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], 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

6. Duke University Population Research Institute, Duke University, Durham, NC 27708, USA

* Correspondence to: XZ (xzhousph@umich.edu)

Abstract

Identifying differentially expressed (DE) genes from RNA sequencing (RNAseq) studies is one of the most common analyses in genomics. DE analysis often represents the first step towards understanding the molecular mechanisms underlying disease susceptibility and phenotypic variation. However, identifying DE genes from RNAseq presents statistical and computational challenges that arise from several unique properties of the sequencing data. Specifically, gene expression estimates in RNAseq experiments are based on read counts that often display over-dispersion. In addition, gene expression levels are heritable and are influenced by the genetic structure of the study samples. Previous count-based methods for identifying DE genes rely on simple hierarchical Poisson models (e.g., negative binomial) to model over-dispersion, which is assumed to be independent among samples. However, these methods fail to account for the gene expression similarity induced by relatedness and/or population structure, which can cause inflation of test statistics and/or loss of power. To address this problem, we present a Poisson mixed model with two random effects terms to account for both independent over-dispersion and sample relatedness in RNAseq DE analysis. To make our method scalable, we develop a novel sampling-based inference algorithm, taking advantage of recently developed innovations in efficient mixed model optimization and a latent variable representation of the Poisson model. With simulations, we show that, in the presence of population structure, our method properly controls for type I error and is more powerful than several widely used approaches. We apply our method to identify DE genes associated with sex in a baboon data set and DE genes associated with type 2 diabetes status as well as fasting glucose levels in a human data set. In both data sets, our method detects at least 40% more DE genes compared with the next best approach while properly controlling for type I error. Our method is implemented in the MACAU software package, 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 towards understanding the molecular basis of disease susceptibility as well as the genetic and environmental basis of gene expression variation. Progress towards this goal requires statistical methods that can handle the complexities of the increasingly large and structurally complex RNAseq data sets that are now being collected from population and family studies (8,9). However, identifying DE genes from such studies presents key statistical and computational challenges, primarily arising from two features of RNAseq data.

The first important feature, and one that has received considerable attention (3,10,11), is that the raw data from an RNAseq experiment come in the form of read counts. Specifically, RNAseq uses the number of reads mapped to a given gene or isoform as an estimate of its expression level. 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 variance across samples, and vice versa. 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 (12-14). More recent methods also take into account over-dispersion across biological replicates (15,16) by replacing Poisson models with negative binomial models (10,11,17-21) or other related approaches (22-26). While non-count based methods are also commonly used (primarily relying on transformation of the count data to more flexible, continuous distributions (27,28)), recent comparisons have highlighted the benefits of modeling RNAseq data using the original counts and accounting for the resulting mean-variance dependence (29-32), consistent with observations from many count data analyses in other statistical settings (33). 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 (18,27,34).

The second important feature of RNAseq data, which has been largely overlooked in DE analysis thus far, is that gene expression levels are heritable. In humans, the narrow-sense heritability of gene expression levels averages from 15%-34% in peripheral blood (35-39) and is about 23% in adipose tissue (35), with a maximum heritability in both tissues as high as 90% (35,36). 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 (ASE) (3,7,40-42). 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 (35) and baboons (7)). Because gene expression levels are heritable, gene expression levels will covary with kinship or population structure. Failure to account for this covariance could lead to spurious associations or reduced power to detect true DE effects. This phenomenon has been extensively documented in genome-wide association studies (43-45) and more recently, in bisulfite sequencing studies (46), 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 population structure. Consequently, even though count-based methods have been shown to be more powerful, recent RNAseq studies have turned to linear mixed models, which are specifically designed for quantitative traits, to deal with the confounding effects of kinship or population structure (7,36,47).

Here, we present a Poisson mixed model (PMM) that can explicitly model both overdispersed count data and genetic structure in RNAseq data for effective DE analysis. To make our model scalable to large data sets, we also develop an accompanying efficient inference algorithm based on an auxiliary variable representation of the Poisson model (48-50) and recent advances in mixed model methods (44,45,51). 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 (46) to the more difficult Poisson model. MACAU works directly on RNAseq count data and introduces two random effects terms to both control for genetic relatedness among individuals and account for additional independent over-dispersion. As a result, MACAU properly controls for type I error in the presence of population structure and, in a variety of settings, is more powerful for identifying DE genes than several commonly used methods. We illustrate the benefits of MACAU with extensive simulations and real data applications to two large-scale RNAseq studies.

Methods and Materials

Methods for Comparison

We compared the performance of seven different methods in the main text: (1) our Poisson mixed model implemented in the MACAU software package (46); (2) the linear model implemented in the *lm* function in R; (3) the linear mixed model implemented in the GEMMA software package (44,45); (4) the Poisson model implemented in the *glm* function in R; (5) the negative binomial model implemented in the *glm.nb* function in R; (6) edgeR implemented in the *edgeR* package in R (18); (7) DESeq implemented in the *DESeq* package in R (10). For edgeR, we used *estimateCommonDisp*, and *exactTest* functions. For DESeq, we used *estimateDispersions* and *nbiomTest* functions and set method="blind", fitType="local", sharingMode="fit-only" in the *estimateDispersions* function. For both DESeq and edgeR, we have tried many parameter settings in addition to this default, and the results presented in the main text are the best we could obtain. All other methods were used with default settings. The performance of each method was evaluated using the area under the curve (AUC) function implemented in the *pROC* package in R (90), a widely used benchmark for RNAseq method comparisons (68).

Both the linear model and the linear mixed model require quantitative phenotypes. Thus, we considered six different transformations of count data to quantitative values: (1) 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); (2) total read count normalization (TRC), where we divided the number of reads mapped to a given gene by the total number of read counts for each individual (without further transformation to a standard normal within genes: (18)); (3) upper quantile normalization (UQ), where we divided the number of reads mapped to a given gene by the upper quantile (0.75-th percentile) of all genes for each individual (91); (4) relative log expression normalization (RLE) (10); (5) trimmed mean of M-values (TMM) (34); and (6) VOOM normalization (27). Simulations showed that TRCQ, VOOM and TRC worked better than the other three methods, with TRCQ showing a small advantage (Figure S13). 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 the real baboon data set 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 = 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 (46). 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 total read count N_i for each individual from a discrete uniform distribution with a minimum (=1,770,083) and a maximum (=9,675,989), which equal to the minimum and maximum total read counts (i.e. summation of read counts across all genes) from the baboon data. We scaled the total read counts to ensure that the coefficient of variation is small (CV = 0.3), moderate (CV = 0.6) or high (CV = 0.9) across individuals (i.e. $N_{new} = \bar{N} + (N - \bar{N})\ CV\ sd(N)\ /\bar{N}$) and then discretized them. In the special case where CV = 0.3 and *n* = 63, we directly used the observed total read counts 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 $\pm\ 0.01$ tolerance) was 0, 0.4 or 0.8, representing no, moderate and highly heritable predictors. Using this procedure, approximately 30 percent of $x's$ generated were retained, with detailed retaining percent varied for different heritability values.

Based on the simulated sample size, total read counts 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 upper quantiles 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 $\log(\lambda)$ (i.e. PVE $= \frac{V(X\beta)}{V(X\beta)+\sigma^2}$; for DE genes). 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 ($\mu = 100$) 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 total read counts and the latent variable λ, or $y_i \sim Poi(N_i \lambda_i)$ for the i'th 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, total read counts and continuous predictor variable, we performed 10 simulation replicates (i.e. 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, 1,000 genes were DE while 9,000 were non-DE.

RNAseq Data Sets

We considered two published RNAseq data sets in this study.

The first RNAseq data set 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) (67). 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 (92,93) using the software COANCESTRY (94). 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 (92,95). Therefore, we used this data set to identify sex-related gene expression variation. In the analysis, we included the top 5 expression PCs as covariates to control for potential batch effects following the original study (7).

The second RNAseq data set was collected from skeletal muscle samples of Finnish individuals (47) as part of the FUSION project (96,97). 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 measured on these 267 individuals and 21,753 genes following the same stringent quality control as in the FUSION study. For genotypes, we excluded SNPs with minor allele frequency (MAF) < 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: 3 pairs of individuals have a kinship coefficient exceeding 1/32 and 6 pairs exceed 1/64. Two predictors from the data were available to us: the oral glucose tolerance test (OGTT) which classifies *n* = 162 individuals as either 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 (47).

For each of these two RNAseq data sets 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 human 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 false discovery rate (FDR) for any given *p*-value threshold (46). This constrained procedure thus differs from the usual unconstrained permutation procedure (every permutation retained) (98) 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 individual relatedness or population structure (98,99). To validate our constrained permutation procedure and test its effectiveness in estimating FDR, we performed a simulation with 1,000 DE genes and 9,000 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 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 (Figure S14). Therefore, we applied the constrained permutation procedure for all real data analysis.

Finally, we investigated whether the methods we compared were sensitive to outliers (25,100,101) in the two data sets. To do so, we first identified genes with potential outliers in the two data sets using BBSeq (22). In total, we identified 8 genes with potential outliers in the baboon data, 130 genes with potential outliers in the human data (n = 267) and 43 genes with potential outliers in the subset of the human data for which we had T2D diagnoses (n = 162). We counted the number of genes with potential outliers in the top 1,000 genes with strong DE association evidence. In the baboon data (Figure S15), 4 genes with potential outliers are in the top 1,000 genes with the strongest sex association determined by various methods: 2 of them by the negative binomial model, 3 of them by the Poisson model, but 0 of them by MACAU, linear model, or GEMMA. In human data, for T2D analysis (Figure S16), 9 genes with potential outliers are in the top 1,000 genes with the strongest T2D association determined by various methods: 1 by MACAU, 3 by negative binomial, 6 by Poisson, 1 by linear, and 1 by GEMMA. For GL analysis (Figure S17), 15 genes with potential outliers are in the top 1,000 genes with the strongest GL association determined by various methods: 2 by MACAU, 7 by negative binomial, 9 by Poisson, 3 by linear, and 3 by GEMMA. Therefore, the influence of outliers on DE analysis is small in the two data sets.

Results

MACAU Overview

Here, we provide a brief overview of the Poisson mixed model (PMM); more details are available in the Supplementary Material. 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

$$y_i \sim Poi(N_i\lambda_i), \qquad i = 1, 2, \cdots, n,$$

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

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

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

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

where $\boldsymbol{w}_i$ 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* relatedness matrix that models the covariance among individuals due to either individual relatedness or population structure; $\boldsymbol{I}$ 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 or genotype (44). When $\boldsymbol{K}$ is standardized to have $\boldsymbol{tr}(\boldsymbol{K})/n = 1$, $h^2 \in [0,1]$ has the usual interpretation of heritability (44), where the $\boldsymbol{tr}(\cdot)$ denotes the trace of a matrix. Importantly, unlike many other popular DE methods (10,18), 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 model two different aspects of over-dispersion. Specifically, $\boldsymbol{g}$ models the fraction of the extra variance that is explained by individual relatedness or population structure while $\boldsymbol{e}$ models the fraction of the extra variance that is independent across samples. 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 individual relatedness and population structure. In addition, our PMM naturally extends the commonly used linear mixed model (LMM) (44,51,52) 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. The PMM belongs to the generalized linear mixed model 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 (53) or Laplace approximation (54,55), 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 (56-62). 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 (63,64). We use MCMC to draw posterior samples but rely on the asymptotic normality of both the likelihood and the posterior distributions (65) to obtain the approximate maximum likelihood estimate $\hat{\beta}_j$ and its standard error se($\hat{\hat{\beta}}_j$). With $\hat{\beta}_j$ and se($\hat{\hat{\beta}}_j$), we can construct approximate Wald test statistics and *p*-values for hypothesis testing (Supplementary Material). Although we use MCMC, our procedure is still frequentist in nature.

At the technical level, our MCMC algorithm is also novel, taking advantage of an auxiliary variable representation of the Poisson likelihood (48-50) and recent linear algebra innovations for fitting linear mixed models (44,45,51). 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 (46). 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 (66) and can be used to analyze hundreds of samples and tens of thousands of genes with a single desktop PC (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 (Figure S2).

Simulations: control for individual relatedness and population structure

We performed a series of simulations to compare the performance of the PMM implemented in MACAU with four other commonly used methods: (1) a linear model; (2) the linear mixed model implemented in GEMMA (44,45); (3) a Poisson model; and (4) a negative binomial model. We used quantile-transformed data for linear model and GEMMA (see Methods and Materials for normalization details and a comparison between various transformations) and used raw count data for the other three methods. To make our simulations realistic, we use parameters inferred from a published RNAseq data set on a population of wild baboons (7,67) to perform simulations (Methods and Materials); this baboon data set contains known related individuals and hence invokes the problem of kinship/population structure outlined above.

Our first set of simulations was performed to evaluate the effectiveness of MACAU and the other four methods in controlling for relatedness and population structure. To do so, we simulated expression levels for 10,000 genes in 63 individuals (the sample size from the baboon data set). 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 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).

Figures 1 and S3 show quantile-quantile plots for analyses using MACAU and the other four methods against the null (uniform) expectation, for $h^2 = 0.6$ and $h^2 = 0.3$, respectively. Because genes are heritable and the predictor variable is also correlated with individual relatedness, the resulting *p*-values from the DE analysis are expected to be uniform only for a method that properly controls for individual relatedness and population structure. If a method fails to control for population structure, then the *p*-values would be inflated, resulting in false positives.

Our results show that, because MACAU controls for population structure, 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 (Figures 1 and S3). 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 towards 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 3 DE genes (when $h^2 = 0.3$) and 1 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 population structure well.

Two important factors influence the severity of population stratification 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 highly heritable predictor variable ($h_x^2 = 0.8$), λ_{gc} (when $n = 1{,}000$) 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).

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 1,000 genes were truly DE and 9,000 were non-DE. For the DE genes, simulated effect sizes corresponded to a fixed proportion of variance explained (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 area under the curve (AUC) (as in (29,68,69)). We also examined several key factors that could influence the relative performance of the alternative methods: (1) gene expression heritability (h^2); (2) correlation between the predictor variable x and genetic relatedness (measured by the heritability of x, or h_x^2); (3) variation of the total read counts across samples (measured by the coefficient of variation, or CV); (4) the over-dispersion parameter (σ^2); (5) the effect size (PVE); and (6) 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.

Our results show that MACAU works as well as or outperforms other methods in a wide range of settings (Figures 3, S4-S7), probably because it both models count data directly and controls for relatedness and population structure. In contrast, the Poisson approach consistently fared the worst across all simulation scenarios, presumably because it fails to account for over-dispersion (Figures 3, S4-S7).

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 works similarly or slightly better than MACAU, most likely because they require fewer model 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 relatedness are much more powerful than the methods that do not. Because almost all genes are influenced by cis-eQTLs (41,42) 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 ideal for routine DE analysis.

The second most important factor in relative model performance was the variation of total read counts 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 linear mixed model fails to account for the mean-variance dependence observed in count data, which is in agreement with previous findings (46,70).

The other three 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) (Figure S4) or large sample sizes (Figure S5), and decreases with large over-dispersion σ^2 (Figure S6) or large h_x^2 (Figure S7).

Finally, in addition to the above comparisons, we also included comparisons with edgeR (18) and DESeq (10), two commonly used methods for DE analysis (32,71). Because edgeR and DESeq were originally designed for discrete predictor valuables, we discretized the continuous predictor x into 0/1 based on the median predictor value across individuals. We then re-simulated the data based on the basic parameter setting (*n* = 63, CV = 0.3, h_x^2 = 0, PVE = 0.25, $h^2 =$ 0.3 and $\sigma^2 =$ 0.25), and applied all methods to the simulated data for comparison. Results are shown in Figure S8. Both edgeR and DESeq do not perform well, presumably because the two methods are specifically designed for small sample comparisons and are not flexible enough for analyses of larger data sets.

Real Data Applications

To gain insight beyond simulation, we applied MACAU and the other four methods to two recently published RNAseq data sets. We did not apply edgeR and DESeq here because they are designed for small data sets.

The first data set we considered is a baboon RNAseq study (7) from the Amboseli Baboon Research Project (ABRP) (67). Expression measurements on 12,018 blood-expressed genes were collected for 63 adult baboons (26 females and 37 males), among which some were relatives. Here, we applied MACAU and the four other methods to identify genes with sex-biased expression patterns. Sex-associated genes are known to be enriched on sex chromosomes (72,73), and we use this enrichment as one of the criteria to compare method performance as in (22). Because the same nominal *p*-value from different methods may correspond to different type I errors, we compared methods based on empirical false discovery rate (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 (Methods and Materials).

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 in the X chromosome enrichment evaluation.

The second data set we considered is a RNAseq study on type II diabetes (T2D) collected as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study (47). Here, the data were collected 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 principal components vaguely depict the three municipalities; Figure S9). Two related phenotypes were available to us: 162 individuals with T2D or NGT (normal glucose tolerance) status (i.e., case/control) based on the oral glucose tolerance test (OGTT) and 267 individuals with the quantitative trait fasting glucose level (GL), a biological relevant trait of T2D.

We performed analyses to identify genes associated with T2D status as well as genes associated with GL. 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).

Several lines of evidence support the biological validity of the genes detected by MACAU. First, we performed Gene Ontology (GO) analysis using LRpath (74) on T2D and GL associated genes identified by MACAU as in the FUSION study (47) (Figure S10). The GO analysis results for T2D and GL are consistent with previous studies (47,75) and are also similar to each other, consistent with 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 (47,75). 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, consistent with a prominent role of mTOR pathway in T2D etiology (76-79).

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 (80-82) and T2D status is determined in part by fasting glucose levels. 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 1,000 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 best, when performing the reciprocal analysis (overlap between genes with the strongest GL-association and the top 1,000 genes that have the strongest T2D-association: Figure 4F). Thus, 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 human 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 data sets, whereas the other methods were not (Figures S11 and S12). Finally, none of the methods compared here are sensitive to outliers in the two data sets (Figures S15-S17): even the Poisson model and negative binomial model only identified a few genes with potential outliers as being strongly DE.

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 population structure and independent over-dispersion. Because of its flexible modeling framework, MACAU controls for type I error in the presence of population structure and 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 two recently published RNAseq data sets.

MACAU is particularly well-suited to data sets that contain related individuals or population structure. Several major population genomic resources contain structure of these kinds. For example, the HapMap population (83), the Human Genome Diversity Panel (84), the 1000 Genomes Project in humans (85) as well as the 1001 Genomes Project in Arabidopsis (86) all contain data from multiple populations or related individuals. Several recent large-scale RNAseq projects also collected individuals from genetically differentiated populations (40). 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 (41,87). 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. Because in practice investigators do not have *a priori* knowledge about population stratification or gene expression heritability levels, we recommend MACAU for general transcriptome analysis.

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. As another example, linear mixed models have been proposed as an effective tool to account for cell type heterogeneity (88) or batch effects (89) in microarray data by using the gene expression covariance matrix. MACAU provides a natural avenue for future extension of linear mixed models to control for these confounding effects in sequencing-based studies. Other future extensions include borrowing information across genes to estimate the over-dispersion parameter (10,15,18), or building in a hierarchical structure to model many genes at once, both of which could confer improvements to power.

Currently, MACAU's biggest limitation is computational speed. 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 (Table S1), it can still take close to 20 hours to analyze a data set of the size of the human data we considered here (267 individuals and 21,753 genes). To speed up computation, we have implemented MACAU with multithreading routines to take advantage of modern CPU architecture. Therefore, with moderate computation resources, MACAU can be easily applied to analyze the largest RNAseq data set currently published (*n* = 922: (42)) within a day. However, new algorithms will be needed to use MACAU for data sets that are orders of magnitude larger.

URLs

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

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

This study was supported by start-up funds from the University of Michigan. XZ is also supported by R01HL117626 (PI Abecasis), R21ES024834 (PI Pierce), and a grant from the Foundation for the National Institutes of Health through the Accelerating Medicines Partnership (BOEH15AMP, co-PIs Boehnke and Abecasis). JT is supported by 1R01GM102562 and R21AG049936. LS is supported by U01DK062370 (PI Boehnke). SS is supported by a scholarship from the China Scholarship Council. We thank Matthew Stephens for insight and support on previous versions of MACAU. We thank the Baylor College of Medicine Human Genome Sequencing Center for access to the current version of the baboon genome assembly (Panu 2.0). We thank the FUSION investigators for access to the FUSION expression data.

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. Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.
10. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, **11**, R106.
11. 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.
12. 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.
13. 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.
14. Langmead, B., Hansen, K.D. and Leek, J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol*, **11**, R83.
15. 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.
16. Auer, P.L. and Doerge, R.W. (2011) A Two-Stage Poisson Model for Testing RNA-Seq Data. *Stat Appl Genet Mol*, **10**.
17. 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**, 1--21.
18. 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.

19. 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.
20. 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**.
21. Wu, H., Wang, C. and Wu, Z.J. (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232-243.
22. 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.
23. 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.
24. Hardcastle, T.J. and Kelly, K.A. (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *Bmc Bioinformatics*, **11**, 1.
25. 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.
26. 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.
27. 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.
28. Zwiener, I., Frisch, B. and Binder, H. (2014) Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. *Plos One*, **9**, e85150.
29. Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *Bmc Bioinformatics*, **14**, 1.
30. 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.
31. 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**.
32. Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G. and Owen-Hughes, T. (2015) Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment. *Bioinformatics*.
33. P. McCullagh, J.A.N.F. (1989) *Generalized Linear Models*. Springer US USA.
34. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**, R25.
35. 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**.
36. 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.
37. 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.

38. 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-U422.
39. 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**, 1.
40. 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.
41. 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.
42. 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.
43. 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. *Nature genetics*, **38**, 904-909.
44. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, **44**, 821-U136.
45. Zhou, X. and Stephens, M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods*, **11**, 407-409.
46. 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 Genetics*, **11**, e1005650.
47. 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. *Nature Communications*, DOI: 10.1038/ncomms11764.
48. 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.
49. Scott, S.L. (2011) Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Stat Pap*, **52**, 87-109.
50. Fruhwirth-Schnatter, S. and Fruhwirth, R. (2010) *Data Augmentation and MCMC for Binary and Multinomial Logit Models*. Springer, New York.
51. 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-U894.
52. 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-U110.
53. 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.

54. Goldstein, H. (1991) Nonlinear Multilevel Models, with an Application To Discrete Response Data. *Biometrika*, **78**, 45-51.
55. Breslow, N.E. and Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Models. *J Am Stat Assoc*, **88**, 9-25.
56. 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.
57. Browne, W.J. and Draper, D. (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal*, **1**, 473-513.
58. 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.
59. Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *J Roy Stat Soc a Sta*, **159**, 505-513.
60. Rodriguez, G. and Goldman, N. (2001) Improved estimation procedures for multilevel models with binary response: a case-study. *J Roy Stat Soc a Sta*, **164**, 339-355.
61. 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-Simul C*, **38**, 692-702.
62. Fong, Y.Y., Rue, H. and Wakefield, J. (2010) Bayesian inference for generalized linear mixed models. *Biostatistics*, **11**, 397-412.
63. Smith, A.F.M. and Roberts, G.O. (1993) Bayesian Computation Via the Gibbs Sampler And Related Markov-Chain Monte-Carlo Methods. *J Roy Stat Soc B Met*, **55**, 3-23.
64. Gelman, A. and Shirley, K. (2011) Inference from Simulations and Monitoring Convergence. *Ch Crc Handb Mod Sta*, 163-174.
65. Schwartz, L. (1965) On Bayes procedures. *Zeitschrift f{\"u}r Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **4**, 10--26.
66. Hadfield, J.D. (2010) MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *J Stat Softw*, **33**, 1-22.
67. Alberts, S. and Altmann, J. (2012) In Kappeler, P. M. and Watts, D. P. (eds.), *Long-Term Field Studies of Primates*. Springer Berlin Heidelberg, pp. 261-287.
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. 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.
70. Chen, H.a.W., Chaolong and Conomos, Matthew P and Stilp, Adrienne M and Li, Zilin and Sofer, Tamar and Szpiro, Adam A and Chen, Wei and Brehm, John M and Celed{\'o}n, Juan C and others. (2016) Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *The American Journal of Human Genetics*, **98**, 653--666.
71. 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.
72. 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.

73. 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-Genes Genom Genet*, **4**, 1-10.

74. 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**.

75. Vamsi K Mootha, C.M.L., Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, and Todd R Golub. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267 - 273.

76. 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.

77. 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.

78. Laplante, M. and Sabatini, D.M. (2012) mTOR Signaling in Growth Control and Disease. *Cell*, **149**, 274-293.

79. Zoncu, R., Efeyan, A. and Sabatini, D.M. (2011) mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat Rev Mol Cell Bio*, **12**, 21-35.

80. 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.

81. 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.

82. 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-U132.

83. 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.

84. 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.

85. 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.

86. Weigel, D. and Mott, R. (2009) The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biol*, **10**, 107.

87. 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.

88. 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-U283.

89. 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.

90. 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**, 74.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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. *P Roy Soc B-Biol Sci*, **281**.

96. 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.

97. 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.

98. Churchill, G.A. and Doerge, R.W. (2008) Naive application of permutation testing leads to inflated type I error rates. *Genetics*, **178**, 609-610.

99. Abney, M. (2015) Permutation Testing in the Presence of Polygenic Variation. *Genet Epidemiol*, **39**, 249-258.

100. Zhou, X.B., Lindsay, H. and Robinson, M.D. (2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res*, **42**, e91.

101. 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**.

102. 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.

103. Venables, W.N.a.R., B. D. (2002) *Modern Applied Statistics with S*. Springer, New York.

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

Statistical method	Directly models counts?	Controls for biological covariates?	Controls for population structure?	Software that implements the method
Linear regression	No	Yes	No	R and many others
Linear mixed model	No	Yes	Yes	GEMMA (44), EMMA (102) and FaSTLMM (51)
Poisson model	Yes	Some methods do	No	GLMP (103) and DEGseq (13)
Negative binomial model	Yes	Some methods do	No	edgeR (18), DESeq (10) and GLMNB(103)
Poisson mixed model	Yes	Yes	Yes	MACAU

Figure 1. QQ-plots comparing expected and observed *p*-value distributions generated by different methods for the null simulations with population structure. In each case, 10,000 non-DE genes were simulated with *n* = 63, CV = 0.3, σ^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 population structure. λ_{gc} is the genomic control factor.

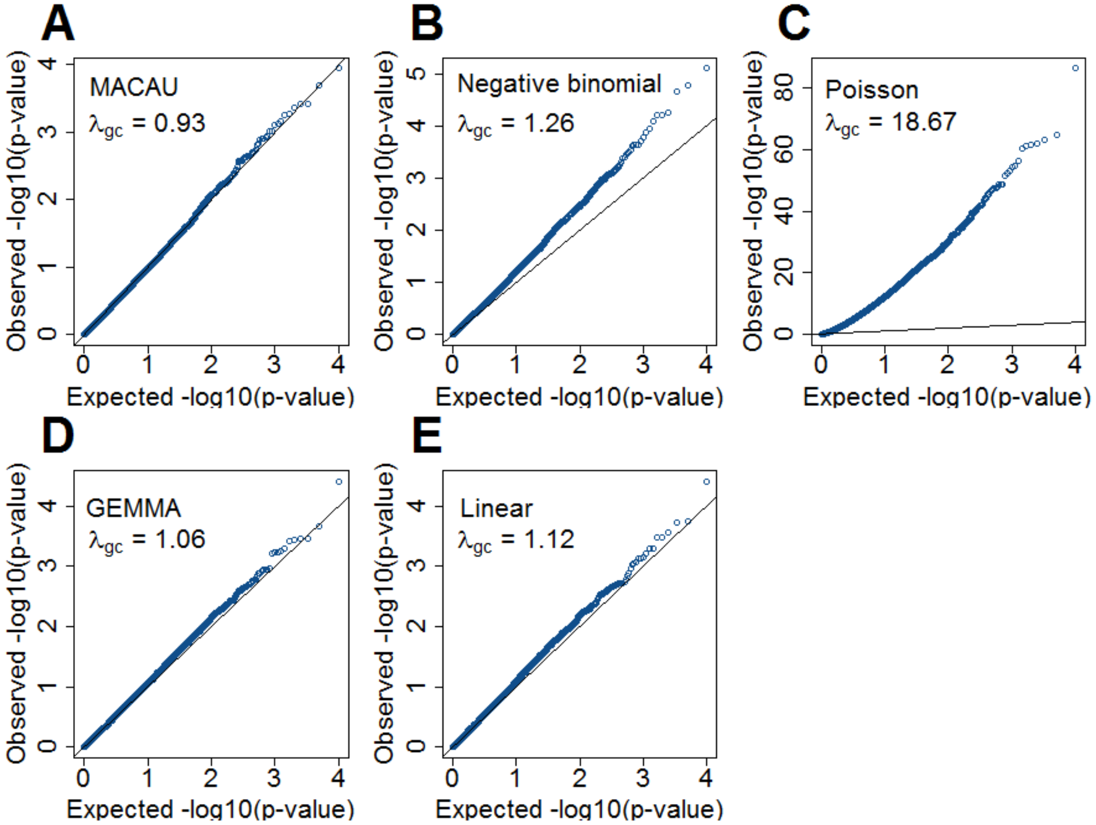


Figure 2. Comparison of the genomic control factor λ_{gc} from different methods for the null simulations with population structure. 10,000 null genes were simulated with CV = 0.3, σ^2 = 0.25, h^2 = 0.6, and (A) $h_x^2 = 0$; (B) $h_x^2 = 0.4$; or (C) $h_x^2 = 0.8$. λ_{gc} (y-axis) changes with sample 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 population structure across a range of settings. λ_{gc} from Poisson exceeds 10 in all settings and is thus not shown.

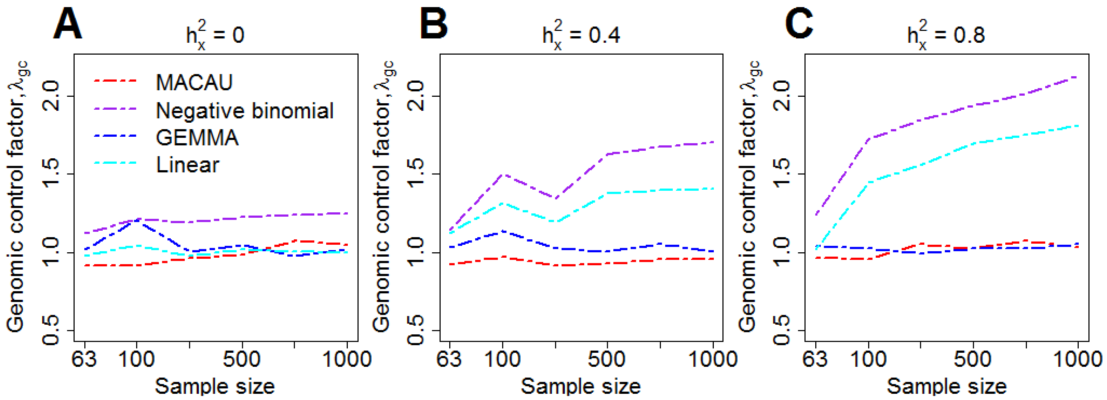


Figure 3. MACAU exhibits increased power to detect true positive DE genes across a range of simulation settings. Area under the curve (AUC) is shown as a measure of performance for MACAU (red), Negative binomial (purple), Poisson (green), GEMMA (blue), and Linear (cyan). Each simulation setting consists 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 = 0.0, PVE = 0.25, σ^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 population structure (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. Insets in the two figures show the rank of different methods, where the top row represents the highest rank.

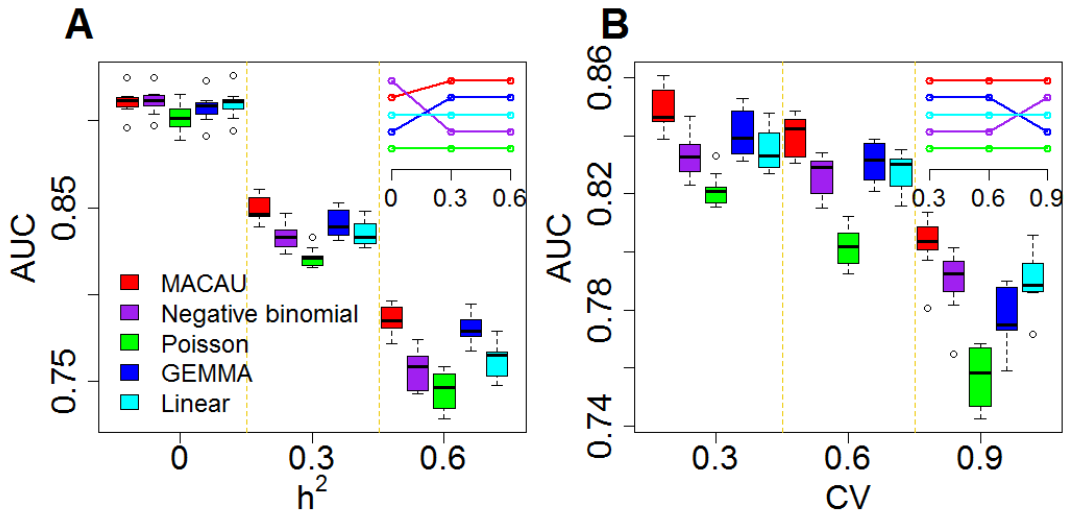
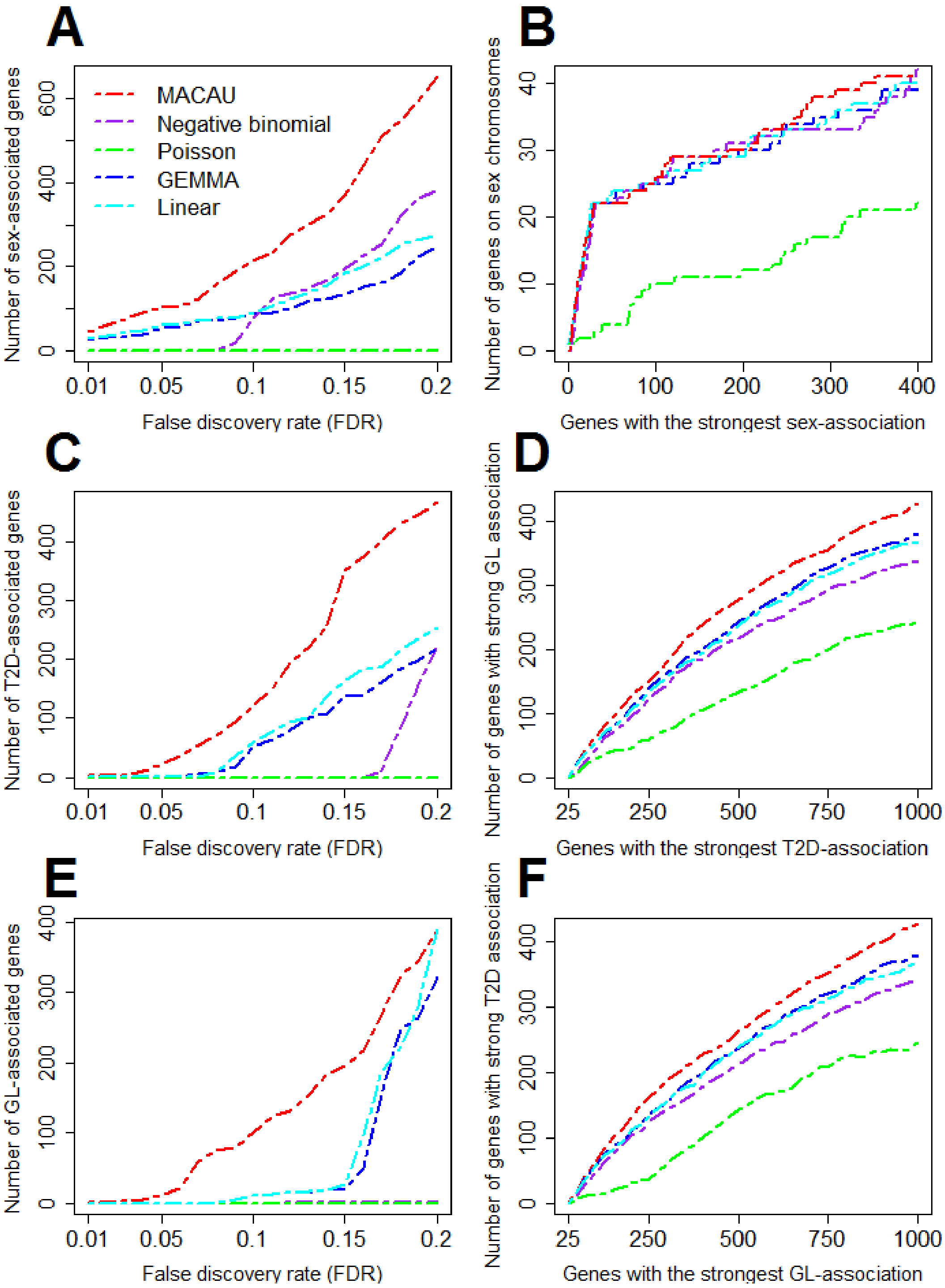


Figure 4. MACAU identifies more differentially expressed genes than other methods in the baboon (panel A and B) and human (panel C, D, E, and F) data sets. Methods for comparison include MACAU (red), Negative binomial (purple), Poisson (green), GEMMA (blue), and Linear (cyan). (A) shows the number of 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.


A
MACAU
Negative binomial
Poisson
GEMMA
Linear
Number of sex-associated genes
0
200
400
600
0.01
0.05
0.1
0.15
0.2
False discovery rate (FDR)
B
Number of genes on sex chromosomes
0
10
20
30
40
0
100
200
300
400
Genes with the strongest sex-association
C
Number of T2D-associated genes
0
100
200
300
400
0.01
0.05
0.1
0.15
0.2
False discovery rate (FDR)
D
Number of genes with strong GL association
0
100
200
300
400
25
250
500
750
1000
Genes with the strongest T2D-association
E
Number of GL-associated genes
0
100
200
300
400
0.01
0.05
0.1
0.15
0.2
False discovery rate (FDR)
F
Number of genes with strong T2D association
0
100
200
300
400
25
250
500
750
1000
Genes with the strongest GL-association