Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in expression mean and variability

Aedan G. K. Roberts[1,*], Daniel R. Catchpoole[1,2] and Paul J. Kennedy[2]

[1] School of Computer Science and Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia
[2] Children's Cancer Research Unit, The Children's Hospital at Westmead, Sydney, Australia
* Corresponding author: aedan.g.roberts@student.uts.edu.au

Abstract

Background: Differential expression analysis of RNA-seq data has advanced rapidly since the introduction of the technology, and methods such as edgeR and DESeq2 have become standard parts of analysis pipelines. However, there is a growing body of research showing that differences in variability of gene expression or overall differences in the distribution of expression values – differential distribution – are also important both in normal biology and in diseases including cancer. Genes whose expression differs in distribution without a difference in mean expression level are ignored by differential expression methods.

Results: We have developed a Bayesian hierarchical model which improves on existing methods for identifying differential dispersion in RNA-seq data, and provides an overall test for differential distribution. We have applied these methods to investigate differential dispersion and distribution in cancer using RNA-seq datasets from The Cancer Genome Atlas. Our results show that differential dispersion and distribution are able to identify cancer-related genes. Further, we find that differential dispersion identifies cancer-related genes that are missed by differential expression analysis, and that differential expression and differential dispersion identify functionally distinct sets of genes.

Conclusion: This work highlights the importance of considering changes beyond differences in mean in the analysis of gene expression data, and suggests that analysis of expression variability may provide insights into genetic aspects of cancer that would not be revealed by differential expression analysis alone. For identification of cancer-related genes, differential distribution analysis allows the identification of genes whose expression is disrupted in terms of either mean or variability.

Background

As RNA sequencing (RNA-seq) has replaced microarray as the leading technology for large-scale gene expression analysis at the whole tissue level, there has been rapid progress in the development of methods for analysing the resulting count data. The focus of most of these methods is differential expression analysis – identifying genes whose mean expression levels differ between groups of interest. However, there is a growing body of evidence to suggest that differences in variability of gene expression are also biologically and medically important. Differences in expression variability have been associated with biological function [1, 2, 3], development [4, 5] and ageing [6, 7, 8, 9]. Changes in expression variability have also been implicated in diseases including schizophrenia [3, 10] and cancer [11, 12, 13, 14]. Genes selected for differences in variability have been demonstrated to have diagnostic and prognostic potential in cancer [12, 13, 15, 16, 14, 17].

Much of the work to date on expression variability has focused on microarray data, and variability has been assessed using empirical measures such as standard deviation [13, 15, 5], the coefficient of variation (CV; defined as the ratio of the standard deviation to the mean) [11, 3, 10], or various measures of deviation of expression values from an expected value defined by a relationship between the mean expression level and its variability [18, 17]. Tests for differences in expression variance between groups have generally been based on established normal distribution-based tests such as the F-test and Bartlett's test, or robust alternatives such as Levene's test or the Brown–Forsyth test [6, 19, 20, 21, 14, 22]. In assuming that the data follows a normal distribution or not assuming any parametric distribution, these tests are likely to be under-powered for all but very large sample sizes. This is particularly the case if they are to be applied to RNA-seq data, which takes the form of counts, and so is discretely, rather than continuously, distributed.

RNA-seq data is most commonly modelled using the negative binomial (NB) distribution, for which the variance is a function of the mean and a dispersion parameter. There is one method published to date that provides a test for differences in variability specifically for bulk RNA-seq data: MDSeq [23], which uses a NB generalised linear model to test for differences in mean and dispersion separately. Generalised additive models for location, shape and scale (GAMLSS) [24] form another family of regression models with potential for identifying differentially dispersed genes. The use of GAMLSS for differential variability analysis for RNA-seq data was recently demonstrated in [25]. BASiCS [26] uses a Bayesian model to test for differences in mean and dispersion for single-cell RNA-seq data which relies on spike-in genes or technical replicates to decompose the observed variability into technical and biological components.

The negative binomial distribution is completely defined by the mean and dispersion parameters. We can therefore define "differential distribution" – any change in the distribution of expression levels between groups – as a difference in either one or both of these parameters. Since there is a clear interest in identifying genes with differences in mean expression, and growing evidence for biological relevance of differences in expression variability, a method that combines both types of analysis into a single test for differential distribution should allow for more complete identification of genes with relevance to a disease or biological state than either method alone.

We have developed a hierarchical Bayesian model based on the negative binomial distribution that provides tests for differential expression, dispersion and distribution for RNA-seq data. The tests for differential expression and dispersion are similar to those implemented in the BASiCS method for single-cell RNA-seq data [26]. The hierarchical model (HM) is incorporated into a hierarchical mixture model (HMM) which provides what is, to the best of our knowledge, the first overall test for differential distribution for RNA-seq data. Testing on simulated and semi-simulated data shows that our method is competitive with state-of-the-art methods for differential expression analysis, and improves on MDSeq for differential dispersion analysis.

Using this model, we report here an investigation into differential dispersion and differential distribution in human cancers. Analysis of eight paired tumour–normal RNA-seq datasets from The Cancer Genome Atlas (TCGA; `https://www.cancer.gov/tcga`) demonstrates that differential dispersion identifies different sets of genes from differential expression. Using lists of genes previously identified as being related to each of these cancers, we provide what is to our knowledge the first demonstration that differential dispersion and differential distribution analyses are able to identify cancer-related genes, and that differential dispersion identifies cancer-related genes that are not identified by differential expression. We further show using gene set enrichment analysis that differential dispersion identifies functionally distinct sets of genes compared to differential expression. Together, these results add to the growing body of literature highlighting the importance of changes in expression variability in cancer independently of changes in mean expression levels, and suggest that differential distribution analysis may provide a more comprehensive way of identifying potential cancer-related genes.

Results

Hierarchical model identifies genes with differences in mean, dispersion or overall distribution

We have developed a Bayesian hierarchical model for RNA-seq data based on the negative binomial distribution, for which a Markov chain Monte Carlo (MCMC) algorithm provides samples from the posterior distributions of the mean μ_j and dispersion ϕ_j for each gene j. These posterior samples are the basis for tests for differences in mean or dispersion between groups, and a two-component mixture model provides the basis for a test for differential distribution. The hierarchical nature of the model allows information to be shared among genes, producing parameter estimates for each gene that are shrunk towards a common value estimated over all genes and thereby allowing stable parameter estimation for small sample sizes. Related information-sharing schemes are common in dispersion estimation procedures for differential expression analysis methods [27, 28, 29], and are at the core of Bayesian methods for differential expression analysis [30, 31], but are not used in MDSeq, the only previously published method for differential dispersion in bulk RNA-seq data. Figure 1 shows a schematic outline of the hierarchical model.

Further details on the hierarchical model, MCMC sampling and posterior inference are given in Methods.

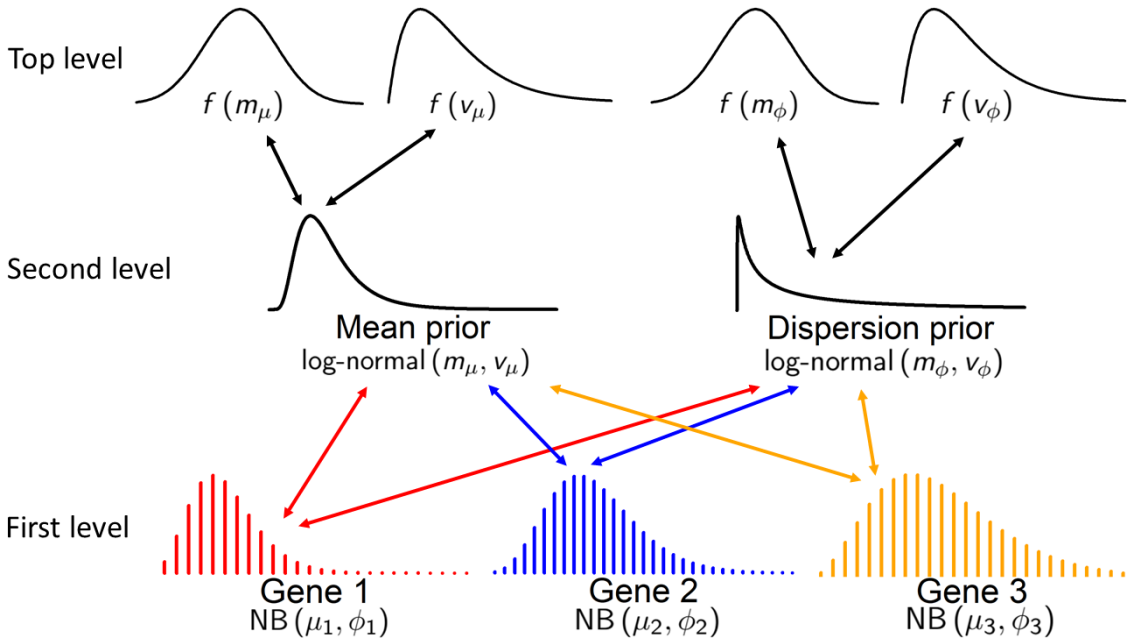

Figure 1: Schematic illustration of the hierarchical model. The distribution of count values for each gene j is assumed to follow a negative binomial distribution with mean and dispersion parameters μ_j and ϕ_j, respectively. The parameters for each gene are related to each other through log-normal priors, allowing information to be shared between genes. The parameters of these priors (hyperparameters: m_μ and v_μ for the mean prior and m_ϕ and v_ϕ for the dispersion prior) are themselves not fixed, but are modelled as random realisations from the hyperpriors $f(m_\mu)$, $f(v_\mu)$, $f(m_\phi)$ and $f(v_\phi)$.

The performance of the hierarchical model was assessed using two approaches: testing on simulated data, and on real RNA-seq data with artificially-induced differences in expression distribution. Using fully simulated data allows assessment of the performance of a model with reference to a known ground truth, but has the disadvantage that simulated data may not adequately reflect the properties of real

data. This is an issue especially when the data is simulated under an assumed parametric model – in this case, the negative binomial distribution. An alternative to using simulated data is to use real data, but, with the exception of spike-in experiments, this has the major disadvantage that the ground truth is unknown.

Here we take a different approach: using real data where there is a reasonable assumption of no differences in distribution between groups, and manipulating the data to artificially introduce differences in distribution for a known subset of genes. The Genotype–Tissue Expression (GTEx) project [32] (`https://gtexportal.org/home/`) provides a source of RNA-seq data from a range of tissues from healthy donors. Randomly splitting samples for a given tissue type into two groups provides a baseline dataset with no expected differences in expression between group. We used GTEx data from two tissues – whole blood and skeletal muscle. For a given tissue type and sample size, we generated datasets with known levels of differential expression and/or differential distribution by randomly selecting samples and altering the counts for a proportion of genes in one group such that each gene's counts reflect a change in mean and/or dispersion from the other group, under the assumption that the counts follow a negative binomial distribution. This was repeated ten times for each tissue type to obtain ten datasets using different samples.

Model performance was assessed on sample sizes from 2 to 50 samples per group. Fifty simulated datasets were used for each sample size, with data simulated for 20,000 genes. Between-group differences in mean only were introduced for 5% of genes, dispersion only for 5%, and both mean and dispersion for another 5%, with a minimum fold change in mean and dispersion of 1.5. Further details are given in the Methods.

Test for differential differential dispersion outperforms existing methods

The differential expression test from HM was compared against edgeR [33], DESeq2 [34], limma-voom [35, 36], baySeq [30] – which also uses a Bayesian hierarchical model – and MDSeq, which provides tests for differential expression and differential dispersion. HM gave similar performance to the best of the other methods (Additional file 1: Section S1).

The HM differential dispersion test was compared against MDSeq and GAMLSS. False discovery curves (Figure 2) show that HM consistently outperforms MDSeq. Compared to GAMLSS, HM returns fewer false discoveries for a given number of discoveries with 5 samples per group, and nearly identical with 50 samples per group. Boxplots of false discovery rates (FDRs) and sensitivity (Figure 3) show that while HM controls the FDR at the 0.05 level better than MDSeq and GAMLSS with 5 samples per group, all three methods have very low power to detect differentially dispersed genes. With 50 samples per group, HM and GAMLSS again show nearly identical performance to each other, and both have greater power than MDSeq while maintaining lower FDRs. False discovery curves for 2, 10 and 20 samples per group are given in Additional file 1: Figure S3, and show similar patterns, as do boxplots of FDR and sensitivity (Additional file 1: Figure S4).

Along with MDSeq, HM has an advantage over GAMLSS in being able to test for differences in dispersion at a given minimum log fold change (LFC). False discovery curves (Additional file 1: Figure S5) and boxplots of FDR and sensitivity (Additional file 1: Figure S6) show that HM clearly outperforms MDSeq for differential dispersion at a minimum LFC of 1.44.

Hierarchical mixture model detects differentially distributed genes

The differential distribution test provided by the hierarchical mixture model was also tested on the simulated and GTEx data. An alternative method of testing for differential distribution was also considered: a "hybrid" method combining separate tests for differential expression and differential dispersion. For the hybrid test, the evidence for differential distribution for each gene is taken as the smaller of the p-values or posterior tail probabilities resulting from separate tests of differential expression and differential dispersion. Differential expression was assessed using edgeR for the simulated

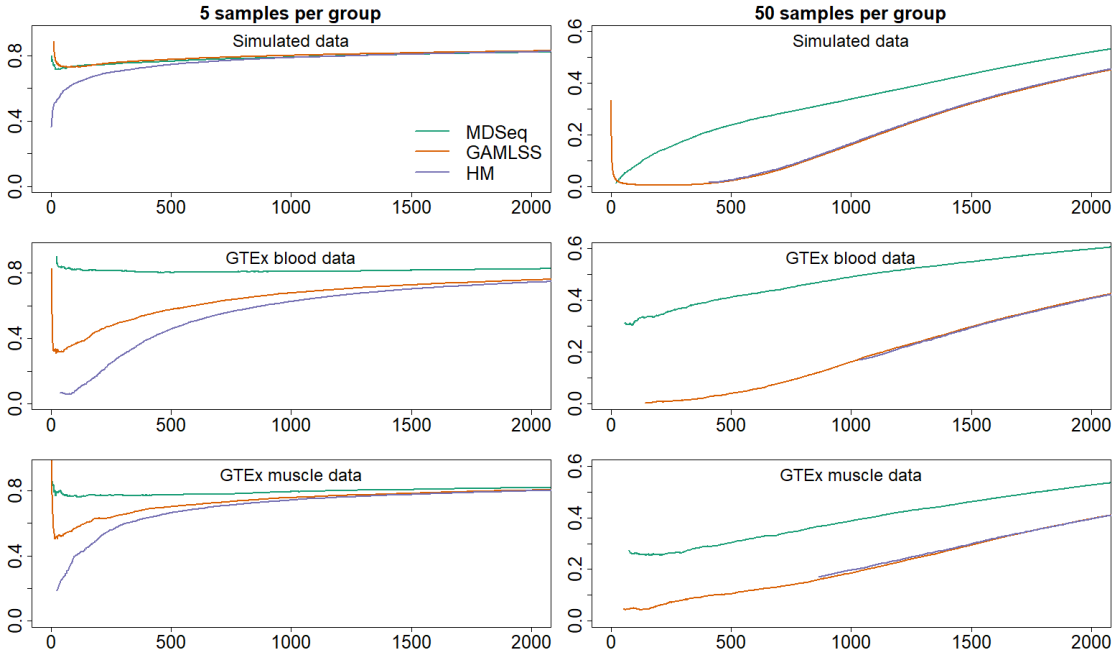

Figure 2: False discovery curves for detection of differential dispersion. Averaged over 50 simulated datasets (top) and 10 datasets generated from GTEx blood (middle) and muscle (bottom) data, with 5 or 50 samples per group.

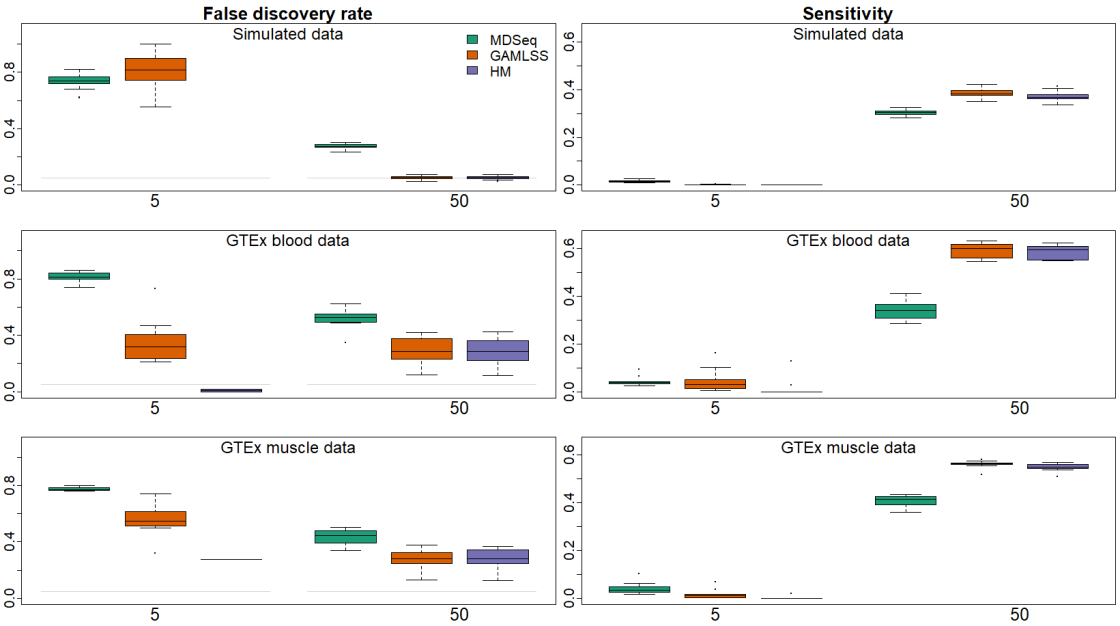

Figure 3: Boxplots of FDR and sensitivity for differential dispersion tests with nominal FDR 0.05. Data from 50 simulated datasets (top) and 10 datasets generated from GTEx blood (middle) and muscle (bottom) data, with 5 or 50 samples per group.

data and limma-voom for the GTEx data as these were the best-performing tests in each situation, and differential dispersion was assessed using the hierarchical model.

Two alternative methods for determining a threshold at which to call a gene as differentially distributed were considered for HMM. The first uses a posterior estimate of the proportion of genes that are differentially distributed, and the second uses a Bayesian estimate of the false discovery rate (BFDR) [37, 38]. It should be noted that while the BFDR aims to control the FDR at a given level, the posterior threshold method is not aimed at FDR control.

While there are no published methods for differential distribution for RNA-seq data, an alternative was considered taking advantage of the fact that under the NB model, the variance is function of both the mean and the dispersion. This means that a test for a difference in variance should effectively act as a test of differential distribution. As such, diffVar was also included as an alternative test of differential distribution.

False discovery curves (Figure 4) show that both HMM and hybrid tests are able to detect differentially distributed genes, and that both are more effective than diffVar, particularly for small sample sizes. With 50 samples per group, HMM and the hybrid method are able to identify up to around 1900 differentially distributed genes while maintaining the FDR below 0.05. Boxplots of FDR and sensitivity for a nominal FDR of 0.05 (Figure 5) show that HMM provides better FDR control than the hybrid method, but is extremely conservative on simulated data, and slightly liberal on the GTEx data. With 5 samples per group, all methods either have very low sensitivity, or reasonable sensitivity but with very poor FDR control. With 50 samples per group, HMM has sensitivity of around 0.7 – lower than the hybrid method, but with better FDR control. False discovery curves for 2, 10 and 20 samples per group are shown in Additional file 1: Figure S7, and boxplots of FDR and sensitivity in Additional file 1: Figure S8.

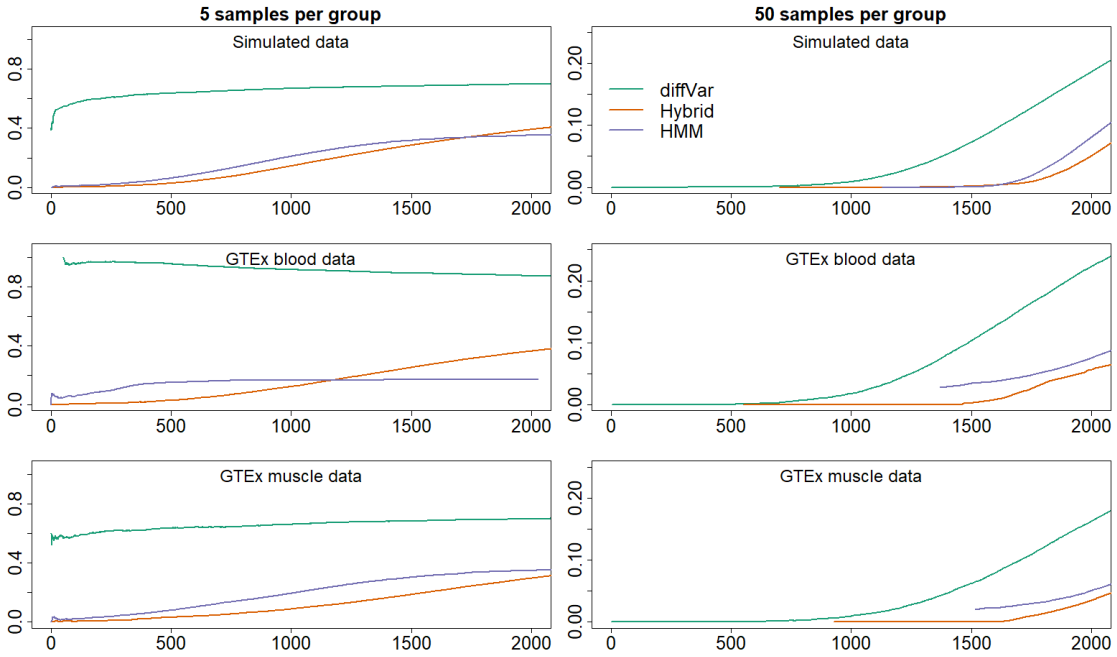

Figure 4: False discovery curves for detection of differential distribution. Averaged over 50 simulated datasets (top) and 10 datasets generated from GTEx blood (middle) and muscle (bottom) data, with 5 or 50 samples per group.

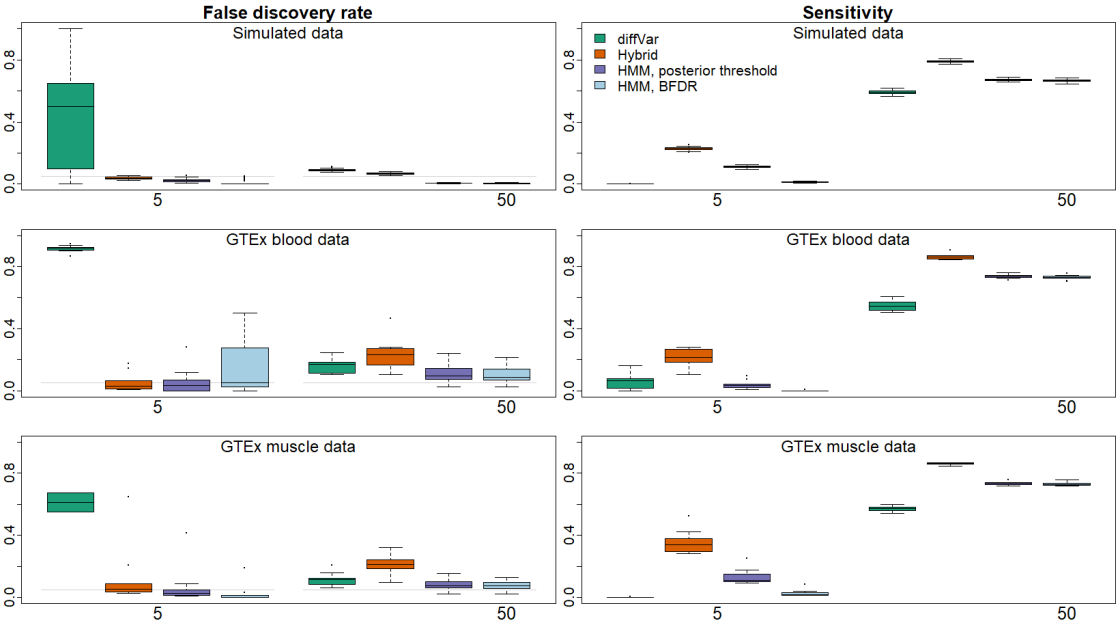

Figure 5: Boxplots of FDR and sensitivity for differential distribution tests with nominal FDR 0.05. Data from 50 simulated datasets (top) and 10 datasets generated from GTEx blood (middle) and muscle (bottom) data, with 5 or 50 samples per group.

Differential expression and differential dispersion identify functionally distinct sets of cancer-related genes

We next applied tests of differential dispersion and distribution to RNA-seq data from TCGA. These analyses were aimed at addressing several important questions around the relevance of differential distribution in cancer. Analyses were performed on paired normal–tumour data from eight cancer types (Table 1). Results are presented here primarily for lung adenocarcinoma and prostate adenocarcinoma, and full results are presented in Additional file 1: Section S4.

Table 1: TCGA tumour–normal sample pairs used

TCGA project ID	TCGA project name	Sample pairs	Reference
BRCA	Breast invasive adenocarcinoma	182	[39]
KIRC	Kidney renal clear cell carcinoma	144	[40]
THCA	Thyroid carcinoma	98	[41]
LUAD	Lung adenocarcinoma	86	[42]
LIHC	Liver hepatocellular carcinoma	98	[43]
LUSC	Lung squamous cell carcinoma	92	[44]
PRAD	Prostate adenocarcinoma	92	[45]
COAD	Colon adenocarcinoma	78	[46]

Differential dispersion and differential distribution identify cancer-related genes

We first tested whether differential dispersion and differential distribution analyses are able to identify cancer-related genes. To do this, we compiled lists of genes that have previously been identified as being related to each of the eight cancer types from five different databases: the Cancer Genome

Census (CGC) [47], DisGeNET [48], IntOGen [49], KEGG [50] and Malacards [51]. We then tested the ability of each method to rank cancer-related genes above other genes using Wilcoxon rank-sum tests. The methods included were differential expression (using limma–voom), differential dispersion using HM, and two alternative differential distribution tests: HMM, and the hybrid method combining limma–voom for differential expression and the HM differential dispersion test.

The resulting p-values are given in Table 2. Cancer-related genes are statistically significantly ranked above other genes at the 0.05 level for all eight cancers by differential expression analysis, and for three out of eight cancers by differential dispersion: thyroid carcinoma, lung adenocarcinoma and lung squamous cell carcinoma. For differential distribution, results were strikingly different between the HMM and hybrid methods: statistically significant associations were identified for all cancers using the hybrid method, and for lung adenocarcinoma, hepatocellular carcinoma and lung squamous cell carcinoma using HMM. This difference can be explained by the extremely small p-values that limma–voom returns for many differential expression tests: around half of all p-values returned by limma–voom are smaller than the minimum tail probability from the HM differential dispersion test, meaning that the hybrid method is effectively biased towards differentially expressed genes.

While differential expression analysis ranks previously identified cancer-related genes above other genes more strongly than differential dispersion or distribution in most cases, this should not be surprising, since differential expression has been one of the main methods for identifying potentially cancer-related genes – it is likely that the majority of the cancer-related genes used here were originally identified as such because they were found to be differentially expressed between normal and cancer tissues. Given this, it is particularly striking that for three of the eight cancer types – thyroid carcinoma, lung adenocarcinoma and lung squamous cell carcinoma – there are similar levels of evidence for association with cancer-related genes for differential dispersion and differential expression.

Table 2: p-values from Wilcoxon rank-sum tests for ranking cancer-related genes above other genes

Dataset	Differential expression	Differential dispersion	Differential distribution (HMM)	Differential distribution (Hybrid)
Breast invasive adenocarcinoma	9×10^{-13}	1.00	0.16	5×10^{-11}
Clear cell renal cell carcinoma	5×10^{-4}	0.94	0.08	3×10^{-4}
Thyroid carcinoma	0.003	0.002	0.28	9×10^{-4}
Lung adenocarcinoma	0.009	1×10^{-4}	0.005	1×10^{-4}
Hepatocellular carcinoma	1×10^{-12}	0.52	0.005	4×10^{-10}
Lung squamous cell carcinoma	3×10^{-4}	3×10^{-4}	0.003	3×10^{-5}
Prostate adenocarcinoma	1×10^{-10}	0.17	0.91	1×10^{-10}
Colon adenocarcinoma	2×10^{-17}	1.00	0.89	4×10^{-11}

Differential dispersion identifies different sets of genes from differential expression

The next question we asked was whether differential dispersion and differential expression identify different sets of genes. This is a critical question, and one that has not been addressed comprehensively in previous studies: if differential dispersion analysis identifies the same sets of genes as differential expression, the method is of little benefit.

To address this, we assessed the correlation between gene lists identified by differential expression and differential dispersion tests. Differential expression was assessed using limma–voom, and differential dispersion using HM. For a given p-value or posterior probability threshold for calling a gene as differentially expressed or dispersed, we created a ranked list of all genes called by both methods and calculated the Spearman correlation – the correlation of the ranks – between the two methods. For a given threshold, a positive correlation means that differential expression and differential dispersion

are producing similarly ranked lists of genes, while zero or negative correlation means that the two methods are producing unrelated or inversely associated gene rankings.

Results are shown in Figure 6a for lung and prostate adenocarcinoma, and in Additional file 1: Figure S9 for all eight cancers. Negative correlations are observed for most of the range of thresholds, and correlations are more strongly negative for the most highly ranked genes. Additional file 1: Figure S9 also shows corresponding p-values from hypothesis tests for negative correlation, with the null hypothesis of zero or positive correlation rejected for all thresholds that are likely to be of practical interest for all eight cancers. These results clearly show that differential expression and differential dispersion analyses identify distinct sets of genes.

These results and those in the previous section demonstrate that, at least for some cancer types, differential dispersion analysis is able to identify cancer-related genes and that it identifies different sets of genes from differential expression. We next looked specifically at whether differential expression and differential dispersion identify different sets of cancer-related genes.

Figure 6b shows, for lung and prostate adenocarcinomas, the number of cancer-related genes identified by each method alone and combined, as the threshold for calling a gene as differentially expressed or dispersed is varied. The results are consistent with the Wilcoxon rank-sum test results: there are more lung adenocarcinoma-related genes identified by differential dispersion than by differential expression, and the opposite for prostate adenocarcinoma. Notably, however, in both cases, even with high thresholds, there are some cancer-related genes that are identified uniquely by each method. For example, even though there is no overall evidence for an association between gene ranking by differential dispersion and prostate adenocarcinoma-related genes, there are around 100 genes that are correctly identified by differential dispersion but not by differential expression. These results provide further support for the idea that neither differential expression nor differential dispersion should be relied on alone to identify cancer-related genes, and that overall assessment of differential distribution of gene expression can more comprehensively identify these genes.

Differential expression and differential dispersion identify genes in different functional categories

To further investigate the types of genes identified by differential expression and differential dispersion, we performed gene set enrichment analysis using the GSEA software [52]. We looked for enriched terms in each of the three Gene Ontology (GO) ontologies – biological process, molecular function and cellular component – among genes ranked by differential expression using limma-voom and differential dispersion using HM. To reduce the number of closely related GO categories returned, the GOSemSim R package [53] was used to measure the semantic similarity between categories, and the rrvgo R package, based on REVIGO [54], was used to discard highly redundant categories.

Tables 3 and 4 show the ten most significantly enriched terms in each of the three ontologies for genes ranked by differential expression and differential dispersion, respectively, for lung adenocarcinoma. Some general themes are evident, most notably the high proportion of terms related to the immune system and signalling for differential expression, and to transcription, translation and intracellular transport for differential dispersion. Results for the other cancer types are given in Additional file 1: Section S4.

GO categories relating to DNA replication, transcription and translation are among the most significantly enriched terms for several other cancer types with genes ranked by differential dispersion. For example, for breast adenocarcinoma, RNA 3'-end processing, gene silencing, DNA recombination, telomere organisation and mRNA processing are among the top 10 terms for biological process; histone binding, single-stranded DNA binding, chromatin binding and basal transcription machinery binding are among the top 10 terms for molecular function; and nuclear chromosome telomeric region and replisome are among the five terms with FDR-adjusted q-values < 0.05 for cellular component. In contrast, there are very few such terms among the most significantly enriched with genes ranked by differential expression.

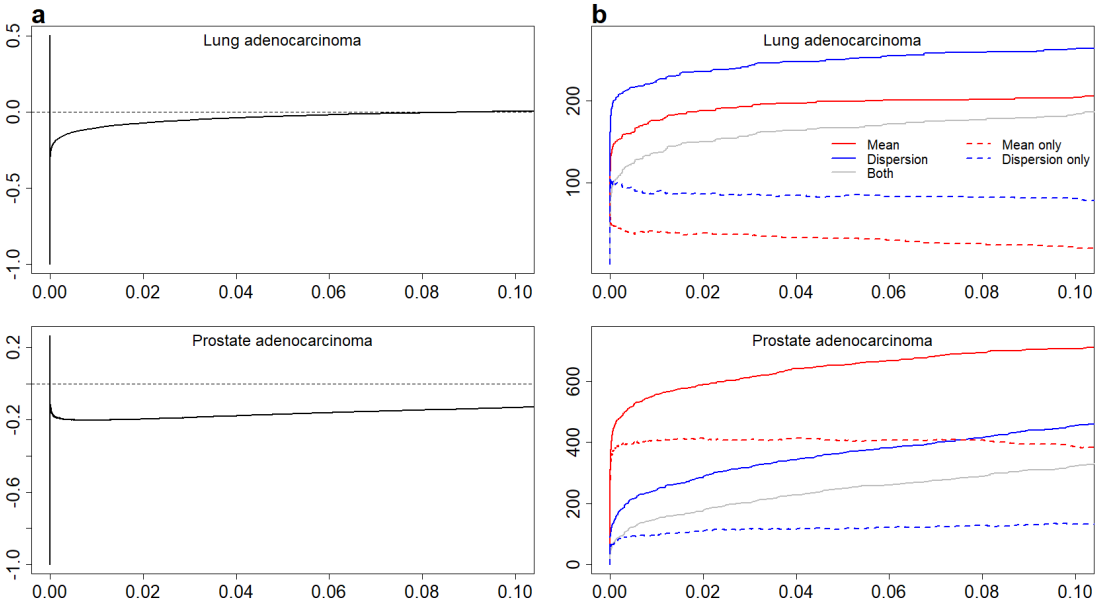

Figure 6: Differential expression and differential dispersion analyses independently identify cancer-related genes. (a) Spearman correlation between gene lists for differential expression and differential dispersion for TCGA lung and prostate adenocarcinoma data, with varying threshold for calling differential expression or dispersion. (b) Number of lung and prostate adenocarcinoma-related genes identified with varying thresholds for calling differential expression or dispersion. Solid lines show the total number of genes identified by differential expression (red), by differential dispersion (blue), and those identified by both methods (grey). Dashed lines show the number of genes uniquely identified by differential expression (red) and by differential dispersion (blue).

Table 3: Top 10 enriched GO terms in each ontology for lung adenocarcinoma based on differential expression

GO term	GO ID	FDR q-value
Biological process		
Phagocytosis recognition	GO:0006910	0.0000
B-cell mediated immunity	GO:0019724	0.0000
Lymphocyte-mediated immunity	GO:0002449	0.0000
Defense response to bacterium	GO:0042742	0.0000
Phagocytosis	GO:0006909	0.0000
Immunoglobulin production	GO:0002377	0.0000
Membrane invagination	GO:0010324	0.0000
FC receptor-mediated stimulatory signaling pathway	GO:0002431	0.0000
Urogenital system development	GO:0001655	0.0000
Regulation of body fluid levels	GO:0050878	0.0000
Molecular function		
Antigen binding	GO:0003823	0.0000
Immunoglobulin receptor binding	GO:0034987	0.0000
Receptor regulator activity	GO:0030545	0.0003
Extracellular matrix structural constituent	GO:0005201	0.0003
Oxygen binding	GO:0019825	0.0003
G-protein coupled receptor activity	GO:0004930	0.0004
Tetrapyrrole binding	GO:0046906	0.0005
Iron ion binding	GO:0005506	0.0006
Carbohydrate binding	GO:0030246	0.0006
Glycosaminoglycan binding	GO:0005539	0.0006
Cellular component		
Immunoglobulin complex circulating	GO:0042571	0.0000
Extracellular matrix	GO:0031012	0.0000
External side of plasma membrane	GO:0009897	0.0000
Collagen trimer	GO:0005581	0.0000
DNA packaging complex	GO:0044815	0.0000
Apical plasma membrane	GO:0016324	0.0000
Blood microparticle	GO:0072562	0.0000
Anchored component of membrane	GO:0031225	0.0001
I band	GO:0031674	0.0001
Basolateral plasma membrane	GO:0016323	0.0002

Table 4: Top 10 enriched GO terms in each ontology for lung adenocarcinoma based on differential dispersion

GO term	GO ID	FDR q-value
Biological process		
Chromatin assembly or disassembly	GO:0006333	0.0000
Glycosylation	GO:0070085	0.0000
DNA conformation change	GO:0071103	0.0000
Protein-DNA complex subunit organization	GO:0071824	0.0000
Homophilic cell adhesion via plasma membrane adhesion molecules	GO:0007156	0.0009
Retrograde vesicle-mediated transport Golgi to endoplasmic reticulum	GO:0006890	0.0027
RNA polyadenylation	GO:0043631	0.0027
Ethanol metabolic process	GO:0006067	0.0028
Telomere organization	GO:0032200	0.0032
Glycoprotein metabolic process	GO:0009100	0.0035
Molecular function		
Oxidoreductase activity acting on the CH-CH group of donors	GO:0016627	0.0008
Ligase activity forming carbon-oxygen bonds	GO:0016875	0.0031
Transferase activity transferring pentosyl groups	GO:0016763	0.0086
Ubiquitin-like protein transferase activity	GO:0019787	0.0087
Cofactor binding	GO:0048037	0.0088
Ubiquitin-like protein binding	GO:0032182	0.0122
Symporter activity	GO:0015293	0.0171
Transferase activity transferring nitrogenous groups	GO:0016769	0.0211
Protein heterodimerization activity	GO:0046982	0.0323
3'-5' exonuclease activity	GO:0008408	0.0325
Cellular component		
DNA packaging complex	GO:0044815	0.0000
Protein-DNA complex	GO:0032993	0.0000
Transport vesicle	GO:0030133	0.0000
Intrinsic component of endoplasmic reticulum membrane	GO:0031227	0.0000
Coated vesicle	GO:0030135	0.0007
Nuclear envelope	GO:0005635	0.0016
Site of polarized growth	GO:0030427	0.0019
Nuclear chromosome telomeric region	GO:0000784	0.0044
Lamellar body	GO:0042599	0.0047
Presynaptic membrane	GO:0042734	0.0075

Discussion

In this study, we have presented a model for RNA-seq data analysis that allows testing for differential expression, dispersion and distribution, and investigated the relationships between these different modes of disruption of normal gene expression patterns in cancer.

Tests for differential dispersion and distribution in RNA-seq data

The hierarchical model presented here is similar to that used by BASiCS, but applied to bulk RNA-seq rather than single-cell, without the need for spike-in genes or technical replicates, and using a fully hierarchical model, where the parameters for the priors on the mean and variance parameters are not fixed, but are also estimated as part of the model. The model is further extended into a mixture model, allowing for an overall test of differential distribution in addition to separate tests for differential expression and differential dispersion.

In tests on simulated data and on real data modified to artificially induce changes in expression, tests for differential dispersion using the hierarchical model outperformed the differential dispersion test of MDSeq. While MDSeq uses a different formulation of the negative binomial distribution, a difference in dispersion under the form used here still equates to a difference in dispersion under the MDSeq model. Unlike the hierarchical model, and most differential expression methods, MDSeq does not share information between genes, instead treating each gene independently. This may partly explain the improvement of the hierarchical model over MDSeq for differential dispersion detection, although any improvement due to information sharing would be expected to decrease with larger sample sizes, and the difference in power between the methods seems to increase with sample size. It may be that with samples even larger than those used here, the difference in performance between the two methods would decrease; here we tested scenarios with up to 50 samples per group, compared with up to 250 in the original MDSeq publication [23]. GAMLSS also outperformed MDSeq, and gave nearly identical performance to HM on larger sample sizes. The improvement of HM over GAMLSS for smaller samples is again likely due to the information sharing provided by the hierarchical model.

It is notable that while HM and GAMLSS both control the FDR well with 50 samples per group on simulated data, both methods fail to control the FDR on the semi-simulated data (Figure 3). This pattern was also seen with all the differential expression methods tested (Additional file 1: Figure S2). These results suggest that care must be taken in assessing the performance of methods on simulated data, particularly when the data is simulated under the same distribution as that assumed by the methods, as is the case here for all tests except for limma–voom. However, we cannot exclude the possibility that this reflects an issue with the method used to introduce artificial differences in expression in the GTEx data.

The hierarchical model also clearly outperformed MDSeq when testing for differential dispersion at a minimum log fold change (Additional file 1: Figures S5 and S6). In this respect, the HM test also has an advantage over GAMLSS, which does not allow for tests at a minimum LFC.

In addition to the hierarchical mixture model, we considered a "hybrid" test for differential distribution that combines separate tests for differential expression and differential dispersion. While both methods were successful in identifying differentially distributed genes, the hybrid method provided better results by some measures than the mixture model. While HMM gave better FDR control at the 0.05 level, false discovery curves showed that, for a given number of discoveries, combining separate tests generally resulted in fewer false discoveries. This may at least in part reflect an issue with inference from MCMC models, where the precision of parameter estimates is limited by the posterior sample size.

This limitation of posterior sampling-based inference is also reflected in the false discovery plots for the methods based on the hierarchical model, for which there are sometimes several hundred, or even up to 1500, top-ranked genes which cannot be separated. This issue could be avoided if a suitable hierarchical model could be identified for which analytical solutions could be found for the posterior parameter distributions. This is likely to mean moving away from the negative binomial model for

RNA-seq count distributions, but the success of limma-voom in identifying differential expression using a method that focuses on appropriately modelling the mean-variance relationship rather than specifying the exact distribution of the counts [36] suggests that such an approach is worth exploring. Another way of avoiding the issue of returning many equally-ranked genes may be to incorporate information on the magnitude of changes in mean or dispersion into gene rankings.

Relatively small sample sizes were used in this study – up to 50 samples per group. The results show that the hierarchical model has advantages over methods that treat each gene independently for small sample sizes, but this is not enough to completely mitigate the difficulty inherent in testing for changes in variability among small datasets. While differential expression analysis can be informative even with very small samples, analysis of differential variability or distribution may be best reserved for situations where larger numbers of samples are available.

Differential dispersion and distribution in cancer

Application to real RNA-seq datasets from TCGA showed that the differential expression and differential dispersion tests on tumour versus normal tissue identify distinct sets of genes. Differential dispersion analysis ranked previously identified cancer-related genes higher than other genes for three out of eight cancer types tested, with strength of evidence for association similar to that found for differential expression. This is particularly remarkable since it can reasonably be assumed that most cancer-related genes identified to date have been identified at least in part because they have been found to differ in mean expression levels between normal and tumour tissues.

As well as identifying different genes, gene set enrichment analysis showed that differential expression and differential dispersion identify functionally distinct sets of genes. GO terms relating to basic cellular processes such as transcription and translation are frequently enriched among genes that with difference in dispersion between normal and tumour samples. For example, ncRNA metabolic process, nucleic acid phosphodiester bond hydrolysis, RNA 3'-end processing and mRNA processing are all significantly enriched among differentially dispersed genes for multiple cancer types. In contrast, GO terms commonly enriched among differential expressed genes are often related to cell structure and migration (for example extracellular structure organisation), signalling (for example G protein-coupled receptor activity) or immune system functions (for example antigen binding).

Functions relating to transcription and translation have previously been identified among low-variability genes [55, 5, 56]. These are processes that need to be tightly regulated for cells to function properly, and so any change in expression – up or down – for genes involved in these processes is likely to disrupt normal cell function. This suggests a possible biological basis for the differences in types of genes identified by differential expression and differential dispersion in cancer. Genes identified by differential expression may be either genes that are normally active and for which loss of expression disrupts normal signalling or cellular transport processes, or genes that are normally expressed only in certain cell types or at certain times and for which constitutive expression similarly disrupts normal cell function. On the other hand, genes identified by differential dispersion or variability may be either genes that under normal circumstances are consistently expressed within a narrow range of levels, the regulation of which is lost in cancer, or genes whose normal function requires changes in expression levels in response to signals, and for which tightening of expression levels therefore disrupts their normal function. This idea is also consistent with previous observations that low variance genes often have "housekeeping" functions, while genes with high expression variability often have functions related to development and response to extracellular signals, for which changes in expression in response to signals are crucial [18]. Altered biological states in cancer may be dependent on different sets of genes being more or less tightly regulated than in healthy tissue, which may result in differences in expression variance between normal and tumour tissues. Decreased variance among a set of genes associated with invasive potential has previously been observed across multiple cancers [11], with the suggestion that tumour progression may be dependent on the precise regulation of these genes.

There is a distinction between cell-to-cell variability in gene expression at the tissue level and individual-to-individual variability at the population level, the former measured using single-cell RNA-

seq and the latter using bulk RNA-seq or microarray. This distinction has not always been made explicit in studies on expression variability, and the arguments above are stronger in the context of differences in expression between cells within a tissue. There have been suggestions that there is some correlation between the different levels of gene expression variation [57, 58], but this has not been demonstrated for multicellular organisms, and it is not clear why levels of variability between cells should correlate with levels of variability measured at a tissue level between individuals. Given this, it is intriguing that studies at the single-cell level [5, 55] and at the tissue level [18, 56] have found similar patterns in the functions of genes with different levels of expression variability, which are also consistent with the GO categories enriched among differentially dispersed genes in this study. This is an active area of research, and there is clearly more work to be done to elucidate the sources and significance of differences in expression variability at the cell-to-cell and individual-to-individual levels. In both cases, care must be taken to distinguish between differences in variability arising as an artifact of hidden heterogeneity and differences in variability within a homogeneous population.

Conclusion

There is a large, and growing, body of evidence that along with differential expression, analysis of differential variability of expression between normal and tumour tissue can identify genes that are involved in the initiation and progression of cancer. While there is a clear interest in elucidating the roles that different patterns of disruption of normal expression play, in terms of identifying cancer-related genes, any difference in the distribution of expression values between groups is of interest. Differential distribution has also been shown to provide improved feature selection for cancer classification compared to differential expression or differential variability alone [17]. An overall analysis of differential distribution may therefore be preferable to separate tests for differential expression and differential variability.

Methods

In this paper, we introduce a method for testing for differences in mean, dispersion or distribution between two groups for RNA-seq data. The performance of our method is assessed against state-of-the-art methods for differential expression, dispersion and distribution. We then apply differential expression, dispersion and distribution analyses to eight paired tumour–normal RNA-seq datasets from TCGA to test: (1) whether differential expression and differential dispersion analyses identify different sets of genes; (2) whether differential dispersion and differential distribution analyses identify cancer-related genes; and (3) whether there are functional differences between the sets of genes identified by differential expression and differential dispersion.

A hierarchical model to detect differential distribution in RNA-seq data

We assume that the observed read count Y_{ij} for gene j in sample i follows a negative binomial distribution, parametrised by a gene-specific mean μ_j and dispersion ϕ_j:

$$Y_{ij} \sim \mathrm{NB}\left(\mu_j, \phi_j\right). \tag{1}$$

We specify log-normal priors for the mean and dispersion parameters:

$$\mu_j \sim \text{log-normal}\left(m_\mu, v_\mu\right) \text{ and } \phi_j \sim \text{log-normal}\left(m_\phi, v_\phi\right). \tag{2}$$

The top level of the hierarchical model consists of hyperpriors on the prior parameters for the means and dispersions. We specify the following normal hyperpriors for the location parameters m_μ

and m_ϕ, and gamma hyperpriors for the scale parameters v_μ and v_ϕ:

$$\begin{aligned} m_\mu &\sim \mathrm{N}\left(2.5, 20\right), \\ v_\mu &\sim \mathrm{gamma}\left(3, 0.8\right), \\ m_\phi &\sim \mathrm{N}\left(-2.5, 2\right) \text{ and} \\ v_\phi &\sim \mathrm{gamma}\left(5, 2\right). \end{aligned} \tag{3}$$

The hyperpriors were chosen to place most density on the regions of the mean and dispersion parameter distributions that are most likely to be observed in real data, but with enough density outside of these regions so as to not overly restrict posterior inference.

All priors are assumed to be independent, and an adaptive MCMC sampling scheme is used to obtain posterior samples of the mean and dispersion parameters, which are the basis for inference from the hierarchical model. Gibbs sampling is used to obtain posterior samples for parameters whose posterior distributions follow standard distributions, and the Metropolis-Hastings algorithm is used for the remaining parameters. While posterior parameter estimates are not used directly in inference of differential expression, dispersion or distribution, parameter estimates – particularly estimates of dispersion – can be obtained as the mean of the posterior sample for each parameter. Full details of the MCMC algorithm are given in Additional file 1: Section S5.

Differential expression and differential dispersion

The posterior samples resulting from the MCMC algorithm are used to form tests for differences in mean and dispersion between two groups, A and B. Given posterior samples for the mean for gene j in each group, μ_{jA} and μ_{jB}, we obtain a posterior sample of the log fold change between groups by taking $\log_2 \mu_{jA} - \log_2 \mu_{jB}$ for each MCMC iteration. Similarly, we obtain a posterior sample of the LFC in dispersion as $\log_2 \phi_{jA} - \log_2 \phi_{jB}$. Tail probabilities from these samples are taken as a measure of the probability that the true difference in mean or dispersion, respectively, between the two groups is not equal to zero or a given minimum LFC. Tail probabilities are obtained by constructing highest posterior density (HPD) intervals – the narrowest range of values that contains a given amount of posterior density – and iteratively finding the amount of density that uniquely defines the narrowest region that does not contain zero. Subtracting these probabilities from one gives a posterior estimate of the probability that these is no difference in mean or dispersion between groups. For tests at a minimum LFC of c, the narrowest region of posterior density that does not include the range $[-c, c]$ is used instead of zero.

Although the tail probabilities are not p-values, it was found that applying the Benjamini–Hochberg false discovery rate procedure [59] worked well to control the FDR in simulated data, and so this method was used where a binary decision on differential expression or dispersion was desired.

Mixture model for differential distribution

To detect differences in distribution between two groups – that is, a difference in mean and/or dispersion – for each gene, the hierarchical model is embedded in a mixture model indexed by a parameter z_j, which indicates which of the two mixture components the data for gene j comes from: $z_j = 0$ if the mean and dispersion for gene j are the same for both groups, and $z_j = 1$ if either the mean or dispersion for gene j differs between the groups, i.e. if there is differential distribution for that gene.

The mixture model is defined by the following model for the distribution of read counts:

$$Y_{ij} \sim \begin{cases} NB(\mu_{j0}, \phi_{j0}) & \text{for all } i \text{ if } z_j = 0 \\ NB(\mu_{jA}, \phi_{jA}) & \text{for } i \in \{1, \ldots n_A\} \text{ if } z_j = 1 \\ NB(\mu_{jB}, \phi_{jB}) & \text{for } i \in \{n_A + 1, \ldots, n\} \text{ if } z_j = 1, \end{cases} \tag{4}$$

where n is the total number of samples and n_A the number of samples in group A, and the following density function:

$$f(y_{ij}|\theta_j, z_j) = f_0(y_{ij}|\theta_{j0})^{1-z_j} f_1(y_{ij}|\theta_{j1})^{z_j}, \tag{5}$$

where

$$\theta_{j0} = (\mu_{j0}, \phi_{j0}) \text{ and } \theta_{j1} = \begin{cases} (\mu_{jA}, \phi_{jA}) & \text{for } i \in \{1, \ldots, n_A\} \\ (\mu_{jB}, \phi_{jB}) & \text{for } i \in \{n_A + 1, \ldots, n\}. \end{cases} \tag{6}$$

We assign a Bernoulli prior on the z_j with a parameter λ representing the probability that $z_j = 1$, that is, the probability of differential distribution. We assign a uniform prior on λ over the range $(0, 1)$, and the posterior distribution of λ provides an estimate of the proportion of genes that are differentially distributed. The posterior mean of λ is taken as an estimate of the proportion of differentially distributed genes, and the posterior means of the z_j are taken as estimates of the probability differential distribution for each gene. These probabilities are used to rank genes by the strength of evidence for differential distribution. A binary decision can be made for each gene by using the posterior estimate of the proportion of differentially distributed genes to set a threshold such that the appropriate number of positive calls are made, or alternatively by using the Bayesian FDR (BFDR) [37, 38].

The R code used to implement the hierarchical model, including the MCMC algorithm and differential expression, dispersion, and distribution tests, is included in Additional file 2. The hierarchical model is implemented in an R package, DiffDist, available at `https://github.com/aedanr/DiffDist` [60].

Datasets

Simulated data

Comparisons with existing methods for differential expression and differential dispersion analysis were made using data simulated using the compcodeR R [61]/Bioconductor [62] package [63], and largely following the methods used in the comprehensive comparison of differential expression methods by [64].

Fifty simulated datasets were generated for each of 2, 5, 10, 20 and 50 samples per group, with differences in mean only for 5% of genes, dispersion only for 5%, and both mean and dispersion for 5%.

Data was simulated for 20,000 genes, using the default compcodeR settings except that a minimum counts per million filter of of 0.5 was applied, and for differentially expressed genes, half were upregulated in the second group and half downregulated. compcodeR samples means and dispersions from estimates obtained from two real datasets [65, 66], and, for differentially expressed genes, multiplies or divides means by a factor of $1.5 + x$, where x a random sample from an exponential distribution with mean 1. The minimum factor is therefore 1.5, and the mean factor is 2.5.

To simulate differential dispersion, a dataset with no differential dispersion and no filtering was first generated, from which dispersion values were extracted and used as baseline dispersions. Differential dispersions were then generated using the same model as for differential expression, and data was simulated specifying these two sets of dispersions.

Artificially introducing differential distributions in GTEx data

Whole blood and skeletal muscle RNA-seq datasets from the GTEx Project [32], processed by the recount2 project [67], were downloaded [68], and gene-level counts obtained using the recount R package. Samples used were limited to those from PAXGene-extracted RNA with RNA integrity number ≥ 6.9, which left 405 whole blood samples and 401 skeletal muscle samples. For each tissue type, samples were randomly selected to create ten datasets with each of 2, 5, 10, 20 and 50 samples per group. Samples were chosen without replacement for 2, 5, 10 and 20 samples per group, so that each of the ten datasets comprised different samples, and with replacement for 50 samples per group since there were insufficient samples to create ten completely distinct datasets.

Counts were then adjusted in one group to introduce changes in mean and dispersion as estimated by the method-of-moments estimators, $\hat{\mu} = \bar{y}$ and $\hat{\phi} = \frac{s^2 - \bar{y}}{\bar{y}^2}$, respectively, where $\bar{y}$ is the sample mean and s^2 is the sample variance. Further details are given in Additional file 1: Section S6. Means and dispersions were altered by factors following the same distributions as for the simulated data: half increased and half decreased, by a factor of 1.5 plus a random number from an exponential distribution with mean 1.

TCGA data

RNA-seq data from TCGA, processed by the recount2 project, were downloaded [68], and gene-level counts obtained using the recount R/Bioconductor package. Matching pairs of primary tumour and solid tissue normal samples were retained. Where there were multiple primary tumour samples for a patient, the sample with the highest mapped read count was selected. Cancer types with at least 40 tumour–normal pairs for which lists of related genes were available from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were selected, and where there were multiple histological diagnoses for a cancer type, only the most common was used: infiltrating ductal breast carcinoma; classical/usual papillary thyroid carcinoma; lung adenocarcinoma not otherwise specified (NOS); hepatocellular carcinoma; lung squamous cell carcinoma NOS; and colon adenocarcinoma. The resulting sample sizes and references for original publications are summarised in Table 1.

Cancer-related genes

Lists of genes previously identified as being related to each of the eight cancer types were compiled by combining lists of related genes from five sources: CGC [47], DisGeNET [48], IntOGen [49], KEGG and MalaCards [51]. Full details are given in the Additional file 1: Section S7.

Comparative assessment of model performance

Differential dispersion performance using the hierarchical model was compared against MDSeq and GAMLSS, and differential expression performance was compared against edgeR [33, 69, 70], DESeq2 [34], limma-voom [36, 71], baySeq [30] and MDSeq. The default settings were used for each method except that any gene filtering that was applied by default was not used as filtering was applied during simulation. Three versions of edgeR were initially tested: quasi-likelihood, likelihood ratio test and exact test. Quasi-likelihood gave the best results, and so only results from this version are reported. Preliminary testing showed that results were nearly identical using edgeR's TMM normalisation [72] and DESeq2's normalisation. TMM was used for all subsequent analyses.

The performance of the hierarchical model in detecting differential distribution was compared against diffVar [73, 74] and a naive method of combining the results from differential expression and differential dispersion methods. Results from separate differential expression and differential dispersion tests were combined by taking the smaller of the p-value or posterior probability of no difference between groups for the two tests for each gene. The resulting values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

The R code used for each method is included in Additional file 3, and can also be found at `https://github.com/aedanr/DiffDist`.

Application to real data

Ability to identify previously-identified cancer-related genes

Wilcoxon rank-sum tests were used to test the ability of differential expression and differential dispersion analyses to identify cancer-related genes. For each method, genes were ranked by the strength of evidence for differential expression or dispersion, and genes previously identified as being related to each cancer type were considered as positive, and all other genes as negative.

Comparison of genes identified by differential expression and differential dispersion

Spearman (rank) correlation was used to assess the similarity between lists of genes identified by differential expression and differential dispersion. For a list of genes for each method defined by a p-value or posterior probability threshold, the Spearman correlation between the union of genes in the two lists was calculated, and a corresponding one-sided correlation hypothesis test performed, with the alternative hypothesis that the correlation was negative.

Gene set enrichment analysis

Gene set enrichment analysis was carried out using the GSEA software [52], version 4.0.3, using the Human_ENSEMBL_Gene_MSigDB.v7.1 chip annotation and with genes ranked by the absolute LFC in mean or dispersion multiplied by $-\log_{10} p$, where p is the p-value or posterior probability of no difference between groups. This means that both the strength of evidence for a difference between normal and tumour and the magnitude of the difference contribute to a gene's ranking, and absolute values are used since a change in either direction is relevant. Analyses were carried out for GO terms in each of the three ontologies – biological process, molecular function and cellular component – with terms with less than two or more than 6 levels of parent terms excluded in order to avoid terms that were too general or too specific to be meaningfully interpreted. GO terms were taken from the org.Hs.eg.db annotation, and relationships between terms were obtained using the GO.db R/Bioconductor package. The resulting terms were then matched to the list of terms for each ontology in Molecular Signatures Database gene sets using the GSEABase package.

To further aid the interpretability of the results of gene set enrichment analysis, similarity matrices for terms in the results lists were identified using the GOSemSim R/Bioconductor package [53], and terms with high redundancy removed using the rrvgo package, which is based on REVIGO [54], with a threshold of 0.5.

Acknowledgements

The authors would like to thank Thomas Lysaght and George Mundackal for their contributions to optimising the MCMC code. The Genotype–Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, USA, and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The results on tumour—normal comparisons are based upon data generated by the TCGA Research Network: `https://www.cancer.gov.tcga`. The authors would like to thank the anonymous specimen donors for their essential contribution to this research.

Funding

This work was enabled by a NSW Health PhD Scholarship and the Australian Government Research Training Program (both AGKR).

Abbreviations

BFDR: Bayesian false discovery rate
CGC: Cancer Genome Census
CV: Coefficient of variation
FDR: False discovery rate
GAMLSS: Generalised additive models for location, shape and scale
GO: Gene ontology
GTEx: Genotype–Tissue Expression
HM: Hierarchical model

HMM: Hierarchical mixture model
HPD: Highest posterior density
KEGG: Kyoto Encyclopedia of Genes and Genomes
LFC: Log fold change
MCMC: Markov chain Monte Carlo
NB: Negative binomial
NOS: Not otherwise specified
RNA-seq: RNA sequencing
TCGA: The Cancer Genome Atlas

Availability of data and materials

The code for the hierarchical model and associated tests is included as Additional file 2, and the model is implemented in an R package, DiffDist, available at `https://github.com/aedanr/DiffDist` [60]. The code used to run each of the differential expression, dispersion and distribution methods for the methods comparisons is included as Additional file 3.

TCGA and GTEx data as processed by the recount2 project [75] were downloaded from `https://jhubiostatistics.shinyapps.io/recount/`. GSEA software was downloaded from `https://www.gsea-msigdb.org/gsea/index.jsp`. Full details of the process for obtaining lists of genes previously identified as being related to each cancer are given in Additional file 1, Section S7. The gene lists obtained from each database are given in Additional file 3, and the compiled lists of cancer-related genes in Additional file 4.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

AGKR, PJK and DRC conceived the study. AGKR developed and implemented the hierarchical model with input from PJK. AGKR designed and performed the analysis, interpreted the results and wrote the manuscript with input from PJK and DRC. All authors read and approved the final manuscript.

Additional Files

Additional file 1 — Supplementary results and methods (.pdf)

Section S1 provides the results of comparative analysis of differential expression tests using the hierarchical model and existing methods. Section S2 shows the results of differential dispersion comparisons with a wider range of sample sizes. Section S3 shows the results of differential distribution comparisons with a wider range of sample sizes. Section S4 provides the results of the investigation into differential

dispersion and distribution on all eight cancer types analysed. Section S5 contains additional details on the hierarchical model, the MCMC algorithm used to generate posterior samples, and the test for differential distribution using the hierarchical mixture model. Section S6 gives further details on introducing changes in mean and dispersion in the GTEx datasets. Section S7 details the process of collating lists of previously-identified cancer-related genes.

Additional file 2 — MCMC algorithm (.pdf)

R code for the MCMC algorithm.

Additional file 3 — Methods comparisons (.pdf)

R code for methods comparisons.

Additional file 4 — Cancer-related gene sources (.xlsx)

Lists of genes obtained from CGC, DisGeNET, IntOGen, KEGG and MalaCards.

Additional file 5 — Cancer-related genes (.xlsx)

Compiled lists of cancer-related genes.

References

[1] Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genetics. 2003 Mar;33(3):422–425.

[2] Komurov K, Ram PT. Patterns of human gene expression variance show strong associations with signaling network hierarchy. BMC Systems Biology. 2010;4:154.

[3] Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, Silburn PA, et al. Variance of gene expression identifies altered network constraints in neurological disease. PLOS Genetics. 2011 Aug;7(8):e1002207.

[4] Kalmar T, Lim C, Hayward P, Muñoz-Descalzo S, Nichols J, Garcia-Ojalvo J, et al. Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells. PLOS Biology. 2009 Jul;7(7):e1000149.

[5] Hasegawa Y, Taylor D, Ovchinnikov DA, Wolvetang EJ, Torrenté Ld, Mar JC. Variability of Gene Expression Identifies Transcriptional Regulators of Early Human Embryonic Development. PLOS Genetics. 2015 Aug;11(8):e1005428.

[6] Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, Dollé MET, et al. Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature. 2006 Jun;441(7096):1011–1014.

[7] Somel M, Khaitovich P, Bahn S, Pääbo S, Lachmann M. Gene expression becomes heterogeneous with age. Current Biology. 2006 May;16(10):R359–R360.

[8] Li Z, Wright FA, Royland J. Age-dependent variability in gene expression in male Fischer 344 rat retina. Toxicological Sciences. 2009 Jan;107(1):281–292.

[9] Viñuela A, Brown AA, Buil A, Tsai PC, Davies MN, Bell JT, et al. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. Human Molecular Genetics. 2018 Feb;27(4):732–741.

[10] Zhang F, Shugart YY, Yue W, Cheng Z, Wang G, Zhou Z, et al. Increased variability of genomic transcription in schizophrenia. Scientific Reports. 2015 Dec;5:17995.

[11] Yu K, Ganesan K, Tan LK, Laban M, Wu J, Zhao XD, et al. A Precisely Regulated Gene Expression Cassette Potently Modulates Metastasis and Survival in Multiple Solid Cancers. PLOS Genetics. 2008 Jul;4(7):e1000129.

[12] Gorlov IP, Byun J, Zhao H, Logothetis CJ, Gorlova OY. Beyond comparing means: the usefulness of analyzing interindividual variation in gene expression for identifying genes associated with cancer development. Journal of Bioinformatics and Computational Biology. 2012 Apr;10(2):1241013.

[13] Corrada Bravo H, Pihur V, McCall M, Irizarry RA, Leek JT. Gene expression anti-profiles as a basis for accurate universal cancer signatures. BMC Bioinformatics. 2012 Dec;13(1):272.

[14] Ecker S, Pancaldi V, Rico D, Valencia A. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. Genome Medicine. 2015 Jan;7(1):8.

[15] Gorlov IP, Yang JY, Byun J, Logothetis C, Gorlova OY, Do KA, et al. How to get the most from microarray data: advice from reverse genomics. BMC Genomics. 2014 Mar;15(1):223.

[16] Dinalankara W, Corrada Bravo H. Gene expression signatures based on variability can robustly predict tumor progression and prognosis. Cancer Informatics. 2015 Jun;2015(14):71–81.

[17] Strbenac D, Mann GJ, Yang JYH, Ormerod JT. Differential distribution improves gene selection stability and has competitive classification performance for patient survival. Nucleic Acids Research. 2016 Jul;44(13):e119.

[18] Alemu EY, Carl JW, Corrada Bravo H, Hannenhalli S. Determinants of expression variability. Nucleic Acids Research. 2014 Apr;42(6):3503–3514.

[19] Prieto C, Rivas MJ, Sánchez JM, López-Fidalgo J, De Las Rivas J. Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. Bioinformatics. 2006 May;22(9):1103–1110.

[20] Ho JWK, Stefani M, Remedios D, G C, Charleston MA. Differential variability analysis of gene expression and its application to human diseases. Bioinformatics. 2008 Jul;24(13):i390–i398.

[21] Bar HY, Booth JG, Wells MT. A mixture-model approach for parallel testing for unequal variances. Statistical Applications in Genetics and Molecular Biology. 2012;11(1):1–21.

[22] Ouyang W, An Q, Zhao J, Qin H. Integrating mean and variance heterogeneities to identify differentially expressed genes. BMC Bioinformatics. 2016 Dec;17(1):497.

[23] Ran D, Daye ZJ. Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. Nucleic Acids Research. 2017 Jul;45(13):e127–e127.

[24] Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005;54(3):507–554.

[25] de Jong TV, Moshkin YM, Guryev V. Gene expression variability: the other dimension in transcriptome analysis. Physiological Genomics. 2019 Mar;51(5):145–158.

[26] Vallejos CA, Richardson S, Marioni JC. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biology. 2016;17:70.

[27] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007 Nov;23(21):2881–2887.

[28] Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013 Apr;14(2):232–243.

[29] Yu D, Huber W, Vitek O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics. 2013 May;29(10):1275–1282.

[30] Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Dec;11(1):422.

[31] van de Wiel MA, Leday GGR, Pardo L, Rue H, van der Vaart AW, van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013 Jan;14(1):113–128.

[32] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013 Jun;45(6):580–585.

[33] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan;26(1):139–140.

[34] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014 Dec;15(12):550.

[35] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. 2004;3(1):1–25.

[36] Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014 Feb;15:R29.

[37] Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics (Oxford, England). 2004 Apr;5(2):155–176.

[38] Ventrucci M, Scott EM, Cocchi D. Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation. Biostatistics. 2011 Jan;12(1):51–67.

[39] The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct;490(7418):61–70.

[40] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013 Jul;499(7456):43–49.

[41] The Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. Cell. 2014 Oct;159(3):676–690.

[42] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014 Jul;511(7511):543–550.

[43] The Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017 Jun;169(7):1327–1341.e23.

[44] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep;489(7417):519–525.

[45] The Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov;163(4):1011–1025.

[46] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul;487(7407):330–337.

[47] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer. 2018 Nov;18(11):696–705.

[48] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research. 2020 Jan;48(D1):D845–D855.

[49] Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods. 2013 Nov;10(11):1081–1082.

[50] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Research. 2019 Jan;47(D1):D590–D595.

[51] Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Research. 2017 Jan;45(D1):D877–D887.

[52] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005 Oct;102(43):15545–15550.

[53] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010 Apr;26(7):976–978.

[54] Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE. 2011 Jul;6(7):e21800.

[55] Kolodziejczyk A, Kim J, Tsang JH, Ilicic T, Henriksson J, Natarajan K, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell. 2015 Oct;17(4):471–485.

[56] Bashkeel N, Perkins TJ, Kærn M, Lee JM. Human gene expression variability and its dependence on methylation and aging. BMC Genomics. 2019 Dec;20(1):941.

[57] Dong D, Shao X, Deng N, Zhang Z. Gene expression variations are predictive for stochastic noise. Nucleic Acids Research. 2011 Jan;39(2):403–413.

[58] Ecker S, Pancaldi V, Valencia A, Beck S, Paul DS. Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity. BioEssays. 2018;40(2):1700148.

[59] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

[60] Roberts AGK. DiffDist. GitHub. 2021. Available from: `https://github.com/aedanr/DiffDist`.

[61] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

[62] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods. 2015 Feb;12(2):115–121.

[63] Soneson C. compcodeR—an R package for benchmarking differential expression methods for RNA-seq data. Bioinformatics. 2014 Sep;30(17):2517–2518.

[64] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013 Mar;14(1):91.

[65] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr;464(7289):768–772.

[66] Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, et al. Polymorphic Cis- and Trans-Regulation of Human Gene Expression. PLOS Biology. 2010 Sep;8(9):e1000480.

[67] Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nature Biotechnology. 2017 Apr;35(4):319.

[68] recount2: analysis-ready RNA-seq gene and exon counts datasets. `https://jhubiostatistics.shinyapps.io/recount/`. Accessed on 28 Nov 2019.

[69] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012 May;40(10):4288–4297.

[70] Lun ATL, Chen Y, Smyth GK. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. In: Mathé E, Davis S, editors. Statistical Genomics. vol. 1418 of Methods in Molecular Biology. New York, NY: Humana Press; 2016. p. 391–416.

[71] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 Apr;43(7):e47–e47.

[72] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010 Mar;11(3):R25.

[73] Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. Genome Biology. 2014 Sep;15:465.

[74] Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016 Jan;32(2):286–288.

[75] Frazee AC, Langmead B, Leek JT. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011 Nov;12(1):449.