Published online 29 March 2017

Nucleic Acids Research, 2017, Vol. 45, No. 11 ***e106***
doi: 10.1093/nar/gkx204

Differential expression analysis for RNAseq using Poisson mixed models

Shiquan Sun[1,2], Michelle Hood[2], Laura Scott[2,3], Qinke Peng[1], Sayan Mukherjee[4], Jenny Tung[5,6] and Xiang Zhou[2,3,*]

[1]Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, P.R. China, [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA, [3]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA, [4]Departments of Statistical Science, Mathematics, and Computer Science, Duke University, Durham, NC 27708, USA, [5]Departments of Evolutionary Anthropology and Biology, Duke University, Durham, NC 27708, USA and [6]Duke University Population Research Institute, Duke University, Durham, NC 27708, USA

Received July 26, 2016; Revised March 2, 2017; Editorial Decision March 14, 2017; Accepted March 17, 2017

ABSTRACT

Identifying differentially expressed (DE) genes from RNA sequencing (RNAseq) studies is among the most common analyses in genomics. However, RNAseq DE analysis presents several statistical and computational challenges, including over-dispersed read counts and, in some settings, sample non-independence. Previous count-based methods rely on simple hierarchical Poisson models (e.g. negative binomial) to model independent over-dispersion, but do not account for sample non-independence due to relatedness, population structure and/or hidden confounders. Here, we present a Poisson mixed model with two random effects terms that account for both independent over-dispersion and sample non-independence. We also develop a scalable sampling-based inference algorithm using a latent variable representation of the Poisson distribution. With simulations, we show that our method properly controls for type I error and is generally more powerful than other widely used approaches, except in small samples (*n* <15) with other unfavorable properties (e.g. small effect sizes). We also apply our method to three real datasets that contain related individuals, population stratification or hidden confounders. Our results show that our method increases power in all three data compared to other approaches, though the power gain is smallest in the smallest sample (*n* = 6). Our method is implemented in MACAU, freely available at www.xzlab.org/software.html.

INTRODUCTION

RNA sequencing (RNAseq) has emerged as a powerful tool for transcriptome analysis, thanks to its many advantages over previous microarray techniques (1–3). Compared with microarrays, RNAseq has increased dynamic range, does not rely on *a priori*-chosen probes, and can thus identify previously unknown transcripts and isoforms. It also yields allelic-specific expression estimates and genotype information inside expressed transcripts as a useful by-product (4–7). Because of these desirable features, RNAseq has been widely applied in many areas of genomics and is currently the gold standard method for genome-wide gene expression profiling.

One of the most common analyses of RNAseq data involves identification of differentially expressed (DE) genes. Identifying DE genes that are influenced by predictors of interest—such as disease status, risk factors, environmental covariates or genotype—is an important first step toward understanding the molecular basis of disease susceptibility as well as the genetic and environmental basis of gene expression variation. Progress toward this goal requires statistical methods that can handle the complexities of the increasingly large and structurally complex RNAseq datasets that are now being collected from population and family studies (8,9). Indeed, even in classical treatment-control comparisons, the importance of larger sample sizes for maximizing power and reproducibility is increasingly well appreciated (10,11). However, identifying DE genes from such studies presents several key statistical and computational challenges, including accounting for ambiguously mapped reads (12), modeling uneven distribution of reads inside a transcript (13) and inferring transcript isoforms (14).

A fundamental challenge shared by all DE analyses in RNAseq, though, is accounting for the count nature of the data (3,15,16). In most RNAseq studies, the number

[*]To whom correspondence should be addressed. Tel: +1 734 764 5722; Fax: +1 734 763 2215; Email: xzhousph@umich.edu

of reads mapped to a given gene or isoform (following appropriate data processing and normalization) is often used as a simple and intuitive estimate of its expression level (13,14,17). As a result, RNAseq data display an appreciable dependence between the mean and variance of estimated gene expression levels: highly expressed genes tend to have high read counts and subsequently high between-sample variance, and *vice versa* (15,18). To account for the count nature of the data and the resulting mean-variance dependence, most statistical methods for DE analysis model RNAseq data using discrete distributions. For example, early studies showed that gene expression variation across technical replicates can be accurately described by a Poisson distribution (19–21). More recent methods also take into account over-dispersion across biological replicates (22,23) by replacing Poisson models with negative binomial models (15,16,24–28) or other related approaches (18,29–32). While non-count based methods are also commonly used (primarily relying on transformation of the count data to more flexible, continuous distributions (33,34)), recent comparisons have highlighted the benefits of modeling RNAseq data using the original counts and accounting for the resulting mean-variance dependence (11,35–37), consistent with observations from many count data analyses in other statistical settings (38). Indeed, accurate modeling of mean-variance dependence is one of the keys to enable powerful DE analysis with RNAseq, especially in the presence of large sequencing depth variation across samples (25,33,39).

A second important feature of many RNAseq datasets, which has been largely overlooked in DE analysis thus far, is that samples often are not independent. Sample non-independence can result from individual relatedness, population stratification or hidden confounding factors. For example, it is well known that gene expression levels are heritable. In humans, the narrow-sense heritability of gene expression levels averages from 15–34% in peripheral blood (40–44) and is about 23% in adipose tissue (40), with a maximum heritability in both tissues as high as 90% (40,41). Similarly, in baboons, gene expression levels are about 28% heritable in the peripheral blood (7). Some of these effects are attributable to nearby, putatively *cis*-acting genetic variants: indeed, recent studies have shown that the expression levels of almost all genes are influenced by cis-eQTLs and/or display allelic specific expression (3,7,45–47). However, the majority of heritability is often explained by distal genetic variants (i.e. *trans*-QTLs, which account for 63–84% of heritability in humans (40) and baboons (7)). Because gene expression levels are heritable, they will covary with kinship or population structure. Besides kinship or population structure, hidden confounding factors, commonly encountered in sequencing studies (48–51), can also induce similarity in gene expression levels across many genes even when individuals are unrelated (52–56). Failure to account for this gene expression covariance due to sample non-independence could lead to spurious associations or reduced power to detect true DE effects. This phenomenon has been extensively documented in genome-wide association studies (9,57–58) and more recently, in bisulfite sequencing studies (59), but is less explored in RNAseq studies. In particular, none of the currently available count-based methods for identifying DE genes in RNAseq can appropriately control for sample non-independence. Consequently, even though count-based methods have been shown to be more powerful, recent RNAseq studies have turned to linear mixed models (LMMs), which are specifically designed for quantitative traits, to deal with the confounding effects of kinship, population structure or hidden confounders (7,41,60).

Here, we present a Poisson mixed model (PMM) that can explicitly model both over-dispersed count data and sample non-independence in RNAseq data for effective DE analysis. To make our model scalable to large datasets, we also develop an accompanying efficient inference algorithm based on an auxiliary variable representation of the Poisson model (61–63) and recent advances in mixed model methods (9,58,64). We refer to the combination of the statistical method and the computational algorithm developed here as MACAU (Mixed model Association for Count data via data AUgmentation), which effectively extends our previous method of the same name on the simpler binomial model (59) to the more difficult Poisson model. MACAU works directly on RNAseq count data and introduces two random effects terms to both control for sample non-independence and account for additional independent over-dispersion. As a result, MACAU properly controls for type I error in the presence of sample non-independence and, in a variety of settings, is more powerful for identifying DE genes than other commonly used methods. We illustrate the benefits of MACAU with extensive simulations and real data applications to three RNAseq studies.

MATERIALS AND METHODS

Methods for comparison

We compared the performance of seven different methods in the main text: (i) our PMM implemented in the MACAU software package (59); (ii) the linear model implemented in the *lm* function in R; (iii) the LMM implemented in the GEMMA software package (9,58,65); (iv) the Poisson model implemented in the *glm* function in R (66); (v) the negative binomial model implemented in the *glm.nb* function in R; (vi) edgeR implemented in the *edgeR* package in R (25); (vii) DESeq2 implemented in the *DESeq2* package in R (24). All methods were used with default settings. The performance of each method in simulations was evaluated using the area under the curve (AUC) function implemented in the *pROC* package in R (67), a widely used benchmark for RNAseq method comparisons (68).

Both the linear model and the LMM require quantitative phenotypes. Here, we considered six different transformations of count data to quantitative values, taking advantage of several methods proposed to normalize RNAseq data (e.g. (12–14,17,22,33,69)): (i) quantile normalization (TRCQ), where we first divided the number of reads mapped to a given gene by the total number of read counts for each individual, and then for each gene, quantile normalized the resulting proportions across individuals to a standard normal distribution (7); (ii) total read count (TRC) normalization, where we divided the number of reads mapped to a given gene by the total number of read counts for each individual (i.e. CPM, counts per million; without further transformation to a standard normal within genes: (25)); (iii) upper quantile (UQ) normalization, where we divided the number of reads mapped to a given gene by the UQ (75th percentile) of all genes for each individual (70); (iv) relative log expression normalization (15); (v) the trimmed mean of M-values (TMM) method (39) where we divided the number of reads mapped to a given gene by the normalization factor output from TMM; and (vi) VOOM normalization (33). Simulation results presented in a supplementary figure (see 'Results' section) showed that TRCQ, VOOM and TRC worked better than the other three methods, with TRCQ showing a small advantage. Therefore, we report results using TRCQ throughout the text.

Simulations

To make our simulations as realistic as possible, we simulated the gene expression count data based on parameters inferred from a real baboon dataset that contains 63 samples (see the next section for a detailed description of the data). We varied the sample size (n) in the simulations (n = 6, 10, 14, 63, 100, 200, 500, 800 or 1000). For n = 63, we used the baboon relatedness matrix $\boldsymbol{K}$ (7). For sample simulations with $n > 63$, we constructed a new relatedness matrix $\boldsymbol{K}$ by filling in its off-diagonal elements with randomly drawn off-diagonal elements from the baboon relatedness matrix following (59). For sample simulations with $n < 63$, we constructed a new relatedness matrix $\boldsymbol{K}$ by randomly sub-sampling individuals from the baboon relatedness matrix. In cases where the resulting $\boldsymbol{K}$ was not positive definite, we used the *nearPD* function in R to find the closest positive definite matrix as the final $\boldsymbol{K}$. In most cases, we simulated the TRC N_i for each individual from a discrete uniform distribution with a minimum (= 1 770 083) and a maximum (= 9 675 989) TRC (i.e. summation of read counts across all genes) equal to the minimum and maximum TRCs from the baboon data. We scaled the TRCs to ensure that the coefficient of variation was small (CV = 0.3), moderate (CV = 0.6) or high (CV = 0.9) across individuals (i.e. $N_{\text{new}} = \bar{N} + (N - \bar{N})\ CV\ sd(N)\ /\bar{N}$) and then discretized them. In the special case where CV = 0.3 and n = 63, we directly used the observed TRCs per individual i (N_i) from the baboon data (which has a CV = 0.33).

We then repeatedly simulated a continuous predictor variable x from a standard normal distribution (without regard to the pedigree structure). We estimated the heritability of the continuous predictor using GEMMA, and retained x if the heritability (h_x^2) estimate (with ± 0.01 tolerance) was 0, 0.4 or 0.8, representing no, moderate and highly heritable predictors. Using this procedure, ~30 percent of x values generated were retained, with different retention percentages for different heritability values.

Based on the simulated sample size, TRCs and continuous predictor variable, we simulated gene expression values using the following procedure. For the expression of each gene in turn, we simulated the genetic random effects $\boldsymbol{g}$ from a multivariate normal distribution with covariance $\boldsymbol{K}$. We simulated the environmental random effects $\boldsymbol{e}$ based on independent normal distributions. We scaled the two sets of random effects to ensure a fixed value of heritability ($h^2 = \frac{V(g)}{V(g)+V(e)}$ 0 or 0.3 or 0.6) and a fixed value of over-dispersion variance ($\sigma^2 = V(g) + V(e) = 0.1,\ 0.25$ or 0.4, close to the lower, median and UQs of the over-dispersion variance inferred from the baboon data, respectively), where the function V(•) denotes the sample variance. We then generated the effect size β of the predictor variable on gene expression. The effect size was either 0 (for non-DE genes) or generated to explain a certain percentage of variance in log(λ) (i.e. PVE = $\frac{V(X\beta)}{V(X\beta)+\sigma^2}$; for DE genes). Proportion of variance explained (PVE) values were 15, 20, 25, 30 or 35% to represent different effect sizes. The predictor effects $X\beta$, genetic effects $\boldsymbol{g}$, environmental effects $\boldsymbol{e}$, and an intercept (= $\log(\frac{100}{\bar{N}})$ to ensure that the expected simulated count is 100) were then summed together to yield the latent variable $\log(\lambda) = \mu + X\beta + \boldsymbol{g} + \boldsymbol{e}$. Note that h^2 does not include the contribution of $X\beta$, which in many cases represent non-genetic effects. Finally, the read counts were simulated based on a Poisson distribution with rate determined by the TRCs and the latent variable λ, or $y_i \sim Poi(N_i\lambda_i)$ for the ith individual.

With the above procedure, we first simulated data for $n =$ 63, CV = 0.3, $h_x^2 = 0$, PVE = 0.25, $h^2 = 0.3$ and $\sigma^2 = 0.25$. We then varied one parameter at a time to generate different scenarios for comparison. In each scenario, conditional on the sample size, TRCs and continuous predictor variable, we performed 10 simulation replicates, where 'replication' is at the level described in the paragraph above. Each replicate consisted of 10 000 genes. For examining type I error control, all 10 000 genes were non-DE. For the power comparison, 1000 genes were DE while 9000 were non-DE.

RNAseq datasets

We considered three published RNAseq datasets in this study, which include small ($n < 15$), medium ($15 \leq n \leq 100$) and large ($n > 100$) sample sizes (based on current RNAseq sample sizes in the literature).

The first RNAseq dataset was collected from blood samples of yellow baboons (7) from the Amboseli ecosystem of southern Kenya as part of the Amboseli Baboon Research Project (ABRP) (71). The data are publicly available on GEO with accession number GSE63788. Read counts were measured on 63 baboons and 12 018 genes after stringent quality control as in (7). As in (7), we computed pairwise relatedness values from previously collected microsatellite data (72,73) using the software COANCESTRY (74). The data contains related individuals: 16 pairs of individuals have a kinship coefficient exceeding 1/8 and 48 pairs exceed 1/16. We obtained sex information for each individual from GEO. Sex differences in health and survival are major topics of interest in medicine, epidemiology and evolutionary biology (72,75). Therefore, we used this dataset to identify sex-related gene expression variation. In the analysis, we included the top five expression principal components (PCs) as covariates to control for potential batch effects following the original study (7).

The second RNAseq dataset was collected from skeletal muscle samples of Finnish individuals (60) as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) project (76,77). The data are publicly available in dbGaP with accession code phs001068.v1.p1. Among the 271 individuals in the original study, we selected 267 individuals who have both genotypes and gene expression measurements. Read counts were obtained on these 267 individuals and 21 753 genes following the same stringent quality control as in the original FUSION RNAseq study. For genotypes, we excluded SNPs with minor allele frequency < 0.05 and Hardy-Weinberg equilibrium P-value $< 10^{-6}$. We used the remaining 5 696 681 SNPs to compute the relatedness matrix using GEMMA. The data contains remotely related individuals (three pairs of individuals have a kinship coefficient exceeding 1/32 and 6 pairs exceed 1/64) and is stratified by the municipality from which samples were collected (see 'Results' section). Two predictors from the data were available to us: the oral glucose tolerance test (OGTT) which classifies $n = 162$ individuals as either type II diabetes (T2D) patient ($n = 66$) or normal glucose tolerance (NGT; i.e. control, $n = 96$); and a T2D-related quantitative trait—fasting glucose levels (GL)—measured on all $n = 267$ individuals. We used these data to identify genes whose expression level is associated with either T2D or GL. In the analysis, we included age, sex and batch labels as covariates following the original study (60).

The third RNAseq dataset was collected from lymphoblastoid cell lines (LCLs) derived from 69 unrelated Nigerian individuals (YRI) (3). The data are publicly available on GEO with accession number GSE19480. Following the original study (3), we aligned reads to the human reference genome (version hg19) using Burrows-Wheeler Aligner (BWA) (78). We counted the number of reads mapped to each gene on either autosomes or the X chromosome using Ensembl gene annotation information obtained from the UCSC genome browser. We then filtered out lowly expressed genes with zero counts in over 90% of individuals. In total, we obtained gene expression measurements on 13 319 genes. Sex is the only phenotype available in the data and we used sex as the predictor variable to identify sex-associated genes. To demonstrate the efficacy of MACAU in small samples, we randomly subsampled individuals from the data to create small datasets with either $n = 6$ (3 males and 3 females) or $n = 10$ (5 males and 5 females) or $n = 14$ individuals (7 males and 7 females). For each sample size n, we performed 20 replicates of subsampling and we evaluated method performance by averaging across these replicates. In each replicate, following previous studies (52–56), we used the gene expression covariance matrix as $\boldsymbol{K}$ (i.e. $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}/\mathbf{p}$, where $\boldsymbol{X}$ is the normalized gene expression matrix and p is the number of genes) and applied MACAU to identify sex-associated genes. Note that the gene expression covariance matrix $\boldsymbol{K}$ contains information on sample non-independence caused by hidden confounding factors (52–56). By incorporating $\boldsymbol{K}$, MACAU can be used to control for hidden confounding factors that are commonly observed in sequencing datasets (48–51).

For each of these RNAseq datasets and each trait, we used a constrained permutation procedure to estimate the empirical false discovery rate (FDR) of a given analytical method. In the constrained permutation procedure, we permuted the predictor across individuals, estimated the heritability of the permuted predictor and retained the permutation only if the permuted predictor had a heritability estimate (h_x^2) similar to the original predictor with ±0.01 tolerance (for the original predictors: $h_x^2 = 0.0002$ for sex in the baboon data; $h_x^2 = 0.0121$ for T2D and $h_x^2 = 0.4023$ for GL in the FUSION data; h_x^2 are all close to zero with small variations depending on the sub-sample size in the YRI data). We then analyzed all genes using the permuted predictor. We repeated the constrained permutation procedure and analysis 10 times, and combined the P-values from these 10 constrained permutations. We used this set of P-values as a null distribution from which to estimate the empirical FDR for any given P-value threshold (59). This constrained procedure thus differs from the usual unconstrained permutation procedure (every permutation retained) (79) in that it constrains the permuted predictor to have the same h_x^2 as the original predictor. We chose to use the constrained permutation procedure here because the unconstrained procedure is invalid under the mixed model assumption: the subjects are not exchangeable in the presence of sample non-independence (individual relatedness, population structure or hidden confounders) (79,80). To validate our constrained permutation procedure and test its effectiveness in estimating FDR, we performed a simulation with 1000 DE genes and 9000 non-DE genes as described above. We considered three predictor variables x with different heritability: $h_x^2 = 0$, $h_x^2 = 0.4$ and $h_x^2 = 0.8$. For each predictor variable and each P-value threshold, we computed the true FDR and then estimated the FDR based on either the constrained or unconstrained permutation procedures. The simulation results presented in a supplementary figure (see 'Results' section) demonstrate that the constrained permutation procedure provides a much more accurate estimate of the true FDR while the unconstrained permutation procedure often under-estimates the true FDR. Therefore, we applied the constrained permutation procedure for all real data analysis.

Finally, we investigated whether the methods we compared were sensitive to outliers (31,81,82) in the first two datasets. To examine outlier sensitivity, we first identified genes with potential outliers using BBSeq (18). In total, we identified 8 genes with potential outliers in the baboon data, 130 genes with potential outliers in the FUSION data ($n = 267$) and 43 genes with potential outliers in the subset of the FUSION data for which we had T2D diagnoses ($n = 162$). We counted the number of genes with potential outliers in the top 1000 genes with strong DE association evidence. In the baboon data, 4 genes with potential outliers are in the top 1000 genes with the strongest sex association determined by various methods: two of them by the negative binomial model, three of them by the Poisson model, but zero of them by MACAU, linear model or GEMMA. In the FUSION data, for T2D analysis, 9 genes with potential outliers are in the top 1000 genes with the strongest T2D association determined by various methods: one by MACAU, three by negative binomial, six by Poisson, one by linear and one by GEMMA. For GL analysis, 15 genes with potential outliers are in the top 1000 genes with the strongest GL association determined by various methods: two by MACAU, seven by negative binomial, nine by Poisson, three by linear and three by GEMMA. All outliers are presented in supplementary figures (see 'Results' section). Therefore, the influence of outliers on DE analysis is small in the real data.

RESULTS

MACAU overview

Here, we provide a brief overview of the PMM; more details are available in the Supplementary Data. To identify DE genes with RNAseq data, we examine one gene at a time. For each gene, we model the read counts with a Poisson distribution

$$y_i \sim Poi\left(N_i \lambda_{\boldsymbol{i}}\right),\ i\ =\ 1,\ 2, \cdots,\ n,$$

where for the i'th individual, y_i is the number of reads mapped to the gene (or isoform); N_i is the TRCs for that individual summing read counts across all genes; and $\lambda_{\boldsymbol{i}}$ is an unknown Poisson rate parameter. We model the log-transformed rate $\lambda_{\boldsymbol{i}}$ as a linear combination of several parameters

$$\log\left(\lambda_{\boldsymbol{i}}\right) = \boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{\alpha} + x_i \beta + g_i + e_i, i = 1, 2, \cdots,\ n,$$

$$\boldsymbol{g} = \left(g_1, g_2, \cdots, g_n\right)^T \sim MVN\left(0, \sigma^2 h^2 \boldsymbol{K}\right),$$

$$\boldsymbol{e} = \left(e_1, e_2, \cdots, e_n\right)^T \sim MVN\left(0, \sigma^2\left(1 - h^2\right) \boldsymbol{I}\right),$$

where $\boldsymbol{w}_{\boldsymbol{i}}$ is a c-vector of covariates (including the intercept); $\boldsymbol{\alpha}$ is a c-vector of corresponding coefficients; x_i represents the predictor variable of interest (e.g. experimental perturbation, sex, disease status or genotype); β is its coefficient; $\boldsymbol{g}$ is an n-vector of genetic effects; $\boldsymbol{e}$ is an n-vector of environmental effects; $\boldsymbol{K}$ is an n by n positive semi-definite matrix that models the covariance among individuals due to individual relatedness, population structure or hidden confounders; $\boldsymbol{I}$ is an n by n identity matrix that models independent environmental variation; $\sigma^2 h^2$ is the genetic variance component; $\sigma^2(1 - h^2)$ is the environmental variance component; and MVN denotes the multivariate normal distribution. In the above model, we assume that $\boldsymbol{K}$ is known and can be computed based on either pedigree, genotype or the gene expression matrix (9). For pedigree/genotype data, when $\boldsymbol{K}$ is standardized to have $\boldsymbol{tr}(\boldsymbol{K})/n =\ 1,\ h^2 \in [0, 1]$ has the usual interpretation of heritability (9), where the $\boldsymbol{tr}(\cdot)$ denotes the trace of a matrix. Importantly, unlike several other DE methods (15,25), our model can deal with both continuous and discrete predictor variables.

Both of the random effects terms $\boldsymbol{g}$ and $\boldsymbol{e}$ model over-dispersion, the extra variance not explained by a Poisson model. However, the two terms $\boldsymbol{g}$ and $\boldsymbol{e}$ model two different aspects of over-dispersion. Specifically, $\boldsymbol{g}$ models the fraction of the extra variance that is explained by sample non-independence while $\boldsymbol{e}$ models the fraction of the extra variance that is independent across samples. For example, let us consider a simple case in which all samples have the same sequencing depth (i.e. $N_i =\ N$) and there is only one intercept term μ included as the covariate. In this case, the random effects term $\boldsymbol{e}$ models the independent over-dispersion: without $\boldsymbol{g}$, $V(y) =\ E(y)(1 + E(y)(e^{\sigma^2} - 1))$ is still larger than the mean $E(y) = Ne^{\mu + \sigma^2/2}$, with the difference between the two increasing with increasing σ^2. In a similar fashion, the random effects term $\boldsymbol{g}$ models the non-independent over-dispersion by accounting for the sample covariance matrix $\boldsymbol{K}$. By modeling both aspects of over-dispersion, our PMM effectively generalizes the commonly used negative binomial model—which only models independent extra variance—to account for sample non-independence. In addition, our PMM naturally extends the commonly used LMM (9,64,83,84) to modeling count data.

Our goal here is to test the null hypothesis that gene expression levels are not associated with the predictor variable of interest, or $H_0 : \beta\ =\ 0$. Testing this hypothesis requires estimating parameters in the PMM (as has previously been done in other settings (85,86), including for modeling uneven RNAseq read distribution inside transcripts (13); details in Supplementary Data). The PMM belongs to the generalized LMM family, where parameter estimation is notoriously difficult because of the random effects and the resulting intractable n-dimensional integral in the likelihood. Standard estimation methods rely on numerical integration (87) or Laplace approximation (88,89), but neither strategy scales well with the increasing dimension of the integral, which in our case equals the sample size. As a consequence, standard approaches often produce biased estimates and overly narrow (i.e. anti-conservative) confidence intervals (90–96). To overcome the high-dimensionality of the integral, we instead develop a novel Markov Chain Monte Carlo (MCMC) algorithm, which, with enough iterations, can achieve high inference accuracy (97,98). We use MCMC to draw posterior samples but rely on the asymptotic normality of both the likelihood and the posterior distributions (99) to obtain the approximate maximum likelihood estimate $\hat{\beta}_j$ and its standard error $\mathrm{se}(\hat{\beta}_j)$. With $\hat{\beta}_j$ and $\mathrm{se}(\hat{\beta}_j)$, we can construct approximate Wald test statistics and P-values for hypothesis testing (Supplementary Material). Although we use MCMC, our procedure is frequentist in nature.

At the technical level, our MCMC algorithm is also novel, taking advantage of an auxiliary variable representation of the Poisson likelihood (61–63) and recent linear algebra innovations for fitting LMMs (9,58,64). Our MCMC algorithm introduces *two* continuous latent variables for each individual to replace the count observation, effectively extending our previous approach of using *one* latent variable for the simpler binomial distribution (59). Compared with a standard MCMC, our new MCMC algorithm reduces the computational complexity of each MCMC iteration from cubic to quadratic with respect to the sample size. Therefore, our method is orders of magnitude faster than the popular Bayesian software MCMCglmm (100) and can be used to analyze hundreds of samples and tens of thousands of genes with a single desktop PC (Supplementary Figure S1). Although our procedure is stochastic in nature, we find the MCMC errors are often small enough to ensure stable P-values across independent MCMC runs (Supplementary Figure S2). We summarize the key features of our method along with other commonly used approaches in Table 1.

Simulations: control for sample non-independence

We performed a series of simulations to compare the performance of the PMM implemented in MACAU with four other commonly used methods: (i) a linear model; (ii) the LMM implemented in GEMMA (9,58); (iii) a Poisson model; and (iv) a negative binomial model. We used quantile-transformed data for linear model and GEMMA

Table 1. Current approaches for identifying differentially expressed genes in RNAseq

Statistical method	Directly models counts?	Controls for biological covariates?	Controls for sample non-independence?	Example software that implements the method
Linear regression	No	Yes	No	R and many others
Linear mixed model	No	Yes	Yes	GEMMA (9) and EMMA (84)
Poisson model	Yes	Some methods do	No	GLMP (66) and DEGseq (20)
Negative binomial model	Yes	Some methods do	No	edgeR (25), DESeq (15) and GLMNB (66)
Poisson mixed model	Yes	Yes	Yes	MACAU

(see 'Materials and Methods' section for normalization details and a comparison between various transformations; Supplementary Figure S3) and used raw count data for the other three methods. To make our simulations realistic, we use parameters inferred from a published RNAseq dataset on a population of wild baboons (7,71) to perform simulations ('Materials and Methods' section); this baboon dataset contains known related individuals and hence invokes the problem of sample non-independence outlined above.

Our first set of simulations was performed to evaluate the effectiveness of MACAU and the other four methods in controlling for sample non-independence. To do so, we simulated expression levels for 10 000 genes in 63 individuals (the sample size from the baboon dataset). Simulated gene expression levels are influenced by both independent environmental effects and correlated genetic effects, where genetic effects are simulated based on the baboon kinship matrix (estimated from microsatellite data (7)) with either zero ($h^2 = 0.0$), moderate ($h^2 = 0.3$), or high ($h^2 = 0.6$) heritability values. We also simulated a continuous predictor variable x that is itself moderately heritable ($h_x^2 = 0.4$). Because we were interested in the behavior of the null in this set of simulations, gene expression levels were not affected by the predictor variable (i.e. no genes were truly DE).

Figure 1, Supplementary Figures S4 and 5 show quantile–quantile plots for analyses using MACAU and the other four methods against the null (uniform) expectation, for $h^2 = 0.6$, $h^2 = 0.3$ and $h^2 = 0.0$ respectively. When genes are heritable and the predictor variable is also correlated with individual relatedness, then the resulting *P*-values from the DE analysis are expected to be uniform only for a method that properly controls for sample non-independence. If a method fails to control for sample non-independence, then the *P*-values would be inflated, resulting in false positives.

Our results show that, because MACAU controls for sample non-independence, the *P*-values from MACAU follow the expected uniform distribution closely (and are slightly conservative) regardless of whether gene expression is moderately or highly heritable. The genomic control factors from MACAU are close to 1 (Figure 1 and Supplementary Figure S4). Even if we use a relatively relaxed *q*-value cutoff of 0.2 to identify DE genes, we do not incorrectly identify any genes as DE with MACAU. In contrast, the *P*-values from negative binomial are inflated and skewed toward low (significant) values, especially for gene expression levels with high heritability. With negative binomial, 27 DE genes (when $h^2 = 0.3$) or 21 DE genes (when $h^2 = 0.6$) are erroneously detected at the *q*-value cutoff of 0.2. The inflation of *P*-values is even more acute in Poisson, presumably because the Poisson model accounts for neither individual relatedness nor over-dispersion. For non-count-based models, the *P*-values from a linear model are slightly skewed towards significant values, with three DE genes (when $h^2 = 0.3$) and one DE gene (when $h^2 = 0.6$) erroneously detected at $q < 0.2$. In contrast, because the LMM in GEMMA also accounts for individual relatedness, it controls for sample non-independence well. Finally, when genes are not heritable, all methods except Poisson correctly control type I error (Supplementary Figure S5).

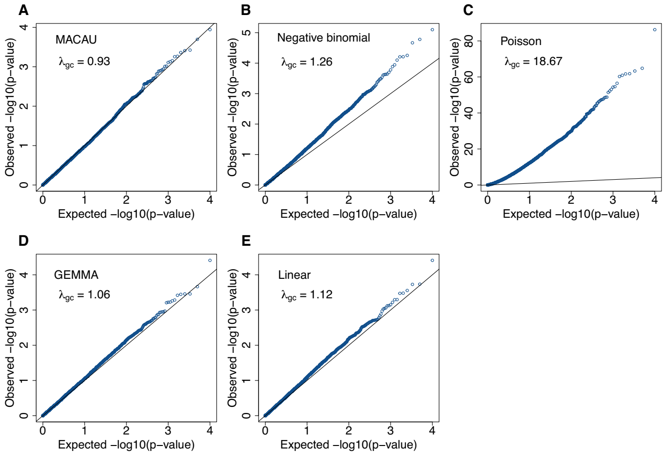

Figure 1. QQ-plots comparing expected and observed *P*-value distributions generated by different methods for the null simulations in the presence of sample non-independence. In each case, 10 000 non-DE genes were simulated with $n = 63$, CV = 0.3, $\sigma^2 = 0.25$, $h^2 = 0.6$, and $h_x{}^2 = 0.4$. Methods for comparison include MACAU (**A**), Negative binomial (**B**), Poisson (**C**), GEMMA (**D**), and Linear (**E**). Both MACAU and GEMMA properly control for type I error well in the presence of sample non-independence. λ_{gc} is the genomic control factor.

Two important factors influence the severity of sample non-independence in RNAseq data (Figure 2). First, the inflation of *P*-values in the negative binomial, Poisson and linear models becomes more acute with increasing sample size. In particular, when $h_x^2 = 0.4$, with a sample size of $n = 1,000$, λ_{gc} from the negative binomial, Poisson and linear models reaches 1.71, 82.28 and 1.41, respectively. In contrast, even when $n = 1,000$, λ_{gc} from both MACAU and GEMMA remain close to 1 (0.97 and 1.01, respectively). Second, the inflation of *P*-values in the three models also becomes more acute when the predictor variable is more correlated with population structure. Thus, for a

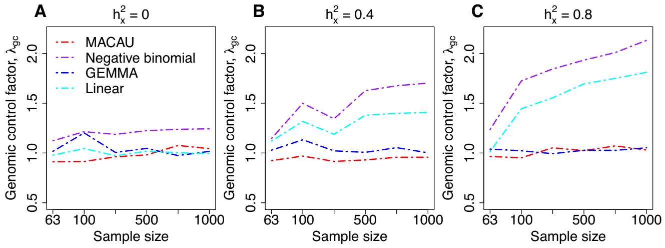

Figure 2. Comparison of the genomic control factor λ_{gc} from different methods for the null simulations in the presence of sample non-independence. 10 000 null genes were simulated with CV = 0.3, $\sigma^2 = 0.25$, $h^2 = 0.6$, and (**A**) $h_x^2 = 0$; (**B**) $h_x^2 = 0.4$; (**C**) $h_x^2 = 0.8$. λ_{gc} (y-axis) changes with sample size n (x-axis). Methods for comparison were MACAU (red), Negative binomial (purple), GEMMA (blue), and Linear (cyan). Both MACAU and GEMMA provide calibrated test statistics in the presence of sample non-independence across a range of settings. λ_{gc} from Poisson exceeds 10 in all settings and is thus not shown.

highly heritable predictor variable ($h_x^2 = 0.8$), λ_{gc} (when $n = 1000$) from the negative binomial, Poisson and linear models increases to 2.13, 101.43 and 1.81, respectively, whereas λ_{gc} from MACAU and GEMMA remains close to 1 (1.02 and 1.05).

We also compared MACAU with edgeR (25) and DESeq2 (15), two commonly used methods for DE analysis (11,101). Because edgeR and DESeq2 were designed for discrete predictor valuables, we discretized the continuous predictor x into 0/1 based on the median predictor value across individuals. We then applied all methods to the same binarized predictor values for comparison. Results are shown in Supplementary Figure S6. For the five methods compared above, the results on binarized values are comparable with those for continuous variables (i.e. Supplementary Figure S6 versus Figure 1). Both edgeR and DESeq2 produce anti-conservative P-values and perform similarly to the negative binomial model in terms of type I error control.

Finally, we explored the use of PCs from the gene expression matrix or the genotype matrix to control for sample non-independence. Genotype PCs have been used as covariates to control for population stratification in association studies (102). However, recent comparative studies have shown that using PCs is less effective than using LMMs (83,103). Consistent with the poorer performance of PCs in association studies (83,103), using the top PCs from either the gene expression matrix or the genotype matrix does not improve type I error control for negative binomial, Poisson, linear, edgeR or DESeq2 approaches (Supplementary Figures S7 and 8).

Simulations: power to identify DE genes

Our second set of simulations was designed to compare the power of different methods for identifying DE genes, again based on parameters inferred from real data. This time, we simulated a total of 10 000 genes, among which 1000 genes were truly DE and 9000 were non-DE. For the DE genes, simulated effect sizes corresponded to a fixed PVE in gene expression levels that ranged from 15 to 35%. For each set of parameters, we performed 10 replicate simulations and measured model performance based on the AUC (as in (35,68,104)). We also examined several key factors that could influence the relative performance of the alternative methods: (i) gene expression heritability (h^2); (ii) correlation between the predictor variable x and genetic relatedness (measured by the heritability of x, or h_x^2); (iii) variation of the TRCs across samples (measured by the CV); (iv) the over-dispersion parameter (σ^2); (v) the effect size (PVE); and (vi) sample size (n). To do so, we first performed simulations using a default set of values ($h^2 = 0.3$, $h_x^2 = 0$, CV = 0.3, $\sigma^2 = 0.25$, PVE = 0.25 and $n = 63$) and then varied them one at a time to examine the influence of each factor on the relative performance of each method.

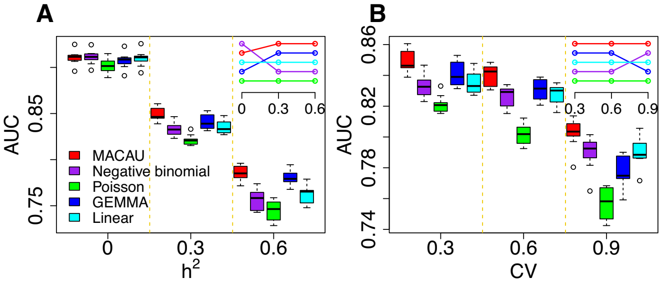

Figure 3. MACAU exhibits increased power to detect true positive DE genes across a range of simulation settings. Area under the curve (AUC) is shown as a measure of performance for MACAU (red), Negative binomial (purple), Poisson (green), GEMMA (blue), and Linear (cyan). Each simulation setting consists of 10 simulation replicates, and each replicate includes 10 000 simulated genes, with 1 000 DE and 9 000 non-DE. We used $n = 63$, $h_x^2 = 0.4$, PVE = 0.25, and $\sigma^2 = 0.25$. In (**A**) we increased h^2 while maintaining CV = 0.3 and in (**B**) we increased CV while maintaining $h^2 = 0.3$. Boxplots of AUC across replicates for different methods show that (A) heritability (h^2) influences the relative performance of the methods that account for sample non-independence (MACAU and GEMMA) compared to the methods that do not (negative binomial, Poisson, linear); (B) variation in total read counts across individuals, measured by the coefficient of variation (CV), influences the relative performance of GEMMA and negative binomial. Insets in the two figures show the rank of different methods, where the top row represents the highest rank.

Our results show that MACAU works either as well as or better than other methods in almost all settings (Figure 3 and Supplementary Figure S9–14), probably because it both models count data directly and controls for sample non-independence. In contrast, the Poisson approach consistently fared the worst across all simulation scenarios, presumably because it fails to account for any sources of overdispersion (Figure 3 and Supplementary Figures S9–14).

Among the factors that influence the relative rank of various methods, the most important factor was heritability (h^2) (Figure 3A). While all methods perform worse with increasing gene expression heritability, heritability disproportionately affects the performance of models that do not account for relatedness (i.e. negative binomial, Poisson and Linear), whereas when heritability is zero ($h^2 = 0$), these approaches tend to perform slightly better. Therefore, for non-heritable genes, linear models perform slightly better than GEMMA, and negative binomial models work similarly or slightly better than MACAU. This observation most likely arises because linear and negative binomial models require fewer parameters and thus have a greater number of degrees of freedom. However, even in this setting, the difference between MACAU and negative binomial is small, suggesting that MACAU is robust to model misspecification and works reasonably well even for non-heritable genes. On the other hand, when heritability is moderate ($h^2 = 0.3$) or high ($h^2 = 0.6$), the methods that account for sample non-independence are much more powerful than the methods that do not. Because almost all genes are influenced by cis-eQTLs (46,47) and are thus likely heritable to some extent, MACAU's robustness for non-heritable genes and its high performance gain for heritable genes make it appealing.

The second most important factor in relative model performance was the variation of TRCs across individuals (CV; Figure 3B). While all methods perform worse with increasing CV, CV particularly affects the performance of GEMMA. Specifically, when CV is small (0.3; as the baboon data), GEMMA works well and is the second best method behind MACAU. However, when CV is moderate (0.6) or high (0.9), the performance of GEMMA quickly decays: it becomes only the fourth best method when CV = 0.9. GEMMA performs poorly in high CV settings presumably because the LMM fails to account for the mean-variance dependence observed in count data, which is in agreement with previous findings (59,105).

The other four factors we explored had small impacts on the relative performance of the alternative methods, although they did affect their absolute performance. For example, as one would expect, power increases with large effect sizes (PVE) (Supplementary Figure S9) or large sample sizes (Supplementary Figure S10), and decreases with large over-dispersion σ^2 (Supplementary Figure S11) or large h_x^2 (Supplementary Figure S12).

Finally, we included comparisons with edgeR (25) and DESeq2 (15). In the basic parameter simulation setting ($n = 63$, CV = 0.3, $h_x^2 = 0$, PVE = 0.25, $h^2 = 0.3$ and $\sigma^2 = 0.25$), we again discretized the continuous predictor x into a binary 0/1 variable based on the median predictor value across individuals. Results for all methods are shown in Supplementary Figure S13A. For the five methods also tested on a continuous predictor variable, the power on binarized values is much reduced compared with the power when the predictor variable is modeled without binarization (e.g. Supplementary Figure S13A versus Figure 3). Further, neither edgeR nor DESeq2 perform well, consistent with the recent move from these methods towards linear models in differential expression analysis (3,7,45–47,106). This result is not contingent on having large sample sizes. In small sample size settings ($n = 6$, $n = 10$ and $n = 14$, with samples balanced between the two classes, 0 or 1), MACAU again outperforms the other methods, though the power difference is much smaller ($n = 10$ and $n = 14$; Supplementary Figures S13C and 13D) and sometimes negligible ($n = 6$, Supplementary Figure S13B).

In summary, the power of MACAU and other methods, as well as the power difference between methods, is influenced in a continuous fashion by multiple factors. Larger sample sizes, larger effect sizes, lower read depth variation, lower gene expression heritability, lower predictor variable heritability and lower over-dispersion all increase power. However, MACAU's power is less diminished by high gene expression heritability and high read depth variability than the non-mixed model methods, while retaining the advantage of modeling the count data directly. In challenging data analysis settings (e.g. when sample size is low *and* effect size is low: Supplementary Figure S13B for $n = 6$), no method stands out and using MACAU results in no or negligible gains in power relative to other methods. When the sample size is low ($n = 6$) and effect sizes are large, however, MACAU consistently outperforms the other methods ($n = 6$, Supplementary Figure S14).

Real data applications

To gain insight beyond simulation, we applied MACAU and the other six methods to three recently published RNAseq datasets.

The first dataset we considered is the baboon RNAseq study (7) used to parameterize the simulations above. Expression measurements on 12 018 blood-expressed genes were collected by the (ABRP) (71) for 63 adult baboons (26 females and 37 males), among which some were relatives. Here, we applied MACAU and the six other methods to identify genes with sex-biased expression patterns. Sex-associated genes are known to be enriched on sex chromosomes (107,108), and we use this enrichment as one of the criteria to compare method performance, as in (18). Because the same nominal *P*-value from different methods may correspond to different type I errors, we compared methods based on empirical FDR. In particular, we permuted the data to construct an empirical null, estimated the FDR at any given *P*-value threshold, and counted the number of discoveries at a given FDR cutoff (see 'Materials and Methods' section for permutation details and a comparison between two different permutation procedures; Supplementary Figure S15).

In agreement with our simulations, MACAU was the most powerful method of those we considered. Specifically, at an empirical FDR of 5%, MACAU identified 105 genes with sex-biased expression patterns, 40% more than that identified by the linear model, the second best method at this FDR cutoff (Figure 4A). At a more relaxed FDR of 10%, MACAU identified 234 sex-associated genes, 47% more than that identified by the negative binomial model, the second best method at this FDR cutoff (Figure 4A). Further, as expected, the sex-associated genes detected by MACAU are enriched on the X chromosome (the Y chromosome is not assembled in baboons and is thus ignored), and this enrichment is stronger for the genes identified by MACAU than by the other methods (Figure 4B). Of the remaining approaches, the negative binomial, linear model and GEMMA all performed similarly and are ranked right after MACAU. The Poisson model performs the worst, and edgeR and DESeq2 fall between the Poisson model and the other methods (Figure 4A and B).

The second dataset we considered is an RNAseq study on T2D collected as part of the FUSION study (60). Here, the data were collected from skeletal muscle samples from 267 individuals with expression measurements on 21 753 genes. Individuals are from three municipalities (Helsinki, Savitaipale and Kuopio) in Finland. Individuals within each municipality are more closely related than individuals between municipalities (e.g. the top genotype PCs generally

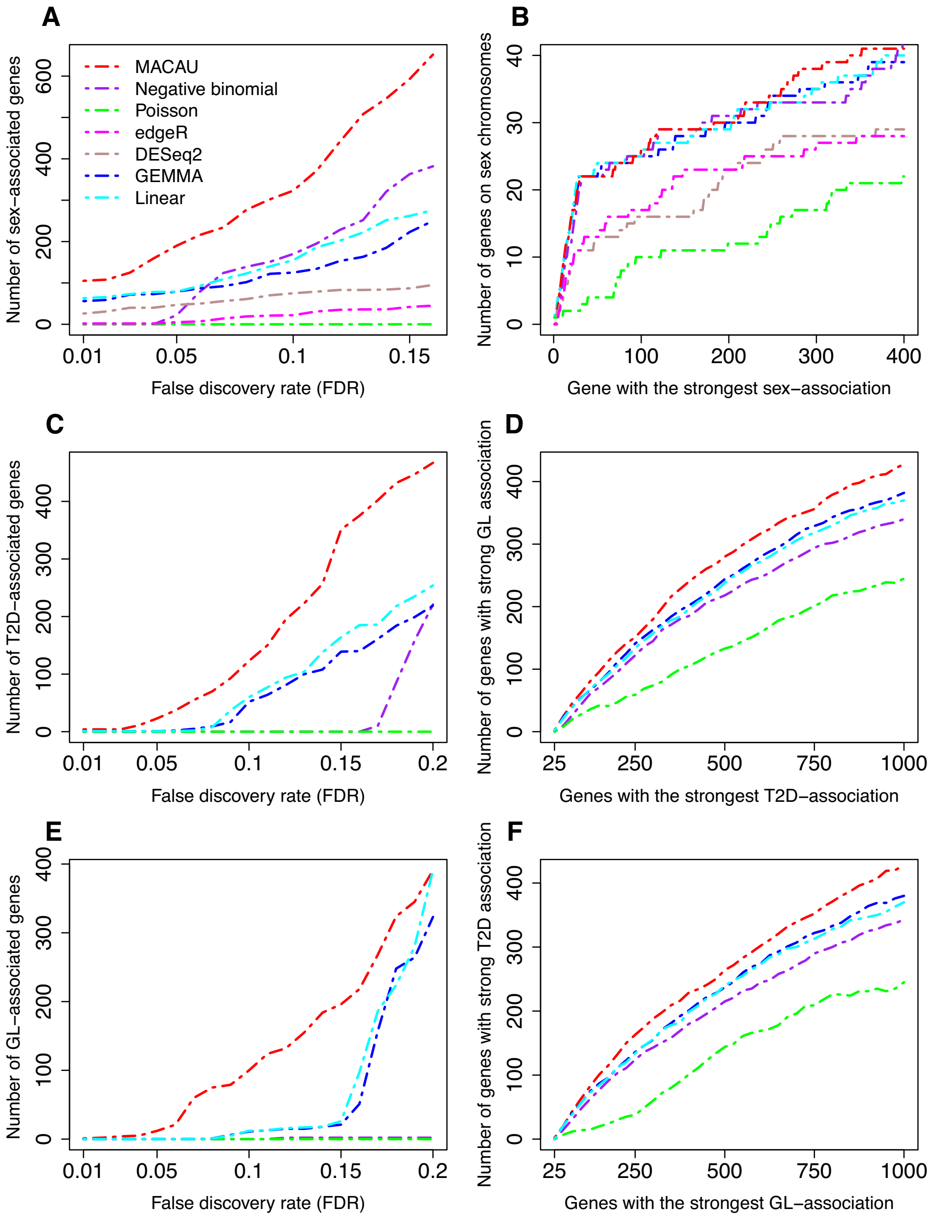

Figure 4. MACAU identifies more differentially expressed genes than other methods in the baboon (**A**-**B**) and FUSION (**C**-**F**) data sets. Methods for comparison include MACAU (red), Negative binomial (purple), Poisson (green), edgeR (magenta), DESeq2 (rosybrown), GEMMA (blue), and Linear (cyan). (A) shows the number of sex-associated genes identified by different methods at a range of empirical false discovery rates (FDRs). (B) shows the number of genes that are on the X chromosome out of the genes that have the strongest sex association for each method (note that the Y chromosome is not assembled in baboons and is thus ignored). For instance, in the top 400 genes identified by MACAU, 41 of them are also on the X chromosome. (C) shows the number of T2D-associated genes identified by different methods at a range of empirical false discovery rates (FDRs). (D) shows the number of genes that are in the list of top 1 000 genes most significantly associated with GL out of the genes that have the strongest association for T2D for each method. For instance, in the top 1 000 genes with the strongest T2D association identified by MACAU, 428 of them are also in the list of top 1 000 genes with the strongest GL association identified by the same method. (E) shows the number of GL-associated genes identified by different methods at a range of FDRs. (F) shows the number of genes that are in the list of top 1 000 genes most significantly associated with T2D out of the genes that have the strongest association for GL for each method. T2D: type II diabetes; GL: fasting glucose level.

correspond to the three municipalities; Supplementary Figure S16). Two related phenotypes were available to us: 162 individuals with T2D or NGT status (i.e. case/control) based on the OGTT and 267 individuals with the quantitative trait fasting GL, a biologically relevant trait of T2D.

We performed analyses to identify genes associated with T2D status as well as genes associated with GL. To accommodate edgeR and DESeq2, we also discretized the continuous GL values into binary 0/1 categories based on the median GL value across individuals. We refer to the resulting values as GL01. Therefore, we performed two sets of analyses for GL: one on the continuous GL values and the other on the discretized GL01 values. Consistent with simulations and the baboon data analysis, MACAU identified more T2D-associated genes and GL-associated genes than other methods across a range of empirical FDR values. For the T2D analysis, MACAU identified 23 T2D-associated genes at an FDR of 5%, while GEMMA and the linear model, the second best methods at this FDR cutoff, identified only 1 T2D-associated gene (Figure 4C). Similarly, at an FDR of 10%, MACAU identified 123 T2D-associated genes, 51% more than that identified by the linear model, the second best method at this FDR cutoff (Figure 4C). For GL analysis, based on an FDR of 5%, MACAU detected 12 DE genes, while the other methods did not identify any DE genes at this FDR cutoff. At an FDR of 10%, MACAU identified 100 GL associated genes, while the second best methods—the linear model and GEMMA—identified 12 DE genes (Figure 4E). For the dichotomized GL01, none of the methods detected any DE genes even at a relaxed FDR cutoff of 20%, highlighting the importance of modeling the original continuous predictor variable in DE analysis.

Several lines of evidence support the biological validity of the genes detected by MACAU. First, we performed gene ontology (GO) analysis using LRpath (109) on T2D and GL associated genes identified by MACAU, as in the FUSION study (60) (Supplementary Figure S17). The GO analysis results for T2D and GL are consistent with previous studies (60,110) and are also similar to each other, as expected given the biological relationship between the two traits. In particular, T2D status and high GL are associated with decreased expression of cellular respiratory pathway genes, consistent with previous observations (60,110). T2D status and GL are also associated with several pathways that are related to mTOR, including generation of precursor metabolites, poly-ubiquitination and vesicle trafficking, in agreement with a prominent role of mTOR pathway in T2D etiology (111–114).

Second, we performed overlap analyses between T2D and GL associated genes. We reasoned that T2D-associated genes are likely associated with GL because T2D shares a common genetic basis with GL (115–117) and T2D status is determined in part by fasting GL. Therefore, we used the overlap between genes associated with T2D and genes associated with GL as a measure of method performance. In the overlap analysis, genes with the strongest T2D association identified by MACAU show a larger overlap with the top 1000 genes that have the strongest GL association than did genes identified by other methods (Figure 4D). For instance, among the top 100 genes with the strongest T2D-association evidence from MACAU, 63 of them also show strong association evidence with GL. In contrast, only 55 of the top 100 genes with the strongest T2D-association identified by GEMMA, the second best method, show strong association evidence with GL. We observed similar results, with MACAU performing the best, when performing the reciprocal analysis (overlap between genes with the strongest GL-association and the top 1000 genes that have the strongest T2D-association: Figure 4F). To include the comparison with edgeR and DESeq2, we further examined the overlap between T2D associated genes and GL01 associated genes for all methods (Supplementary Figure S18). Again, MACAU performs the best, followed by GEMMA and the linear model, and neither edgeR nor DESeq2 perform well in this context (Supplementary Figure S18). Therefore, MACAU appears to both confer more power to identify biologically relevant DE genes and be more consistent across analyses of related phenotypes.

To assess the type I error rate of various methods, we permuted the trait data from the baboon and the FUSION studies. Consistent with our simulation results, the *P*-values from MACAU and GEMMA under the permuted null were close to uniformly distributed (slightly conservative) in both datasets, whereas the other methods were not (Supplementary Figures S19 and 20). In addition, none of the methods compared here are sensitive to outliers in the two datasets (Supplementary Figures S21–23).

Finally, although large, population-based RNAseq datasets are becoming more common, MACAU's flexible PMM modeling framework allows it to be applied to DE analysis in small datasets with unrelated individuals as well. In this setting, MACAU can use the gene expression covariance matrix as the $\boldsymbol{K}$ matrix to control for hidden confounding effects that are commonly observed in sequencing studies (48–51). Hidden confounders can induce similarity in gene expression levels across many genes even though individuals are unrelated (52–56), similar to the effects of kinship or population structure. Therefore, by defining $\boldsymbol{K}$ using a gene expression (instead of genetic) covariance matrix, MACAU can effectively control for sample non-independence induced by hidden confounders, thus extending the LMM widely used to control for hidden confounders in array based studies (52–56) to sequencing count data.

To illustrate this application, we analyzed a third dataset on LCLs derived from 69 unrelated Nigerian individuals (YRI) (3) from the HapMap project (118), with expression measurements on 13 319 genes. We also aimed to identify sex-associated genes in this dataset. To demonstrate the effectiveness of MACAU in small samples, we randomly subsampled individuals from the data to create small datasets with either $n = 6$ (3 males and 3 females), $n = 10$ (5 males and 5 females) or $n = 14$ individuals (7 males and 7 females). For each sample size n, we performed 20 replicates of random subsampling and then evaluated method performance by averaging across replicates. In each replicate, we used the gene expression covariance matrix as $\boldsymbol{K}$ and compared MACAU's performance against other methods. Because of the small sample size, none of the methods were able to identify DE genes at an FDR cutoff of 10%, consistent with recent arguments that at least 6–12 biological replicates are needed to ensure sufficient power and replicability in DE analysis (11). We therefore used enrichment of genes on the sex chromosomes to compare the performance of different methods (Supplementary Figure S24). The enrichment of top ranked sex-associated genes on sex chromosomes has previously been used for method comparison and is especially suitable for comparing methods in the presence of batch effects and other hidden confounding factors (119).

In this comparison, MACAU performs the best of all methods when the sample size is either $n = 10$ or $n = 14$, and is ranked among the best (together with the negative binomial model) when $n = 6$. For instance, when $n = 6$, among the top 50 genes identified by each method, the number of genes on the sex chromosomes for MACAU, negative binomial, Poisson, edgeR, DESeq2, GEMMA and Linear are 3.3, 2.7, 3.1, 1.8, 3.0, 2.0 and 2.4, respectively. The advantage of MACAU becomes larger when the sample size increases: for example, when $n = 14$, an average of 10.6 genes in the top 50 genes from MACAU are on the sex chromosomes, which is again larger than that from the negative binomial (8.3), Poisson (6.0), edgeR (6.65), DESeq2 (8.8), GEMMA (9.8) or Linear (8.05). These results suggest that MACAU can also perform better than existing methods in relatively small sample study designs with unrelated individuals by controlling for hidden confounders. However, MACAU's power gain is much smaller in this setting than in the first two datasets we considered (the baboon and Fusion data). In addition, MACAU's power gain is negligible in the case of $n = 6$ when compared with the second best method, though its power gain over the commonly used edgeR and DESeq2 is still substantial. MACAU's small power gain in this data presumably stems from both the small sample size and the small effect size of sex in the data, consistent with previous reports for blood cell-derived gene expression (3,7,120).

DISCUSSION

Here, we present an effective Poisson mixed effects model, together with a computationally efficient inference method and software implementation in MACAU, for identifying DE genes in RNAseq studies. MACAU directly models count data and, using two random effects terms, controls for both independent over-dispersion and sample non-independence. Because of its flexible modeling framework, MACAU controls for type I error in the presence of individual relatedness, population structure and hidden confounders, and MACAU achieves higher power than several other methods for DE analysis across a range of settings. In addition, MACAU can easily accommodate continuous predictor variables and biological or technical covariates. We have demonstrated the benefits of MACAU using both simulations and applications to three recently published RNAseq datasets.

MACAU is particularly well-suited to datasets that contain related individuals or population structure. Several major population genomic resources contain structure of these kinds. For example, the HapMap population (118), the Human Genome Diversity Panel (121), the 1000 Genomes Project in humans (122) as well as the 1001 Genomes Project in Arabidopsis (123) all contain data from multiple populations or related individuals. Several recent large-scale RNAseq projects also collected individuals from genetically differentiated populations (45). MACAU is also well-suited to analyzing genes with moderate to high heritability. Previous studies in humans have shown that, while heritability varies across genes, many genes are moderately or highly heritable, and almost all genes have detectable eQTL (46,124). Analyzing these data with MACAU can reduce false positives and increase power. Notably, even when genes exhibit zero heritability, our results show that MACAU incurs minimal loss of power compared with other approaches.

While we have mainly focused on illustrating the benefits of MACAU for controlling for individual relatedness and population stratification, we note that MACAU can be used to control for sample non-independence occurred in other settings as we have demonstrated with the third real data application. For example, cell type heterogeneity (54) or other hidden confounding factors (52) are commonly observed in sequencing studies and can induce gene expression similarity even when individuals are unrelated (48–51). Because the gene expression covariance matrix $\boldsymbol{K}$ contains information on sample non-independence caused by hidden confounding factors (52–56), MACAU could be applied to control for hidden confounding effects by using the gene expression covariance as the $\boldsymbol{K}$ matrix. Therefore, MACAU provides a natural avenue for extending the commonly used mixed effects model for controlling for hidden confounding effects (52–55) in array-based studies to sequencing studies. In addition, although we have designed MACAU for differential expression analysis, we note that MACAU may also be effective in other common settings. For example, MACAU could be readily applied in QTL mapping studies to identify genetic variants that are associated with gene expression levels estimated using RNAseq or related high-throughput sequencing methods.

In the present study, we have focused on demonstrating the performance of MACAU in three published RNAseq datasets with sample sizes ranging from small ($n = 6$) to medium ($n = 63$) to large ($n = 267$), relative to the size of most current RNAseq studies. Compared with small sample studies, RNAseq studies with medium or large sample sizes are better powered and more reproducible and are thus becoming increasingly common in genomics (10,11). For example, a recent comparative study makes explicit calls for medium to large sample RNAseq studies performed with at least 12 replicates per condition (i.e. $n \geq 24$) (11). However, we recognize that many RNAseq studies are still carried out with a small number of samples (e.g. 3 replicates per condition). As our simulations make clear, the power of all analysis methods is dramatically reduced with decreasing sample size, conditional on fixed values of other factors that influence power (e.g., effect size, gene expression heritability). Thus, MACAU's advantage is no longer obvious in simulated data with only three replicates per condition when the effect size is also small (although its advantage becomes apparent when the simulated effect size increases: Supplementary Figures S13B and 14). In addition, MACAU's advantage is much smaller and sometimes negligible in the small real dataset when compared with the medium and large datasets analyzed here. Furthermore, because MACAU requires estimating one more parameter than other existing methods, MACAU requires at least five samples to run while existing DE methods require at least four. Therefore, MACAU may not confer benefits to power in some settings, and is especially likely (like all methods) to be underpowered in very small sample sizes with small effect sizes. Future extensions of MACAU are likely needed to ensure its robust performance in small as well as moderate to large samples. For example, further power improvements could be achieved by borrowing information across genes to estimate the over-dispersion parameter (15,22,25) or building in a hierarchical structure to model many genes at once.

Like other DE methods (24,25), MACAU requires data pre-processing to obtain gene expression measurements from raw sequencing read files. This data pre-processing step may include read alignment, transcript assembly, alternative transcript quantification, transcript measurement and normalization. Many methods are available to perform these tasks (12,14,68,125–130) and different methods can be differentially advantageous across settings (68,125,131). Importantly, MACAU can be paired with any pre-processing method that retains the count nature of the data. While we provide a preliminary comparison of several methods here (see 'Materials and Methods' section; Supplementary Figure S3), a full analysis of how different data pre-processing choices affect MACAU's performance in alternative study designs is beyond the scope of this paper. Notably, recent results suggest that a recommended approach is to incorporate data pre-processing and DE analysis into the same, joint statistical framework (132), which represents an important next step for the MACAU software package.

We note that, like many other DE methods (15,25), we did not model gene length in MACAU. Because gene length does not change from sample to sample, it does not affect differential expression analysis on any particular gene (15,25). However, gene length will affect the power of DE analysis across different genes: genes with longer length tend to have a larger number of mapped reads and more accurate expression measurements, and as a consequence, DE analysis on these genes tends to have higher statistical power (2,70,133). Gene length may also introduce sample-specific effects in certain datasets (134). Therefore, understanding the impact of, and taking into account gene length effects, in MACAU DE analysis represents another possible future extension.

Currently, despite the newly developed computationally efficient algorithm, applications of MACAU can still be limited by its relatively heavy computational cost. The MCMC algorithm in MACAU scales quadratically with the number of individuals/samples and linearly with the number of genes. Although MACAU is two orders of magnitude faster than the standard software MCMCglmm for fitting Poisson mixed effects models (Supplementary Table S1), it can still take close to 20 h to analyze a dataset of the size of the FUSION data we considered here (267 individuals and 21 753 genes). Therefore, new algorithms will be needed to use MACAU for datasets that are orders of magnitude larger.

URLs

The software implementation of MACAU is freely available at: www.xzlab.org/software.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Matthew Stephens for insight and support on previous versions of MACAU. We thank Baylor College of Medicine Human Genome Sequencing Center for access to the current version of the baboon genome assembly (Panu 2.0). We thank FUSION investigators for access to the FUSION expression data.

FUNDING

National Institutes of Health [R01HG009124 to X.Z.; R01GM102562 to J.T.; R21AG049936 to J.T.; U01DK062370 to L.S.]; China Scholarship Council (to S.S.). Funding for open access charge: University of Michigan.

Conflict of interest statement. None declared.

REFERENCES

1. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
2. Mortazavi,A., Williams,B.A., Mccue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
3. Pickrell,J.K., Marioni,J.C., Pai,A.A., Degner,J.F., Engelhardt,B.E., Nkadori,E., Veyrieras,J.B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.
4. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
5. Oshlack,A., Robinson,M.D. and Young,M.D. (2010) From RNA-seq reads to differential expression results. *Genome Biol.*, **11**, 220.
6. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
7. Tung,J., Zhou,X., Alberts,S.C., Stephens,M. and Gilad,Y. (2015) The genetic architecture of gene expression levels in wild baboons. *Elife*, **4**, e04729.
8. Bennett,B.J., Farber,C.R., Orozco,L., Kang,H.M., Ghazalpour,A., Siemers,N., Neubauer,M., Neuhaus,I., Yordanova,R., Guan,B. *et al.* (2010) A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res.*, **20**, 281–290.
9. Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.
10. Wood,A.R., Esko,T., Yang,J., Vedantam,S., Pers,T.H., Gustafsson,S., Chu,A.Y., Estrada,K., Luan,J., Kutalik,Z. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.
11. Schurch,N.J., Schofield,P., Gierlinski,M., Cole,C., Sherstnev,A., Singh,V., Wrobel,N., Gharbi,K., Simpson,G.G., Owen-Hughes,T. *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–851.
12. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
13. Hu,M., Zhu,Y., Taylor,J.M.G., Liu,J.S. and Qin,Z.H.S. (2012) Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics*, **28**, 63–68.
14. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**, 323.
15. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
16. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
17. Li,J., Jiang,H. and Wong,W.H. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, R50.

18. Zhou,Y.H., Xia,K. and Wright,F.A. (2011) A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, **27**, 2672–2678.
19. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
20. Wang,L.K., Feng,Z.X., Wang,X., Wang,X.W. and Zhang,X.G. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.
21. Langmead,B., Hansen,K.D. and Leek,J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, **11**, R83.
22. Li,J., Witten,D.M., Johnstone,I.M. and Tibshirani,R. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538.
23. Auer,P.L. and Doerge,R.W. (2011) A two-stage poisson model for testing RNA-Seq data. *Stat. Appl. Genet. Mol.*, **10**, 1–26.
24. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
25. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
26. McCarthy,D.J., Chen,Y.S. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
27. Di,Y.M., Schafer,D.W., Cumbie,J.S. and Chang,J.H. (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol.*, **10**, 1–28.
28. Wu,H., Wang,C. and Wu,Z. (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232–243.
29. Van De Wiel,M.A., Leday,G.G.R., Pardo,L., Rue,H., Van Der Vaart,A.W. and Van Wieringen,W.N. (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, **14**, 113–128.
30. Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
31. Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.*, **22**, 519–536.
32. Tarazona,S., Garcia-Alcalde,F., Dopazo,J., Ferrer,A. and Conesa,A. (2011) Differential expression in RNA-seq: A matter of depth. *Genome Res.*, **21**, 2213–2223.
33. Law,C.W., Chen,Y.S., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
34. Zwiener,I., Frisch,B. and Binder,H. (2014) Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One*, **9**, e85150.
35. Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.
36. Kvam,V.M., Lu,P. and Si,Y.Q. (2012) A comparison of statistical methods for detecting differentially expressed genes from Rna-Seq data. *Am. J. Bot.*, **99**, 248–256.
37. Zhang,Z.H., Jhaveri,D.J., Marshall,V.M., Bauer,D.C., Edson,J., Narayanan,R.K., Robinson,G.J., Lundberg,A.E., Bartlett,P.F., Wray,N.R. *et al.* (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS One*, **9**, e103207.
38. McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman and Hall/CRC, London.
39. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
40. Price,A.L., Helgason,A., Thorleifsson,G., McCarroll,S.A., Kong,A. and Stefansson,K. (2011) Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.*, **7**, e1001317.
41. Wright,F.A., Sullivan,P.F., Brooks,A.I., Zou,F., Sun,W., Xia,K., Madar,V., Jansen,R., Chung,W.I., Zhou,Y.H. *et al.* (2014) Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.*, **46**, 430–437.
42. Monks,S.A., Leonardson,A., Zhu,H., Cundiff,P., Pietrusiak,P., Edwards,S., Phillips,J.W., Sachs,A. and Schadt,E.E. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, **75**, 1094–1105.
43. Emilsson,V., Thorleifsson,G., Zhang,B., Leonardson,A.S., Zink,F., Zhu,J., Carlson,S., Helgason,A., Walters,G.B., Gunnarsdottir,S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
44. Yang,S.J., Liu,Y.Y., Jiang,N., Chen,J., Leach,L., Luo,Z.W. and Wang,M.H. (2014) Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics*, **15**, 13.
45. Lappalainen,T., Sammeth,M., Friedlander,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzalez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
46. Ardlie,K.G., DeLuca,D.S., Segre,A.V., Sullivan,T.J., Young,T.R., Gelfand,E.T., Trowbridge,C.A., Maller,J.B., Tukiainen,T., Lek,M. *et al.* (2015) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
47. Battle,A., Mostafavi,S., Zhu,X.W., Potash,J.B., Weissman,M.M., McCormick,C., Haudenschild,C.D., Beckman,K.B., Shi,J.X., Mei,R. *et al.* (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, **24**, 14–24.
48. Stegle,O., Parts,L., Piipari,M., Winn,J. and Durbin,R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500–507.
49. Leek,J.T. (2014) Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, **42**, e161.
50. Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.
51. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.
52. Kang,H.M., Ye,C. and Eskin,E. (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.
53. Listgarten,J., Kadie,C., Schadt,E.E. and Heckerman,D. (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 16465–16470.
54. Zou,J., Lippert,C., Heckerman,D., Aryee,M. and Listgarten,J. (2014) Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods*, **11**, 309–311.
55. Rahmani,E., Zaitlen,N., Baran,Y., Eng,C., Hu,D.L., Galanter,J., Oh,S., Burchard,E.G., Eskin,E., Zou,J. *et al.* (2016) Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods*, **13**, 443-445.
56. McGregor,K., Bernatsky,S., Colmegna,I., Hudson,M., Pastinen,T., Labbe,A. and Greenwood,C.M.T. (2016) An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol.*, **17**, 84.
57. Price,A.L., Patterson,N.J., Plenge,R.M., Weinblatt,M.E., Shadick,N.A. and Reich,D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
58. Zhou,X. and Stephens,M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.
59. Lea,A.J., Alberts,S.C., Tung,J. and Zhou,X. (2015) A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet.*, **11**, e1005650.
60. Scott,L.J., Erdos,M.R., Huyghe,J.R., Welch,R.P., Beck,A.T., Boehnke,M., Collins,F.S. and Parker,S.C.J. (2016) The genetic regulatory sigature of type 2 diabetes in human skeletal muscle. *Nat. Commun.*, **7**, 11764.
61. Fruhwirth-Schnatter,S. and Wagner,H. (2006) Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika*, **93**, 827–841.

62. Scott,S.L. (2011) Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Stat. Pap.*, **52**, 87–109.
63. Fruhwirth-Schnatter,S. and Fruhwirth,R. (2010) Data Augmentation and MCMC for Binary and Multinomial Logit Models. *Statistical Modelling and Regression Structures*. Springer, NY.
64. Lippert,C., Listgarten,J., Liu,Y., Kadie,C.M., Davidson,R.I. and Heckerman,D. (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.
65. Zhou,X., Carbonetto,P. and Stephens,M. (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.*, **9**, e1003264.
66. Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S*. Springer, NY.
67. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S plus to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
68. Teng,M., Love,M.I., Davis,C.A., Djebali,S., Dobin,A., Graveley,B.R., Li,S., Mason,C.E., Olson,S., Pervouchine,D. *et al.* (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biol.*, **17**, 74.
69. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y.F., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
70. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
71. Alberts,S. and Altmann,J. (2012) In: Kappeler,PM and Watts,DP (eds). *Long-Term Field Studies of Primates*. Springer, Berlin Heidelberg, pp. 261–287.
72. Alberts,S.C., Buchan,J.C. and Altmann,J. (2006) Sexual selection in wild baboons: from mating opportunities to paternity success. *Anim. Behav.*, **72**, 1177–1196.
73. Buchan,J.C., Alberts,S.C., Silk,J.B. and Altmann,J. (2003) True paternal care in a multi-male primate society. *Nature*, **425**, 179–181.
74. Altmann,J., Altmann,S. and Hausfater,G. (1981) Physical maturation and age estimates of yellow baboons, Papio-Cynocephalus, in Amboseli National-Park, Kenya. *Am. J. Primatol.*, **1**, 389–399.
75. Archie,E.A., Tung,J., Clark,M., Altmann,J. and Alberts,S.C. (2014) Social affiliation matters: both same-sex and opposite-sex relationships predict survival in wild female baboons. *Proc. R. Soc. B.*, **281**, 20141261.
76. Valle,T., Ehnholm,C., Tuomilehto,J., Blaschak,J., Bergman,R.N., Langefeld,C.D., Ghosh,S., Watanabe,R.M., Hauser,E.R., Magnuson,V. *et al.* (1998) Mapping genes for NIDDM—design of the finland united states investigation of NIDDM Genetics (FUSION) study. *Diabetes Care*, **21**, 949–958.
77. Vaatainen,S., Keinanen-Kiukaanniemi,S., Saramies,J., Uusitalo,H., Tuomilehto,J. and Martikainen,J. (2014) Quality of life along the diabetes continuum: a cross-sectional view of health-related quality of life and general health status in middle-aged and older Finns. *Qual. Life Res.*, **23**, 1935–1944.
78. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
79. Churchill,G.A. and Doerge,R.W. (2008) Naive application of permutation testing leads to inflated type I error rates. *Genetics*, **178**, 609–610.
80. Abney,M. (2015) Permutation testing in the presence of polygenic variation. *Genet. Epidemiol.*, **39**, 249–258.
81. Zhou,X., Lindsay,H. and Robinson,M.D. (2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.*, **42**, e91.
82. George,N.I., Bowyer,J.F., Crabtree,N.M. and Chang,C.W. (2015) An iterative leave-one-out approach to outlier detection in RNA-seq data. *PLoS One*, **10**, e0125224.
83. Kang,H.M., Sul,J.H., Service,S.K., Zaitlen,N.A., Kong,S.Y., Freimer,N.B., Sabatti,C. and Eskin,E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
84. Kang,H.M., Zaitlen,N.A., Wade,C.M., Kirby,A., Heckerman,D., Daly,M.J. and Eskin,E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.
85. Tempelman,R.J. and Gianola,D. (1996) A mixed effects model for overdispersed count data in animal breeding. *Biometrics*, **52**, 265–279.
86. Tempelman,R.J. (1998) Generalized linear mixed models in dairy cattle breeding. *J. Dairy Sci.*, **81**, 1428–1444.
87. Pinheiro,J.C. and Chao,E.C. (2006) Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J. Comput. Graph. Stat.*, **15**, 58–81.
88. Goldstein,H. (1991) Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, **78**, 45–51.
89. Breslow,N.E. and Clayton,D.G. (1993) Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.*, **88**, 9–25.
90. Breslow,N.E. and Lin,X.H. (1995) Bias correction in generalized linear mixed models with a single-component of dispersion. *Biometrika*, **82**, 81–91.
91. Browne,W.J. and Draper,D. (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.*, **1**, 473–513.
92. Lin,X.H. and Breslow,N.E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Stat. Assoc.*, **91**, 1007–1016.
93. Goldstein,H. and Rasbash,J. (1996) Improved approximations for multilevel models with binary responses. *J. R. Stat. Soc. A*, **159**, 505–513.
94. Rodriguez,G. and Goldman,N. (2001) Improved estimation procedures for multilevel models with binary response: a case-study. *J. R. Stat. Soc. A*, **164**, 339–355.
95. Jang,W. and Lim,J. (2009) A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects. *Commun. Stat.*, **38**, 692–702.
96. Fong,Y.Y., Rue,H. and Wakefield,J. (2010) Bayesian inference for generalized linear mixed models. *Biostatistics*, **11**, 397–412.
97. Smith,A.F.M. and Roberts,G.O. (1993) Bayesian computation via the gibbs sampler and related markov-chain monte-carlo methods. *J. R. Stat. Soc. B*, **55**, 3–23.
98. Gelman,A. and Shirley,K. (2011) Inference from simulations and monitoring convergence. *Handbook of Markov Chain Monte Carlo*, 163–174.
99. Schwartz,L. (1965) On Bayes procedures. *Zeitschrift fűr Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **4**, 10–26.
100. Hadfield,J.D. (2010) MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J. Stat. Softw.*, **33**, 1–22.
101. Seyednasrollah,F., Laiho,A. and Elo,L.L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform.*, **16**, 59–70.
102. Patterson,N., Price,A.L. and Reich,D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, 2074–2093.
103. Yang,J., Zaitlen,N.A., Goddard,M.E., Visscher,P.M. and Price,A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.
104. Rapaport,F., Khanin,R., Liang,Y.P., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D. and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq datas. *Genome Biol.*, **14**, R95.
105. Chen,H., Wang,C., Conomos,M.P., Stilp,A.M., Li,Z., Sofer,T., Szpiro,A.A., Chen,W., Brehm,J.M., Celedón,J.C. *et al.* (2016) Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.*, **98**, 653–666.
106. Zhou,X., Cain,C.E., Myrthil,M., Lewellen,N., Michelini,K., Davenport,E.R., Stephens,M., Pritchard,J.K. and Gilad,Y. (2014) Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol.*, **15**, 547.
107. Vawter,M.P., Evans,S., Choudary,P., Tomita,H., Meador-Woodruff,J., Molnar,M., Li,J., Lopez,J.F., Myers,R., Cox,D. *et al.* (2004) Gender-specific gene expression in post-mortem human brain: Localization to sex chromosomes. *Neuropsychopharmacol*, **29**, 373–384.
108. Lemos,B., Branco,A.T., Jiang,P.P., Hartl,D.L. and Meiklejohn,C.D. (2014) Genome-wide gene expression effects of sex chromosome imprinting in Drosophila. *G3*, **4**, 1–10.

109. Kim,J.H., Karnovsky,A., Mahavisno,V., Weymouth,T., Pande,M., Dolinoy,D.C., Rozek,L.S. and Sartor,M.A. (2012) LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC Genomics*, **13**, 526.
110. Mootha,V.K., Lindgren,C.M., Eriksson,K.-F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstråle,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
111. Leibowitz,G., Cerasi,E. and Ketzinel-Gilad,A. (2008) The role of mTOR in the adaptation and failure of beta-cells in type 2 diabetes. *Diabetes Obes. Metab.*, **10**, 157–169.
112. Ost,A., Svensson,K., Ruishalme,I., Brannmark,C., Franck,N., Krook,H., Sandstrom,P., Kjolhede,P. and Stralfors,P. (2010) Attenuated mTOR signaling and enhanced autophagy in adipocytes from obese patients with type 2 diabetes. *Mol. Med.*, **16**, 235–246.
113. Laplante,M. and Sabatini,D.M. (2012) mTOR signaling in growth control and disease. *Cell*, **149**, 274–293.
114. Zoncu,R., Efeyan,A. and Sabatini,D.M. (2011) mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat. Rev. Mol. Cell Biol.*, **12**, 21–35.
115. Matthews,D.R., Hosker,J.P., Rudenski,A.S., Naylor,B.A., Treacher,D.F. and Turner,R.C. (1985) Homeostasis model assessment—insulin resistance and beta-cell function from fasting plasma-glucose and insulin concentrations in man. *Diabetologia*, **28**, 412–419.
116. Lyssenko,V., Nagorny,C.L.F., Erdos,M.R., Wierup,N., Jonsson,A., Spegel,P., Bugliani,M., Saxena,R., Fex,M., Pulizzi,N. *et al.* (2009) Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat. Genet.*, **41**, 82–88.
117. Dupuis,J., Langenberg,C., Prokopenko,I., Saxena,R., Soranzo,N., Jackson,A.U., Wheeler,E., Glazer,N.L., Bouatia-Naji,N., Gloyn,A.L. *et al.* (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, **42**, 105–116.
118. Gibbs,R.A., Belmont,J.W., Hardenbol,P., Willis,T.D., Yu,F.L., Yang,H.M., Ch'ang,L.Y., Huang,W., Liu,B., Shen,Y. *et al.* (2003) The International HapMap Project. *Nature*, **426**, 789–796.
119. Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.
120. Powell,J.E., Henders,A.K., McRae,A.F., Wright,M.J., Martin,N.G., Dermitzakis,E.T., Montgomery,G.W. and Visscher,P.M. (2012) Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.*, **22**, 456–466.
121. Cann,H.M., de Toma,C., Cazes,L., Legrand,M.F., Morel,V., Piouffre,L., Bodmer,J., Bodmer,W.F., Bonne-Tamir,B., Cambon-Thomsen,A. *et al.* (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
122. Landi,M.T., Wang,Y.F., Mckay,J.D., Rafnar,T., Wang,Z.M., Timofeeva,M., Broderick,P., Stefansson,K., Risch,A., Chanock,S.J. *et al.* (2014) Imputation from the 1000 Genomes Project identifies rare large effect variants of BRCA2-K3326X and CHEK2-I157T as risk factors for lung cancer; a study from the TRICL consortium. *Cancer Res.*, **74**, 942–942.
123. Weigel,D. and Mott,R. (2009) The 1001 genomes project for arabidopsis thaliana. *Genome Biol.*, **10**, 107.
124. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
125. Kanitz,A., Gypas,F., Gruber,A.J., Gruber,A.R., Martin,G. and Zavolan,M. (2015) Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.*, **16**, 150.
126. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
127. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
128. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
129. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
130. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
131. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szczesniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
132. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
133. Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, 14.
134. Hansen,K.D., Irizarry,R.A. and Wu,Z.J. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216.