>%
  select(-1) %>%  # remove the first column
  rename(libID = libid)
```

```
## New names:
## Rows: 49 Columns: 6
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): libid dbl (4): ...1, median_cv_coverage, mapped_reads_pct,
## fastq_total_reads lgl (1): passQC
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
#head(metric)
```

1.2: Sample and library metadata

1.3: Counts table

```
# Read the raw gene counts data from the CSV file
count_raw <- read_csv("data_raw/230124-P512-1_RawGeneCounts.csv")
```

```
## Rows: 58362 Columns: 53
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): geneId, geneName, description, geneBiotype
## dbl (49): BR16130001 - lib90126, BR16130002 - lib90127, BR16130003 - lib9012...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
count <- count_raw %>%
  # Remove unused columns (from geneName to geneBiotype)
  select(-c(geneName:geneBiotype)) %>%
  # Rename the 'geneId' column to match the key 'ensembl_gene_id'
  rename(ensembl_gene_id = geneId) %>%
  # Clean up the library ID by removing the unnecessary prefix 'BR16[0-9]{6} - '
  rename_all(~gsub("BR16[0-9]{6} - ", "", .))

#count[1:3, 1:5]
```

1.4: Combine metadata

Combine the libID extracted from the counts table with the metrics, sample, and patient data into a single metadata object.

```
meta <- data.frame(
  # Extract column names from 'count.raw' and exclude the first four columns
```

```
  name = colnames(count_raw)[-c(1:4)]
) %>%
  # Split the 'name' column into two columns: 'accession_no' and 'libID'
  separate(name, into = c("accession_no", "libID"), sep=" - ") %>%
  # Join the metadata dataframe with the 'sample' dataframe
  left_join(sample) %>%
  # Join the resulting dataframe with the 'patient' dataframe
  left_join(patient) %>%
  # Join the resulting dataframe with the 'metric' dataframe
  left_join(metric)
```

```
## Joining with `by = join_by(accession_no)`
## Joining with `by = join_by(FULLIDNO)`
## Joining with `by = join_by(libID)`
```

2. Quality-filter data

2.1: Filter poor-quality libraries

We assess sample quality using several metrics including:

- `fastq_total_reads`: total reads passing quality control (QC)
- `mapped_reads_pct`: percent alignment
- `median_cv_coverage`: median coefficient of variation of coverage i.e. how variable sequence coverage is across the genome

Ideal libraries have high total sequences, high percent alignment, and low CV coverage.

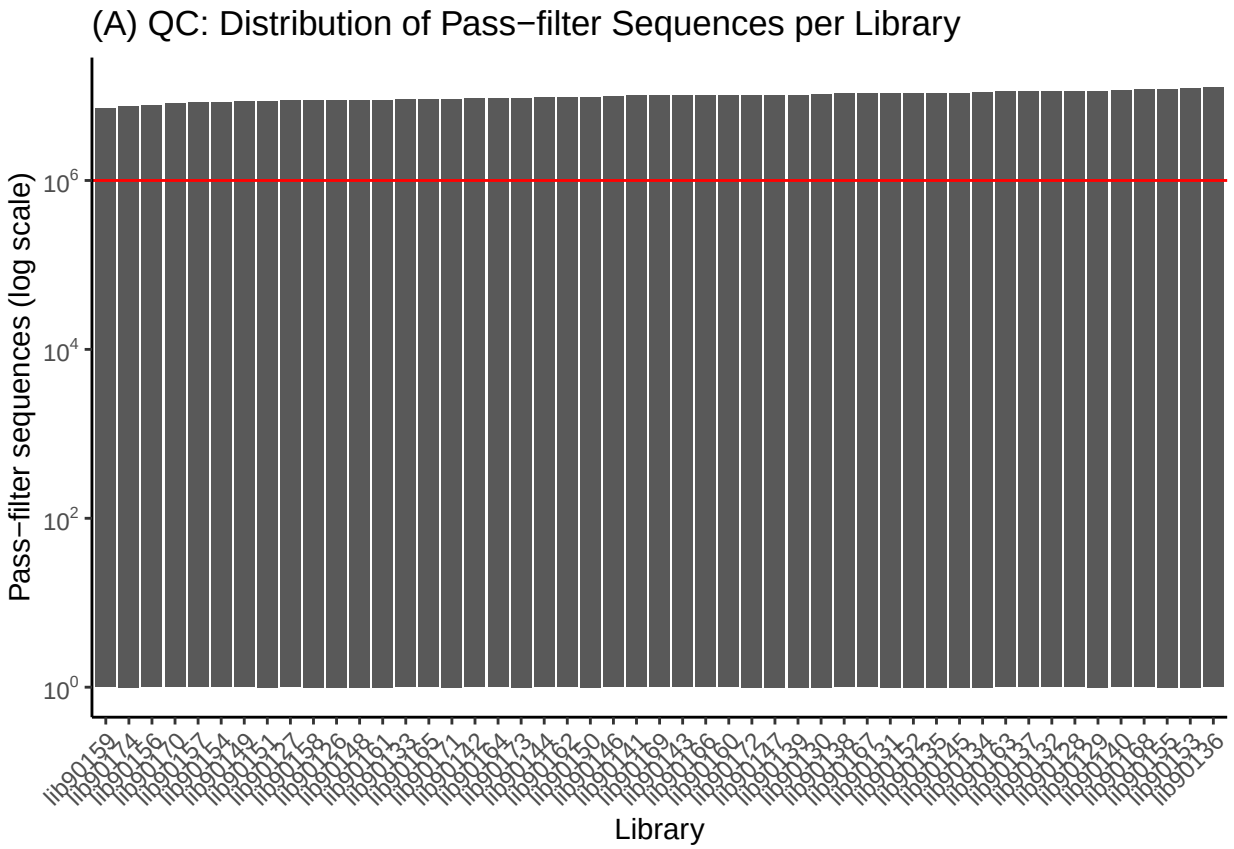
- reads $>$ 1,000,000
- CV coverage $<$ 1
- alignment $>$ 75%

Ideal libraries have high total reads, high percent alignment to reference genome, and low CV coverage which together indicate high quality and reliable gene expression measurements.

```
## Set cutoff values
seq_cutoff <- 1E6
cv_cutoff <- 1
align_cutoff <- 75
```

Check if any libraries are below the seq cutoff looking at libraries below the red line. No libraries fail the total seq cutoff.

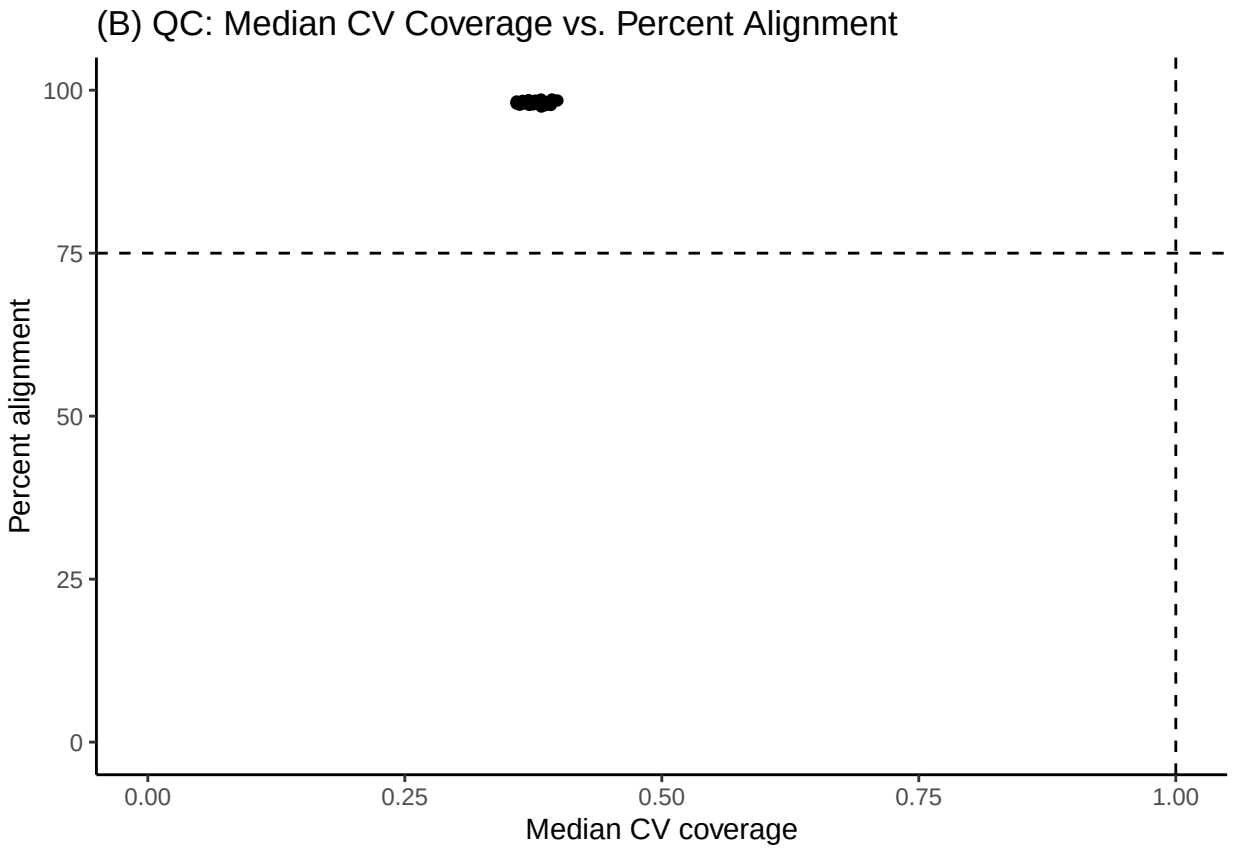
```
p1 <- ggplot(meta, aes(x = reorder(libID, fastq_total_reads), y = fastq_total_reads)) +
  # Add a column geom to represent the 'fastq_total_reads' values
  geom_col() +
  # Add a red horizontal line to indicate the 'seq_cutoff' threshold
  geom_hline(yintercept = seq_cutoff, color = "red") +
  # Apply a log10 transformation to the y-axis scale
  scale_y_continuous(trans = 'log10',
                     breaks = trans_breaks("log10", function(x) 10^x),
                     labels = trans_format("log10", math_format(10^.x))) +
  theme_classic() +
  labs(x = "Library", y = "Pass-filter sequences (log scale)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("(A) QC: Distribution of Pass-filter Sequences per Library")

p1
```



Check if any libraries are below CV or percent aligment cutoffs looking at data cluster in upper left quadrant. No libraries fail CV or percent alignment cutoffs.

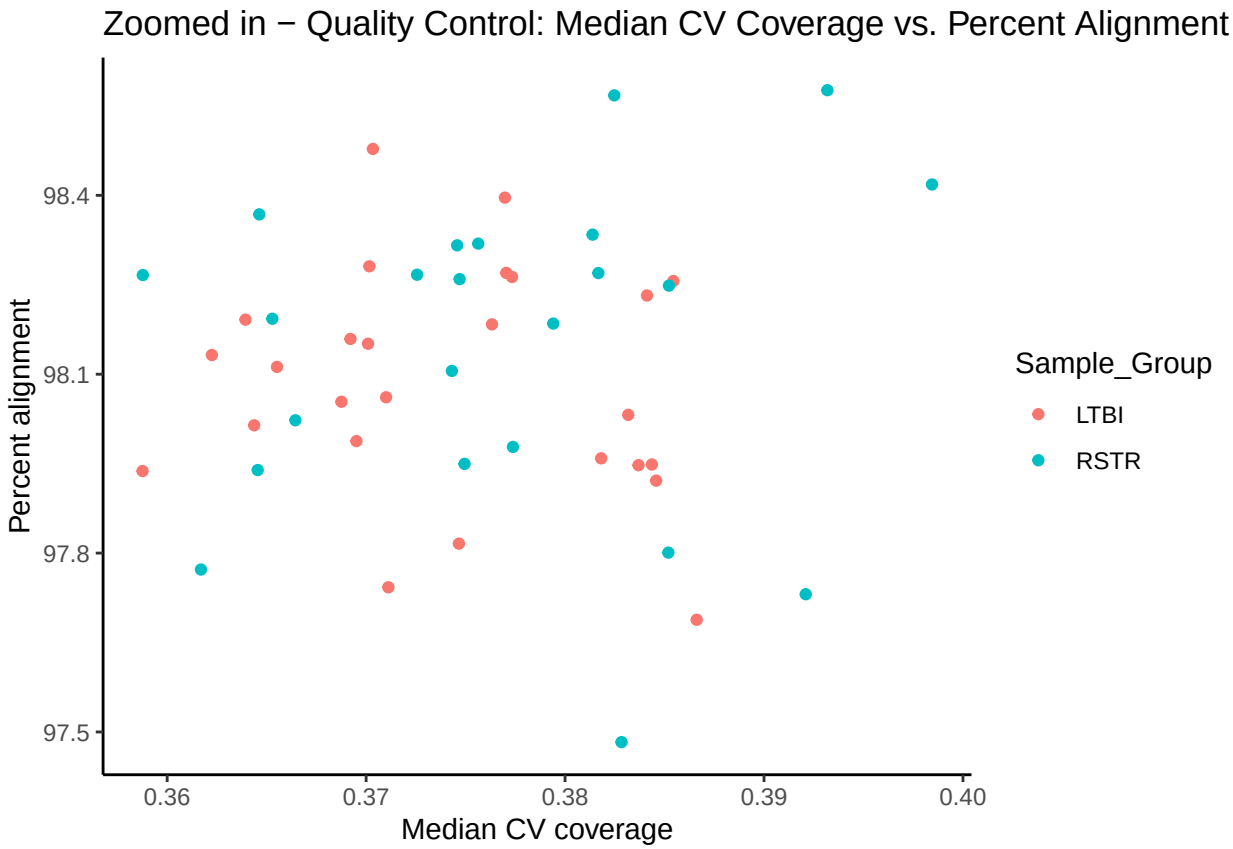
```
p2 <- ggplot(meta, aes(x = median_cv_coverage, y = mapped_reads_pct)) +
  geom_point() +
  # Set the axis limits for x and y
  lims(x = c(0, 1), y = c(0, 100)) +
  # Add dashed horizontal and vertical lines to indicate cutoff thresholds
```

```
  geom_hline(yintercept = align_cutoff, linetype = "dashed") +
  geom_vline(xintercept = cv_cutoff, linetype = "dashed") +
  geom_text_repel(data = filter(meta,
                                median_cv_coverage > cv_cutoff |
                                  mapped_reads_pct < align_cutoff),
                  aes(label = libID), show.legend = FALSE,
                  max.overlaps = Inf, box.padding = 1) +
  theme_classic() +
  labs(x = "Median CV coverage", y = "Percent alignment") +
  ggtitle("(B) QC: Median CV Coverage vs. Percent Alignment")

p2
```



Zoom in on the library data cluster.

```
ggplot(meta, aes(x = median_cv_coverage,
                 y = mapped_reads_pct)) +
  geom_point(aes(color = Sample_Group)) +
  theme_classic() +
  labs(x = "Median CV coverage", y="Percent alignment") +
  ggtitle("Zoomed in - Quality Control: Median CV Coverage vs. Percent Alignment")
```




```
meta.filter <- meta %>%
  filter(median_cv_coverage < cv_cutoff &
           fastq_total_reads > seq_cutoff &
           mapped_reads_pct > align_cutoff)

count.filter <- count %>%
  select(1, all_of(meta.filter$libID))
```

This section for filtering low quality libraries removes **0 libraries**

2.2: Filter non-protein-coding genes

`biomaRt` used to annotate gene biotypes. Next filter to retain only protein coding genes.

```
ensembl <- biomaRt::useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl", mirror = "useast")

# Retrieve the gene key from the Ensembl database, including gene IDs, biotypes, and other information
key <- biomaRt::getBM(attributes = c("ensembl_gene_id", "entrezgene_id", "hgnc_symbol", "gene_biotype",
                                     "chromosome_name", "start_position", "end_position"), mart = ensembl) %>%
  # Filter for protein-coding genes only
  filter(gene_biotype == "protein_coding")

# Filter the gene key for genes that are present in the counts table
key.filter <- key %>%
  # Filter for genes in the counts table
  filter(ensembl_gene_id %in% count$ensembl_gene_id) %>%
  # Collapse multiple annotations into a single row per gene
  group_by(ensembl_gene_id, hgnc_symbol, gene_biotype, chromosome_name, start_position, end_position) %>%
  summarise(entrezgene_id = list(unique(entrezgene_id)), .groups = "drop") %>%
  # Group by additional identifiers and summarize
  group_by(ensembl_gene_id, entrezgene_id, gene_biotype, chromosome_name, start_position, end_position) %>%
  summarise(symbol = list(unique(hgnc_symbol)), .groups = "drop")
```

```
count.filter.pc <- count.filter %>%
  filter(ensembl_gene_id %in% key.filter$ensembl_gene_id)
```

This removes 38630 genes from this data set.

2.3: Filter PCA outliers

PCA is used to identify sample libraries as potential outliers defined as more than 3 standard deviations away from the mean on PC1 and/or PC2. One library lib90167 is an outlier.

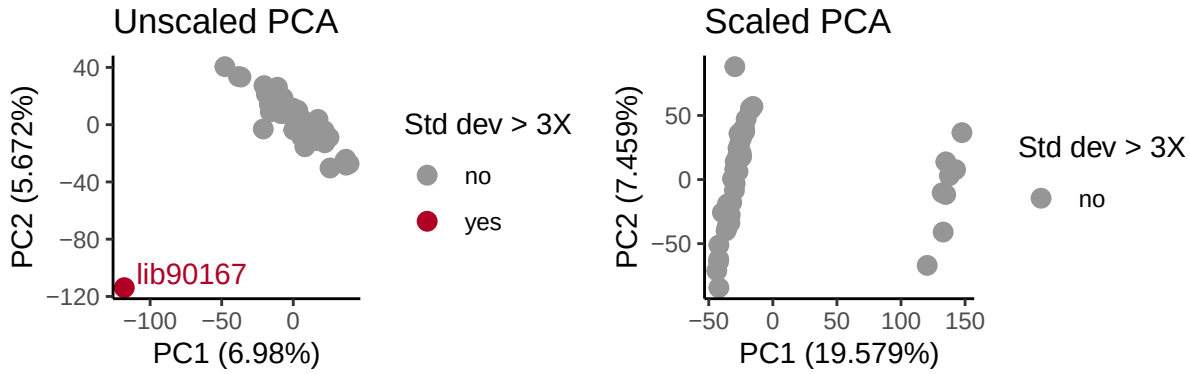
```
pca3a <- BIGpicture::plot_pca(count.filter.pc, meta=meta.filter,
                              scale = FALSE,
                              vars="outlier", transform_logCPM=TRUE,
                              outlier_sd = 3)
```

```
## Joining with `by = join_by(libID)`
```

```
pca3b <- BIGpicture::plot_pca(count.filter.pc, meta=meta.filter,
                              scale = TRUE,
                              vars="outlier", transform_logCPM=TRUE,
                              outlier_sd = 3)
```

```
## Joining with `by = join_by(libID)`
```

```
pca3a$outlier + labs(title = "Unscaled PCA") +
  pca3b$outlier + labs(title = "Scaled PCA")
```




```
p3 <- pca3a$outlier + labs(title = "(C) Unscaled PCA")
```

Next we remove outlier from the data.

```
not.outlier <- pca3a$outlier$data %>%
  filter(col.group == "no")
```

```
meta.filter.out <- meta.filter %>%
  filter(libID %in% not.outlier$libID)

count.filter.pc.out <- count.filter.pc %>%
  select(1, all_of(meta.filter.out$libID))
```

Here, this removes 1 library(ies) since we more than std dev > 3X.

2.4: Create DGEList

Next we create create DGEList while also making sure data in correct format.

```
#Order metadata by library ID
meta.filter.out.ord <- meta.filter.out %>%
  arrange(libID)

#Order counts by library ID and gene ID
count.filter.pc.out.ord <- count.filter.pc.out %>%
  select(1, all_of(meta.filter.out.ord$libID)) %>%
  arrange(ensembl_gene_id)

#Order gene key by gene ID
key.filter.ord <- key.filter %>%
  arrange(ensembl_gene_id)

#check libraries
identical(meta.filter.out.ord$libID,
          colnames(count.filter.pc.out.ord)[-1])
```

```
## [1] TRUE
```

```
#Check genes
identical(key.filter.ord$ensembl_gene_id,
          count.filter.pc.out.ord$ensembl_gene_id)
```

```
## [1] TRUE
```

Next, we pass data into the `edgeR` format.

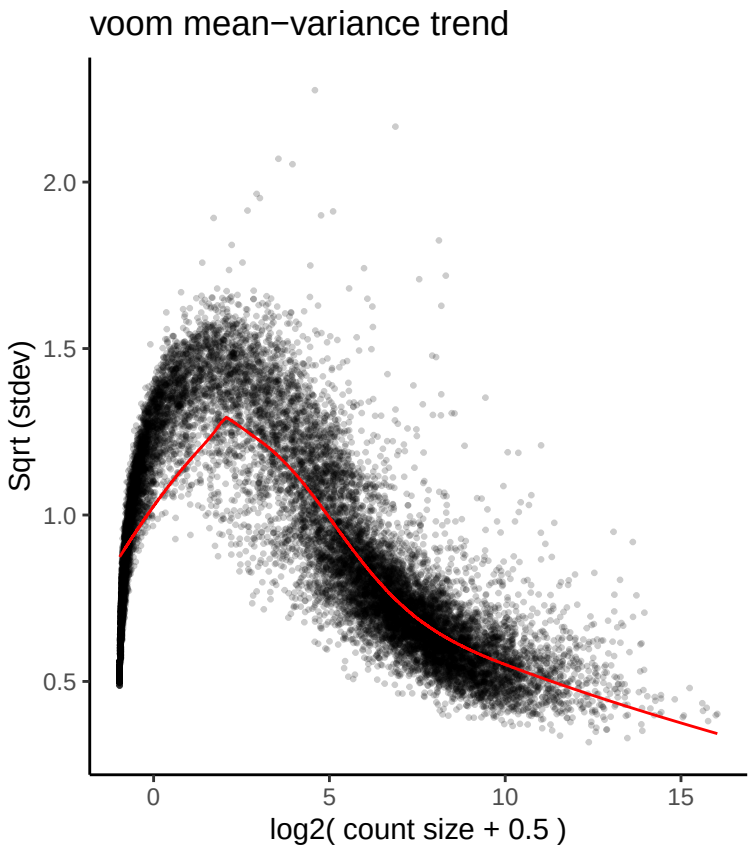
```
# Move gene names from a variable in the df to rownames and format to matrix
count.filter.pc.out.ord.mat <-
  count.filter.pc.out.ord %>%
  column_to_rownames("ensembl_gene_id") %>%
  as.matrix()

dat <- DGEList(
  #count table
  counts=count.filter.pc.out.ord.mat,
  #metadata
  samples=meta.filter.out.ord,
  #gene info
  genes=key.filter.ord)
```

2.5: Filter low abundance genes

Low abundance (small counts) and zero count genes are removed from the data. Our goal is to remove genes in the lower left of the mean-variance plot because counts (x) and variance (y) are low *e.g.* these genes break the red mean variance trend. We also want to filter to lower multiple testing penalty for genes likely to be absent in both groups.

```
mv_model <- "~ Sample_Group"
```

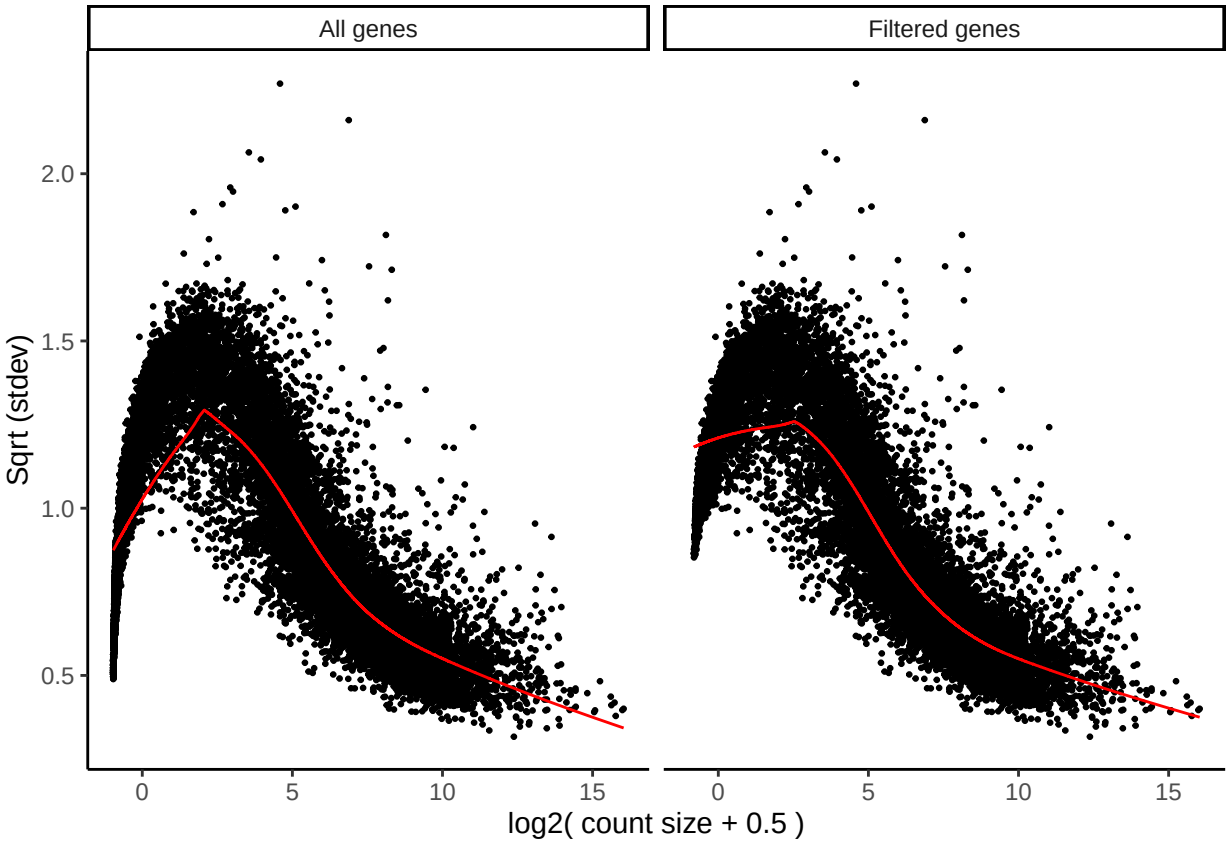
```
BIGpicture::plot_mv(dat, design = mv_model)
```



Filter to retain only genes that are at least `min.CPM` counts per million in at least `min.sample` number of samples. Here, we use 0.5 CPM in at least 3 samples.

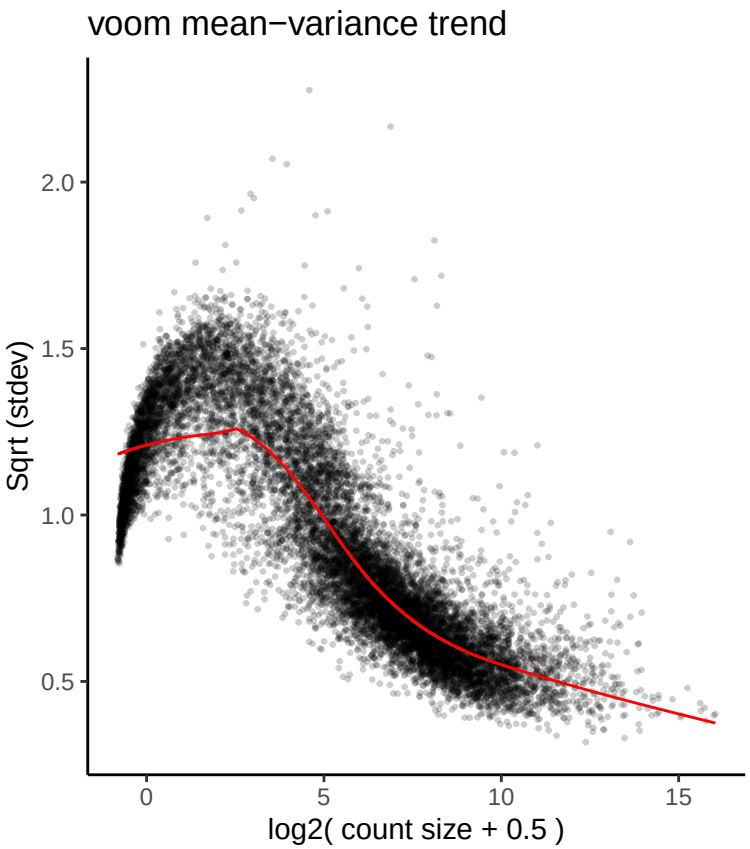
```
p4 <- dat.abund <- RNAetc::filter_rare(dat, min.CPM = 0.5,
                                       min.sample = 3,
                                       gene.var="ensembl_gene_id",
                                       plot = TRUE)
```

```
## 5924 (30.02%) of 19732 genes removed.
```



Looking at the filtered plot, we see the red trend line is improved.

```
BIGpicture::plot_mv(dat.abund, design = mv_model)
```



Since rare genes are filtered out, here we save them as a `csv` for other potential analysis.

```
rare <- as.data.frame(cpm(dat$counts)) %>%
  # Filter out genes that were removed in dat.abund
  rownames_to_column("ensembl_gene_id") %>%
  filter(!(ensembl_gene_id %in% rownames(dat.abund$counts))) %>%
  # Add gene symbols
  left_join(dat$genes, by = "ensembl_gene_id") %>%
```

```
  # Remove unwanted columns
  select(-c(chromosome_name:end_position)) %>%
  # Pivot data to long format
  pivot_longer(-c(ensembl_gene_id, gene_biotype, symbol, entrezgene_id)) %>%
  # Calculate summary statistics for each gene
  group_by(ensembl_gene_id, gene_biotype, symbol) %>%
  summarise(mean.CPM = mean(value, na.rm=TRUE),
            min.CPM = min(value, na.rm=TRUE),
            max.CPM = max(value, na.rm=TRUE),
            express.in.libs = length(value[value > 0]),
            .groups="drop") %>%
  # Convert gene symbols to character type
  mutate(symbol = as.character(symbol))

write_csv(rare, file="data_clean/rare_genes.csv")
```

3. Normalize data

3.1: Trimmed mean of M (TMM)

Calculate TMM scaling factors.

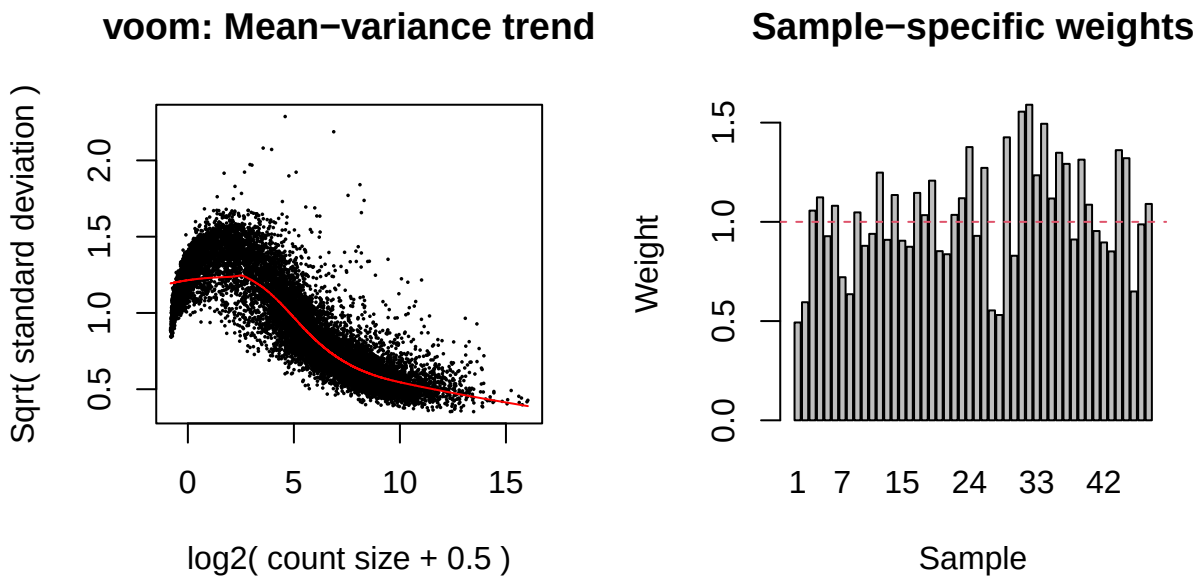
```
dat.abund.norm <- calcNormFactors(dat.abund, method = "TMM")
```

3.2: voom aka log2 counts per million (CPM)

Gene counts converted to log2 CPM within each sample with `voom`. This accounts for differential sampling depth and some of the skewed nature of the counts data. Next calculate gene-level weights for linear modeling of gene expression data.

```
dat.abund.norm.voom <-
  voomWithQualityWeights(
    dat.abund.norm,
```

```
    design=model.matrix(as.formula(mv_model),
                        data=dat.abund.norm$samples),
    plot=TRUE)
```



4. Summary and save

Filtering Summarize libraries removed from analysis.

```
##       libID      filter
## 1  lib90126 pass-filter
## 2  lib90127 pass-filter
```

```
## 3  lib90128 pass-filter
## 4  lib90129 pass-filter
## 5  lib90130 pass-filter
## 6  lib90131 pass-filter
## 7  lib90132 pass-filter
## 8  lib90133 pass-filter
## 9  lib90134 pass-filter
## 10 lib90135 pass-filter
## 11 lib90136 pass-filter
## 12 lib90137 pass-filter
## 13 lib90138 pass-filter
## 14 lib90139 pass-filter
## 15 lib90140 pass-filter
## 16 lib90141 pass-filter
## 17 lib90142 pass-filter
## 18 lib90143 pass-filter
## 19 lib90144 pass-filter
## 20 lib90145 pass-filter
## 21 lib90146 pass-filter
## 22 lib90147 pass-filter
## 23 lib90148 pass-filter
## 24 lib90149 pass-filter
## 25 lib90150 pass-filter
## 26 lib90151 pass-filter
## 27 lib90152 pass-filter
## 28 lib90153 pass-filter
## 29 lib90154 pass-filter
## 30 lib90155 pass-filter
## 31 lib90156 pass-filter
## 32 lib90157 pass-filter
## 33 lib90158 pass-filter
## 34 lib90159 pass-filter
## 35 lib90160 pass-filter
## 36 lib90161 pass-filter
## 37 lib90162 pass-filter
## 38 lib90163 pass-filter
```

```
## 39 lib90164 pass-filter
## 40 lib90165 pass-filter
## 41 lib90166 pass-filter
## 42 lib90167 PCA outlier
## 43 lib90168 pass-filter
## 44 lib90169 pass-filter
## 45 lib90170 pass-filter
## 46 lib90171 pass-filter
## 47 lib90172 pass-filter
## 48 lib90173 pass-filter
## 49 lib90174 pass-filter
```

This leaves the following total samples

```
dat.abund.norm.voom$targets %>%
  count(Sample_Group)
```

```
##   Sample_Group  n
## 1         LTBI 26
## 2         RSTR 22
```

```
pca4 <- BIGpicture::plot_pca(dat.abund.norm.voom,
                             vars="outlier", scale = FALSE,
                             outlier_sd = 3)$outlier +
  labs(title = "Unscaled PCA")
```

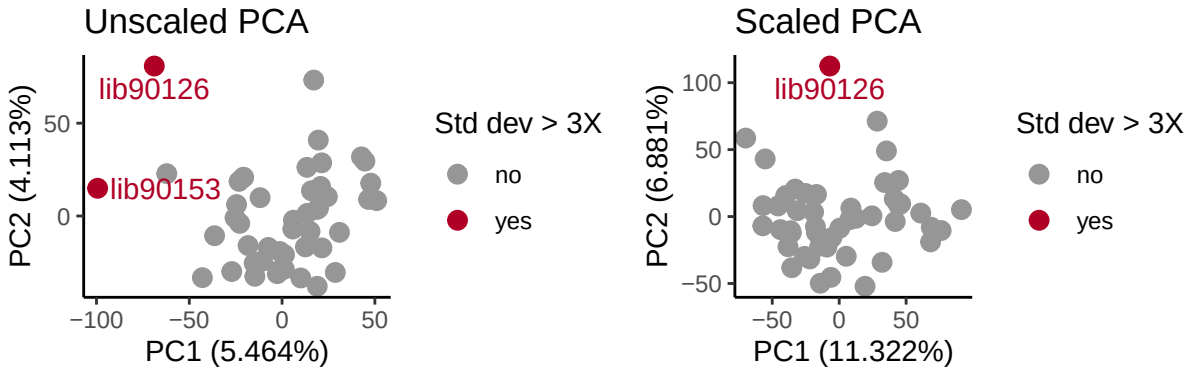
Re-check PCA outliers

```
## Joining with `by = join_by(libID)`
```

```
pca5 <- BIGpicture::plot_pca(dat.abund.norm.voom,
                             vars="outlier", scale = TRUE,
                             outlier_sd = 3)$outlier +
  labs(title = "Scaled PCA")
```

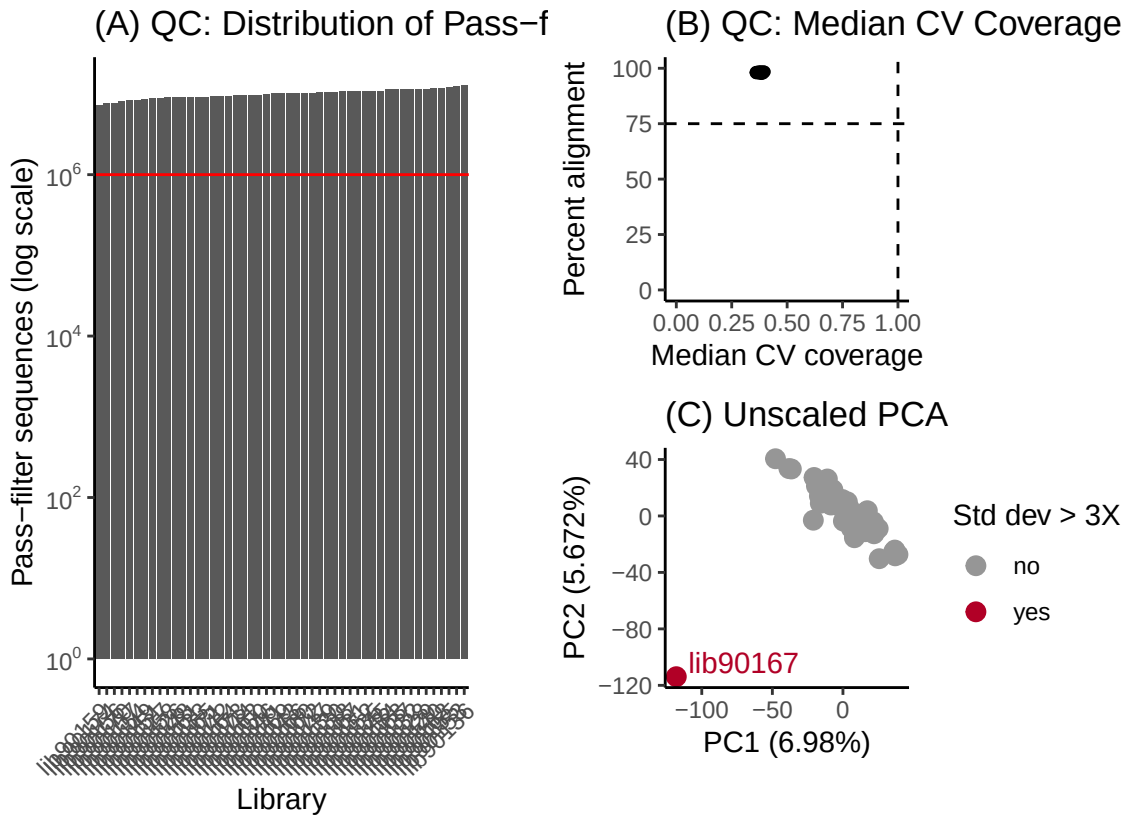
```
## Joining with `by = join_by(libID)`
```

```
pca4 + pca5
```



Combined plots Combine QC plots -

```
qc_plots <- p1 | p2 / p3
qc_plots
```



Save Save plots

```
ggsave(paste0("results/", "qc_plots.png"), qc_plots)
```

```
## Saving 6.5 x 4.5 in image
```

Save data as `RData`

```
save(dat.abund.norm, file = "data_clean/P512_dat.RData")
save(dat.abund.norm.voom, file = "data_clean/P512_voom.RData")
```

Save gene counts as `csv`

```
as.data.frame(dat.abund.norm$counts) %>%
  rownames_to_column("ensembl_gene_id") %>%
  write_csv("data_clean/P512_counts.csv")

as.data.frame(dat.abund.norm.voom$E) %>%
  rownames_to_column("ensembl_gene_id") %>%
  write_csv("data_clean/P512_counts_voom.csv")
```

R session

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
##  [1] tinytex_0.45     patchwork_1.1.2  ggrepel_0.9.3    scales_1.2.1
##  [5] BIGpicture_1.1.0 edgeR_3.40.2     limma_3.54.2     RNAetc_1.0.0
##  [9] janitor_2.2.0    readxl_1.4.2     lubridate_1.9.2  forcats_1.0.0
## [13] stringr_1.5.0    dplyr_1.1.2      purrr_1.0.1      readr_2.1.4
## [17] tidyr_1.3.0      tibble_3.2.1     ggplot2_3.4.2    tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-7           bit64_4.0.5            filelock_1.0.2
##  [4] progress_1.2.2         httr_1.4.6             GenomeInfoDb_1.34.9
##  [7] tools_4.2.2            utf8_1.2.3             R6_2.5.1
## [10] DBI_1.1.3              BiocGenerics_0.44.0    colorspace_2.1-0
## [13] withr_2.5.0            tidyselect_1.2.0       prettyunits_1.1.1
## [16] bit_4.0.5              curl_5.0.0             compiler_4.2.2
## [19] textshaping_0.3.6      cli_3.6.1              Biobase_2.58.0
## [22] xml2_1.3.4             labeling_0.4.2         rappdirs_0.3.3
## [25] systemfonts_1.0.4      digest_0.6.31          rmarkdown_2.22
## [28] XVector_0.38.0         pkgconfig_2.0.3        htmltools_0.5.5
## [31] dbplyr_2.3.2           fastmap_1.1.1          highr_0.10
## [34] rlang_1.1.1            rstudioapi_0.14        RSQLite_2.3.1
## [37] farver_2.1.1           generics_0.1.3         vroom_1.6.3
## [40] RCurl_1.98-1.12        magrittr_2.0.3         GenomeInfoDbData_1.2.9
## [43] Rcpp_1.0.10            munsell_0.5.0          S4Vectors_0.36.2
## [46] fansi_1.0.4            lifecycle_1.0.3        stringi_1.7.12
## [49] yaml_2.3.7             snakecase_0.11.0       zlibbioc_1.44.0
## [52] BiocFileCache_2.6.1    grid_4.2.2             blob_1.2.4
## [55] parallel_4.2.2         crayon_1.5.2           lattice_0.21-8
## [58] Biostrings_2.66.0      hms_1.1.3              KEGGREST_1.38.0
## [61] locfit_1.5-9.7         knitr_1.43             pillar_1.9.0
## [64] codetools_0.2-19       biomaRt_2.54.1         stats4_4.2.2
## [67] XML_3.99-0.14          glue_1.6.2             evaluate_0.21
## [70] png_0.1-8              vctrs_0.6.2            tzdb_0.4.0
## [73] foreach_1.5.2          cellranger_1.1.0       gtable_0.3.3
## [76] cachem_1.0.8           xfun_0.39              ragg_1.2.5
## [79] iterators_1.0.14       AnnotationDbi_1.60.2   memoise_2.0.1
## [82] IRanges_2.32.0         timechange_0.2.0
```

Uganda RSTR Alveolar Macrophages RNAseq

Part 2: Gene Expression Linear Statistical Modeling

Contents

Overview

This analysis performs linear statistical modeling to model gene expression data accounting for covariates age, sex, kinship and exposure risk score to explore potential confounding.

AIC metrics are used to compare statistical model fit and pick the best statistical model for downstream analysis.

Differentially expressed genes (DEGs) are visualized with volcano plots.

0. Setup

Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(BIGverse)
```

```
## -- Attaching packages ----------------------------------------- BIGverse 1.0.0 --
## v kimma      1.4.4     v BIGpicture 1.1.0
## v RNAetc     1.0.0     v SEARchways 1.0.0
## -- Conflicts ---------------------------------------- BIGverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(limma)
library(edgeR)
library(patchwork)
library(ggpubr)
set.seed(651)
```

1. Load data

All counts, gene, and sample metadata are contained in a single object from the `limma` package.

```
#Extract and rename data object
attach("data_clean/P512_voom.RData")
dat <- dat.abund.norm.voom
```

Can load the precomputed statistical model results since they are compute intensive.

```
load("results/lm_data.Rdata")
```

Load kinship data

```
load('data_clean/kin.filter.Rdata')
```

2. Linear Models

2.1 Base Model, no covariates

`Sample_Group`

Run the model -

```
lm_rstr <- kmFit(dat = dat,
                         model = "~ Sample_Group",
                         patientID = "FULLIDNO",
                         run.lm = TRUE, use.weights = TRUE,
                         metrics = TRUE)
```

Check for DEGs -

```
#summarise_kmFit(fdr = lm_rstr$lm, fdr.cutoff = c(0.05,0.5))
summarise_kmFit(fdr = lm_rstr$lm)
```

```
## Error in summarise_kmFit(fdr = lm_rstr$lm): No significant genes. Please increase FDR or P-value cutoffs.
```

Check for lowest FDR DEG -

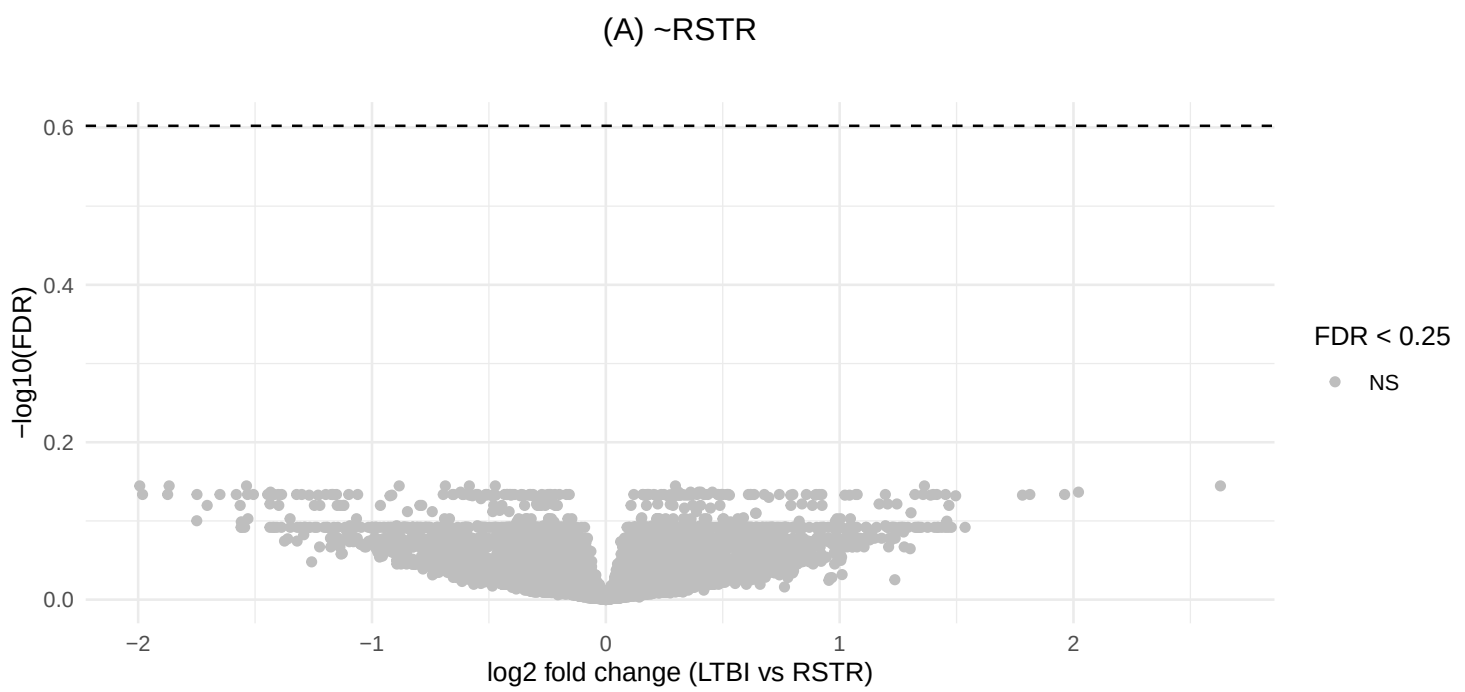
```
lowest_FDR <- lm_rstr$lm %>%
  filter(variable == 'Sample_GroupRSTR') %>%
  summarise(FDR = min(FDR))

# print(paste("Lowest FDR w/ no covariate is", round(lowest_FDR, digits=2)))
```

For the base model only DEGs w/ FDR > 0.72.

Plot volano plot -

```
base_volcano <- plot_volcano(model_result = lm_rstr,
             model = "lm", variables = "Sample_GroupRSTR",
             y.cutoff = 0.25, label = "all",
             genes = dat$genes, genes_label = "symbol")


base_volcano <- base_volcano +
```

```
  labs(title = "(A) ~RSTR", x = "log2 fold change (LTBI vs RSTR)") +

  theme(plot.title = element_text(hjust = 0.5),
        strip.text = element_text(color = NA))

base_volcano
```



Plot shows there are no DEGs at FDR cutoff of 0.25.

2.2 Base Model + Age, Sex

```
Sample_Group + age + sex
```

```
lm_rstr_age_sex <- kmFit(dat = dat,
                         model = "~ Sample_Group + KCHCA_AGE_YR_CURRENT + MO_KCVSEX",
                         patientID = "FULLIDNO",
                         run.lm = TRUE, use.weights = TRUE,
                         metrics = TRUE)
```

Check for DEGs -

```
summarise_kmFit(fdr = lm_rstr_age_sex$lm, fdr.cutoff = c(0.05,0.3))
```

```
## # A tibble: 3 x 3
##   variable             fdr_0.05 fdr_0.3
##   <fct>                   <int>   <int>
## 1 KCHCA_AGE_YR_CURRENT        3    1046
## 2 MO_KCVSEXM                 97     476
## 3 total (nonredundant)      100    1454
```

Check for lowest FDR DEG -

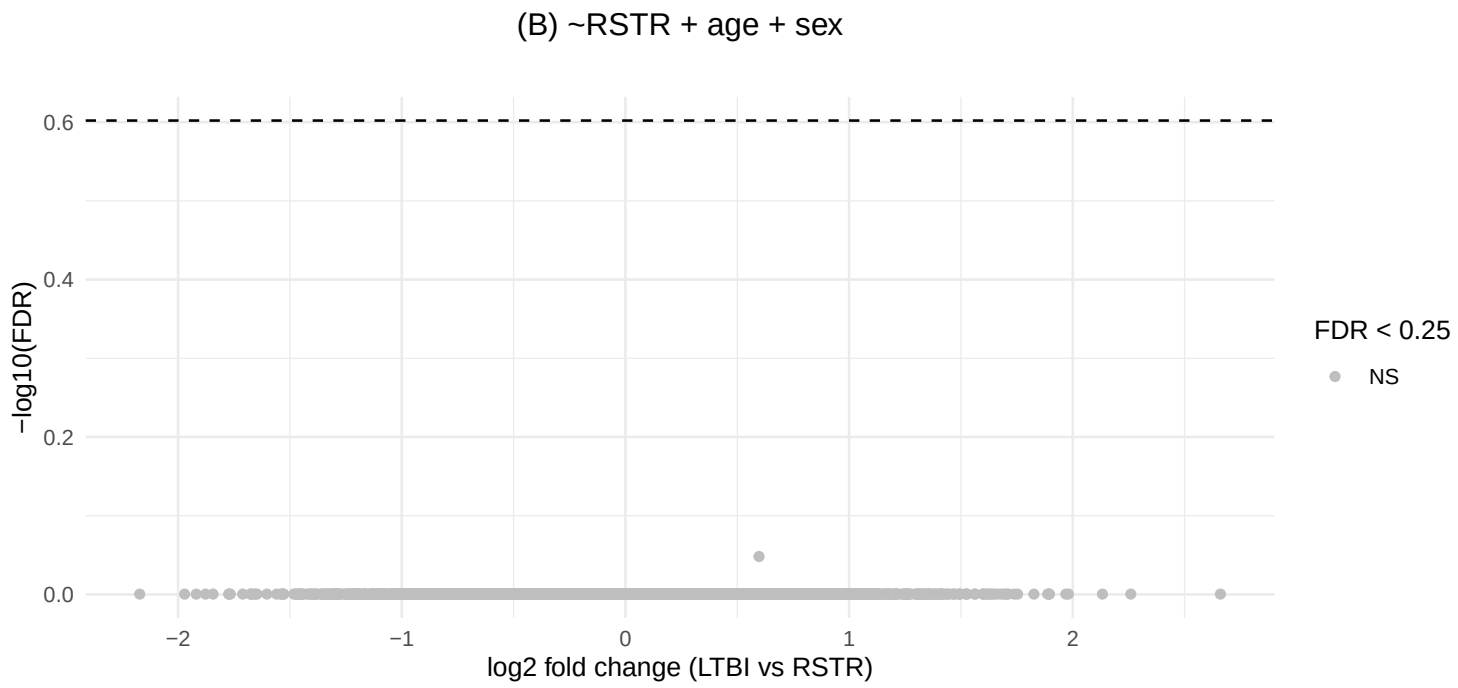
```
lowest_FDR <- lm_rstr_age_sex$lm %>%
  filter(variable == 'Sample_GroupRSTR') %>%
  summarise(FDR = min(FDR))

#print(paste("Lowest FDR wtih age and sex is", round(lowest_FDR, digits=2)))
```

For the model adjusted for age and sex only DEGs w/ FDR > 0.89.

Volcano Plot -

```
base_age_sex_volcano <- plot_volcano(model_result = lm_rstr_age_sex,
             model = "lm", variables = "Sample_GroupRSTR",
             y.cutoff = 0.25, label = "all",
             genes = dat$genes, genes_label = "symbol")
```

```
base_age_sex_volcano <- base_age_sex_volcano +
  labs(title = "(B) ~RSTR + age + sex", x = "log2 fold change (LTBI vs RSTR)") +
  theme(plot.title = element_text(hjust = 0.5),
        strip.text = element_text(color = NA))


base_age_sex_volcano
```



Plot shows there are no DEGs at FDR cutoff of 0.25.

Compare model fits -

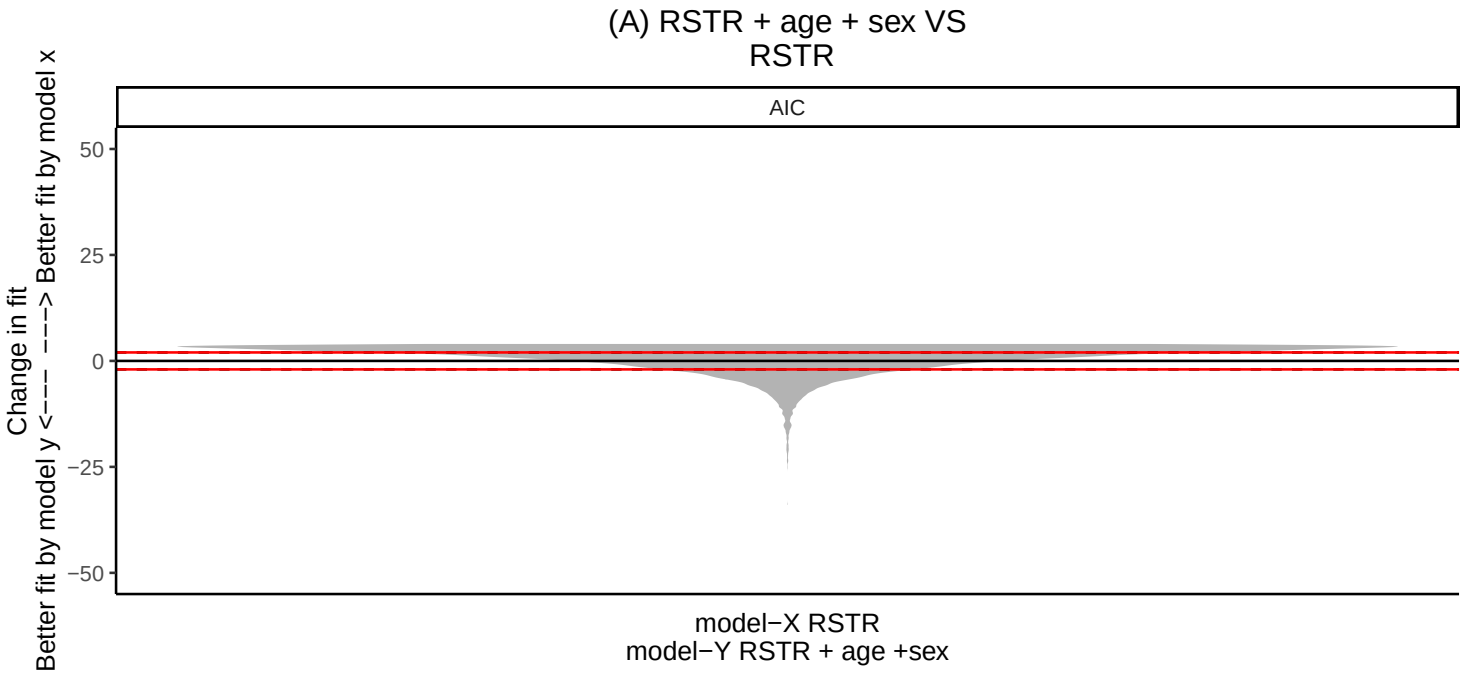
```r
aic_1 <- plot_fit2(model_result = lm_rstr, x="lm",
          model_result_y = lm_rstr_age_sex, y="lm",
          metrics = "AIC")
```

Summary

```
##                                                Best fit Metric Total genes
## 1                                  lm_Sample_GroupRSTR    AIC        9757
## 2 lm_Sample_GroupRSTR_KCHCA_AGE_YR_CURRENT_MO_KCVSEXM    AIC        4051
##   Mean delta Stdev delta
## 1   2.427805    1.107453
## 2   3.620894    7.767820
```

```r
aic_1 <- aic_1[[1]] + ylim(-50, 50) + geom_hline(yintercept = c(-2, 2), color = "red", linetype = "solid")


aic_1 <- aic_1 + labs(title = "(A) RSTR + age + sex VS\n RSTR",
                      x = "model-X RSTR\nmodel-Y RSTR + age +sex") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(margin = margin(t = 5, b = 5)))

print(aic_1)
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_ydensity()`).
```



2.3 Base Model + Age, Sex, Kin

```
~ Sample_Group + age + sex + kin
```

Load kinship data.

```
kin <- read_csv("data_raw/2023.02.01_kinship_Hawn_all.csv")
```

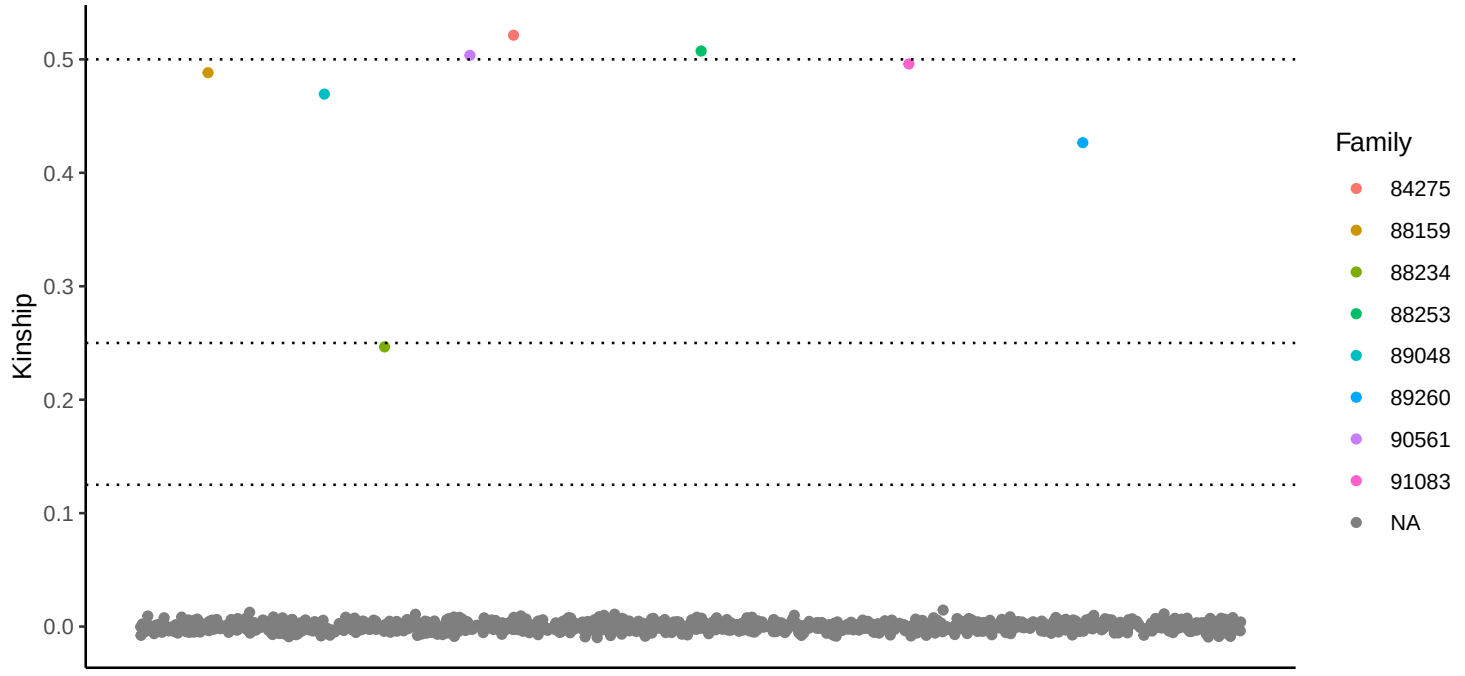
```
## Rows: 774 Columns: 775
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr   (1): rowname
## dbl (774): 84165-1-03, 84165-1-06, 84182-1-02, 84183-1-02, 84183-1-05, 84186...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
overlap <- intersect(colnames(kin), dat$targets$FULLIDNO)

kin.filter <- kin %>%
  #filter samples in the rnaseq data
  filter(rowname %in% overlap) %>%
  column_to_rownames("rowname") %>%
  select(all_of(overlap))
```

In total, 48 of 48 individuals have kinship data. Of these, 8 pairs are third degree related or more.

```
## Warning: Expected 1 pieces. Additional pieces discarded in 1128 rows [1, 2, 3, 4, 5, 6,
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
## Expected 1 pieces. Additional pieces discarded in 1128 rows [1, 2, 3, 4, 5, 6,
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```



Run the model -

```
lme_rstr_age_sex_kin <- kmFit(dat = dat, kin = kin.filter,
    model = "~ Sample_Group + KCHCA_AGE_YR_CURRENT + MO_KCVSEX + (1|FULLIDNO)",
                          patientID = "FULLIDNO",
                          run.lmerel = TRUE, use.weights = TRUE,
                          metrics = TRUE)
```

Check for DEGs -

```
summarise_kmFit(fdr = lme_rstr_age_sex_kin$lmerel, fdr.cutoff = c(0.05,0.3))
```

```
## # A tibble: 5 x 3
```

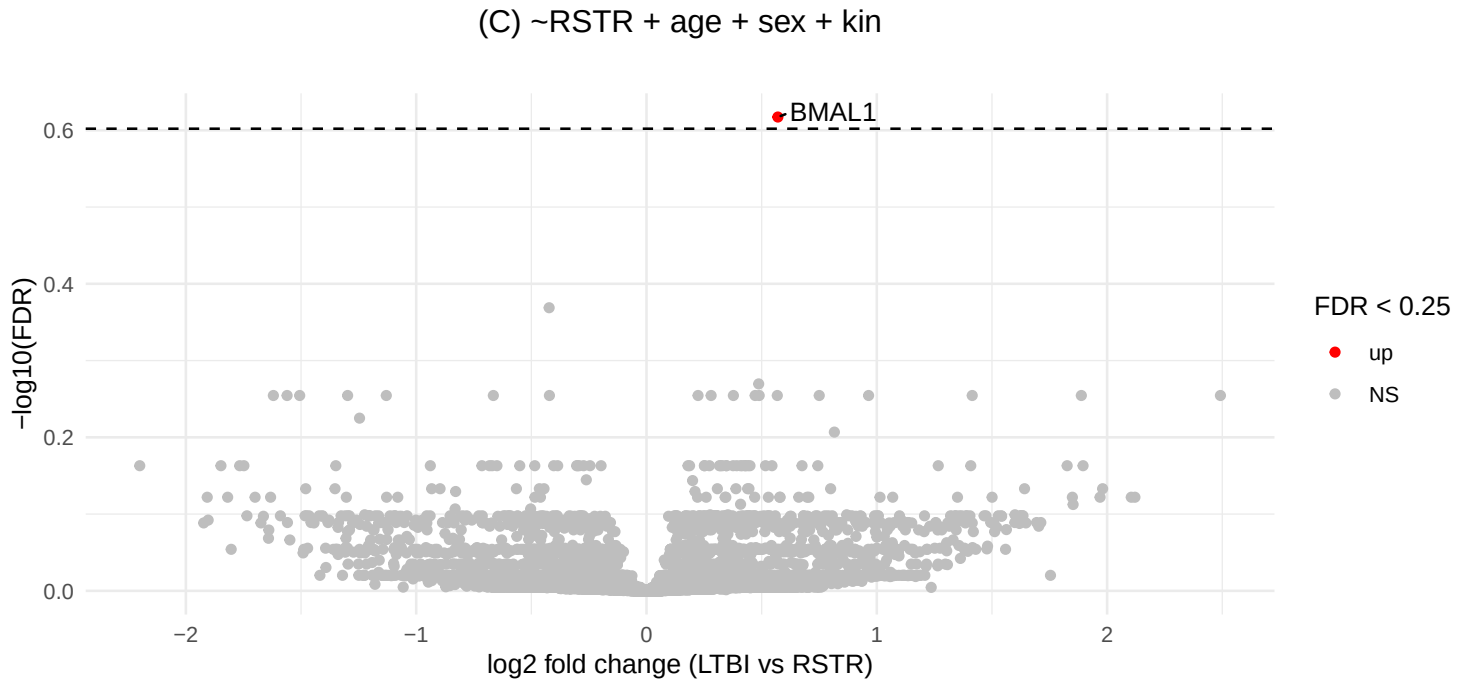
```
##   variable             fdr_0.05 fdr_0.3
##   <fct>                   <int>   <int>
## 1 (1 | FULLIDNO)             NA       2
## 2 KCHCA_AGE_YR_CURRENT       34    1761
## 3 M0_KCVSEX                 180     859
## 4 Sample_Group               NA       1
## 5 total (nonredundant)      214    2445
```

```
lowest_FDR <- lme_rstr_age_sex_kin$lmerel %>%
  filter(variable == 'Sample_Group') %>%
  summarise(FDR = min(FDR))

#print(paste("Lowest FDR wtih age, sex and kin is", round(lowest_FDR, digits=2)))
```

For the model adjusted for age, sex, kinship only DEGs w/ FDR > 0.24.

Volcano Plot -

```
base_age_sex_kin_volcano <- plot_volcano(model_result = lme_rstr_age_sex_kin,
             model = "lmerel", variables = "Sample_Group",
             y.cutoff = 0.25, label = "all",
             genes = dat$genes, genes_label = "symbol")

base_age_sex_kin_volcano <-  base_age_sex_kin_volcano +
  labs(title = "(C) ~RSTR + age + sex + kin", x = "log2 fold change (LTBI vs RSTR)") +
  theme(plot.title = element_text(hjust = 0.5),
        strip.text = element_text(color = NA))

base_age_sex_kin_volcano
```



Plot shows there is 1 DEGs at FDR cutoff of 0.25, the gene BMAL1 which is up in the RSTR group at about 0.5 log2 fold change which is equivalent to roughly 1.5x increase in the number of gene transcripts. (a log2 fold change of 1 means a 2x increase in expression count)

Compare model fits (`Base Model + Age, Sex`, Kin VS `Base Model`)

```
aic_2 <- plot_fit2(model_result = lm_rstr, x="lm",
                model_result_y = lme_rstr_age_sex_kin, y="lmerel",
                metrics = "AIC")
```

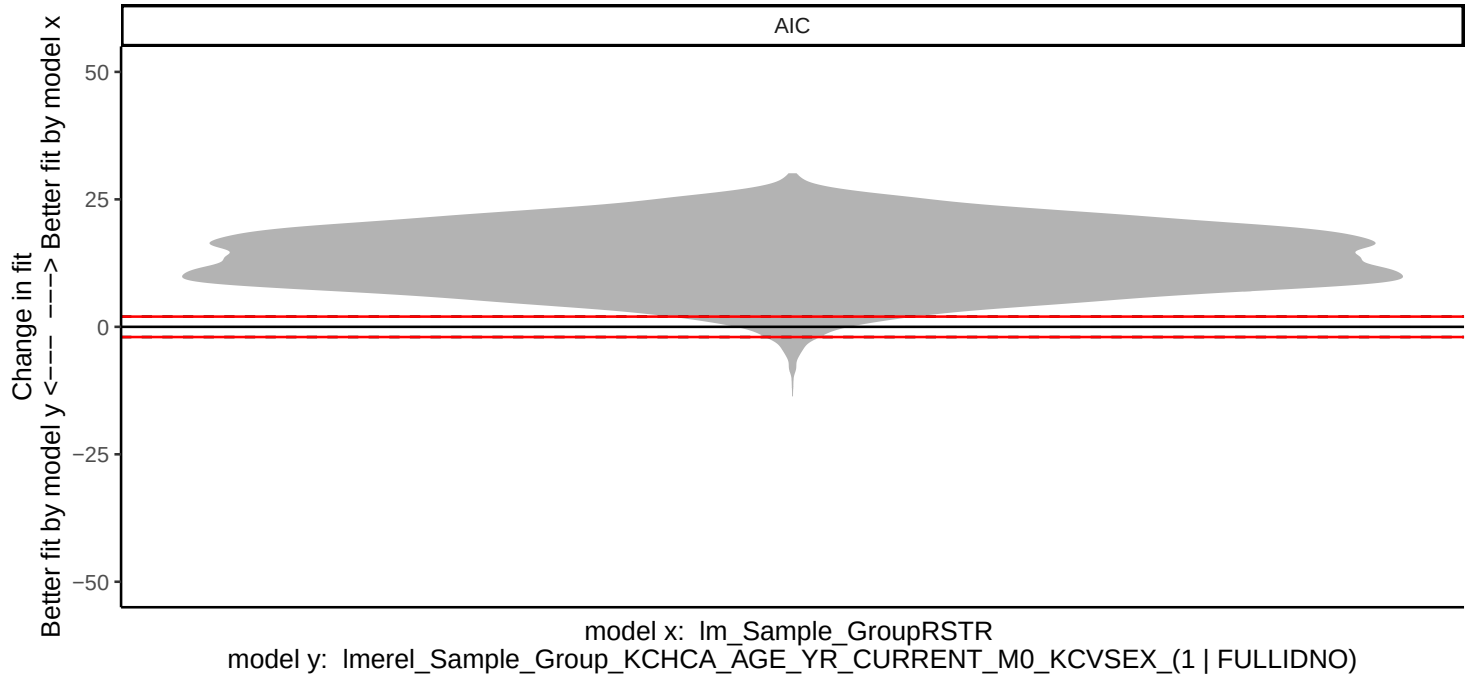
```
## Summary
```

```
##                                                  Best fit Metric
```

```
## 1                                                lm_Sample_GroupRSTR    AIC
## 2 lmerel_Sample_Group_KCHCA_AGE_YR_CURRENT_M0_KCVSEX_(1 | FULLIDNO)    AIC
##   Total genes Mean delta Stdev delta
## 1       13619  13.642242    5.934001
## 2         189   8.298702   29.852645
```

```
aic_2 <- aic_2[[1]] + ylim(-50, 50) + geom_hline(yintercept = c(-2, 2), color = "red", linetype = "solid")
aic_2
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_ydensity()`).
```



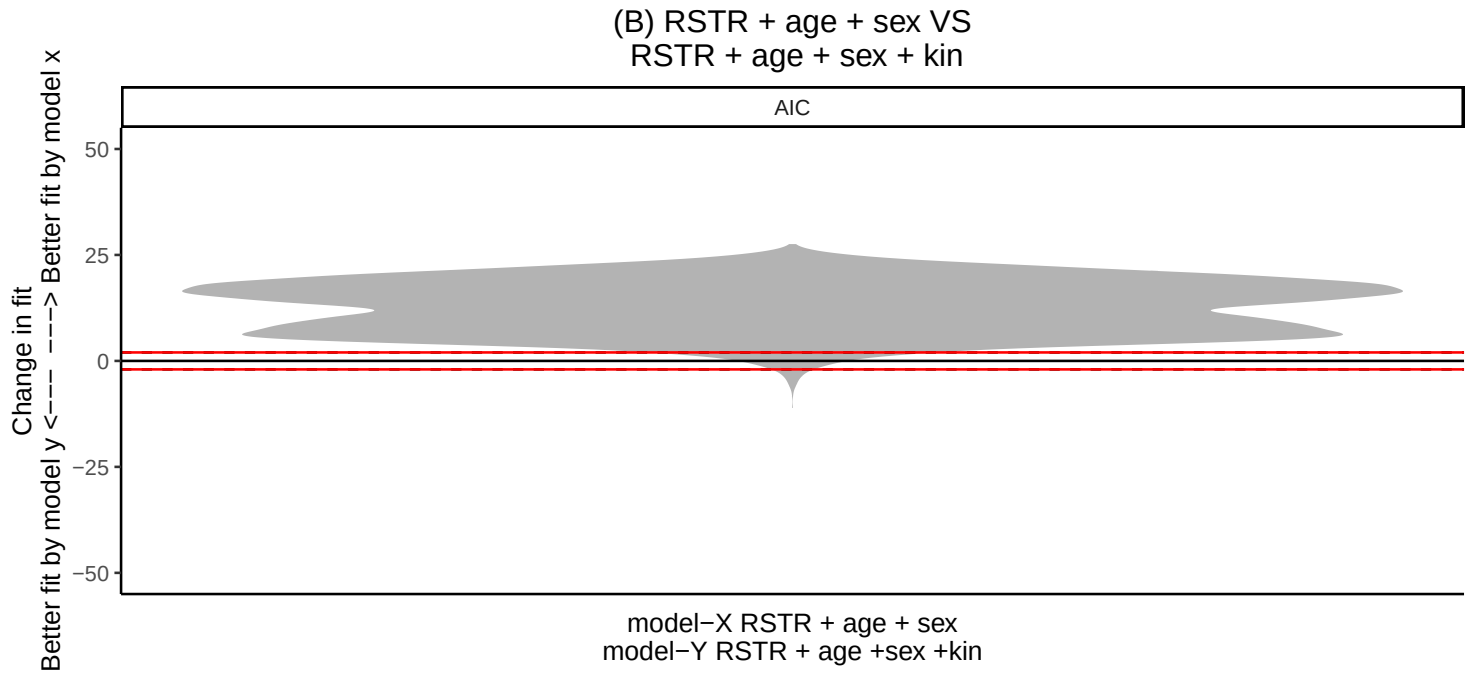
Next, compare model fits (`Base Model + Age, Sex, Kin` VS `Base Model + Age, Sex`)

```r
aic_3 <- plot_fit2(model_result = lm_rstr_age_sex, x="lm",
                model_result_y = lme_rstr_age_sex_kin, y="lmerel",
                metrics = "AIC")
```

```
## Summary
```

```
##                                                          Best fit Metric
## 1               lm_Sample_GroupRSTR_KCHCA_AGE_YR_CURRENT_MO_KCVSEXM    AIC
## 2 lmerel_Sample_Group_KCHCA_AGE_YR_CURRENT_MO_KCVSEX_(1 | FULLIDNO)    AIC
##   Total genes Mean delta Stdev delta
## 1       13667  12.839519    5.931427
## 2         141   1.931337    1.859044
```

```r
aic_3 <- aic_3[[1]] + ylim(-50, 50) + geom_hline(yintercept = c(-2, 2), color = "red", linetype = "solid")

aic_3 <- aic_3 + labs(title = "(B) RSTR + age + sex VS\n RSTR + age + sex + kin",
                      x = "model-X RSTR + age + sex\nmodel-Y RSTR + age +sex +kin") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(margin = margin(t = 5, b = 5)))

aic_3
```



Here we see inclusion of kinship does not improve model fit overall with only small improvement in about 150 genes.

We can also look at BMAL1 specifically to find which model is a better fit.

ENSG00000133794 - BMAL1

```
lm_rstr_age_sex$lm.fit %>%
  filter(gene == "ENSG00000133794")
```

```
## # A tibble: 1 x 7
##   model  gene            sigma   AIC   BIC   Rsq adj_Rsq
##   <chr>  <chr>           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 lm.fit ENSG00000133794 0.811  56.6  66.0 0.391   0.349
```

```
lme_rstr_age_sex_kin$lmerel.fit %>%
  filter(gene == 'ENSG00000133794')
```

```
## # A tibble: 1 x 7
##   model      gene            sigma   AIC   BIC   Rsq adj_Rsq
##   <chr>      <chr>           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 lmerel.fit ENSG00000133794 0.104  72.0  83.3 0.194   0.252
```

BMAL1 is better fit 56 vs 72 in age + sex model. Difference between models 72 - 56.6 = 15.4, therefore 15.4 can be interpreted as high evidence of improved model fit since AIC >10.

...

2.4 Base Model + RiskScore

~ Sample_Group + RiskScore

```
lm_rstr_risk_score <- kmFit(dat = dat,
                        model = "~ Sample_Group + RISK_SCORE",
                        patientID = "FULLIDNO",
                        run.lm = TRUE, use.weights = TRUE,
                        metrics = TRUE)
```

Check for DEGs -

```
summarise_kmFit(fdr = lm_rstr_risk_score$lm)
```

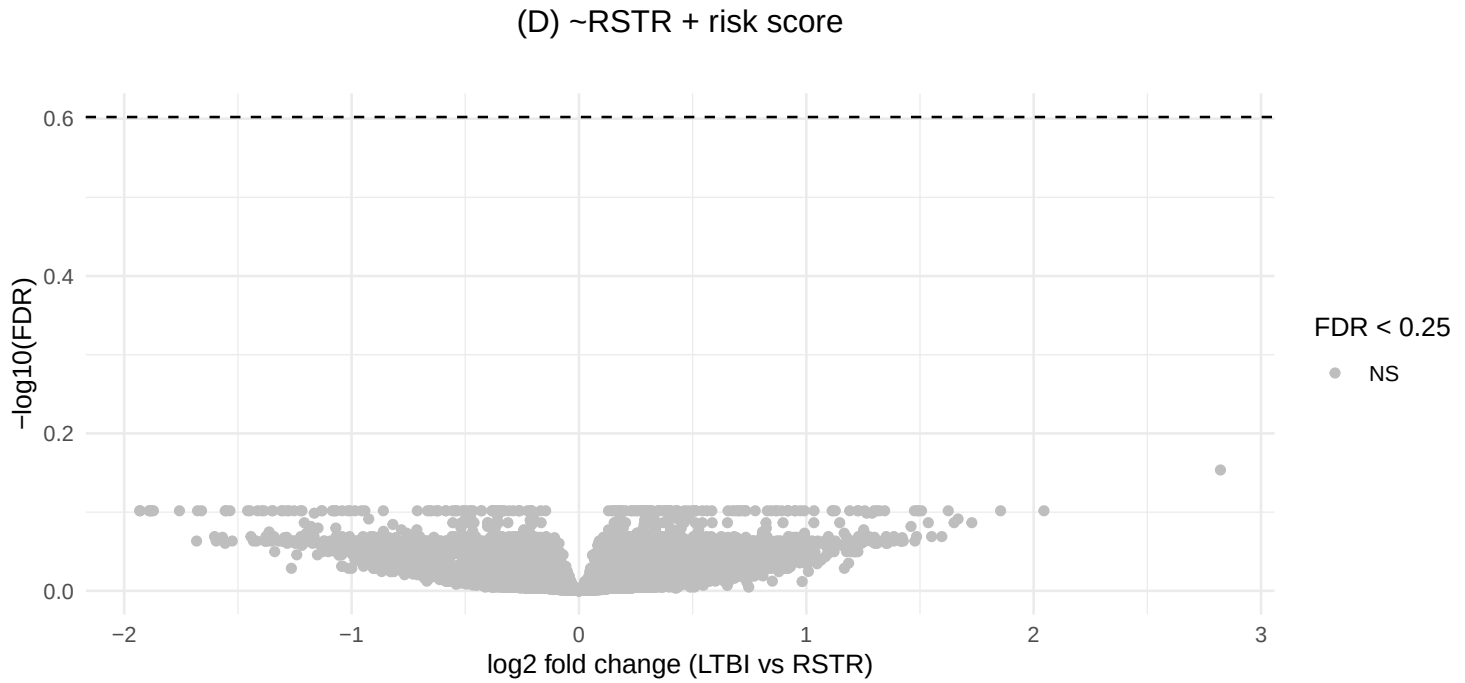
```
## # A tibble: 2 x 7
##   variable             fdr_0.05 fdr_0.1 fdr_0.2 fdr_0.3 fdr_0.4 fdr_0.5
##   <fct>                   <int>   <int>   <int>   <int>   <int>   <int>
## 1 RISK_SCORE                  1       3       8     287     811    1533
## 2 total (nonredundant)        1       3       8     287     811    1533
```

```
lowest_FDR <- lm_rstr_risk_score$lm %>%
  filter(variable == 'Sample_GroupRSTR') %>%
  summarise(FDR = min(FDR))

#print(paste("Lowest FDR with risk score is", round(lowest_FDR, digits=2)))
```

For the model adjusted for exposure risk score only DEGs w/ FDR > 0.7.

Volcano Plot -

```
base_risk_volcano <-  plot_volcano(model_result = lm_rstr_risk_score,
             model = "lm", variables = "Sample_GroupRSTR",
             y.cutoff = 0.25, label = "all",
             genes = dat$genes, genes_label = "symbol")

base_risk_volcano <-  base_risk_volcano +
  labs(title = "(D) ~RSTR + risk score", x = "log2 fold change (LTBI vs RSTR)") +
  theme(plot.title = element_text(hjust = 0.5),
        strip.text = element_text(color = NA))

base_risk_volcano
```

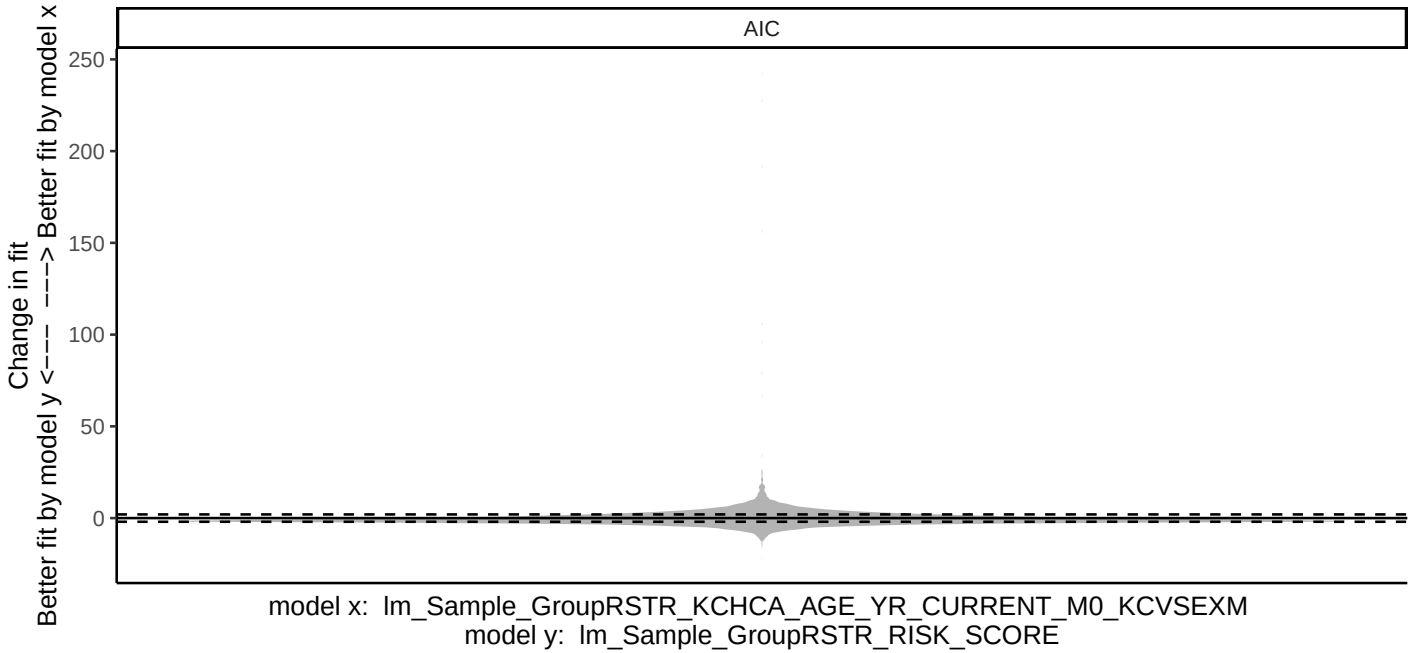


Compare model fits -

```
plot_fit2(model_result = lm_rstr_age_sex, x="lm",
                model_result_y = lm_rstr_risk_score, y="lm",
                metrics = "AIC")
```

```
## Summary

##                                              Best fit Metric Total genes
## 1 lm_Sample_GroupRSTR_KCHCA_AGE_YR_CURRENT_M0_KCVSEXM    AIC        5004
## 2                      lm_Sample_GroupRSTR_RISK_SCORE    AIC        8804
##   Mean delta Stdev delta
## 1   3.244342    7.052639
```

```
## 2   2.146043    1.712110
```



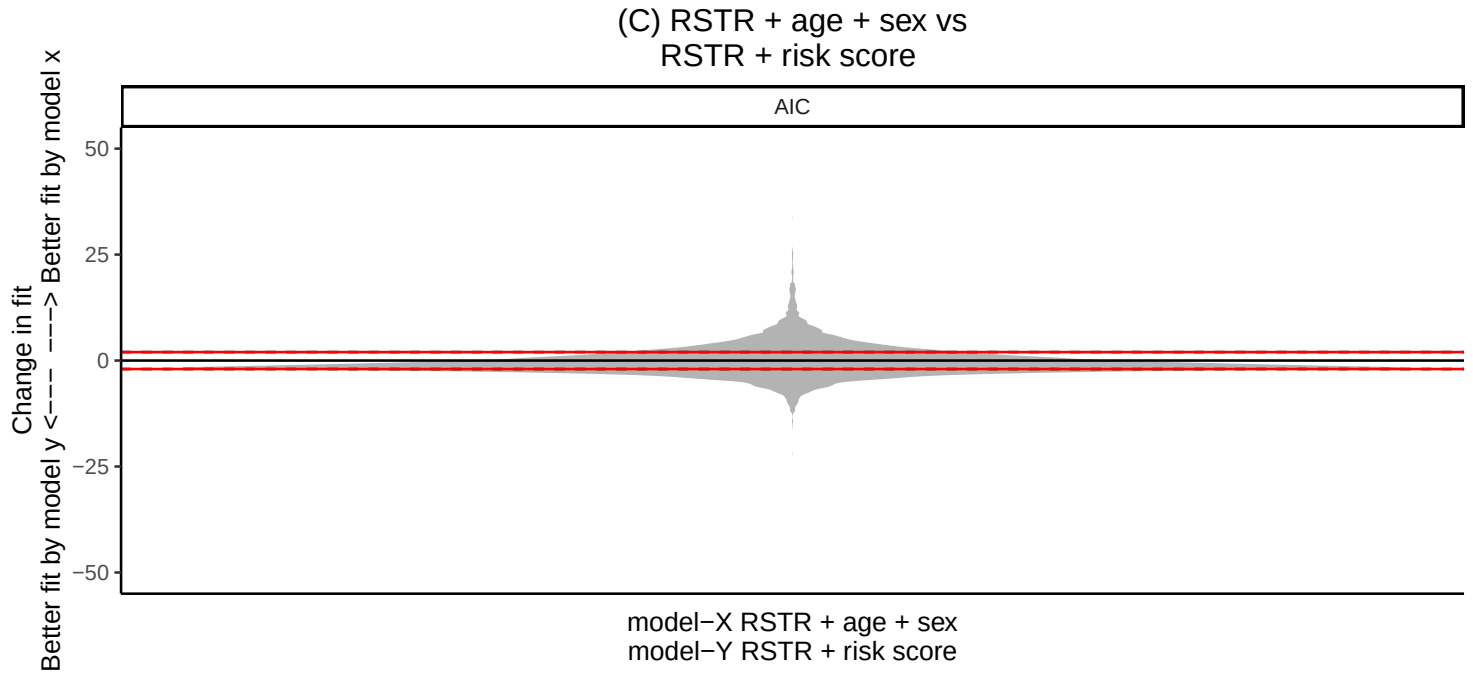

```
aic_4 <- plot_fit2(model_result = lm_rstr_age_sex, x="lm",
                   model_result_y = lm_rstr_risk_score, y="lm",
                   metrics = "AIC")
```

```
## Summary

##                                                Best fit Metric Total genes
## 1 lm_Sample_GroupRSTR_KCHCA_AGE_YR_CURRENT_M0_KCVSEXM    AIC        5004
## 2                      lm_Sample_GroupRSTR_RISK_SCORE    AIC        8804
##   Mean delta Stdev delta
```

```
## 1   3.244342    7.052639
## 2   2.146043    1.712110
```

```
aic_4 <- aic_4[[1]] + ylim(-50, 50) + geom_hline(yintercept = c(-2, 2), color = "red", linetype = "solid")

aic_4 <- aic_4 + labs(title = "(C) RSTR + age + sex vs\n RSTR + risk score  ",
                      x = "model-X RSTR + age + sex\nmodel-Y RSTR + risk score") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(margin = margin(t = 5, b = 5)))

print(aic_4)
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_ydensity()`).
```

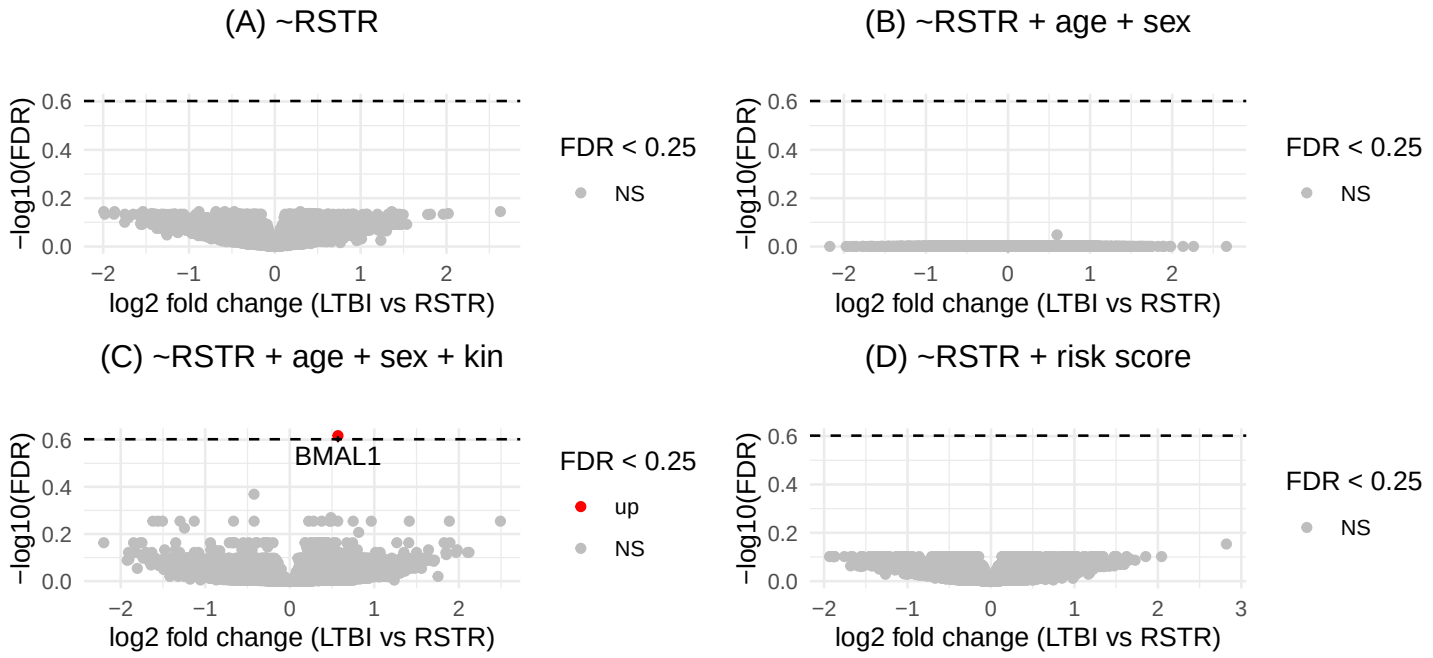


We see risk score does not improve model fit compared to modeling and adjusting for age and sex.

3. Combined plots and Save data

Combine DEG plots for all linear model -

```
DEGs <- base_volcano + base_age_sex_volcano + base_age_sex_kin_volcano + base_risk_volcano

DEGs
```



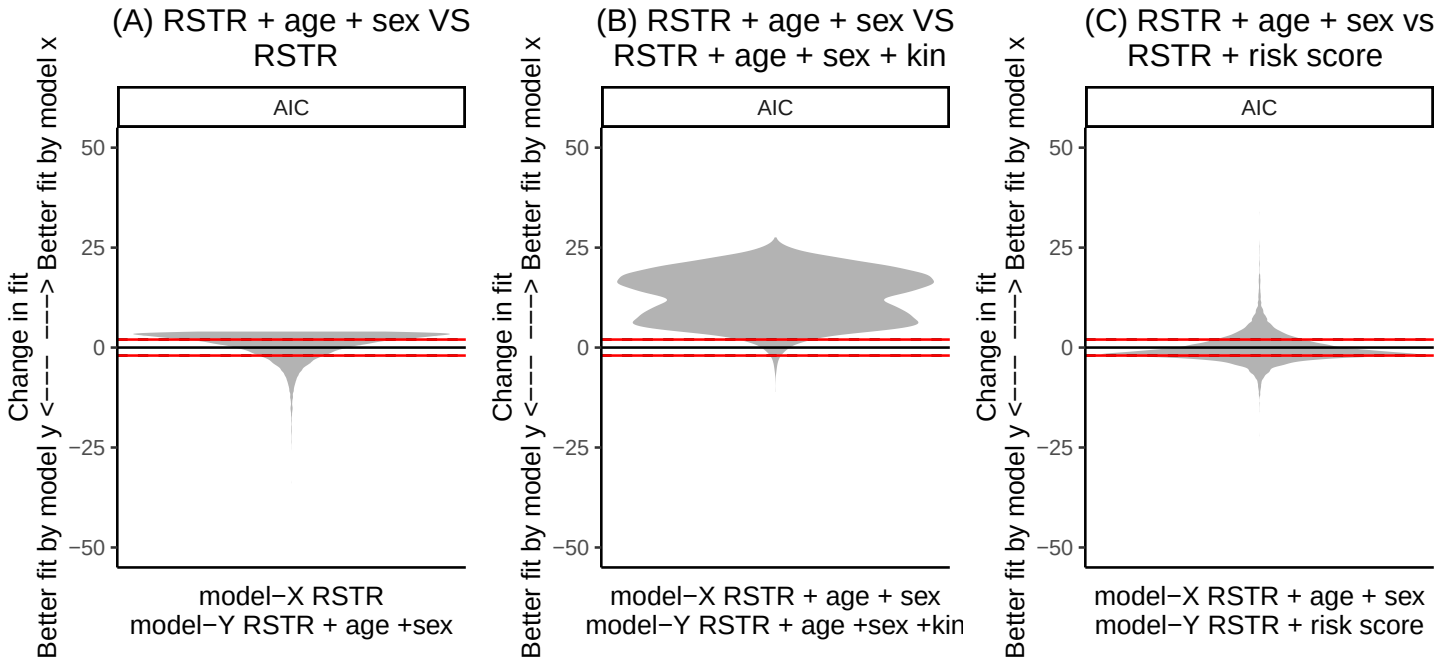
```
ggsave(paste0("results/", "combined_DEGs.png"), DEGs)
```

```
## Saving 8.5 x 4 in image
```

Combined AIC plots for all linear models -

```
AICs <- aic_1 + aic_3 + aic_4

AICs
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_ydensity()`).
## Removed 8 rows containing non-finite values (`stat_ydensity()`).
```



```
ggsave(paste0("results/", "combined_AICs.png"), AICs)
```

```
## Saving 8.5 x 4 in image
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_ydensity()`).
## Removed 8 rows containing non-finite values (`stat_ydensity()`).
```

```
# Get the list of all objects
all_objects <- ls()

# Filter the objects starting with 'lm'
lm_objects <- all_objects[startsWith(all_objects, "lm")]

# Save all lm_objects into a single .Rdata file
do.call(save, c(lm_objects, list(file="results/lm_data.Rdata")))
```

R session

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
```

```
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggpubr_0.6.0     patchwork_1.1.2  edgeR_3.40.2     limma_3.54.2
##  [5] SEARchways_1.0.0 BIGpicture_1.1.0 RNAetc_1.0.0     kimma_1.4.4
##  [9] BIGverse_1.0.0   lubridate_1.9.2  forcats_1.0.0    stringr_1.5.0
## [13] dplyr_1.1.2      purrr_1.0.1      readr_2.1.4      tidyr_1.3.0
## [17] tibble_3.2.1     ggplot2_3.4.2    tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] ggrepel_0.9.3     Rcpp_1.0.10       locfit_1.5-9.7    lattice_0.21-8
##  [5] digest_0.6.31     foreach_1.5.2     utf8_1.2.3        R6_2.5.1
##  [9] backports_1.4.1   evaluate_0.21     httr_1.4.6        highr_0.10
## [13] pillar_1.9.0      rlang_1.1.1       rstudioapi_0.14   car_3.1-2
## [17] rmarkdown_2.22    textshaping_0.3.6 labeling_0.4.2    bit_4.0.5
## [21] munsell_0.5.0     broom_1.0.4       compiler_4.2.2    xfun_0.39
## [25] systemfonts_1.0.4 pkgconfig_2.0.3   htmltools_0.5.5   tidyselect_1.2.0
## [29] codetools_0.2-19  fansi_1.0.4       crayon_1.5.2      tzdb_0.4.0
## [33] withr_2.5.0       grid_4.2.2        gtable_0.3.3      lifecycle_1.0.3
## [37] magrittr_2.0.3    scales_1.2.1      cli_3.6.1         stringi_1.7.12
## [41] vroom_1.6.3       carData_3.0-5     farver_2.1.1      ggsignif_0.6.4
## [45] ragg_1.2.5        generics_0.1.3    vctrs_0.6.2       iterators_1.0.14
## [49] tools_4.2.2       bit64_4.0.5       glue_1.6.2        hms_1.1.3
## [53] parallel_4.2.2    abind_1.4-5       fastmap_1.1.1     yaml_2.3.7
## [57] timechange_0.2.0  colorspace_2.1-0  rstatix_0.7.2     knitr_1.43
```

Uganda RSTR Alveolar Mac RNAseq

Part 3: Gene Set Enrichment Analysis

Contents

Overview

This analysis performs Gene Set Enrichment Analysis (GSEA) using the Broad MSigDB molecular signatures database.

The analysis will look for:

- HALLMARK gene sets
- C1-C7 Broad gene targets
- GO terms
- KEGG
- Xue Modules

Linear Modeling Covariate Dictionary: - `0`: No covariates - `A`: Age - `S`: Sex - `K`: Kin - `R`: Risk Score

More information on GSEA can be found at Broad Institute GSEA page.

0. Setup

Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(BIGverse)
```

```
## -- Attaching packages ---------------------------------------- BIGverse 1.0.0 --
## v kimma      1.4.4     v BIGpicture 1.1.0
## v RNAetc     1.0.0     v SEARchways 1.0.0
## -- Conflicts ------------------------------------------ BIGverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork)
library(readxl)
```

Load data

Load results from the precomputed linear modeling analysis

```
load("results/lm_data.Rdata")
```

```
attach("data_clean/P512_voom.RData")
dat <- dat.abund.norm.voom
```

Gene set enrichment analysis (GSEA)

Data formatting

```
# Base model no covar
gene_0 <- lm_rstr$lm %>%
  filter(variable %in% c("Sample_GroupRSTR")) %>%
  mutate(group="0") %>%
  select(group, gene, estimate)
```

```
# +age+sex
gene_AS <- lm_rstr_age_sex$lm %>%
  filter(variable %in% c("Sample_GroupRSTR")) %>%
  mutate(group="AS") %>%
  select(group, gene, estimate)

# +age+sex+kin
gene_ASK <- lme_rstr_age_sex_kin$lmerel %>%
  filter(variable %in% c("Sample_Group")) %>%
  mutate(group="ASK") %>%
  select(group, gene, estimate)

# +risk score
gene_R <- lm_rstr_risk_score$lm %>%
  filter(variable %in% c("Sample_GroupRSTR")) %>%
  mutate(group="R") %>%
  select(group, gene, estimate)

# Combine
gene_all <- bind_rows(gene_0, gene_AS, gene_ASK, gene_R)
```

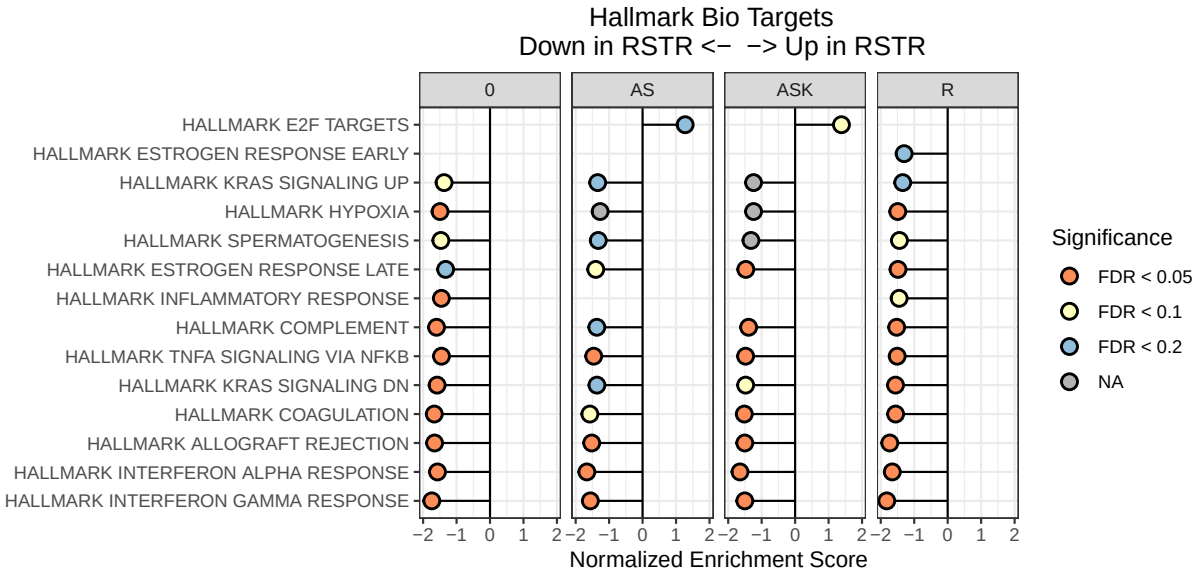
1. Hallmark

HALLMARK gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "H")
```

```
## 0

## AS

## ASK

## R
```

```
p_h <- plot_gsea(gsea.h,
          fdr.cutoff = 0.25)

p_h <- p_h + ggtitle('Hallmark Bio Targets  \n Down in RSTR <-  -> Up in RSTR') + theme(plot.title = el

p_h
```




```
# Format to save E2F leading edge genes
gsea.h %>% filter(pathway=="HALLMARK_E2F_TARGETS") %>%
  pull(leadingEdge)
```

```
## [[1]]
##  [1] "ENSG00000129173" "ENSG00000122966" "ENSG00000152253" "ENSG00000105173"
##  [5] "ENSG00000119969" "ENSG00000131747" "ENSG00000139734" "ENSG00000131470"
##  [9] "ENSG00000090889" "ENSG00000100916" "ENSG00000165304" "ENSG00000146263"
## [13] "ENSG00000159147" "ENSG00000135473" "ENSG00000156802" "ENSG00000170836"
## [17] "ENSG00000111665" "ENSG00000136982" "ENSG00000128050" "ENSG00000100162"
## [21] "ENSG00000071539" "ENSG00000133119" "ENSG00000100941" "ENSG00000091651"
## [25] "ENSG00000111581" "ENSG00000076382" "ENSG00000083642" "ENSG00000186185"
## [29] "ENSG00000205339" "ENSG00000118007" "ENSG00000113522"
##
## [[2]]
##  [1] "ENSG00000152253" "ENSG00000131747" "ENSG00000119969" "ENSG00000090889"
##  [5] "ENSG00000156970" "ENSG00000136982" "ENSG00000122966" "ENSG00000076382"
##  [9] "ENSG00000129173" "ENSG00000156802" "ENSG00000165304" "ENSG00000105173"
## [13] "ENSG00000100162" "ENSG00000100916" "ENSG00000100941" "ENSG00000111665"
## [17] "ENSG00000159147" "ENSG00000131470" "ENSG00000146263" "ENSG00000170836"
## [21] "ENSG00000133119" "ENSG00000035928" "ENSG00000135473" "ENSG00000128050"
## [25] "ENSG00000111581" "ENSG00000117399" "ENSG00000139618" "ENSG00000138778"
## [29] "ENSG00000130816" "ENSG00000108848" "ENSG00000205339" "ENSG00000113456"
## [33] "ENSG00000071539" "ENSG00000138376" "ENSG00000111276" "ENSG00000113522"
## [37] "ENSG00000139734" "ENSG00000101057"
##
## [[3]]
##  [1] "ENSG00000090889" "ENSG00000152253" "ENSG00000131747" "ENSG00000119969"
##  [5] "ENSG00000156970" "ENSG00000129173" "ENSG00000136982" "ENSG00000122966"
##  [9] "ENSG00000076382" "ENSG00000156802" "ENSG00000165304" "ENSG00000105173"
## [13] "ENSG00000100162" "ENSG00000100941" "ENSG00000111665" "ENSG00000159147"
## [17] "ENSG00000131470" "ENSG00000146263" "ENSG00000133119" "ENSG00000170836"
## [21] "ENSG00000128050" "ENSG00000035928" "ENSG00000135473" "ENSG00000113456"
## [25] "ENSG00000100916" "ENSG00000111581" "ENSG00000167900" "ENSG00000139618"
```

```
## [29] "ENSG00000138778" "ENSG00000130816" "ENSG00000111276" "ENSG00000108848"
## [33] "ENSG00000205339" "ENSG00000071539" "ENSG00000239672" "ENSG00000138376"
## [37] "ENSG00000113522" "ENSG00000139734" "ENSG00000101057" "ENSG00000148773"
## [41] "ENSG00000075131" "ENSG00000076242" "ENSG00000117399" "ENSG00000010292"
## [45] "ENSG00000083642"
##
## [[4]]
##  [1] "ENSG00000122966" "ENSG00000152253" "ENSG00000129173" "ENSG00000119969"
##  [5] "ENSG00000131747" "ENSG00000146263" "ENSG00000105173" "ENSG00000090889"
##  [9] "ENSG00000139734" "ENSG00000136982" "ENSG00000100916" "ENSG00000156802"
## [13] "ENSG00000165304" "ENSG00000111665" "ENSG00000091651" "ENSG00000100941"
## [17] "ENSG00000159147" "ENSG00000131470" "ENSG00000170836" "ENSG00000135473"
## [21] "ENSG00000128050" "ENSG00000113522" "ENSG00000133119" "ENSG00000071539"
## [25] "ENSG00000111581" "ENSG00000083642" "ENSG00000010292" "ENSG00000205339"
## [29] "ENSG00000076382" "ENSG00000130816" "ENSG00000100162" "ENSG00000138778"
```

```
E2F_leadingEdge <- gsea.h %>%
  unnest(leadingEdge) %>%
  filter(pathway=="HALLMARK_E2F_TARGETS") %>%
  filter(group == "AS") %>%
  left_join(dat$genes, by=c("leadingEdge"="ensembl_gene_id")) %>%
  unnest(symbol)
```

```
## Loading required package: limma
```

2. Broad MSigDB C2-C8 computional targets

Next we plot Broad C2 targets curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C2")
```

```
## 0
```

```
## AS
```

```
## ASK
```

```
## R
```

```
p_c2 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.15)

p_c2 <- p_c2 + ggtitle('2') + theme(plot.title = element_text(hjust = 0.5))

p_c2
```

Next we plot Broad C3 targets regulatory target gene sets based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C3")
```

```
## 0

## AS

## ASK

## R
```

```
p_c3 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.05)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.05): No gene sets are significant. Please increase fdr_cut
```

```
p_c3 <- p_c3 + ggtitle('C1') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_c3' not found
```

```
p_c3
```

```
## Error in eval(expr, envir, enclos): object 'p_c3' not found
```

Next we plot Broad C4 -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C4")
```

```
## 0

## AS

## ASK

## R
```

```
p_c4 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.05)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.05): No gene sets are significant. Please increase fdr_cut
```

```
p_c4 <- p_c4 + ggtitle('C4') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_c4' not found
```

```
p_c4
```

```
## Error in eval(expr, envir, enclos): object 'p_c4' not found
```

Next we plot Broad C5 -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C5")
```

```
## 0

## AS

## ASK

## R
```

```
p_c5 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.05)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.05): No gene sets are significant. Please increase fdr_cut
```

```
p_c5 <- p_c5 + ggtitle('C5') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_c5' not found
```

```
p_c5
```

```
## Error in eval(expr, envir, enclos): object 'p_c5' not found
```

Next we plot Broad C6 -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C6")
```

```
## 0
```

```
## AS
```

```
## ASK
```

```
## R
```

```
p_c6 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.05)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.05): No gene sets are significant. Please increase fdr_cut
```

```
p_c6 <- p_c6 + ggtitle('C6') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_c6' not found
```

```
p
```

```
## Error in eval(expr, envir, enclos): object 'p' not found
```

Next we plot Broad C7 immunologic signature gene sets represent cell states and perturbations within the immune system.

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C7")
```

```
## 0
```

```
## AS

## ASK

## R
```

```
p_c7 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.05)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.05): No gene sets are significant. Please increase fdr_cut
```

```
p_c7 <- p_c7 + ggtitle('C7') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_c7' not found
```

```
p_c7
```

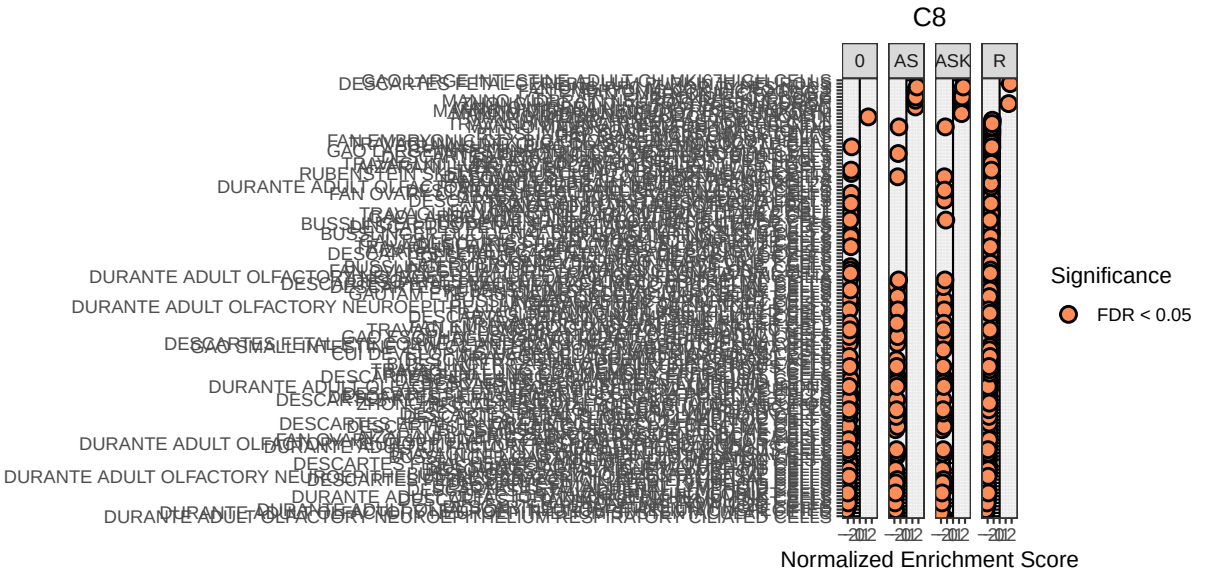
```
## Error in eval(expr, envir, enclos): object 'p_c7' not found
```

Next we plot Broad C8

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                   ID = "ENSEMBL",
                   category = "C8")
```

```
## 0

## AS

## ASK

## R
```

```
p_c8 <- plot_gsea(gsea.h,
          fdr.cutoff = 0.05)

p_c8 <- p_c8 + ggtitle('C8') + theme(plot.title = element_text(hjust = 0.5))

p_c8
```



3. GO biological pathways

Plot GO Biological Proceses (BP) -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C5",
                subcategory = "GO:BP")
```

```
## 0

## AS

## ASK

## R
```

```
p_bp <- plot_gsea(gsea.h,
          fdr.cutoff = 0.3)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.3): No gene sets are significant. Please increase fdr_cutoff
```

```
p_bp <- p_bp + ggtitle('GO Bio Pathways') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_bp' not found
```

```
p_bp
```

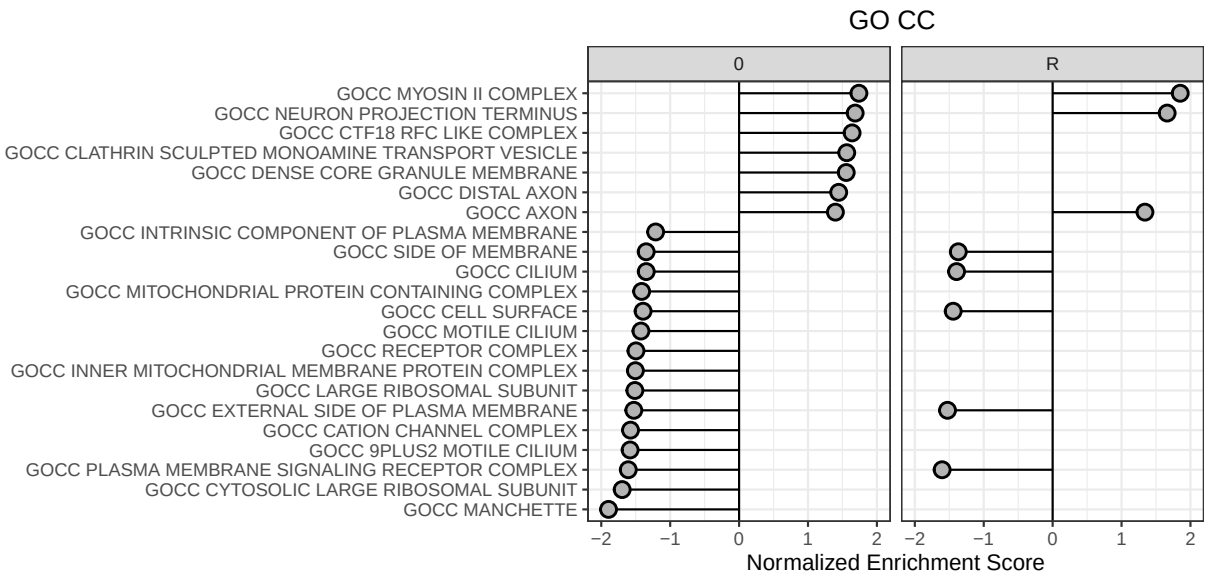
```
## Error in eval(expr, envir, enclos): object 'p_bp' not found
```

Next we plot GO Cellular Component (CC) -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C5",
                subcategory = "GO:CC")
```

```
## 0

## AS

## ASK

## R
```

```
p_cc <- plot_gsea(gsea.h,
          fdr.cutoff = 0.3)

p_cc <- p_cc + ggtitle('GO CC ') + theme(plot.title = element_text(hjust = 0.5))

p_cc
```



Next we plot GO Molecuar Function (MF) -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C5",
                subcategory = "GO:MF")
```

```
## 0

## AS

## ASK

## R
```

```
p_mf <- plot_gsea(gsea.h,
          fdr.cutoff = 0.3)
```

```
## Error in plot_gsea(gsea.h, fdr.cutoff = 0.3): No gene sets are significant. Please increase fdr_cuto
```

```
p_mf <- p_mf + ggtitle('GO MF ') + theme(plot.title = element_text(hjust = 0.5))
```

```
## Error in eval(expr, envir, enclos): object 'p_mf' not found
```

```
p_mf
```

```
## Error in eval(expr, envir, enclos): object 'p_mf' not found
```

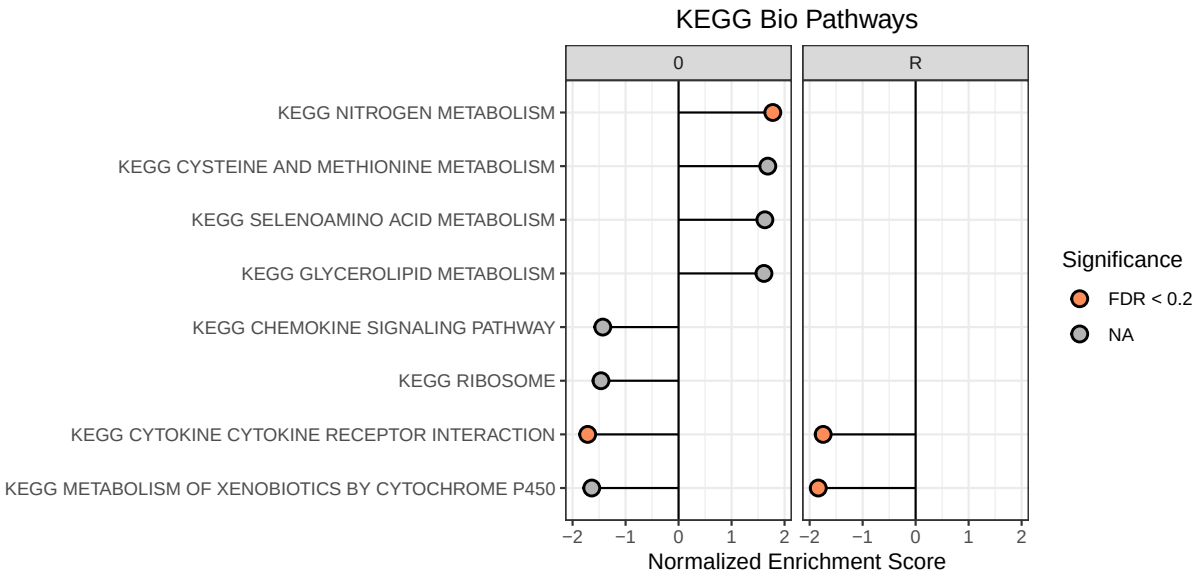
4. KEGG

Next plot KEGG -

```
set.seed(345)

gsea.h <- BIGsea(gene_df = gene_all,
                 ID = "ENSEMBL",
                 category = "C2",
                subcategory = "CP:KEGG")
```

```
## 0

## AS

## ASK

## R
```

```
p_kegg <- plot_gsea(gsea.h,
          fdr.cutoff = 0.3)

p_kegg <- p_kegg + ggtitle('KEGG Bio Pathways') + theme(plot.title = element_text(hjust = 0.5))

p_kegg
```



5. Xue weighted gene correlation netwoks analsys (WGCNA) modules

```
# Load genes from Xue modules
xue <- read_excel("data_raw/Xue/mmc3.xlsx", sheet="Table S2B", skip=3) %>%
  pivot_longer(-Genes, values_to = "symbol") %>%
  mutate(module=ifelse(as.numeric(name) <= 9,
                       paste("0", name, sep=""),name),
         module=paste("module", module, sep="_")) %>%
  dplyr::select(module, symbol) %>%
  drop_na(symbol)

# Format to list like Broad GMT
xue.ls <- list()

for(mod in unique(xue$module)){
  temp <- xue %>%
    filter(module == mod) %>%
    distinct(symbol) %>% unlist(use.names = FALSE)
  xue.ls[[mod]] <- temp
}
```

Calculate the gene list from linear model -

```
gene_AS <- lm_rstr_age_sex$lm %>%
  filter(variable %in% c("Sample_GroupRSTR")) %>%
  select(variable, gene, estimate)
```

Prepare dataframes into correct format. Also note initial run did not work since Xue modules use diff gene names.

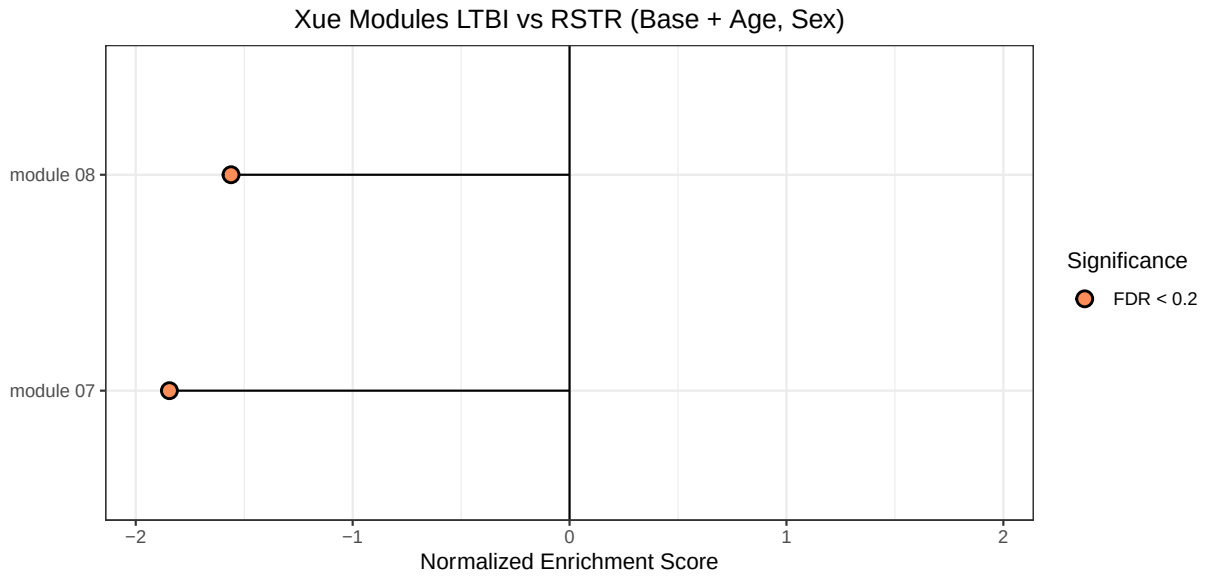
```r
# Use the human gene database
ensembl <- biomaRt::useMart("ensembl", dataset = "hsapiens_gene_ensembl")

# Get the gene symbol for your list of Ensembl gene IDs
gene_symbols <- biomaRt::getBM(
  filters = "ensembl_gene_id",
  attributes = c("ensembl_gene_id", "hgnc_symbol"),
  values = gene_AS$gene,
  mart = ensembl
)

gene_ASX <- merge(gene_AS, gene_symbols, by.x = "gene", by.y = "ensembl_gene_id")

gene_ASX <- gene_ASX %>% dplyr::select(variable, hgnc_symbol, estimate)

gene_ASX <- rename(gene_ASX, gene = hgnc_symbol)

# Get gene list into correct formar for BIGSea
gene_ASX_vector <- setNames(gene_ASX$estimate, gene_ASX$gene)
gene_ASX_list <- list(RSTR = gene_ASX_vector)
```

Plot enrichment in RSTR vs LTBI using Xue modules.

```r
set.seed(123)

gsea.all <- BIGsea(gene_ASX_list, db = xue)
```

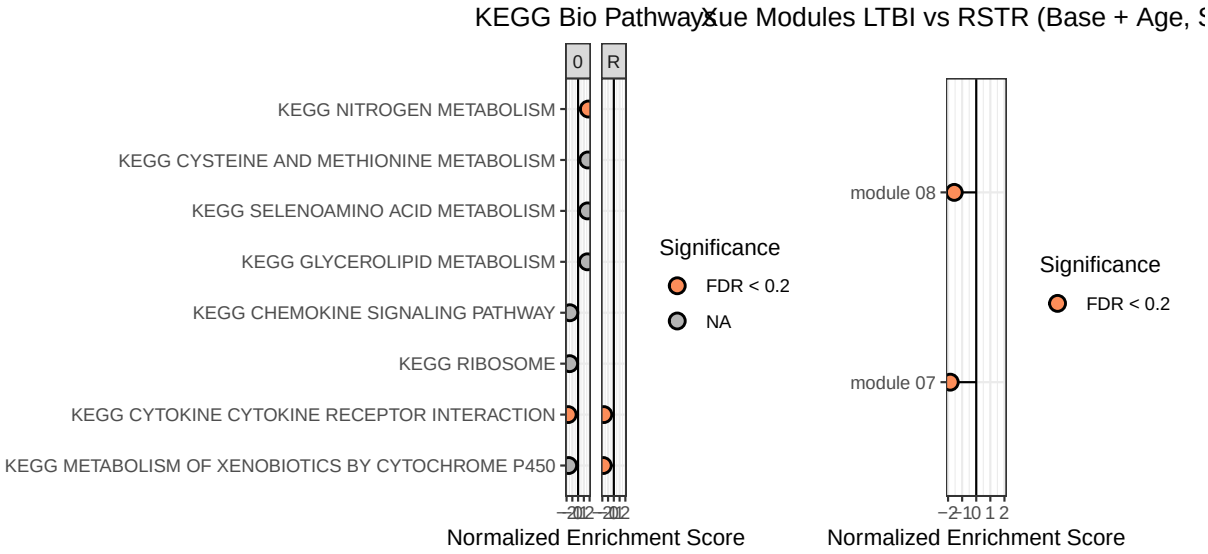
```
## RSTR

## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize, gseaParam, : There are ties in
## The order of those tied genes will be arbitrary, which may produce unexpected results.

## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize,
## gseaParam, : There are duplicate gene names, fgsea may produce unexpected
## results.
```

```r
p_xue <- plot_gsea(gsea.all,
          fdr.cutoff = 0.25)

p_xue <- p_xue + ggtitle('Xue Modules LTBI vs RSTR (Base + Age, Sex)') + theme(plot.title = element_tex

print(p_xue)
```



Combine plots

```
GSEA_plots <-  p_kegg | p_xue
GSEA_plots
```



Save

Save leading edge genes for E2F -

```
# Write E2F leading edge genes to CSV
write_csv(data.frame(gene = E2F_leadingEdge$symbol),
          file = "data_clean/E2F_LeadingEdge_Genes.csv")
```

```
# Save E2F GSEA plot
ggsave(paste0("results/", 'e2f_gsea.png'), p_h)
```

```
## Saving 8.5 x 4 in image
```

R session

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] limma_3.54.2     readxl_1.4.2     patchwork_1.1.2  SEARchways_1.0.0
##  [5] BIGpicture_1.1.0 RNAetc_1.0.0     kimma_1.4.4      BIGverse_1.0.0
##  [9] lubridate_1.9.2  forcats_1.0.0    stringr_1.5.0    dplyr_1.1.2
## [13] purrr_1.0.1      readr_2.1.4      tidyr_1.3.0      tibble_3.2.1
## [17] ggplot2_3.4.2    tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-7           bit64_4.0.5            filelock_1.0.2
##  [4] progress_1.2.2         RColorBrewer_1.1-3     httr_1.4.6
##  [7] GenomeInfoDb_1.34.9    tools_4.2.2            utf8_1.2.3
## [10] R6_2.5.1               DBI_1.1.3              BiocGenerics_0.44.0
## [13] colorspace_2.1-0       withr_2.5.0            prettyunits_1.1.1
## [16] tidyselect_1.2.0       curl_5.0.0             bit_4.0.5
## [19] compiler_4.2.2         textshaping_0.3.6      cli_3.6.1
## [22] Biobase_2.58.0         xml2_1.3.4             labeling_0.4.2
## [25] scales_1.2.1           rappdirs_0.3.3         systemfonts_1.0.4
## [28] digest_0.6.31          rmarkdown_2.22         XVector_0.38.0
## [31] pkgconfig_2.0.3        htmltools_0.5.5        dbplyr_2.3.2
## [34] fastmap_1.1.1          highr_0.10             rlang_1.1.1
## [37] rstudioapi_0.14        RSQLite_2.3.1          farver_2.1.1
## [40] generics_0.1.3         vroom_1.6.3            BiocParallel_1.32.6
## [43] RCurl_1.98-1.12        magrittr_2.0.3         GenomeInfoDbData_1.2.9
## [46] Matrix_1.5-4.1         Rcpp_1.0.10            munsell_0.5.0
## [49] S4Vectors_0.36.2       fansi_1.0.4            babelgene_22.9
## [52] lifecycle_1.0.3        stringi_1.7.12         yaml_2.3.7
## [55] zlibbioc_1.44.0        BiocFileCache_2.6.1    grid_4.2.2
```

```
## [58] blob_1.2.4             parallel_4.2.2         crayon_1.5.2
## [61] lattice_0.21-8         Biostrings_2.66.0      cowplot_1.1.1
## [64] msigdbr_7.5.1          hms_1.1.3              KEGGREST_1.38.0
## [67] knitr_1.43             pillar_1.9.0           fgsea_1.24.0
## [70] codetools_0.2-19       biomaRt_2.54.1         stats4_4.2.2
## [73] fastmatch_1.1-3        XML_3.99-0.14          glue_1.6.2
## [76] evaluate_0.21          data.table_1.14.8      vctrs_0.6.2
## [79] png_0.1-8              tzdb_0.4.0             foreach_1.5.2
## [82] cellranger_1.1.0       gtable_0.3.3           cachem_1.0.8
## [85] xfun_0.39              ragg_1.2.5             iterators_1.0.14
## [88] AnnotationDbi_1.60.2   memoise_2.0.1          IRanges_2.32.0
## [91] timechange_0.2.0
```