

An empirical bayesian approach for testing gene expression fold change and its application in detecting global dosage effects

Zhenxing Guo^{1,†}, Ying Cui^{1,†}, Xiaowen Shi², James A. Birchler², Igor Albizua³,
Stephanie L. Sherman³, Zhaohui S. Qin^{1,*} and Tieming Ji^{4,5,*}

¹Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA, ²Division of Biological Sciences, University of Missouri at Columbia, Columbia, MO 65211, USA, ³Department of Human Genetics, Emory University, Atlanta, GA 30322, USA, ⁴Department of Statistics, University of Missouri at Columbia, Columbia, MO 65211, USA and ⁵Roche Diagnostics, Santa Clara, CA 95050, USA

Received March 08, 2020; Revised July 27, 2020; Editorial Decision August 21, 2020; Accepted August 29, 2020

ABSTRACT

We are motivated by biological studies intended to understand global gene expression fold change. Biologists have generally adopted a fixed cutoff to determine the significance of fold changes in gene expression studies (e.g. by using an observed fold change equal to two as a fixed threshold). Scientists can also use a *t*-test or a modified differential expression test to assess the significance of fold changes. However, these methods either fail to take advantage of the high dimensionality of gene expression data or fail to test fold change directly. Our research develops a new empirical Bayesian approach to substantially improve the power and accuracy of fold-change detection. Specifically, we more accurately estimate gene-wise error variation in the log of fold change. We then adopt a *t*-test with adjusted degrees of freedom for significance assessment. We apply our method to a dosage study in Arabidopsis and a Down syndrome study in humans to illustrate the utility of our approach. We also present a simulation study based on real datasets to demonstrate the accuracy of our method relative to error variance estimation and power in fold-change detection. Our developed R package with a detailed user manual is publicly available on GitHub at <https://github.com/cuiyingbeicheng/Foldseq>.

INTRODUCTION

Detecting genes that express differently between case and control conditions is a fundamental problem in functional

genomic studies. There are two different ways to quantify differential expressions, either by mean gene expression difference or by fold change. These two criteria lead to two different ranked lists of significant genes. Top ranked genes selected by the expression difference method often have higher mean expression levels but relatively smaller fold change, and the opposite is true when using the fold change criterion. The vast majority of available statistical methods and software packages test significant gene expression difference between treatment groups. However, there are scenarios where testing expression fold change is a more natural question to inquire. This calls for development of sophisticated data analysis methods to satisfy the need.

Aneuploidy, or the presence of an abnormal number of individual chromosomes in the genome, is a well-known phenomenon found in many different organisms. Aneuploidy originates during cell division when sister chromosomes do not separate properly between two cells. An extra or missing copy of a chromosome is a common cause of genetic disorders, including human birth defects such as Down syndrome (DS) and Edwards syndrome. DS, otherwise known as trisomy 21, arises from an extra copy of chromosome 21. English physician John Langdon Down (1) first documented the condition in 1866, hence its name 'Down syndrome.' The additional genetic materials from chromosome 21 is thought to alter the course of normal development and thus lead to characteristics associated with DS. Similarly, Edwards syndrome (also known as trisomy 18) features an extra copy of chromosome 18 due to errors in cell division, known as meiotic nondisjunction. Different from DS, the developmental problems resulting from trisomy 18 pose greater threats to health in the early months and years of life; among babies with the condition who are carried to term, approximately only half will be born alive.

*To whom correspondence should be addressed. Tel: +1 573 882 4075; Fax: +1 573 884 5524; Email: tjinmsuedu@gmail.com
Correspondence may also be addressed to Zhaohui S. Qin. Tel: +1 404 712 9576; Fax: +1 404 727 1370; Email: zhaohui.qin@emory.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Aneuploidy has long been known to produce severe phenotypic consequences. Similar phenomena have been observed and documented in eukaryotes, including yeast, protozoa, vertebrates and especially the plant kingdom (2–7); more specifically, the addition or subtraction of a single chromosome in the entire set is more detrimental than altering the dosage of the complete complement. To better understand global gene expression differences in aneuploidy, a comprehensive analysis of such differences for all five trisomies and in diploids, triploids and tetraploids of *Arabidopsis thaliana* was conducted (8). Results indicated that, in general, gene expression on the varied chromosome ranged from compensation to a dosage effect, whereas genes from the remainder of the genome exhibited anywhere from no effect to reduced expression approaching the inverse level of chromosomal imbalance. In addition, gene functional analysis indicated that ribosomal, proteasomal and gene body methylated genes were less modulated compared with all gene classes, whereas transcription factors, signal transduction components and organelle-targeted protein genes were more tightly inversely affected (8). These findings suggest that the addition or subtraction of a single chromosome will invoke global gene expression changes, including in the varied chromosome and the remainder of the genome. Such altered regulatory stoichiometry is a major contributor to genetic imbalance.

An important task when analyzing aneuploid data compared to normal diploid data is to identify gene groups (both *cis*- and *trans*-genes) that exhibit a dosage effect, compensation effect, or inverse dosage effect. Statistically, this task requires testing the fold change of gene expression in comparison between aneuploidy and diploid data. At present, no sophisticated data analysis methods or ready-to-use software package is available for detecting genome-wide fold changes; researchers typically use computational tools designed to detect differentially expressed (DE) genes. For example, in a recent trisomy study on *Arabidopsis* (8) and transcriptomic analysis of DS (9), edgeR and DESeq (10–12) were used to identify genes with an expression level greater or less than normal; however, this strategy is sub-optimal because edgeR and DESeq are intended to test gene expression differences rather than to test fold-change statistics directly. In this paper, we demonstrate that testing fold change directly can identify more fold change-relevant genes than testing expression differences in search of fold change.

Specifically, we propose a hierarchical model on the log of fold change between two comparison groups. In an aneuploid experiment or observational study, the comparison groups include an aneuploid group (disease group) and diploid group (control/healthy group). For the constructed hierarchical model, we put an inverse gamma prior on the error variance for each gene. An empirical Bayesian approach is adopted to produce gene-specific error variance estimation, which generates more accurate estimates than currently available approaches as information is borrowed across observations from all genes. We then use the improved error variance estimates to test null hypotheses of fold change. We also construct a *t*-type test on the log of fold change statistics using our improved error variance estimate and its corresponding adjusted degrees of freedom. To

evaluate performance, we adopt public RNA-seq datasets along with simulation studies based on real datasets. We demonstrate that, compared to the usual *t*-test, edgeR and DESeq methods, error variance estimation is substantially improved by our new approach. The power of detecting true fold change while controlling the false discovery rate (FDR) is also greatly enhanced. We call our new method the Foldseq approach, and our developed R package is publicly available on GitHub at <https://github.com/cuiyingbeicheng/Foldseq>.

MATERIALS AND METHODS

Data model

To ground our method, we will first introduce relevant notations. Let X_{ij}^g denote the normalized gene expression of gene g of replicate j in treatment group i , where $i = 1$ for the control group and $i = 2$ for the treatment group, $j = 1, \dots, n_i$ and $g = 1, \dots, G$. Our proposed method can be easily extended to a multiple-group comparison scenario. To illustrate our approach, we use a two-group comparison as an example here.

Suppose the control group and the treatment group have the population mean expression μ_1^g and μ_2^g , respectively. The objective of our study is to detect whether the ratio of μ_2^g and μ_1^g (i.e. fold change) is within or outside a region of interest, such as

$$\mu_2^g/\mu_1^g \geq d_0, \text{ or } \mu_2^g/\mu_1^g \leq d_0, \text{ or } d_1 \leq \mu_2^g/\mu_1^g \leq d_2, \quad (1)$$

among others, with strong statistical evidence.

As the mean gene expression ratio is of interest for our purposes, we define variable $Y_j^g = \log((X_{2j}^g + 0.5)/(\bar{X}_1^g + 0.5))$, where $\bar{X}_1^g = \sum_{j=1}^{n_1} X_{1j}^g/n_1$ and $g = 1, \dots, G$. This logarithmic transformation permits the fold-change variable to be modeled on the entire real space. Typically, the log of fold change uses base 2. We retain this conventional approach and thus use base 2 in our method. The 0.5's in the numerator and denominator are intended to avoid extreme observations when taking the log transformation.

We model that $Y_j^g \sim N(c_g, \sigma_g^2)$, where c_g and σ_g^2 denote the gene-specific mean and variance of the log fold change, respectively. Often there are only a few replicates (i.e. two or three) in gene expression studies due to the expenses associated with biological replicate acquisition and sequencing experiments. A usual *t*-test that is only based on Y_j^g , c_g , and σ_g^2 is unreliable because gene-wise estimation of the variance parameter σ_g^2 is not accurate when using only a few replicates. To improve the estimation of σ_g^2 , we adopt a Bayesian approach to borrow information across genes. Specifically, we assume $\sigma_g^2 \sim \text{IG}(\alpha_0, \beta_0)$, which is a commonly used conjugate prior for variance parameters.

Let \mathbf{y}^g denote all observations from gene g ; that is, $\mathbf{y}^g = (Y_1^g, \dots, Y_{n_2}^g)^T$. We derive that the posterior distribution of σ_g^2 given \mathbf{y}^g is

$$\sigma_g^2 | \mathbf{y}^g \sim \text{IG} \left(\alpha = \alpha_0 + \frac{n_2}{2}, \beta = \beta_0 + \frac{1}{2} \sum_{j=1}^{n_2} (Y_j^g - c_g)^2 \right). \quad (2)$$

We propose using the posterior mean of σ_g^2 , $\hat{\sigma}_g^2$, as the estimate of σ_g^2 , where

$$\hat{\sigma}_g^2 = E(\sigma_g^2 | \mathcal{Y}^g) = \frac{\beta_0 + \sum_{j=1}^{n_2} (Y_j^g - c_g)^2 / 2}{\alpha_0 + n_2 / 2 - 1}. \quad (3)$$

To test our hypotheses of interest, such as the null hypotheses in (1), we construct the following t -statistic

$$t_g = \frac{\bar{Y}^g - \log d_0}{\sqrt{\hat{\sigma}_g^2 / n_2}} = \frac{\sum_{j=1}^{n_2} Y_j^g / n_2 - \log d_0}{\sqrt{\hat{\sigma}_g^2 / n_2}}, \quad (4)$$

where d_0 can be replaced by d_1 and d_2 to test regions with two-sided boundaries. At last, we use Storey's method (13) to control the FDR at a desired level.

Hyperparameter estimation

To implement our method, we still need to estimate hyperparameters α_0 and β_0 . A number of methods have been proposed to infer the gene-wise hyperparameters using Empirical Bayes (EB) methods. For microarray data, Newton *et al.* proposed to use Gamma and log-normal distributions to model microarray expression data and applied EB models to estimate hyperparameters (14). Smyth popularized the hierarchical model proposed by Lönnstedt and Speed (15) and applied EB method to estimate hyperparameters. More EB methods have been developed for analyzing sequencing-based transcriptomic data. In particular, Hardcastle and Kelly developed baySeq, in which the authors used EB method to determine the prior distributions (16). Robinson *et al.* developed edgeR, where an EB method was to moderate the degree of overdispersion across transcripts (10). Wu *et al.* presented an improved EB shrinkage estimate of dispersion parameters and demonstrated improved DE detection (17). Love *et al.* developed DESeq2 (18), where an EB method was used for dispersion estimation.

Inspired by the method in *limma* (19), we propose the following estimation approach. Specifically, we compute the sample estimate of gene-wise variance, denoted by s_g^2 , where $s_g^2 = \sum_{j=1}^{n_2} (Y_j^g - \bar{Y}^g)^2 / (n_2 - 1)$. We can show that $s_g^2 | \alpha_0, \beta_0 \sim \beta_0 / \alpha_0 F(n_2 - 1, 2\alpha_0)$, which is a scaled F-distribution. Let z_g denote $\log s_g^2$ with a natural base, such that z_g is distributed as a constant plus Fisher's z distribution as demonstrated in (19). The distribution of z_g is roughly normal with

$$E(z_g) = \log\left(\frac{\beta_0}{\alpha_0}\right) + \psi\left(\frac{n_2 - 1}{2}\right) - \psi(\alpha_0) + \log\left(\frac{2\alpha_0}{n_2 - 1}\right), \quad \text{and} \quad (5)$$

$$\text{Var}(z_g) = \psi'\left(\frac{n_2 - 1}{2}\right) + \psi'(\alpha_0), \quad (6)$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ denote the digamma and trigamma functions, respectively.

Let $e_g = z_g - \psi((n_2 - 1)/2) + \log((n_2 - 1)/2)$, in which case we have $E(e_g) = \log(\beta_0/\alpha_0) - \psi((n_2 - 1)/2) + \log((n_2 - 1)/2)$, and $E\{(e_g - \bar{e})^2 G / (G - 1) - \psi'((n_2 -$

$1)/2)\} \approx \psi'(\alpha_0)$ where $\bar{e} = \sum e_g / G$. We therefore estimate α_0 by solving $\psi'(\alpha_0) = \text{mean}\{(e_g - \bar{e})^2 G / (G - 1) - \psi'((n_2 - 1)/2)\}$. To solve this equation, we use Newton's iteration. Specifically, let a denote $\text{mean}\{(e_g - \bar{e})^2 G / (G - 1) - \psi'((n_2 - 1)/2)\}$. In the initial iteration, we set $\alpha_0^{(0)} = 0.5 + 1/a$. In the k th iteration, we let $\alpha^{(k+1)} = \alpha^{(k)} + \psi'(\alpha^{(k)})\{1 - \psi'(\alpha^{(k)})/a\} / \psi''(\alpha^{(k)})$. The iteration stops once $|\alpha^{(k+1)} - \alpha^{(k)}| / \alpha^{(k)} < \epsilon$, where ϵ is a small positive number. After convergence, the estimate of α_0 is denoted by $\hat{\alpha}_0$. We estimate $\hat{\beta}_0 = \hat{\alpha}_0 \exp\{\bar{e} + \psi(\hat{\alpha}_0) - \log(\hat{\alpha}_0)\}$.

RESULTS

Simulation study

Simulation setting. To ensure that our simulation results are reproducible in real data analysis, we simulated data based on real datasets in (8); we simulated 18 000 genes in total. Each gene has n_1 replicates in the control group and n_2 replicates in the treatment group. We let $n_1 = 3$ based on the number of replicates in (8) for the dosage effect analysis in *Arabidopsis*. To assess the parameter estimation and detection power of our proposed method, we let n_2 be 3, 5, 7 and 9, respectively. To simulate observations for the control group, we randomly selected 18,000 genes from the control group of trisomy 1 in the *Arabidopsis* analysis and directly used their observations as X_{11} , X_{12} and X_{13} . We first simulated the log fold change (with base 2); that is, Y_j^g 's, and then simulate observations for the treatment group. Specifically, we follow the procedure as described below:

- (i) Set α_0 and β_0 based on *Arabidopsis* trisomy 1 dataset. Then simulate $\sigma_g^2 \stackrel{iid}{\sim} \text{IG}(\alpha_0, \beta_0)$.
- (ii) Randomly select certain proportion p^* (e.g. $p^* = 30\%$) genes to have significant fold change (i.e. DE genes) of c_0 with $c_g = \log_2(c_0)$. For the other $(1-p)$ genes, we set them to have no change, i.e. $c_g = \log_2(1) = 0$. Simulate $Y_j^g \stackrel{iid}{\sim} \text{N}(c_g, \sigma_g^2)$.
- (iii) Calculate the observations in the treatment group with $X_{2j}^g = (\bar{X}_1^g + 0.5) \cdot 2^{Y_j^g} - 0.5$.

We consider three different parameter settings for α and β , respectively, with $\alpha = 3, 4, 5$ and $\beta = 0.15, 0.18, 0.20$ and three choices for the magnitude of fold change c_0 : 1.3, 1.5 and 2.0. For each of these 27 settings, we repeated 100 times.

Simulation results. We compared our proposed Foldseq method with five competing methods: (i) the usual t -test without a Bayesian shrinkage prior; (ii) the edgeR method; (iii) the DESeq2 method; (iv) the voom method (20); and (v) a two-group EB t -test method, which modifies the one group EB t -test in our proposed Foldseq method. By comparing our method to the naive t -test, we examined the power gain of using a Bayesian prior. We chose the edgeR, DESeq2 and voom algorithms for comparison because they are the most commonly used methods for testing DE genes.

Figure 1 shows the absolute biases of α_0 estimates (left panel) and absolute biases of β_0 (right panel); specifically, $|\hat{\alpha}_{0,k} - \alpha_0|$ and $|\hat{\beta}_{0,k} - \beta_0|$, where $k = 1, \dots, K$ and $K = 100$

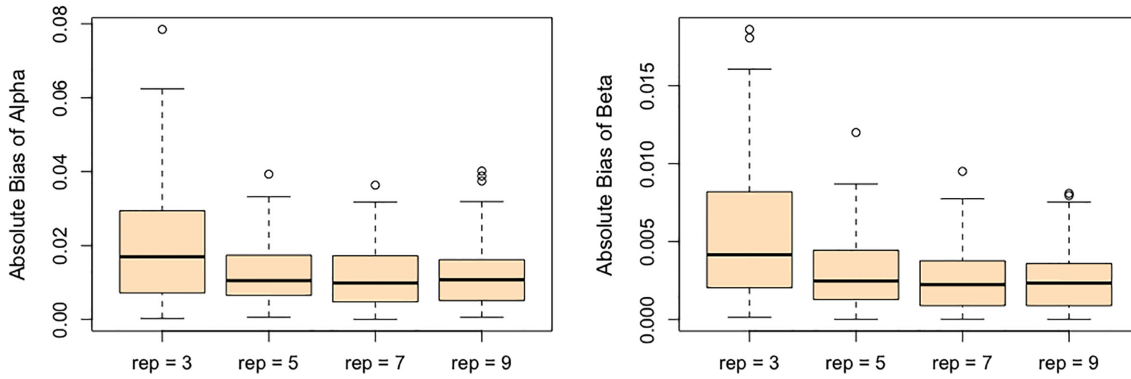


Figure 1. Absolute biases of estimates with 3, 5, 7 and 9 treatment replicates, respectively. Left panel shows absolute biases of estimates of α_0 ; right panel shows absolute biases of estimates of β_0 .

Table 1. Average bias and MSE across 100 simulations

	True	rep = 3		rep = 5		rep = 7		rep = 9	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
α	1.221	0.020	0.001	0.013	0.000	0.012	0.000	0.012	0.000
β	0.189	0.005	0.000	0.003	0.000	0.003	0.000	0.003	0.000

index the 100-round simulation. Figure 1 suggests that with $n_1 = n_2 = 3$, the absolute biases of α_0 and β_0 estimates are small. When $n_1 = 3$ and n_2 increases to 5, 7 and 9, respectively, the absolute biases of the estimates of α_0 and β_0 decline as the sample size increases. Table 1 summarizes the averaged bias ($\sum_{k=1}^K (\hat{\alpha}_{0,k} - \alpha_0) / K$) and mean squared error (MSE) ($\sum_{k=1}^K (\hat{\alpha}_{0,k} - \alpha_0)^2 / K$) across 100 simulations. The results in Table 1 suggest satisfactory and stable estimation results across repeated simulations.

We next compared the performance of error variance estimate, $\hat{\sigma}_g^2$'s, of Foldseq, with the usual sample estimates that are used in t -tests, s_g^2 's. We set $n_1 = n_2 = 3$; Figure 2 plots the MSEs, variances and biases of s_g^2 's and $\hat{\sigma}_g^2$'s with respect to the true σ_g^2 's, which shows that our variance estimates $\hat{\sigma}_g^2$'s are closer to the truth on average than sample estimates obtained without Bayesian shrinkage.

Additionally, we compared the power of fold-change detection between Foldseq and the five competing methods. To be precise, the null hypothesis considered by edgeR, DESeq2 and voom methods is $\mu_1^g - \mu_2^g = 0$; whereas for t -test, the two group test and Foldseq, the hypothesis is $\mu_2^g / \mu_1^g = 1$. The two null hypotheses are equivalent, but the former focuses on the mean difference whereas the later focuses on the mean ratio (i.e. fold change). Thus, we expect the tests using different statistics (either difference or ratio) to exhibit different power, and the one using the observed fold-change statistics to detect fold change will have higher power than methods using mean-difference statistics. Figure 3 summarizes mean AUCs ($n_1 = n_2 = 3$) of all these methods in comparison across 100-round simulation for each simulation setting. As expected, Foldseq demonstrated a larger AUC compared to other methods, suggesting improvement in the power of fold-change detection as well as better gene ranking.

Furthermore, we examined the top 5000 most significant genes declared by each method. Averaging across 100-round simulation, 82.6% of these genes selected by Foldseq were true positives. The proportion of true positives were 58.1, 78.0, 63.7 and 76.1%, respectively, for DESeq2, usual t -test, voom and the two group test. Figure 4 shows a Venn diagram illustrating the overlapping pattern among the top 5000 significant genes selected by each method in four-round simulation when $\alpha = 3$, $\beta = 0.18$ and true positives have fold change 1.3. The Venn diagram suggests that our proposed Foldseq method has the most number of overlapping genes than any other method in comparison.

Finally, we examined the true FDRs produced by these methods when controlling empirical FDRs at 0.2, 0.1, 0.05 and 0.01, respectively. The results for simulation setting with $\alpha = 3$, $\beta = 0.18$ and fold change of 1.3 are presented in Table 2. Table 2 indicates that only Foldseq, the usual t -test and DESeq2 control empirical FDRs under true FDRs. In addition, the usual t -test is more conservative than both Foldseq and DESeq2. The edgeR, voom and two group methods are liberal under most FDR thresholds.

Analysis of dosage effects and dosage compensation in *Arabidopsis thaliana*

Changes in the dosage of part of the genome have long been known to produce much more severe phenotypic consequences than changes in the number of whole genomes (2–4). Therefore, a simple assumption emerged that this phenotypic effect resulted from varied genes showing a dosage effect; in other words, the gene expression fold change matches the gene copy number change for genes on the varied chromosome. However, other evidence in maize and *Drosophila* has revealed the presence of global genome-wide cascading modulations (21–25), presumably because transcription factors and signal transduction components are dosage-sensitive (26–30) and their targets would be modulated regardless of the chromosomal locations of the latter.

To more comprehensively understand the molecular basis of genetic imbalance characterized by aneuploidies, researchers (8) gathered a complete set of primary trisomies of *Arabidopsis* and compared their gene expression fold changes with the normal diploid. Specifically, mature leaf tissues from all five trisomies and the normal diploid of *A.*

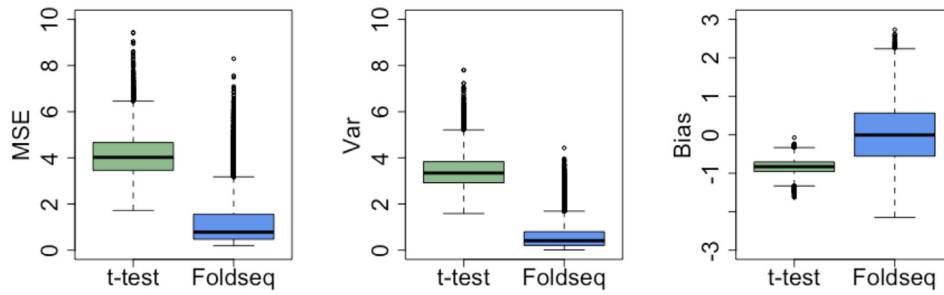


Figure 2. MSE, variance and bias for naive variance estimator and Bayesian estimator. In each plot, the left box plot shows the result for the naive estimator, and the right box plot shows the result for the Bayesian estimator.

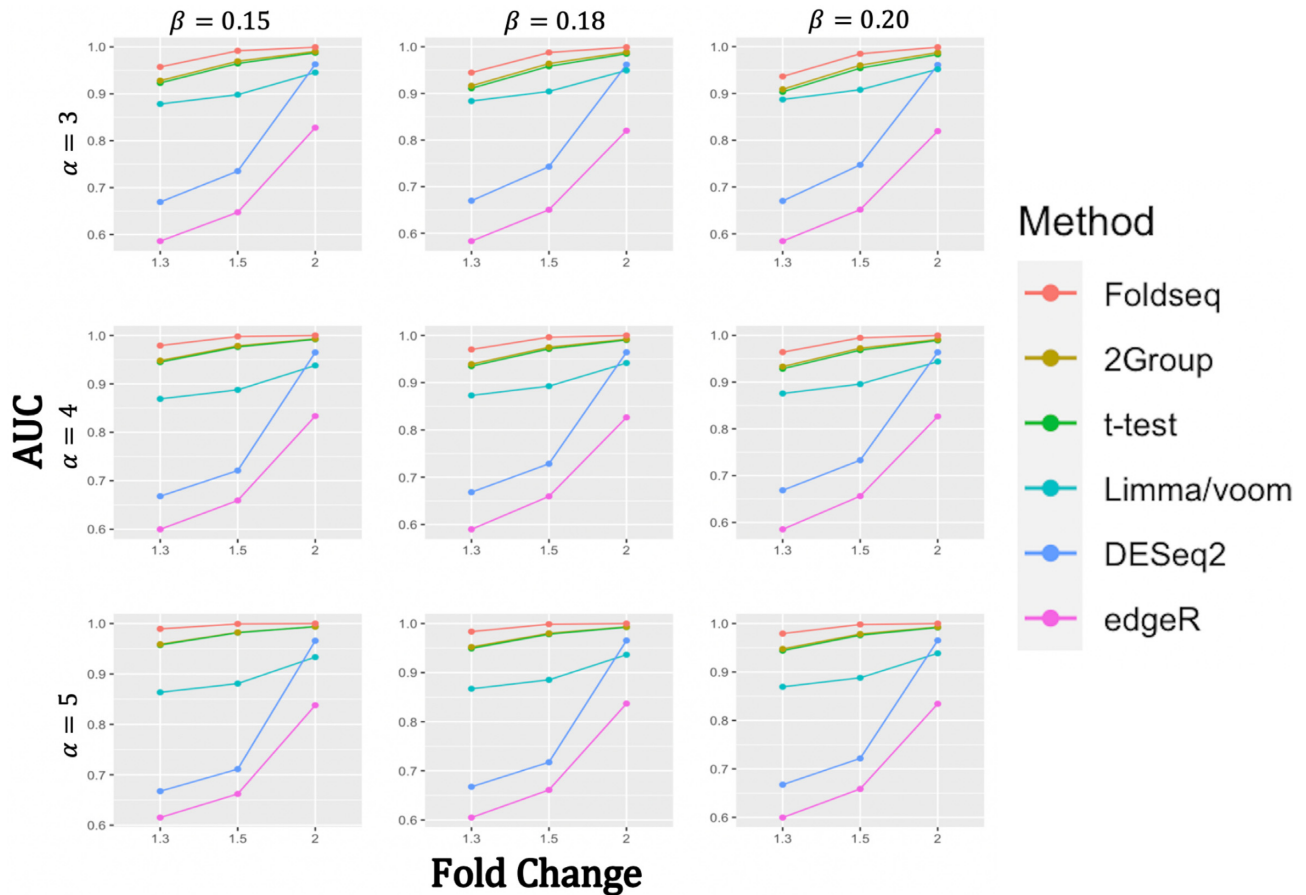


Figure 3. Mean AUC across 100 simulations for our proposed Foldseq method, the two group test, the usual *t*-test, voom, DESeq2 and edgeR for each simulation setting.

thaliana were collected. There were two biological replicates for each trisomy and three biological replicates for the normal diploid. RNA-seq experiments were conducted for each replicate to measure gene expressions. Raw gene expression data were then normalized using an External RNA Controls Consortium (ERCC) spike-in followed by polyA RNA isolation to detect any transacting effects of aneuploidy on global gene expression (31).

We retrieved normalized gene expression data published in (8) and reanalyzed them using our new statistical method. Figure 5 plots the sample averages of fold changes, indicated by $\bar{X}_2^g / \bar{X}_1^g$'s across biological replicates, respectively,

for *cis*- and *trans*-genes for each trisomy. Figure 5 presents a wide spread in gene expression ratio distributions comparing trisomies with the normal diploid. We further examined the ratio distributions for two gene categories: genes present on the varied chromosome (*cis*) and genes in the remainder of the genome (*trans*). For each trisomy, the ratio distribution of the *cis*-genes suggests a mode between 1.0 (dosage compensation) and 1.5 (dosage effect) with a wide spread that extends above and below these levels. For each trisomy, the ratio distribution of the *trans*-genes suggests a mode between 2/3 (inverse dosage effect) and 1.0 (no modulation), also with a wide spread. The fact that the *cis* and *trans* ef-

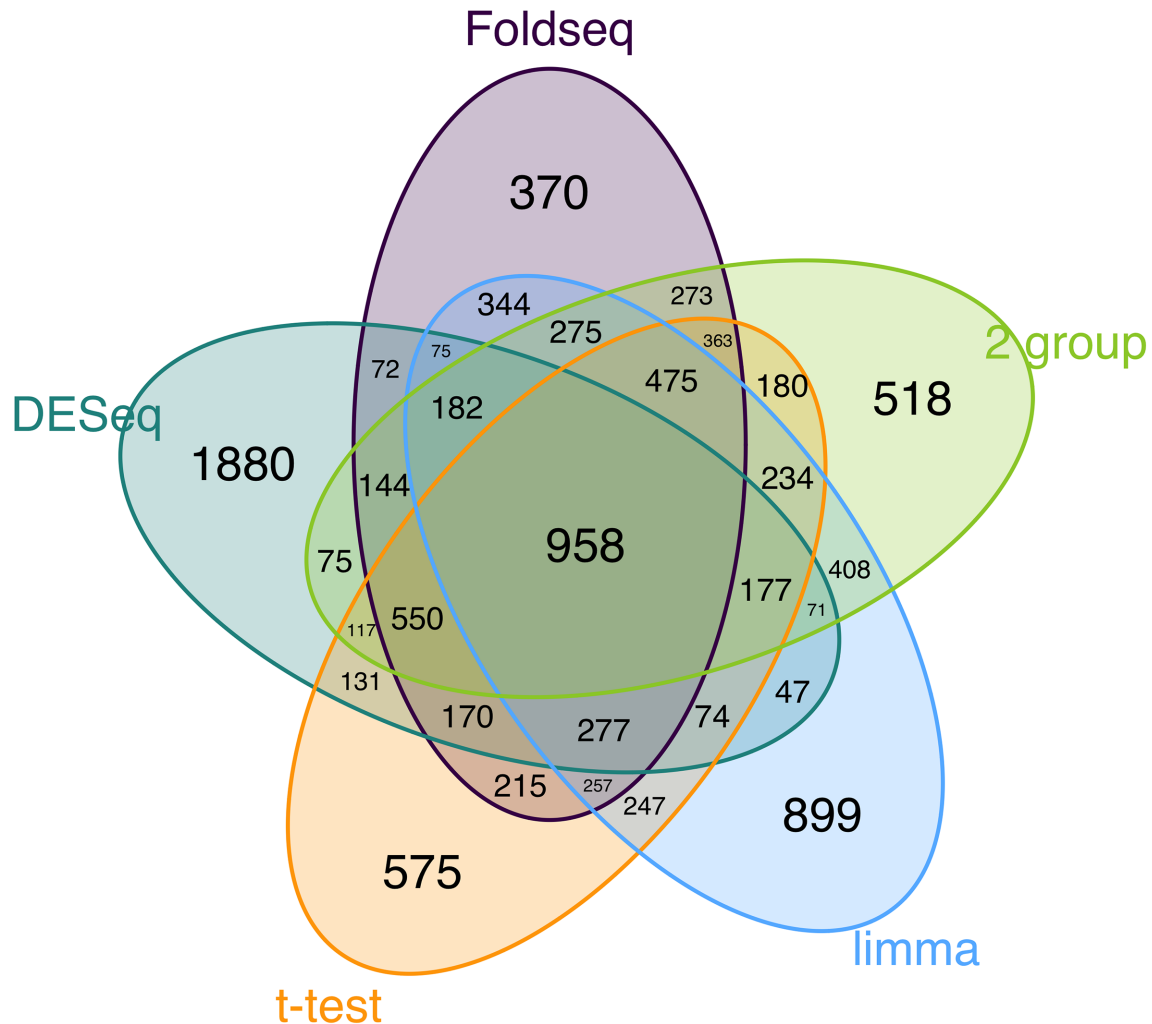


Figure 4. Venn diagram showing 4 times of simulation results of top 5000 most significant genes selected selected by Foldseq, DESeq2, *t*-test, limma/voom and the two group test where $\alpha = 3$, $\beta = 0.18$ and fold change of 1.3.

Table 2. True FDRs when controlling empirical FDRs at 0.2, 0.1, 0.05 and 0.01 for the simulation setting with $\alpha = 3$, $\beta = 0.18$ and fold change as 1.3 for true positive genes

FDR	Foldseq	2Group	<i>t</i> -test	voom	DESeq2	edgeR
0.2	0.020	0.228	0.000	0.532	0.030	0.065
0.1	0.010	0.176	0.000	0.431	0.010	0.050
0.05	0.010	0.150	0.000	0.345	0.010	0.030
0.01	0.000	0.126	0.000	0.000	0.010	0.020

fects are modulated coordinately to some extent suggests a common influence of the aneuploid state on gene expression on the varied and unvaried chromosomes.

To statistically test the above conclusions, we applied our method to these trisomies. We estimated hyperparameters using the method illustrated in the ‘Materials and Methods’ section. For instance, to compare trisomy 1 and the diploid, our estimate of α_0 is 1.22 and of β_0 is 0.19. Then, for each *cis*-gene, we tested $H_0 : \mu_2^g/\mu_1^g = 1.5$. We controlled the FDR at the 0.05 level. By rejecting the null hypothesis,

we further clustered genes with $\mu_2^g/\mu_1^g > 1.5$ and genes with $\mu_2^g/\mu_1^g < 1.5$. For each *trans*-gene, we tested $H_0 : \mu_2^g/\mu_1^g = 1$. By rejecting the null hypothesis, depending on the fold-change direction, we also clustered genes with $\mu_2^g/\mu_1^g > 1$ and genes with $\mu_2^g/\mu_1^g < 1$. We summarized the number of genes in each category using our inference procedure for all five trisomies compared to the diploid; findings appear in Table 3.

Our test results substantiate the observations in (8) and provide detailed numerical evidence of the global gene expression effects of the additional chromosome in each trisomy. For instance, for *cis*-genes in trisomy 1, we found 2736 genes with a fold change < 1.5 and only 50 genes with a fold change > 1.5 with strong statistical support. This pattern reinforces the observations that the *cis*-genes’ distribution has a median between a dosage effect (1.5 fold change) and dosage compensation (no fold change). Our finding also suggests that numerically, $\sim 45\%$ of the *cis*-genes $((2736 + 50)/(2736 + 3455 + 50))$ and approximately one-fifth of the *trans*-genes are being modulated in trisomy 1. Similar conclusions can be drawn for all these five tri-

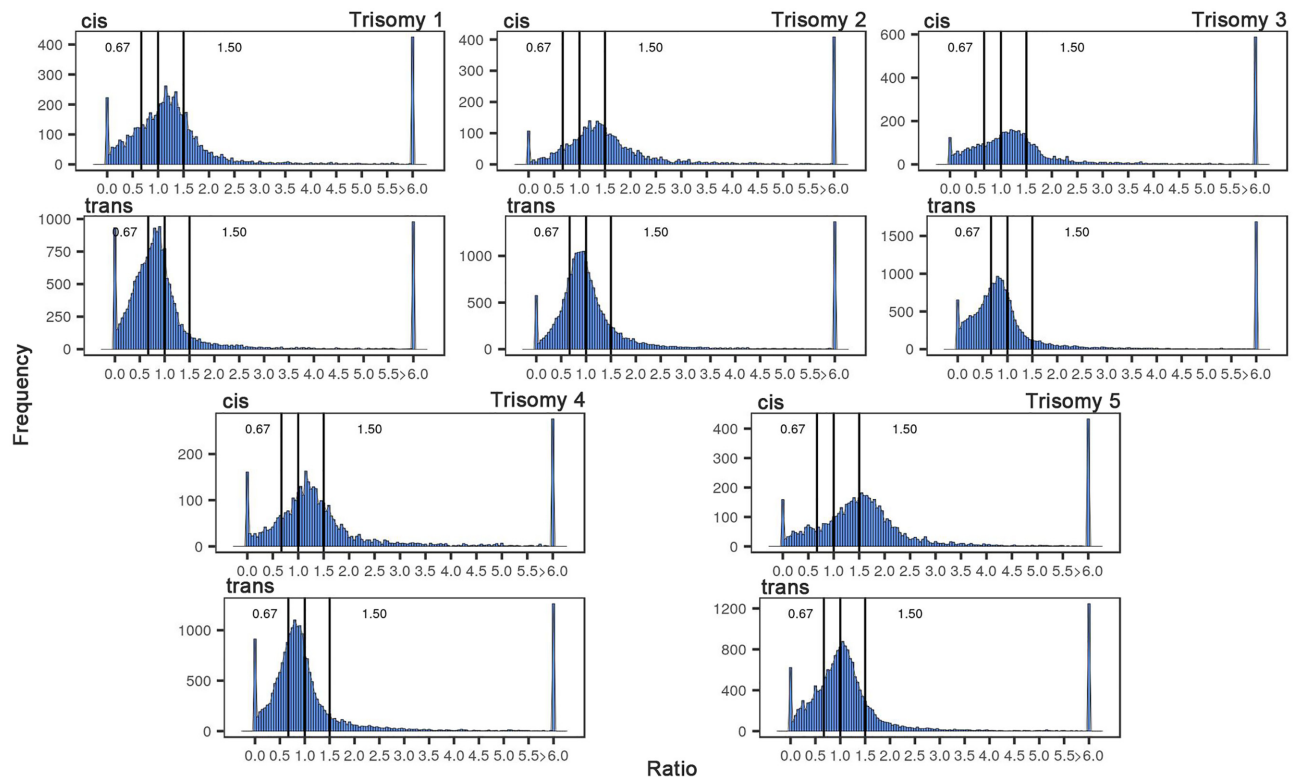


Figure 5. Observed sample average fold-change distributions of gene expression in each trisomy compared with diploid. A ratio of 1.50 represents a gene-dosage effect in *cis*, whereas 1.00 represents dosage compensation. A ratio of 0.67 represents the inverse ratio of gene expression in *trans*. Ratio values are demarcated with labeled vertical lines at 0.67, 1.00 and 1.50.

Table 3. Fold-change detection results for *Arabidopsis* trisomies

Trisomy	<i>cis</i> -genes			<i>trans</i> -genes		
	$\mu_2^g/\mu_1^g < 1.5$	$\mu_2^g/\mu_1^g = 1.5$	$\mu_2^g/\mu_1^g > 1.5$	$\mu_2^g/\mu_1^g < 1$	$\mu_2^g/\mu_1^g = 1$	$\mu_2^g/\mu_1^g > 1$
1	2736	3455	50	2984	14 217	409
2	2302	1379	206	4315	13 448	2595
3	3892	744	338	8506	8474	2741
4	1496	2159	26	3008	16 733	530
5	2122	3389	127	1872	15 331	1439

somies. Specifically, higher percentages of *cis*-genes in trisomies 2 and 3 are modulated than for *cis*-genes in trisomies 1, 4 and 5; and higher percentages of *trans*-genes in trisomies 2 and 3 are modulated compared to those in trisomies 1, 4 and 5.

Analysis of trisomy 21 Down syndrome gene expression data

Trisomy 21, most often caused by the non-disjunction of chromosome 21 in oocytes, leads to DS and is the most common autosomal aneuploidy among live births. Due to its high frequency, DS has become an important model for exploring the consequences of trisomy in humans. Several studies have focused on elucidating the molecular consequences attributable to the presence of an extra copy of the chromosome 21 (9,32–34). Among them, a study in 2018 by Gonzales *et al.* (9) analyzed gene expression levels via RNA-seq data analysis from trisomic- and disomic-induced pluripotent stem cells (iPSCs) and differentiated cortical neurons derived from the same individual with DS. In the

original study, researchers (9) applied DESeq to identify differentially expressed genes. To evaluate Foldseq, here we re-analyze their data from iPSC cultures using our proposed Foldseq algorithm. To ensure a fair comparison, we only considered genes that appeared in both analyses: 217 genes from chromosome 21 and 621 genes in the rest of the genome. Between DESeq and Foldseq, when using the same FDR level of 0.1, Foldseq identified eight upregulated genes compared to 13 identified by DESeq, five of which were identified by both methods. We found that the actual expression levels of the three genes identified solely by Foldseq (*S100B*, *EVA1C* and *OLIG2*) were low and their variance was relatively high, which may explain why DESeq did not identify them as differentially expressed. By contrast, the variance between the log fold change was stable and relatively moderate, which enabled these genes to be identified by our method. Interestingly, *S100B* is a gene known to have increased expression levels in DS and is associated with Alzheimer’s disease, a pathology that manifests early in adults with DS (35–38). Also, *OLIG2*, a gene known to

have altered expression in DS, is associated with key consequences during brain development (39,40). At last, *EVA1C*, a slit receptor located in the DS critical region, has been shown to be expressed in axons and to serve a key function in neural circuit formation in the developing nervous system of mice (41).

The differences between methods are even more striking upon examining downregulated genes on chromosome 21: our method identified the four genes deemed downregulated by Gonzales *et al.* (*CBR3*, *COL6A1*, *DSCAM* and *TRPM2*) along with six new genes not previously considered downregulated by DESeq. Among them, we found *BACE2*, a gene identified in previous investigations for its influence in Alzheimer's disease and DS. Although the exact role of *BACE2* has not been clearly identified (42–44), (38) pointed out that this gene may have a protective role in the development of Alzheimer's disease in adults with DS. The downregulation of this chromosome 21 gene identified by our method may provide insight into the gene's association with DS-related Alzheimer's disease, further highlighting the importance of Foldseq.

Analysis of genome-wide expression values for 25 621 genes indicated that Foldseq detected no significant fold changes in expression levels between trisomic and disomic cells in 23 567 genes (91.98%), whereas DESeq found 24 700 genes (96.4%) with no significant change. Strikingly, Foldseq found 1528 genes (5.96%) that were statistically upregulated in trisomic cells compared to only 429 genes (1.67%) found by DESeq, a difference of 1099 genes. At last, the number of downregulated genes identified by both methods was similar: 526 genes (2.05%) and 492 genes (1.92%) for Foldseq and DESeq, respectively. We believe that these differences, particularly in the high number of upregulated genes solely identified by Foldseq, could influence enrichment analysis. To test this supposition, we used DAVID (45), a Gene Ontology (GO) term (46) enrichment analysis method, to identify enriched pathways, similar to (9).

Our results show a similar pattern to that described by (9), wherein iPS trisomic cells exhibit increased transcript levels of genes involved in neurogenesis and neuronal function. The main difference is that certain genes, which Foldseq identified as significantly upregulated in trisomic cells and which have known functions in the neuronal pathway, were undetected by DESeq. For example, *CNTN2*, a known gene with an important function in axon elongation or axonal guidance (47), was considered upregulated by Foldseq but found to have no statistical significance by DESeq. This gene was in turn detected by GO analysis in different pathways related to axon guidance (GO:0007411) or central nervous system development (GO:0007417). Among other genes considered differentially expressed by Foldseq and not DESeq, we found additional key genes involved in axonal growth and guidance; examples include *RET* (48), neuroligin Y (*NLGN4Y*) and neuroligin X (*NLGN4X*), which are synaptic cell-adhesion molecules that connect postsynaptic neurons at synapses. Along with neuroligins, these proteins are key components of the molecular machinery that controls synaptic transmission (49) and *SPTBN4*, a spectrin protein known as β IV with key functions at the axon initial segments and nodes of Ranvier (50,51).

Table 4. Computational time (in seconds) of FoldSeq for different number of genes and sample size in each treatment group

No. of genes	3 replicates	5 replicates	7 replicates	9 replicates
10 000	0.58	0.57	0.52	0.59
12 000	0.56	0.68	0.52	0.68
15 000	0.32	0.57	0.54	0.75
18 000	0.41	0.53	0.58	0.84
20 000	0.42	0.53	0.60	0.74

We believe that these results demonstrate the benefit of using Foldseq to analyze gene expression data from aneuploid genomes. As we discovered, Foldseq can detect statistically significant gene expression fold changes (that went unnoticed by DESeq) with known and proven functions in the pathology of DS, the organization and development of the nervous system, and axonal growth. These findings show that Foldseq could provide a more accurate picture of genome-wide variations in genomic expression occurring in aneuploidies.

DISCUSSION

Tests of differential expression and fold change involve two related hypotheses. Many gene expression experiments concern the former hypothesis while others concern the latter. Several sophisticated statistical methods have been developed to address differential expression. The vast biological community tends to directly apply methods of differential expression to test fold change, although others have applied the naive *t*-test. We used two real datasets to demonstrate that, if data analysis concerns fold change, then a test based on fold-change statistics provides more accurate and powerful results. In addition, a Bayesian prior that shares information across genes can substantially improve inferential results compared to a naive method without priors.

Computational time is an important factor for algorithms analyzing high throughput genomic data. For Foldseq, we conducted a comprehensive survey on the computational time cost for various combinations of gene numbers and sample sizes on a commodity compute server. The results are presented in Table 4. The survey results show that Foldseq's computational time is <1 s for all of these settings.

CONCLUSION

In this paper, we have proposed a new approach to detect gene expression fold change between two comparison groups. Our method could be generalized to any high-throughput microarray and sequencing experiment. Specifically, we constructed a model on the log of fold change for each gene, where the error variance of each gene was modeled as a random sample from an inverse gamma distribution. Hyperparameters in the inverse gamma prior were estimated by observations from all genes. Due to the large number of genes in a typical gene expression experiment, our hyperparameter estimation was accurate. A *t*-type test with adjusted degrees of freedom was developed for hypothesis testing. Real data analysis and simulation data analysis that mimicked actual datasets revealed our new approach, the Foldseq method, to be superior to existing methods for

detecting gene expression fold change; specifically, our approach demonstrated improved power while controlling the FDR at a desired level. Our R package is freely available on GitHub at <https://github.com/cuiyingbeicheng/Foldseq>.

DATA AVAILABILITY

Sequencing gene expression data for the *A. thaliana* study described in this paper have been deposited in the Gene Expression Omnibus (GEO) database with accession no. GSE79676. Sequencing gene expression data for the DS study described in this paper have been deposited in GEO with accession no. GSE101942. The Foldseq R package is available on GitHub at <https://github.com/cuiyingbeicheng/Foldseq>.

ACKNOWLEDGEMENTS

We would like to thank the reviewers, Associate Editor and Editor for their invaluable contributions to the peer review process.

FUNDING

National Science Foundation [1615789, 1853556 to T.J.].
Conflict of interest statement. None declared.

REFERENCES

- Van Robys, J. (2016) John Langdon Down (1828-1896). *Facts Views Vis. Obgyn.*, **8**, 131–136.
- Blakeslee, A.F., Belling, J. and Farnham, M.E. (1920) Chromosomal duplication and Mendelian phenomena in *Datura* mutants. *Science*, **52**, 388–390.
- Blakeslee, A.F. (1934) New Jimson weeds from old chromosomes. *J. Hered.*, **25**, 81–108.
- Bridges, C.B. (1925) Sex in relation to chromosomes and genes. *Am. Nat.*, **59**, 127–137.
- Tasdighian, S., Van Bel, M., Li, Z., Van de Peer, Y., Carretero-Paulet, L. and Maere, S. (2017) Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *Plant Cell*, **29**, 2766–2785.
- Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.*, **60**, 433–453.
- Emery, M., Willis, M.M.S., Hao, Y., Barry, K., Oakgrove, K., Peng, Y., Schmutz, J., Lyons, E., Pires, J.C., Edger, P.P. *et al.*, (2018) Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genet.*, **14**, e1007267.
- Hou, J., Shi, X., Chen, C., Islam, M.S., Johnson, A.F., Kanno, T., Huettel, B., Yen, M.R., Hsu, F.M., Ji, T. *et al.*, (2018) Global impacts of chromosomal imbalance on gene expression in *Arabidopsis* and other taxa. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E11321–E11330.
- Gonzales, P.K., Roberts, C.M., Fonte, V., Jacobsen, C., Stein, G.H. and Link, C.D. (2018) Transcriptome analysis of genetically matched human induced pluripotent stem cells disomic or trisomic for chromosome 21. *PLoS One*, **13**, e0194581.
- 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.
- McCarthy, D.J., Chen, Y. and Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
- Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.
- Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.
- Lönnstedt, I. and Speed, T.P. (2002) Replicated microarray data. *Stat. Sin.*, **12**, 31–46.
- Hardcastle, T.J. and Kelly, K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
- 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.
- 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.
- Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
- Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
- Birchler, J.A. (1979) A study of enzyme activities in a dosage series of the long arm of chromosome one in maize. *Genetics*, **92**, 1211–1229.
- Birchler, J.A. and Newton, K.J. (1981) Modulation of protein levels in chromosomal dosage series of maize: the biochemical basis of aneuploid syndromes. *Genetics*, **99**, 247–266.
- Guo, M. and Birchler, J.A. (1994) Trans-acting dosage effects on the expression of model gene systems in maize aneuploids. *Science*, **266**, 1999–2002.
- Sun, L., Johnson, A.F., Donohue, R.C., Li, J., Cheng, J. and Birchler, J.A. (2013) Dosage compensation and inverse effects in triple X metafemales of *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 7383–7388.
- Sun, L., Johnson, A.F., Li, J., Lambdin, A.S., Cheng, J. and Birchler, J.A. (2013) Differential effect of aneuploidy on the X chromosome and genes with sex-biased expression in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 16514–16519.
- Birchler, J.A., Bhadra, U., Bhadra, M.P. and Auger, D.L. (2001) Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev. Biol.*, **234**, 275–288.
- Seidman, J.G. and Seidman, C. (2002) Transcription factor haploinsufficiency: when half a loaf is not enough. *J. Clin. Invest.*, **109**, 451–455.
- Kondrashov, F.A. and Koonin, E.V. (2004) A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.*, **20**, 287–290.
- Boell, L., Pallares, L.F., Brodski, C., Chen, Y., Christian, J.L., Kousa, Y.A., Kuss, P., Nelsen, S., Novikov, O., Schutte, B.C. *et al.*, (2013) Exploring the effects of gene dosage on mandible shape in mice as a model for studying the genetic basis of natural variation. *Dev. Genes Evol.*, **223**, 279–287.
- Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., De Grassi, A., Lee, C. *et al.*, (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.
- Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I. and Young, R.A. (2012) Revisiting global gene expression analysis. *Cell*, **151**, 476–482.
- Lockstone, H., Harris, L.W., Swatton, J.E., Wayland, M.T., Holland, A.J. and Bahn, S. (2007) Gene expression profiling in the adult Down syndrome brain. *Genomics*, **90**, 647–660.
- Vilardell, M., Rasche, A., Thormann, A., Maschke-Dutz, E., Prez-Jurado, L.A., Lehrach, H. and Herwig, R. (2011) Meta-analysis of heterogeneous Down Syndrome data reveals consistent genome-wide dosage effects related to neurological processes. *BMC Genomics*, **12**, 229.
- Pelleri, M.C., Cattani, C., Vitale, L., Antonaros, F., Strippoli, P., Locatelli, C., Cocchi, G., Piovesan, A. and Caracausi, M. (2018) Integrated quantitative transcriptome maps of human trisomy 21 tissues and cells. *Front. Genet.*, **9**, 125.
- Netto, C.B., Portela, L.V., Ferreira, C.T., Kieling, C., Matte, U., Felix, T., da Silveira, T.R., Souza, D.O., Goncalves, C.A. and Giugliani, R. (2005)

- Ontogenetic changes in serum S100B in Down syndrome patients. *Clin. Biochem.*, **38**, 433–435.
36. Shapiro, L.A., Bialowas-McGoey, L.A. and Whitaker-Azmitia, P.M. (2010) Effects of S100B on serotonergic plasticity and neuroinflammation in the hippocampus in down syndrome and Alzheimer's disease: studies in an S100B overexpressing mouse model. *Cardiovas. Psychiatry Neurol.*, **2010**, 153657.
 37. Lu, J., Esposito, G., Scuderi, C., Steardo, L., Delli-Bovi, L.C., Hecht, J.L., Dickinson, B.C., Chang, C.J., Mori, T. and Sheen, V. (2011) S100B and APP promote a gliocentric shift and impaired neurogenesis in down syndrome neural progenitors. *PLoS One*, **6**, e22126.
 38. Webb, R.L. and Murphy, M.P. (2012) β -Secretases, Alzheimer's disease, and down syndrome. *Curr. Gerontol. Geriatr. Res.*, **2012**, 362839.
 39. Lu, J., Lian, G., Zhou, H., Esposito, G., Steardo, L., Delli-Bovi, L.C., Hecht, J.L., Lu, Q.R. and Sheen, V. (2012) OLIG2 over-expression impairs proliferation of human Down syndrome neural progenitors. *Hum. Mol. Genet.*, **21**, 2330–2340.
 40. Xu, R., Brawner, A.T., Li, S., Liu, J.J., Kim, H., Xue, H., Pang, Z.P., Kim, W.Y., Hart, R.P., Liu, Y. *et al.*, (2019) OLIG2 drives abnormal neurodevelopmental phenotypes in human iPSC-Based organoid and chimeric mouse models of down syndrome. *Cell Stem Cell*, **24**, 908–926.
 41. James, G., Foster, S.R., Key, B. and Beverdam, A. (2013) The expression pattern of EVA1C, a novel slit receptor, is consistent with an axon guidance role in the mouse nervous system. *PLoS One*, **8**, e74115.
 42. Sun, X., He, G. and Song, W. (2006) BACE2, as a novel APP θ -secretase, is not responsible for the pathogenesis of Alzheimer's disease in Down syndrome. *FASEB J.*, **20**, 1369–1376.
 43. Fluhrer, R., Capell, A., Westmeyer, G., Willem, M., Hartung, B., Condrón, M.M., Teplow, D.B., Haass, C. and Walter, J. (2002) A non-amyloidogenic function of BACE-2 in the secretory pathway. *J. Neurochem.*, **81**, 1011–1020.
 44. Azkona, G., Amador-Arjona, A., Obradors-Tarragó, C., Varea, E., Arque, G., Pinacho, R., Fillat, C., De La Luna, S., Estivill, X. and Dierssen, M. (2010) Characterization of a mouse model overexpressing beta-site APP-cleaving enzyme 2 reveals a new role for BACE2. *Genes Brain Behavior*, **9**, 160–172.
 45. Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
 46. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.*, (2018) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
 47. Masuda, T. (2017) Contactin-2/TAG-1, active on the front line for three decades. *Cell Adh. Migr.*, **11**, 524–531.
 48. Enomoto, H., Crawford, P.A., Gorodinsky, A., Heuckeroth, R.O., Johnson, E.M. and Milbrandt, J. (2001) RET signaling is essential for migration, axonal growth and axon guidance of developing sympathetic neurons. *Development*, **128**, 3963–3974.
 49. Südhof, T.C. (2008) Neuroligins and neuexins link synaptic function to cognitive disease. *Nature*, **455**, 903–911.
 50. Berghs, S., Aggujaro, D., Dirckx, R., Maksimova, E., Stabach, P., Hermel, J.M., Zhang, J.P., Philbrick, W., Slepnev, V., Ort, T. *et al.*, (2000) β IV spectrin, a new spectrin localized at axon initial segments and nodes of ranvier in the central and peripheral nervous system. *J. Cell Biol.*, **151**, 985–1002.
 51. Wang, C.C., Ortiz-Gonzalez, X.R., Yum, S.W., Gill, S.M., White, A., Kelter, E., Seaver, L.H., Lee, S., Wiley, G., Gaffney, P.M. *et al.*, (2018) β IV spectrinopathies cause profound intellectual disability, congenital hypotonia, and motor axonal neuropathy. *Am. J. Hum. Genet.*, **102**, 1158–1168.