Direct inference and control of genetic population structure from RNA sequencing data

Muhamad Fachrul[1,2,3,*], Abhilasha Karkey[4,5], Mila Shakya[4,5], Louise M. Judd[6], Taylor Harshegyi[6], Kar Seng Sim[7], Susan Tonks[8], Sabina Dongol[4,5], Rajendra Shrestha[5], Agus Salim[9,10,11], STRATAA study group, Stephen Baker[12], Andrew J. Pollard[8], Chiea Chuen Khor[7], Christiane Dolecek[13,14], Buddha Basnyat[4,13], Sarah J. Dunstan[15], Kathryn E. Holt[6,16], Michael Inouye[1,2,17,18,19,20,*]

[1] Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[2] Department of Clinical Pathology, University of Melbourne, Parkville, Victoria, Australia
[3] School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia
[4] Oxford University Clinical Research Unit, Patan Academy of Health Sciences, Kathmandu, Nepal
[5] Patan Academy of Health Sciences, Patan Hospital, Lalitpur, Nepal
[6] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia
[7] Genome Institute of Singapore, Singapore
[8] Oxford Vaccine Group, Department of Paediatrics, University of Oxford, and the NIHR Oxford Biomedical Research Centre, Oxford, UK
[9] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia
[10] School of Mathematics and Statistics, The University of Melbourne, Melbourne, Victoria, Australia.
[11] Department of Population Health, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[12] Department of Medicine, University of Cambridge, Cambridge, UK
[13] Nuffield Department of Medicine, Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK
[14] Mahidol Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok, Thailand
[15] The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC, Australia
[16] Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, UK
[17] Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
[18] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
[19] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[20] Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge UK

*Correspondence: M.F. (mfachrul@student.unimelb.edu.au) and M.I. (mi336@medschl.cam.ac.uk)

Abstract

RNAseq data can be used to infer genetic variants, yet its use for estimating genetic population structure remains underexplored. Here, we construct a freely available computational tool (RGStraP) to estimate RNAseq-based genetic principal components (RG-PCs) and assess whether RG-PCs can be used to control for population structure in gene expression analyses. Using whole blood samples from understudied Nepalese populations and the Geuvadis study, we show that RG-PCs had comparable results to paired array-based genotypes, with high genotype concordance and high correlations of genetic principal components, capturing subpopulations within the dataset. In differential gene expression analysis, we found that inclusion of RG-PCs as covariates reduced test statistic inflation. Our paper demonstrates that genetic population structure can be directly inferred and controlled for using RNAseq data, thus facilitating improved retrospective and future analyses of transcriptomic data.

Introduction

RNA sequencing (RNAseq) has revolutionized our understanding of the transcriptome, offering both accurate quantification method for gene expression as well as identification of specific alternative splicing sites and cell-type-specific transcripts[1,2]. Its application extends to the clinical setting, allowing us to further elucidate complex diseases and identify prospective biomarkers in both communicable and non-communicable diseases[3].

Yet, studies using RNAseq rarely consider the germline genetic variation also contained within RNAseq read sets. Studies which do not leverage this information may be vulnerable to bias and confounding, such as population stratification, which may affect transcription between groups[4–7]. To overcome this issue, researchers have typically relied on genome-wide array or whole genome sequence (WGS) data matched for the same individuals with RNAseq. This allows researchers to deploy approaches to control for population stratification, such as the calculation of genetic principal components (PCs) and their use as covariates in subsequent statistical association models[8–10]. The genetic PCs are taken to represent the latent genetic structure within and between populations, which introduce confounding due to differences in social environment[11] or (in the case of differential gene expression) due to heterogeneity of quantitative trait loci between groups. However, the need for genome-wide array or WGS to match with RNAseq data is potentially unnecessary and indeed may not be possible in settings where resources are limited, such as Low and Low Middle Income Countries (LMICs) with highly diverse and understudied populations.

It has been demonstrated that genotype calls can be made from RNAseq data using tools such as GATK[12–14]. The approach of utilizing RNAseq data to capture genetic structure has been applied for livestock and agricultural purposes[15–18], for example to investigate the population structure, history and adaptation of domesticated barley (*Hordeum vulgare*)[17].

While proof-of-concept and subsequent utility of RNAseq-based genotypes have been demonstrated such as for tissue-specific variants[19], its application to infer human population structure shows promise yet remains relatively underexplored[20].

The aims of this study are to (i) demonstrate that RNAseq-based genotypes can capture the genetic population structure of a diverse yet understudied human population, and (ii) show that use of RNAseq-based genetic principal components (RG-PCs) can effectively control for population structure in association analysis. Here, we recruited and generated whole blood RNAseq data of 376 individuals from Nepal, a landlocked country situated in the Himalayas with over 125 ethnic groups[21,22]. We developed an RNAseq analysis pipeline (RGStraP) to calculate genetic principal components directly from RNAseq data, then validated RGStraP's performance with genome-wide array genotype data from the same Nepalese individuals. We also tested the pipeline on samples from the Geuvadis consortium, which contains 465 samples with paired genotype-RNAseq data from five of the 1000 Genomes populations[23]. Finally, we show the validity of adjusting for RG-PCs in an association analysis to identify sex-specific gene expression. Overall, our study establishes that human population structure, particularly from an understudied but diverse population, can be effectively captured and controlled directly using RNAseq data.

Results

In this study, we constructed the RGStraP pipeline to calculate RG-PCs from genetic variants called from RNAseq data. RGStraP relies on GATK for its variant calling suite, as well as PLINK and flashPCA to filter the SNPs and calculate genetic principal components from them, respectively (**Methods**). We make RGStraP available to the community via github (https://github.com/fachrulm/RGStraP).

RNAseq-based variant calling captures comparable population structure to paired array genotypes

We collected whole blood samples from 376 individuals recruited in Latlipur, Nepal as part of the STRATAA study, then performed whole blood RNAseq using Illumina Novaseq (**Methods**). The cohort included individuals with and without confirmed S. Typhi infection; for the purpose of this study, the disease groups were used as a regression covariate to adjust for gene expression during downstream analyses. Self-reported ethnicity showed individuals belonging to 6 broad ethnic / caste groups.

We ran the RGStraP pipeline **(Figure 1)** on 376 whole blood RNAseq samples, of which 362 passed QC after genetic variant calling. A total of 4,921,472 genetic variants were called across all samples (**Methods**). We found that sequencing depth was a median of 92,782,803 reads per sample (range 21,545,569 to 182,140,303) and correlated moderately (r=0.487) with the total genetic variants called per sample **(Figure S1A).**

To determine the efficiency of estimating genetic PCs from RNAseq data, we investigated the effects of minor allele frequency (MAF), linkage disequilibrium (LD) and the use of a pre-specified set of genetic variants. We found that the selection of a MAF threshold of >0.05 and a pairwise LD threshold of r^2<0.05 struck the optimal balance of offering the most variants for analysis and the highest correlation between RNAseq- and array-based genetic PCs (**Figure S2**). From the total of 4,921,472 genetic variants, 152,072 SNPs passed the MAF filter (MAF>0.05), and 36,440 SNPs further passed the LD filter (LD<0.05). Genetic variants from paired genomic data are available for 299 out of the initial 376 individuals; a total of 552,758 SNPs were identified and passed initial quality control filters (**Methods**), of which 315,615 SNPs and 29,943 SNPs then passed MAF>0.05 and further LD<0.05 filters, respectively. Out of the 299 samples with both RNAseq and paired array genotypes, 280 of them passed quality control and were used for further downstream analyses.

Among the 280 samples with matched array and RNAseq-based genotypes, we found 7,343 overlapping SNPs between the MAF-filtered RNAseq and array SNP sets based on their exact chromosome positions. Genetic concordance was then calculated from the common SNPs based on matching allele genotypes for each position, taking into consideration strand flipping. Most RNAseq samples were found concordant with their respective paired array genotypes, with mean concordance of 0.925 for all samples and 232 sample-pairs (82.8%) having concordance higher than 0.90 **(Figure 2A)**. We found that high RNAseq depth was positively correlated with high genetic concordance with paired array genotypes; however, outliers of low genetic concordance were also present at high depths (ρ=0.151; **Figure S1**).

When looking at the correlations between array-based genetic PCs and RG-PCs, we found that merely filtering based on MAF and LD was not adequate, as the main RG-PCs (specifically PCs 1 and 2) did not represent the genetic structure found in the array-based PCs, and meaningful correlation was only found from RG-PC3 and on (**Figure S3**). We found that the common approach of subsetting the genotype calls to the variants in HapMap3[24] offered higher quality genotype calls and improved correlation between RNAseq and array-based genetic PCs (**Figure S4**). An overlap of 23,227 well-defined SNPs was found between HapMap3 and the MAF-filtered (MAF>0.05) variants, of which 4,887 passed the LD filter (LD<0.05) and were used to calculate RG-PCs. We also calculated genetic PCs from the 29,943 paired genotype array SNPs as a measure of true genetic structure to be compared against RG-PCs. To assess the consistency of inferred population structure between the two approaches, we calculated Spearman correlation between genetic PCs from paired genotype array SNPs and the RG-PCs. PC1 of both RNAseq and array sets correlated strongly with each other ($|\rho|$=0.93), followed by RG-PC3 and PC2 from array data ($|\rho|$=0.61) and RG-PC3 and PC2 from array data ($|\rho|$=0.6) (**Figure S4**). As expected, the genetic PCs of one approach do not exclusively correspond to only one PC of the other

approach, as can be seen with significant correlations of a single array PC with several RG-PCs. To investigate this further, we performed canonical correlation analysis between the top 10 array PCs and the RG-PCs and found that the RG-PCs fully explained the variance of the top 10 array PCs (**Figure 2B**).

PCA of both array and RG-PCs showed visible clustering by self-reported patient ethnicity. Array PC1 vs PC2 captured the clustering of Janajati-Hill, Newar, and Madhesi groups, with array PC3 showing a clear distinction between the Newar samples and other samples **(Figure 3).** Consistent with their Spearman correlations, RG-PC1, RG-PC2, and RG-PC3 also captured the clustering of the groups shown by the array data yet lacked some distance between the groups primarily due to array PC2.

RNAseq-based variant calling can differentiate the genetic structure of distinct groups between and within populations akin to paired array genotypes

We also tested RGStraP's performance on a dataset of 465 samples from the Geuvadis consortium, spread across 5 different populations: British in England and Scotland (GBR), Utah residents with Northern and Western European ancestry (CEU), Finnish in Finland (FIN), Toscani in Italy (TSI), and Yoruba in Ibadan, Nigeria (YRI)[23]. A total of 463 samples passed the downstream filtering as part of the variant calling process (**Methods**). Clustering by each population can be seen in the main PCs from both paired array and RNAseq data: PC1 separates the European (EUR) and African (AFR) samples, while PC2 separates the EUR samples, with more distinct clusters seen for FIN and TSI samples (**Figure S5A**). Canonical correlation analysis between the top 10 array PCs and RG-PCs showed that RG-PCs fully explained the variance of the array PCs, with CV1 from RG-PCs representing 0.903 proportion of variance shared and the first 3 CVs (R_{c1}=0.994, R_{c2}=0.942, R_{c3}=0.752) reaching a cumulative proportion of shared variance of 0.998 (**Figure 4A**).

To project the Nepalese samples into PC space with other populations, we performed PCA of the Nepalese and Geuvadis samples together. Akin to the array PCs, results from RGStraP were able to distinguish the broad ancestry groups; RG-PC1 was able to separate between AFR and other samples, whereas RG-PC2 distinguished the EUR and Nepal (in this case, representing South Asian) samples (**Figure 4B**). Separation within the EUR samples was also visible in RG-PC4, showing distinct clusters of FIN and TSI samples (**Figure 4B**). When projected with the Geuvadis samples, clusters of the Nepalese self-reported ethnicity groups were also still visible (**Figure S5B**). This analysis further validates the performance of RGStraP in capturing genetic structure comparable to array genotypes across diverse populations, which is also supported by the canonical correlation analysis results (**Figure S5C**).

RNAseq-based genetic PCs in differential gene expression analysis

To assess the extent to which genetic PCs control population stratification in differential gene expression analysis (DGE) of these individuals, we performed DGE on the 280 Nepal samples to identify sex-specific gene expression, with and without adjustment for genetic PCs using either array PCs or RG-PCs (**Methods**). Prior to using edgeR for DGE analysis, we filtered out lowly expressed genes based on counts per million (CPM>0.05) to account for differences in sequencing depths. Only autosomal genes were included in the analyses.

A systematic reduction was evident when adjusting for population structure in DGE analysis. A total of 3,038 (p-value < 0.05) and 325 (FDR < 0.05) genes were differentially expressed when adjusting only for age and disease groups, whereas the number declined after adjusting with genetic PCs; inclusion of array PCs reduced the numbers of differentially expressed genes to 2,585 (p-value < 0.05) and 144 (FDR < 0.05), while the number declined

to 2,778 (p-value < 0.05) and 272 (FDR < 0.05) when including RG-PCs. The majority of the differentially expressed genes identified without considering genetic PCs were still found after including either array PCs or RG-PCs (2,478 and 2,381 at p-value < 0.05, respectively; 138 and 213 at FDR < 0.05, respectively). The majority of DE genes were shared between the results using array PCs and RG-PCs (2,175 at p-value < 0.05 and 130 at FDR < 0.05). When taking into consideration log-fold change, 4 genes passed the filter (FDR < 0.05, |logFC|>1) in the set without considering genetic PCs, and the number decreased to 3 when including either array or RG-PCs. This demonstrates how RG-PCs control for population stratification in downstream RNAseq analysis similar to the genetic PCs calculated from paired array genotypes, reducing significant associations that reflected variations in population structure instead of the biology of interest.

Effects of population structure were visible in quantile-quantile (QQ) plots, with attenuated test statistics across the transcriptome in the analysis with RG-PCs **(Figure 5A)**. This is quantitatively supported by lower systematic inflation (m); comparing the ratio of medians of the chi-square statistics between DGE results without genetic PCs and with RG-PCs as covariates, we found a slight systematic reduction of test statistics after including RG-PCs (m=0.935). A similar reduction is also found when using array PCs as covariates (m=0.92; **Figure S6**). Finally, we assessed a mixed linear model (MLM) with a genomic relationship matrix (GRM) constructed from the RNAseq-based SNPs (**Methods**). Similarly, we found that RG-PCs were able to control for population structure in the RNAseq-based GRM (m=0.985) **(Figure 5B)**.

Discussion

Population structure is typically captured via genotype arrays, which are not always done for projects focused on gene expression analyses. Genotyping may not be practical when

resources are limited, or when analyzing existing and/or publicly available RNAseq datasets which nearly always do not offer access to original samples. In this study, we demonstrated how SNPs acquired solely from RNAseq variant calling were able to capture population structure comparable to the results from array data. We further showed that RNAseq-based genetic principal components (RG-PCs) were able to control for population structure in differential gene expression analysis, and that mixed linear model analysis using a genetic relatedness matrix based on RNAseq genotype calls was able to achieve similar. To facilitate the use of RG-PCs, we also construct and make freely available the tool RGStraP (https://github.com/fachrulm/RGStraP) for the wider research community.

Our study enables the control of genetic population structure in analyses of current and historic transcriptomic datasets, which frequently do not come with matched genotypes. This is particularly important in low resource settings, in particular LMICs (e.g. Nepal) with highly diverse and structured populations or for traits where fine-scale genetic population structure may be confounding.

Our study has several limitations. The main challenge of constructing genetic PCs from RNAseq variant calling results is how to properly curate the SNPs to include, as MAF and LD filtering are not enough to remove uninformative SNPs, resulting in the main PCs not corresponding to self-reported ethnicity. This was addressed by using the overlapping SNPs between the RNAseq variant calling results and HapMap3 variants, allowing us to acquire a set of well-established SNPs; however, this SNP set may still miss important genetic structure, e.g. low-frequency and rare variants. Due to the nature of RNAseq platforms only capturing variants in transcribed regions, rarer variants that fall in crucial functional regions such as promoter and enhancer regions go unaccounted for[24]. These rare, non-transcribed variants are mostly linked to increased risk of various diseases and are enriched within expressed quantitative trait loci (eQTLs)[25–27].

Variants called from RNAseq data using a comparable GATK method have previously been reported to be reliable for tissue-specific eQTL mapping and allele-specific expression (ASE) analyses[19]. However, the approach heavily relied on genotype imputation using a reference panel that remains predominantly European, rendering the approach still suboptimal for eQTL analysis in other populations, especially those that are understudied. Thus, the method described in our study should not be utilized as a one-to-one replacement for conventional genotyping array or whole genome sequencing.

In summary, we have developed an approach and tool for inferring genetic population structure directly from RNAseq data across diverse populations, then demonstrated its use in differential gene expression analysis to control for genetic structure where genotyping data are unavailable. We hope that our results enable better control for confounding in RNAseq analyses and facilitate more rigorous retrospective and meta-analyses of RNAseq data.

Methods

Human ethics

The STRATAA study was approved by the Nepal Health Research Council (NHRC, ref 306/2015) and OxTREC (Oxford Tropical Research Ethics Committee, ref 39-15). All participants provided informed consent for human genetic tests. Blood and nucleic acid samples, and associated data, were de-identified by the STRATAA team in Nepal prior to being sent to overseas for analysis**.**

Sample processing, library preparation, and sequencing conditions

This study utilized blood samples collected in Lalitpur, Nepal as part of the Strategic Typhoid Alliance across Africa and Asia (STRATAA) study, which included passive surveillance for enteric fever and a population-based serosurvey[28,29]. Blood was collected from febrile

participants recruited into the passive surveillance study, specifically those presenting >38.5°C temperature or history of fever for >72 hours. From the total blood sample volumes (≤16 mL for patients >16 years of age, ≤7 mL for ≤16 years), aliquots were subjected to (i) bacteriological culture to identify presence of *Salmonella enterica* serovars Typhi (*S.* Typhi); (ii) storage in PAXgene tubes for later RNA extraction; and (iii) DNA extraction and subsequent human genotyping. Blood was also collected from healthy participants in the serosurvey (≤8 mL for patients >16 years of age, ≤7 mL for ≤16 years), from which aliquots were also subjected to (i) serological analysis; (ii) PAXgene storage for RNA analysis; and (iii) DNA extraction.

We analysed RNA from 49 *S.* Typhi culture-positive participants and 275 *S.* Typhi culture-negative participants from passive surveillance, and 52 healthy controls from the serosurvey. PAXgene tubes were sent to Monash University in Melbourne, Australia, where RNA was extracted using PAXgene Blood RNA Kit. An aliquot of RNA (60 ul, 10-20 ng/ul) for each sample was sent to the Wellcome Sanger Institute in Hinxton, England for sequencing. Library preparation was done using NEBNext Ultra II RNA custom kit performed on an Agilent Bravo WS automation platform, pulling down poly(A) tails. After polymerase chain reaction (PCR) cycles (14 standard cycles), plates were purified using Agencourt AMPure XP SPRI beads and libraries were then quantified using Biotium Accuclear Ultra high sensitivity dsDNA Quantitative kit. Pooled libraries were quantified using Agilent bioanalyser and normalised to 2.8nM. Samples were then subjected to globin depletion using the KAPA RNA HyperPrep with RiboErase (HMR) Globin kit. Libraries were subjected to 2x 100 bp paired-end sequencing using the Illumina Novaseq platform. Depending on the level of multiplexing, each library was sequenced up to three times to generate an average of up to 80 million reads per sample.

For 299 of the Nepalese participants with RNAseq data, human DNA was successfully extracted (at Patan Hospital in Nepal) using the QIAamp DNA blood midi kit (QIAGEN) and DNA aliquot was shipped to the Genome Institute of Singapore for SNP genotyping. Genotyping was performed using the Illumina Infinium Global Screening Array-24 Kit (GSAMD-24v2_A1 chip). Stringent quality control filters were used to remove poorly performing SNP markers. SNPs were removed based on the following parameters; SNPs with a call rate of <95%; SNPs with differential call-rate between cases vs control P value <0.001; SNPs with a Hardy-Weinberg equilibrium P value <1e-07 in controls; SNPs with a Hardy-Weinberg equilibrium P value <1e-11 in cases. Samples with a call rate of <95%, those with a lower call rate sample from each sample pair having first- or second-degree relationship from identity by descent (IBD), and those that were PCA outlier samples were removed.

We also used a validation dataset from the Geuvadis consortium, which contains lymphoblastoid cell line samples from five of the 1000 Genomes populations: British in England and Scotland (GBR), Utah residents with Northern and Western European ancestry (CEU), Finnish in Finland (FIN), Toscani in Italy (TSI), and Yoruba in Ibadan, Nigeria (YRI). Library preparation and sequencing details are described in the original publication[23]. The available paired genotype data was obtained from the 1000 Genomes Project Phase 3 dataset[30].

Quality control and sequencing data processing

Computational analyses were performed on an HPC cluster with 6 nodes, each equipped with 2 x 16 cores (32 threads) CPUs and 512GB memory. It is advised to run RGStraP on a high-performance computing (HC) cluster as the alignment step alone requires a minimum RAM of 32GB. We recommend at least 100GB of memory, a multi-core processor with at least 8 cores, and ample storage depending on the number of samples processed. A

workload management system such as SLURM is also recommended for better resource management.

Quality control (QC) was done to the sequencing data according to the *FastQC* (bioinformatics.babraham.ac.uk/projects/fastqc) readouts. Illumina adapters were then trimmed using *Trim Galore!* (bioinformatics.babraham.ac.uk/projects/trim_galore), and optical duplicates were removed using bbmap's Clumpify. We then followed GATK's best practices pipeline for RNAseq short variant discovery, which commenced with a 2-pass mapping process to human genome GRCh38 using STAR[31], from which the resulting sequencing alignment map (BAM) files from the different runs of each sample were merged using Picard[32]. Low quality reads (MAPQ<20) were filtered out from the merged files using SAMtools[33], followed by further QC, variant calling and filtering steps with GATK4[12]. The analysis-ready variants were then filtered using PLINK[34,35], removing duplicated and palindromic variants and only keeping autosomal single nucleotide polymorphisms (SNPs). We excluded individuals with more than 20% missing genotypes and further filtered RNAseq-based SNPs, after removing variants in the HLA region, based on minor allele frequency (MAF; *maf=0.01, hwe=0, geno=0.1, mind=0.2*) as well as linkage disequilibrium (LD; *win=1000, step=50, r2=0.05*). RG-PCs were successfully captured for 351 and 463 individuals in the Nepal and Geuvadis cohorts, respectively. The same SNP filters were also applied to the array-based SNPs, and for the SNP-level correlation analyses we converted genome annotations of the array SNPs from GRCh37 to GRCh38 (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Merging of the SNP sets from the Nepal and Geuvadis samples was done using PLINK[34,35].

Calculating concordance and visualizing stratification

To calculate SNP-level concordance, we took overlapping SNPs between MAF-filtered array and RNAseq results based on chromosome position and calculated paired sample

concordance based on matching genotype (considering heterozygosity and possible strand flipping) for each SNP in both results. A total of 280 Nepalese samples with RG-PCs and matching array genotype was used for this analysis.

To keep only meaningful SNPs from the RNAseq variant calling results for generating genetic principal components (PCs), we took overlapping SNPs (identical chromosome positions and matching genotypes taking into consideration possible strand flipping) between MAF-filtered RNAseq SNPs and HapMap3 variants to get a set of well-defined SNPs, after which LD filtering was performed on the set. We then generated genetic PCs for the LD-filtered RNAseq and array SNP sets separately using flashPCA[36], from which PCA plots were created with ggplot2[37] and Spearman correlations between the two sets of PCs were calculated using the function from the *stats* package of R[38]. We also calculated canonical correlation coefficients[39–41] between the array PCs and RG-PCs to assess how well the latter captured genetic structure information presented by the former, as a single array PC may be represented by multiple RG-PCs.

Differential gene expression analysis between phenotypes

Sequencing count data of the Nepalese samples were extracted from the aligned sequence files using *featureCounts* (http://bioinf.wehi.edu.au/featureCounts/), from which we kept only autosomal genes and genes with CPM>0.05 in at least 20% of the samples from the analyses. Differential gene expression (DGE) analyses was done contrasting males and females using edgeR[42,43], taking into account age, disease group, and sequencing batches; we ran the analyses with and without populations structure PCs as an additional covariate to then compare how genetic structure may stratify gene expression. From both results, we also plotted the Q-Q plot and calculated the systematic inflation (m), which is the ratio of the median of the empirically observed chi-squared test statistics (in our case, results of DGE analysis with RG-PCs) to the expected median chi-squared test statistics (results of DGE

analysis without RG-PCs), to quantify the stratification due to population structure in gene expression data.

We also created a genetic relationship matrix (GRM) and ran a mixed linear model (MLM) on the RNAseq SNPs with and without population structure PCs as random effects with fastGWA[44]. From the two results, a quantile-quantile (Q-Q) plot was created and systematic inflation (m) was then calculated to quantify the effect of populations structure PCs on the genomic data[45].

References

1. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, pdb.top084970 (2015).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet. 2008 101* **10**, 57–63 (2009).
3. Byron, S. A., Keuren-Jensen, K. R. Van, Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet. 2016 175* **17**, 257–271 (2016).
4. Storey, J. D. *et al.* Gene-Expression Variation Within and Among Human Populations. *Am. J. Hum. Genet.* **80**, 502–509 (2007).
5. Thami, P. K. & Chimusa, E. R. Population structure and implications on the genetic architecture of HIV-1 phenotypes within Southern Africa. *Front. Genet.* **10**, (2019).
6. Li, J., Liu, Y., Kim, T., Min, R. & Zhang, Z. Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility. *PLOS Comput. Biol.* **6**, e1000910 (2010).
7. Jovov, B. *et al.* Differential Gene Expression between African American and European American Colorectal Cancer Patients. *PLoS One* **7**, (2012).
8. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet. 2006 388* **38**, 904–909 (2006).
9. Kryvokhyzha, D. *et al.* The influence of population structure on gene expression and flowering time variation in the ubiquitous weed Capsella bursa-pastoris (Brassicaceae). *Mol. Ecol.* **25**, 1106–1121 (2016).

10. Sillanpää, M. J. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Hered. 2011 1064* **106**, 511–519 (2010).
11. Tian, C., Gregersen, P. K. & Seldin, M. F. Accounting for ancestry: Population substructure and genome-wide association studies. *Hum. Mol. Genet.* **17**, R143 (2008).
12. GATK Team. RNAseq short variant discovery (SNPs + Indels). 1–4 (2020).
13. Brouard, J. S., Schenkel, F., Marete, A. & Bissonnette, N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *J. Anim. Sci. Biotechnol.* **10**, 44 (2019).
14. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).
15. Rogier, O. *et al.* Accuracy of RNAseq based SNP discovery and genotyping in Populus nigra. *BMC Genomics* **19**, 909 (2018).
16. Bakhtiarizadeh, M. R. & Alamouti, A. A. RNA-Seq based genetic variant discovery provides new insights into controlling fat deposition in the tail of sheep. *Sci. Rep.* **10**, 13525 (2020).
17. Takahagi, K. *et al.* Analysis of single nucleotide polymorphisms based on RNA sequencing data of diverse bio-geographical accessions in barley. *Sci. Rep.* **6**, 1–11 (2016).
18. Wang, W. *et al.* Genetic structure of six cattle populations revealed by transcriptome-wide SNPs and gene expression. *Genes and Genomics* **40**, 715–724 (2018).
19. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7**, 1–13 (2015).
20. Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinón-Torres, F. & Salas, A. Ancestry patterns inferred from massive RNA-seq data. *RNA* **27**, 857–868 (2019).
21. Ministry of Foreign Affairs Nepal. Nepal Profile - Ministry of Foreign Affairs Nepal MOFA. Available at: https://mofa.gov.np/about-nepal/nepal-profile/. (Accessed: 12th March 2021)
22. United Nations Population Fund. Population Situation Analysis of Nepal (With Respect to Sustainable Development). *Unfpa Nepal* 47 (2017).
23. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nat. 2013 5017468* **501**, 506–511 (2013).
24. Consortium, T. I. H. 3. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
25. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal

cancer. *Nat. Genet. 2018 511* **51**, 76–87 (2018).

26. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nat. 2017 5507675* **550**, 239–243 (2017).

27. Lange, M., Begolli, R. & Giakountis, A. Non-Coding Variants in Cancer: Mechanistic Insights and Clinical Potential for Personalized Medicine. *Non-Coding RNA 2021, Vol. 7, Page 47* **7**, 47 (2021).

28. Darton, T. C. *et al.* The STRATAA study protocol: A programme to assess the burden of enteric fever in Bangladesh, Malawi and Nepal using prospective population census, passive surveillance, serological studies and healthcare utilisation surveys. *BMJ Open* **7**, e016283 (2017).

29. Meiring, J. E. *et al.* Burden of enteric fever at three urban sites in Africa and Asia: a multicentre population-based study. *Lancet Glob. Heal.* **9**, e1688–e1696 (2021).

30. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

31. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013). doi:10.1093/bioinformatics/bts635

32. Broad Institute. Picard Toolkit. *Broad Institute, GitHub Repository* (2019). Available at: https://github.com/broadinstitute/picard. (Accessed: 30th July 2021)

33. P, D. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

34. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559 (2007).

35. Purcell, S. & Chang, C. PLINK 1.9. https://www.cog-genomics.org/plink2 (2015).

36. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).

37. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

38. R Core Team. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* **Vienna**, (2018).

39. Rozeboom, W. W. Linear correlations between sets of variables. *Psychom. 1965 301* **30**, 57–71 (1965).

40. Alpert, M. I. & Peterson, R. A. On the Interpretation of Canonical Analysis. *J. Mark. Res.* **9**, 187 (1972).

41. Sherry, A. & Henson, R. K. Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. *J. Pers. Assess.* **84**, 37–48 (2005).

42. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids*

Res. (2012). doi:10.1093/nar/gks042

43. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

44. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet. 2019 5112* **51**, 1749–1755 (2019).

45. Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997–1004 (1999).

Availability of Data and Materials

The RGStraP pipeline developed in this study is available at: https://github.com/fachrulm/RGStraP.

Raw sequence data will be deposited or public access following the submission of the manuscript for peer-review.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We acknowledge the contributions of individuals and organizations who have arranged and taken part in the studies as well as the laboratory and field teams at the site, including the STRATAA Study Group and the Nepal Family Development Foundation team. We thank the Sanger sequencing teams.

This research was funded in whole, or in part, by the Wellcome Trust [STRATAA, 106158/Z/14/Z and Sanger, 098051]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This research was also funded by NHMRC [project grant APP1101728] and supported by core funding from the British Heart Foundation (RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. *The views expressed are those of the author(s) ad not necessarily those of the NIHR or the Department of Health and Social Care.

MI was supported by the Munz Chair of Cardiovascular Prediction and Prevention and the Cambridge Baker Systems Genomics Initiative. MF was supported by a Melbourne Research Scholarship from The University of Melbourne jointly funded by the Baker Heart and Diabetes Institute. This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. This study was also supported by the Victorian Government's Operational Infrastructure Support (OIS) program.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed in this manuscript are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

The core team of STRATAA study group also includes John Clemens, Gordon Dougan, Melita Gordon, Robert Heyderman, Farhana Khanam, James Meiring, Virginia Pitzer, Firdausi Qadri, and Merryn Voysey.

Figures

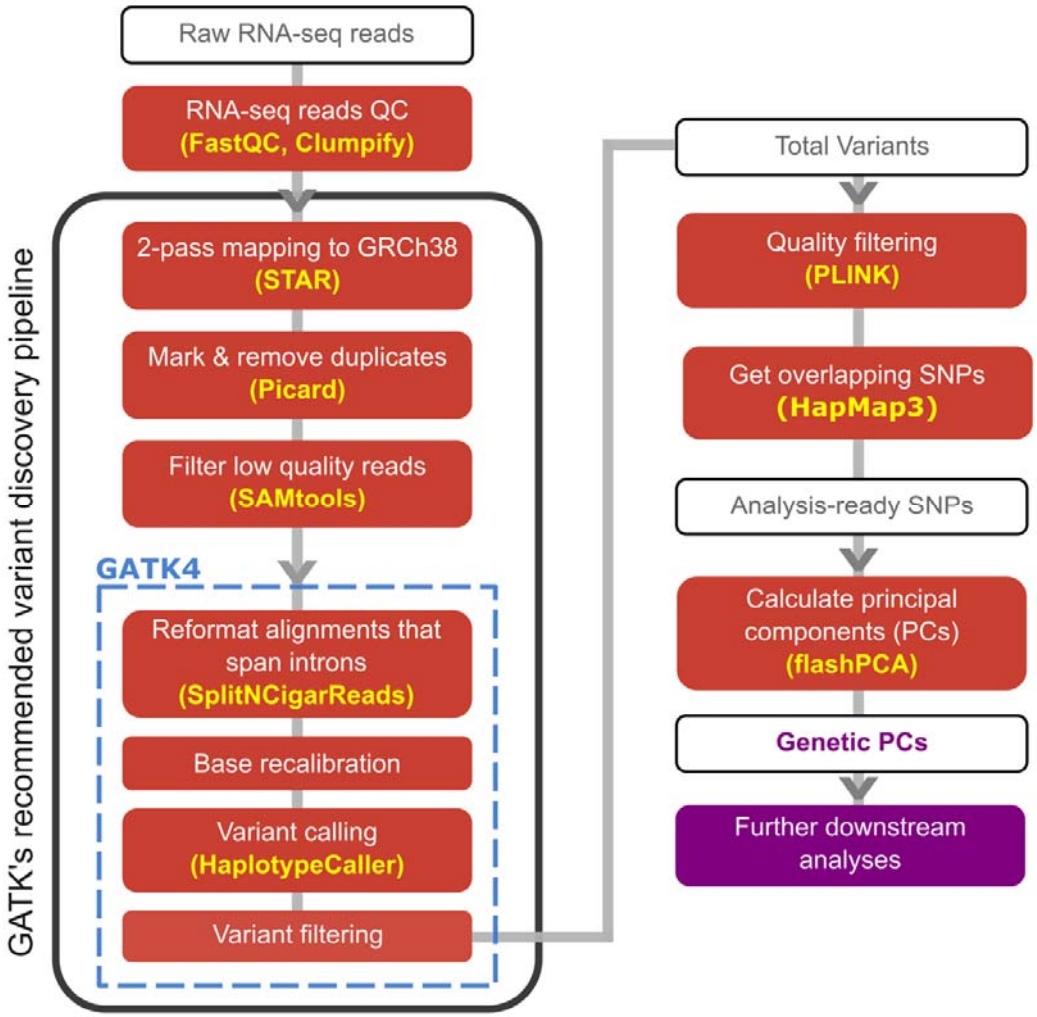

Figure 1. The analysis pipeline used in this study, mainly following GATK's recommended variant discovery pipeline to call genetic variants from RNAseq samples. Further filtering based on missingness, MAF, and LD was then done before generating principal components (PCs) representing the population structure.

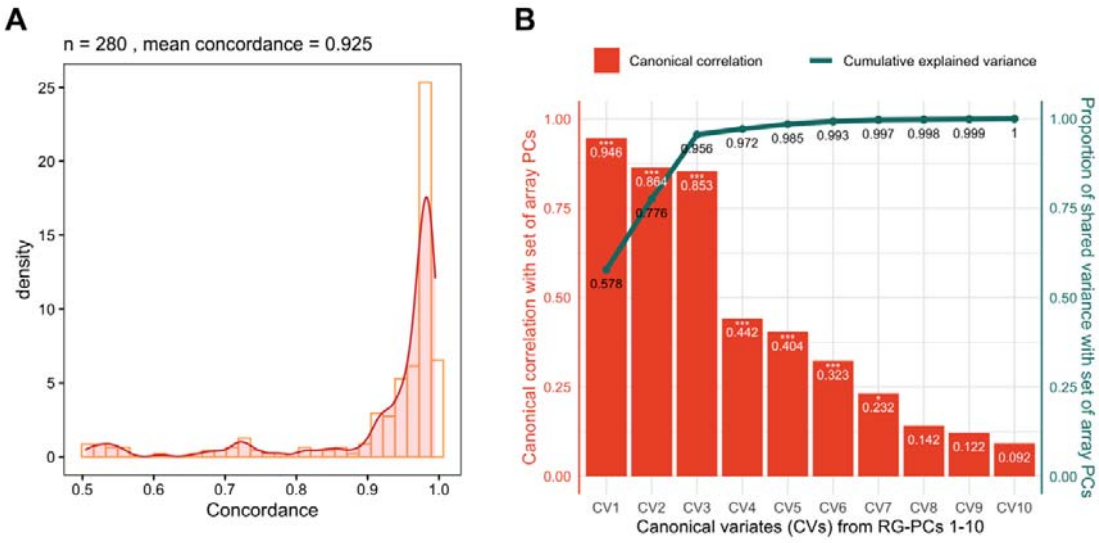

Figure 2. RGStraP was able to capture population structure in the Nepal cohort comparable to paired array genotype. **(A)** Genotype concordance of common SNPs between array and RNAseq samples was found to be high, with most samples (232 out of 280) reaching >0.90 concordances. **(B)** Canonical correlation analysis between ten RG-PCs and ten array PCs showed significant (Wilks' Lambda, p-value<0.05) correlations for the first 7 canonical variates (CVs) between the two sets. The first 3 CVs from 10 RG-PCs strongly captured the genetic information from array PCs (R_{c1}=0.946, R_{c2}=0.864, R_{c3}=0.853), in which the cumulative proportion of shared variance between the two sets reached up to 0.956 from just the 3 CVs.

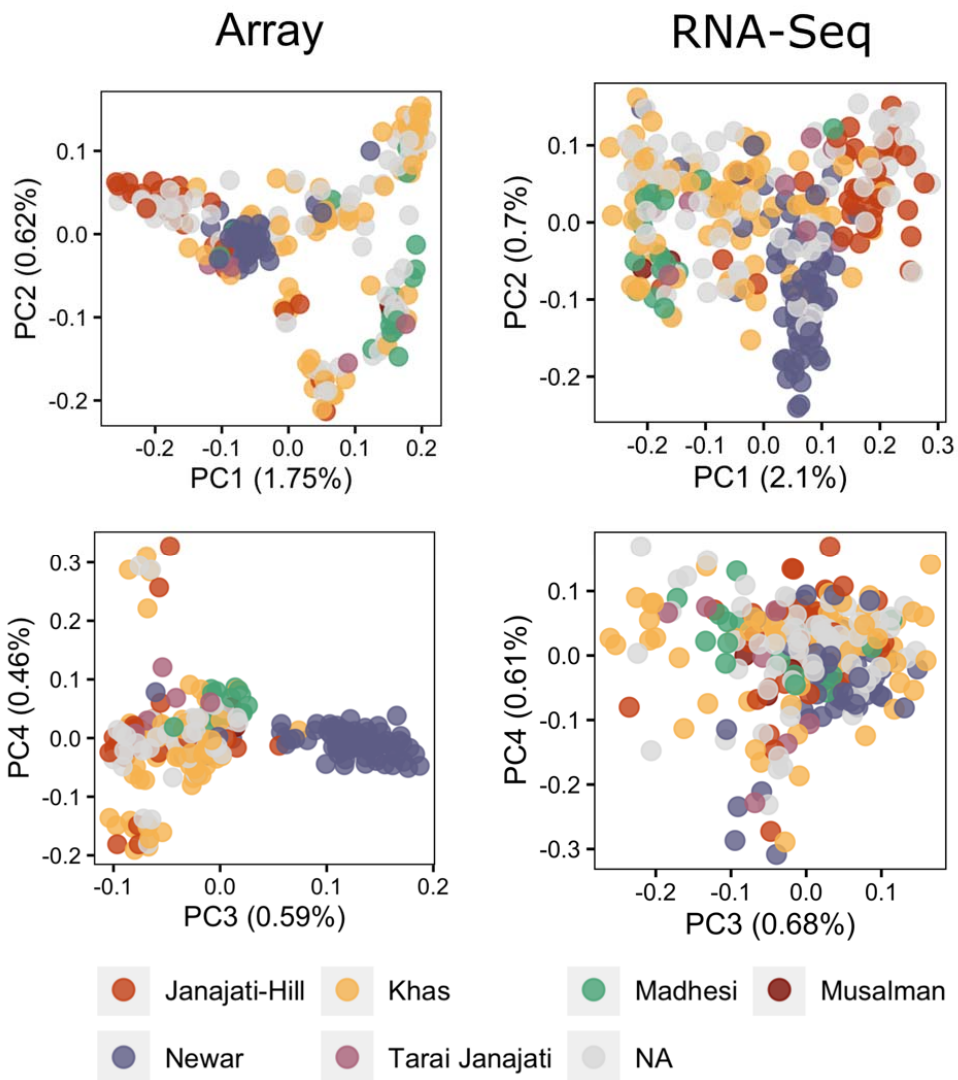

Figure 3. SNPs from the overlap of RNAseq variant calling results and HapMap3 were able to capture genetic structure separating self-reported ethnicities in Nepal when compared to the paired array genotypes, shown by the groupings in the PCA plots.

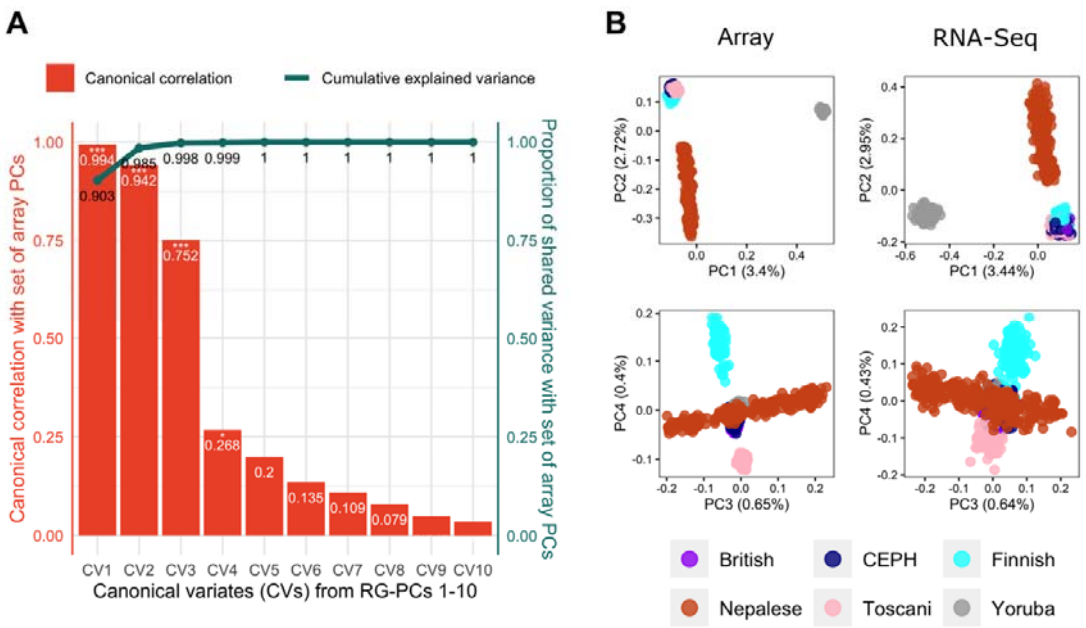

Figure 4. RGStraP was also tested on samples from the Geuvadis cohort, which includes 5 populations of European and African ancestries. **(A)** Canonical correlation analysis between ten RG-PCs and ten array PCs of the Geuvadis samples showed significant (Wilks' Lambda, p-value<0.05) correlations for the first 4 canonical variates (CVs) between the two sets. The first 3 CVs from 10 RG-PCs strongly captured the genetic information from array PCs (R_{c1}=0.994, R_{c2}=0.942, R_{c3}=0.752). The cumulative proportion of shared variance between the three sets reached 0.998 from just the 3 CVs, 0.903 of them represented by CV1. **(B)** PCA plots of the Nepal and Geuvadis samples showing comparable population structure between the array-based PCs and RG-PCs, separating the ancestry groups (European, African, and South Asian) in the main PCs.

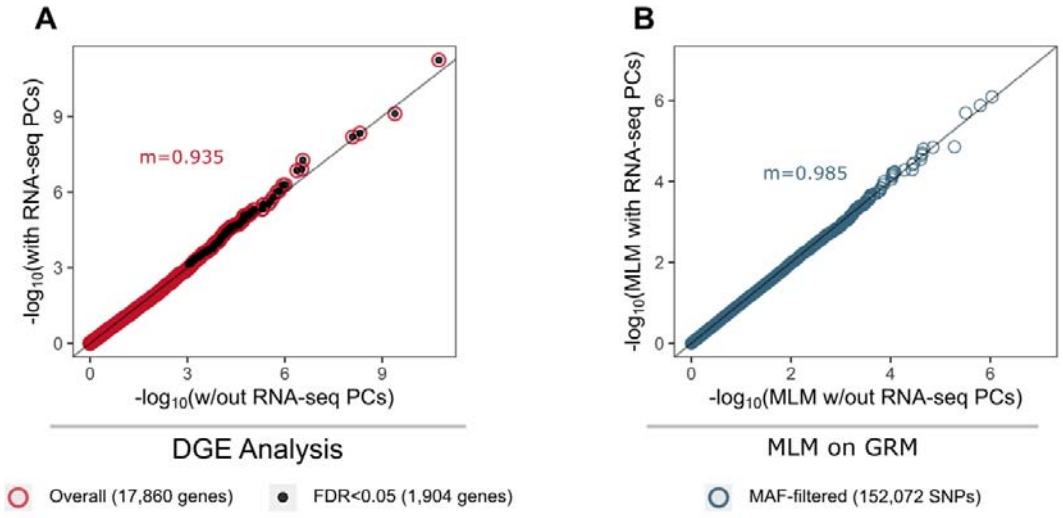

Figure 5. Quantile-quantile (Q-Q) plots showing the distribution of the probabilities between analyses with and without including genetic PCs. Differential gene expression analysis results between samples of different sex show **(A)** systematic reduction in test statistics when including RNAseq-based genetic PCs as covariates compared to without, demonstrated by the low systematic inflation metric (m); **(B)** Distribution of probabilities after mixed linear model (MLM) on a genetic relationship matrix (GRM) was also found to be slightly deflated in the analysis with genetic PCs compared to the one without, though not to the extent of the DGE analysis.