Published online 14 January 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqab124

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

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

[1]School of Computer Science and Australian Artificial Intelligence Institute, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia and [2]Children's Cancer Research Unit, The Children's Hospital at Westmead, Locked Bag 4001, Westmead, NSW 2145, Australia

Received July 25, 2021; Revised November 19, 2021; Editorial Decision December 10, 2021; Accepted December 16, 2021

ABSTRACT

There is increasing evidence that changes in the variability or overall distribution of gene expression are important both in normal biology and in diseases, particularly cancer. Genes whose expression differs in variability or distribution without a difference in mean are ignored by traditional differential expression-based analyses. Using a Bayesian hierarchical model that provides tests for both differential variability and differential distribution for bulk RNA-seq data, we report here an investigation into differential variability and distribution in cancer. Analysis of eight paired tumour–normal datasets from The Cancer Genome Atlas confirms that differential variability and distribution analyses are able to identify cancer-related genes. We further demonstrate that differential variability identifies cancer-related genes that are missed by differential expression analysis, and that differential expression and differential variability identify functionally distinct sets of potentially cancer-related genes. These results suggest that differential variability analysis may provide insights into genetic aspects of cancer that would not be revealed by differential expression, and that differential distribution analysis may allow for more comprehensive identification of cancer-related genes than analyses based on changes in mean or variability alone.

INTRODUCTION

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 (1–3), development (4,5) and ageing (6–9). Changes in expression variability have also been implicated in diseases including schizophrenia (3,10) and cancer (11–14). Genes selected for differences in variability have been demonstrated to have diagnostic and prognostic potential in cancer (12–17).

Work on expression variability to date has often focused on microarray data, and variability has been assessed using empirical measures such as standard deviation (5,13,15), the coefficient of variation (CV; the ratio of the standard deviation to the mean) (3,10,11), or measures of deviation from expected expression values determined by a relationship between the mean and variability (17,18). Tests for changes in variability 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 (6,14,19–22). 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. Because of this relationship, variability in RNA-seq data is generally assessed using the dispersion rather than the variance. There is one method published to date that provides a test for differences in dispersion specifically for bulk RNA-seq data: MDSeq (23), which uses an NB generalised linear model to test for differences in mean and dispersion separately. Generalised additive models for location, scale and shape (GAMLSS) (24) form another family of regression models with potential for identifying differentially variable genes,

*To whom correspondence should be addressed. Tel: +61 421 778 363; Email: aedan.r@gmail.com

as has recently been demonstrated (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 NB distribution is completely defined by the mean and dispersion. 'Differential distribution'—any change in the distribution of expression levels between groups—can therefore be defined as a difference in either one or both of these parameters. Since there is a clear interest in identifying genes with differences in mean, and growing evidence for biological relevance of differences in 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 NB 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 BASiCS. The hierarchical model (HM) is incorporated into a hierarchical mixture model (HMM) which provides an overall test for differential distribution.

Using this model, we report here an investigation into differential variability and 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 variability identifies different sets of genes from differential expression. Using lists of genes previously identified as being related to each of these cancers, we provide a demonstration that differential variability and differential distribution analyses are able to identify cancer-related genes, and that differential variability identifies cancer-related genes that are not identified by differential expression. We further show using gene set enrichment analysis that differential variability 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, and suggest that differential distribution analysis may provide a more comprehensive way of identifying potential cancer-related genes.

MATERIALS AND METHODS

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:

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

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

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

The top level of the hierarchical model consists of hyperpriors on the prior parameters for the means and dispersions. We specify normal hyperpriors for the location parameters m_μ and m_ϕ, and gamma hyperpriors for the scale parameters v_μ and v_ϕ. 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 Markov chain Monte Carlo (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. 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 Supplementary File 1, Section S5.

Tests for differential expression and 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 (LFC) between groups by taking $\log_2\mu_{jA} - \log_2\mu_{jB}$ for each MCMC iteration, and similarly for dispersion. 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 there 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 (27) worked well to control the false discovery rate (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, 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:

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

where n is the total number of samples and n_A the number of samples in group A.

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). 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 of 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) (28,29).

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

Computational resources

Whole blood and skeletal muscle RNA-seq data from the Genotype–Tissue Expression (GTEx) project (30) (https://gtexportal.org/home/) and RNA-seq data from TCGA (https://www.cancer.gov/tcga), processed by the recount2 project (31), were downloaded from https://jhubiostatistics.shinyapps.io/recount/.

Lists of genes that have previously been identified as being related to each of the eight cancer types considered were obtained from five different databases: the Cancer Gene Census (CGC) (32), DisGeNET (33), IntOGen (34), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (35) and Malacards (36). Full details are given in Supplementary File 1, Section S7. The lists of genes obtained from each database are given in Supplementary File 3, and the compiled lists of cancer-related genes in Supplementary File 4.

Datasets

Simulated data. Data was simulated using the compcodeR R (37)/Bioconductor (38) package (39). 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 (40,41), and, for differentially expressed genes, multiplies or divides means by a factor of 1.5 + x, where x is 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 expression 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.

Table 1. TCGA tumour–normal sample pairs used

TCGA ID	TCGA project name	Sample pairs	Ref.
BRCA	Breast invasive adenocarcinoma	182	(86)
KIRC	Kidney renal clear cell carcinoma	144	(87)
THCA	Thyroid carcinoma	98	(88)
LUAD	Lung adenocarcinoma	86	(89)
LIHC	Liver hepatocellular carcinoma	98	(90)
LUSC	Lung squamous cell carcinoma	92	(91)
PRAD	Prostate adenocarcinoma	92	(92)
COAD	Colon adenocarcinoma	78	(93)

Artificially introducing differential distributions in GTEx data. Gene-level counts for the GTEx data were obtained using the recount R/Bioconductor 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. Full details are given in Supplementary File 1, Section S6.

TCGA data. RNA-seq data from TCGA were downloaded and processed as for the GTEx data. 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 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.

Comparative assessment of model performance

Differential dispersion performance using HM was compared against MDSeq and GAMLSS, and differential expression performance was compared against edgeR

(42–44), DESeq2 (45), limma–voom (46,47), baySeq (48) 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 (49) and DESeq2's normalisation. TMM was used for all subsequent analyses.

The performance of the HMM differential distribution test was compared against diffVar (50,51) and a naive method combining the results from separate differential expression and differential dispersion tests. Differential expression was assessed using edgeR for the simulated data and limma–voom for the GTEx data as these were the best-performing tests in each situation, and differential dispersion using HM. Results from the two tests were combined by taking the smaller of the *P*-value or posterior probability of no difference between groups 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 Supplementary File 5, and can also be found at https://github.com/aedanr/DiffDist.

Analysis of tumour–normal comparisons

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. Where genes were ranked equally by the HM differential dispersion test, the LFC in dispersion was used to separate them, genes with a greater change in dispersion being ranked higher.

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 was performed 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 Gene Ontology (GO) terms in each of the three ontologies (biological process, molecular function and cellular component), with terms with less than two or more than six 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, semantic 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.

Statistical analyses

Relative performances of differential expression, dispersion and distribution methods were informally assessed, primarily using FDR curves: plots of FDR against the number of discoveries as the threshold for declaring a discovery is varied. The FDR and number of discoveries were averaged over the 50 (for simulated data) or 10 (for GTEx data) datasets used for each comparison. Other measures of performance used were boxplots of area under the receiver operating characteristic curve (AUC), FDR, and true positive rate (TPR; also known as sensitivity, power or recall).

Wilcoxon rank-sum tests were used to test whether differential expression and differential dispersion analyses ranked cancer-related genes above other genes. This method allows the overall ranking of genes to be assessed, avoiding the need to rely on an arbitrary threshold to declare a gene as differentially expressed or dispersed. Tests were performed using the wilcox.test function in R.

Tests for correlation between gene lists identified by different methods were performed using the cor.test function in R, with method='spearman' and alternative='less', to test the null hypothesis of zero or positive correlation against the alternative hypothesis of negative correlation.

RESULTS

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

We perform differential dispersion and distribution analyses using a Bayesian hierarchical model for RNA-seq data based on the NB distribution. An MCMC algorithm provides samples from the posterior distributions of the mean and dispersion for each gene. 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 (55–57), and are at the core of Bayesian methods for differential expression analysis (48,58), but are not used in MDSeq, the only previously published method for differential dispersion in bulk RNA-seq data. Figure 1A shows a schematic out-

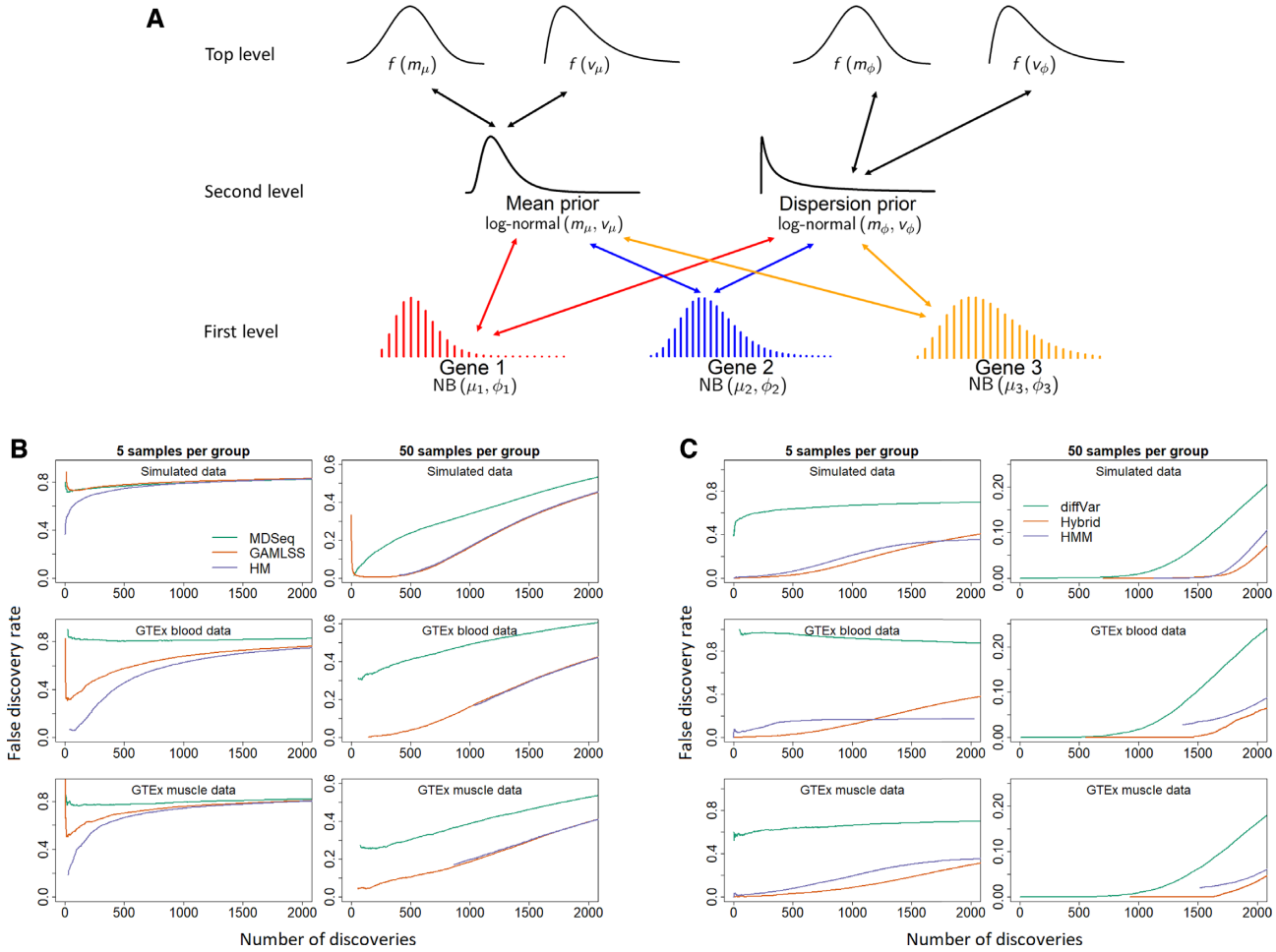

Figure 1. Hierarchical model for differential expression, dispersion and distribution. (**A**) Schematic illustration of the model. The count for each gene is assumed to follow a negative binomial distribution (first level), with mean and dispersion parameters that are modelled as random realisations from log-normal prior distributions (second level). The parameters of the log-normal priors are themselves modelled as random realisations from a set of fixed hyperpriors (top level). The mean and dispersion parameters for each gene are therefore related to the parameters for all other genes; this is the information sharing property of the hierarchical model that allows stable parameter estimates to be obtained even with small sample sizes. (**B**) False discovery curves for detection of differential dispersion using the hierarchical model (HM), MDSeq and GAMLSS, averaged over 50 simulated datasets (top) and 10 datasets generated from GTEx blood (middle) and muscle (bottom) data, with no minimum log fold change. (**C**) False discovery curves for differential distribution using the hierarchical mixture model (HMM), diffVar and a hybrid method combining separate tests for differential expression and differential dispersion, averaged over 50 simulated datasets (top) and 10 datasets generated from GTEx blood (middle) and muscle (bottom) data.

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

Model performance was assessed using two approaches: testing on simulated data, and on real RNA-seq data with artificially-induced differences in expression. Using simulated data allows assessment of the performance of a model with reference to a known ground truth, but has the disadvantage that the 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 NB 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. GTEx (30) (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 groups. Using data from two tissues—whole blood and skeletal muscle—we generated datasets with known levels of differential expression and/or dispersion by randomly selecting samples and altering the counts for a proportion of genes in one group to reflect a change in mean and/or dispersion, under the assumption that the counts follow an NB distribution. Full details are given in Materials and Methods.

The HM differential expression test was compared against edgeR, DESeq2, limma–voom, baySeq—which also uses a Bayesian hierarchical model—and MDSeq. HM provided similar performance to the best of the other methods (Supplementary File 1, Section S1).

The HM differential dispersion test was compared against MDSeq and GAMLSS. False discovery curves (Figure 1B) show that HM consistently outperforms MDSeq, and outperforms GAMLSS with five samples per group while providing similar performance with 50 samples per group. False discovery curves for 2, 10 and 20 samples per group (Supplementary File 1, Figure S3) show similar patterns, as do boxplots of FDR and sensitivity (Supplementary File 1, Figure S4). A weakness of HM is that, since inference is based on a posterior sample, there is a limit to the degree to which the most significantly altered genes can be separated. This is evident particularly for the GTEx data with 50 samples per group. This issue can be mitigated by using additional information to rank genes, such as the magnitude of the change in the parameter of interest. This method is used in the analysis of TCGA data below.

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

Along with the HMM differential distribution test, an alternative method of testing for differential distribution was also considered: a 'hybrid' method combining separate tests for differential expression and differential dispersion (see Materials and Methods). While there are no published methods for differential distribution for RNA-seq data, another alternative was considered taking advantage of the fact that under the NB model, the variance is function of 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 1C) 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 identify up to around 1900 differentially distributed genes while maintaining the FDR below 0.05. False discovery curves for 2, 10 and 20 samples per group are shown in Supplementary File 1, Figure S7, and boxplots of FDR and sensitivity in Supplementary File 1, Figure S8.

Differential variability and differential distribution identify cancer-related genes

To test whether differential variability and distribution analyses are able to identify cancer-related genes, we applied tests of differential expression, dispersion and distribution to paired normal–tumour RNA-seq data from eight cancer types from TCGA, and compiled lists of genes that have previously been identified as being related to each cancer from five different databases: CGC, DisGeNET, IntOGen, KEGG and Malacards. 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 differential distribution using both 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, and for three out of eight by differential dispersion: thyroid carcinoma, lung adenocarcinoma and lung squamous cell carcinoma. For differential distribution, 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 used to identify cancer-related genes. Given this, it is particularly striking that for three of the eight cancer types, there are similar levels of evidence for association with cancer-related genes for differential dispersion and differential expression.

Differential variability identifies different sets of genes from differential expression

We next asked whether differential variability and differential expression identify different sets of genes. This is a critical question, and one that has not been formally tested in previous studies: if differential variability 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. 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 2A for lung and prostate adenocarcinoma, and in Supplementary 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. Supplementary 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 differ-

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}

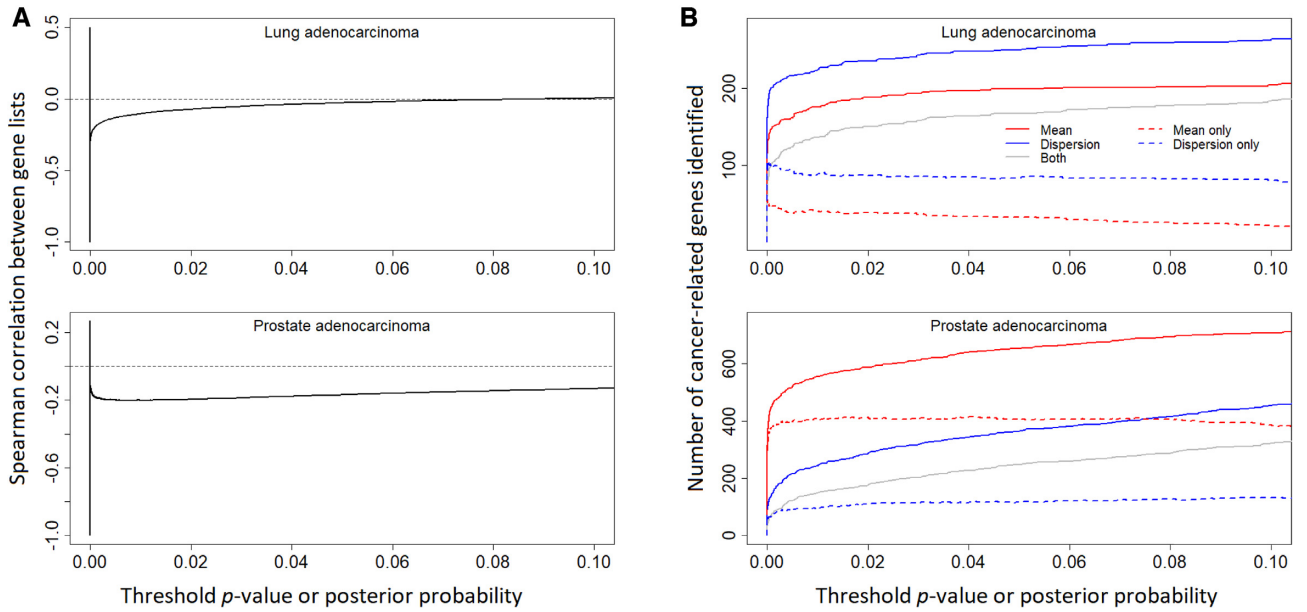

Figure 2. Differential expression and differential dispersion analyses independently identify cancer-related genes. (**A**) Spearman correlation between gene lists for differential expression and differential dispersion for TCGA lung and prostate adenocarcinoma data, with varying threshold for calling 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).

ential expression and differential dispersion identify distinct sets of genes.

These results and those in the previous section demonstrate that, at least for some cancer types, differential variability 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 variability identify different sets of cancer-related genes.

Figure 2B 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. Results for all eight cancers are given in Supplementary File 1, Figure S9. 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 variability should be relied on alone to identify cancer-related genes, and that overall assessment of differential distribution can more comprehensively identify these genes.

Differential expression and differential variability identify genes in different functional categories

To further investigate the types of genes identified by differential expression and differential variability, we performed gene set enrichment analysis. We looked for enriched terms in each of the three GO ontologies—biological process, molecular function and cellular component—among genes

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

ranked by differential expression and differential dispersion.

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 Supplementary File 1, Section S4, and in Supplementary File 6.

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. To assess the consistency of the GO terms identified using differential variability analysis, we repeated this analysis using GAMLSS instead of the hierarchical model. These results are given in Supplementary File 6, and show very high concordance between the two differential variability methods.

Table 5 gives a deeper insight into the top genes driving the GSEA results for differential variability analysis on lung adenocarcinoma. Differential variability and differential expression results are shown for the top 20 ranked genes among the top 10 GO terms in each ontology. Twelve of these genes code for histone proteins, and others are also involved in basic cellular processes, including protein degradation (CAND1 (59)) and transcriptional regulation (TAF7L (60,61)). Several of these genes have altered expression or demonstrated functional roles in cancers, including non-small cell lung cancer (NSCLC) and lung adenocarcinoma specifically: expression of CAND1 is altered in NSCLC (62); TAF7L is mutated in colorectal cancer (63) and has reduced expression in acute myeloid

Table 5. Differential dispersion and expression results for the top 20 genes ranked by differential dispersion between normal and lung adenocarcinoma tissue, considering only genes that are in the top 10 enriched GO terms in each of the three ontologies

	Differential dispersion			Differential expression		
Gene	LFC	FDR	Rank	LFC	FDR	Rank
CAND1	6.2	2×10^{-4}	24	0.7	3×10^{-6}	6547
H4C5	6.1	2×10^{-4}	26	2.2	2×10^{-5}	3434
H4C3	5.6	2×10^{-4}	55	1.3	0.04	9400
H2BC17	5.5	2×10^{-4}	70	1.9	2×10^{-6}	3294
TAF7L	5.5	2×10^{-4}	72	1.6	3×10^{-5}	4490
H4C4	5.4	2×10^{-4}	78	1.3	5×10^{-3}	7811
H2CA12	5.2	2×10^{-4}	90	1.5	4×10^{-4}	5615
H1-3	5.2	2×10^{-4}	94	2.9	6×10^{-6}	2391
SYP	5.1	2×10^{-4}	98	−1.8	9×10^{-12}	1837
PCCA	5.1	2×10^{-4}	101	−0.2	0.15	16721
DDHD2	5.0	2×10^{-4}	109	0.0	0.91	22 439
H2BC7	5.0	2×10^{-4}	112	2.6	1×10^{-9}	1529
H2BC15	4.9	2×10^{-4}	125	1.3	4×10^{-6}	4731
H4C8	4.9	2×10^{-4}	135	1.2	3×10^{-5}	5500
FOLR1	4.9	2×10^{-4}	138	−2.9	3×10^{-8}	1594
H2BC13	4.8	2×10^{-4}	142	1.8	8×10^{-5}	4422
H2CA21	4.8	2×10^{-4}	143	1.6	2×10^{-3}	6439
SFTPC	4.8	2×10^{-4}	144	−8.4	5×10^{-16}	20
OS9	4.8	2×10^{-4}	154	0.0	0.56	21103
AKR1C1	4.8	2×10^{-4}	157	−0.8	0.13	12784

Table 6. Differential dispersion and expression results for the top 20 genes ranked by differential dispersion between normal and lung adenocarcinoma tissue

	Differential dispersion			Differential expression		
Gene	LFC	FDR	Rank	LFC	FDR	Rank
WDR74	7.6	2×10^{-4}	1	1.2	6×10^{-4}	6614
RNU5A-1	7.0	2×10^{-4}	2	0.8	0.08	11 976
DEFA8P	6.8	2×10^{-4}	3	0.1	0.19	18 199
SNORA53	6.7	2×10^{-4}	4	0.7	0.06	12 371
SNORD13	6.7	2×10^{-4}	5	0.6	0.04	12 142
SNORA49	6.7	2×10^{-4}	6	0.8	0.01	9958
RNU11	6.6	2×10^{-4}	7	0.4	0.17	17 823
RN7SKP90	6.6	2×10^{-4}	8	0.7	0.21	14 248
RPL7A pseudogene	6.6	2×10^{-4}	9	0.3	0.70	19 181
ST8SIA6-AS1	6.6	2×10^{-4}	10	−0.3	0.73	19 542
RP11-388K2.1	6.6	2×10^{-4}	11	−0.4	0.28	15 904
DEFA9P	6.5	2×10^{-4}	12	0.3	0.13	15 358
Y RNA	6.5	2×10^{-4}	13	1.0	0.02	9659
RNU11	6.5	2×10^{-4}	14	0.1	0.44	19 217
RNVU1-2	6.5	2×10^{-4}	15	0.7	0.06	11 977
SCARNA1	6.4	2×10^{-4}	16	0.6	0.04	12 104
7SK	6.4	2×10^{-4}	17	0.8	0.01	10 266
RN7SL296P	6.4	2×10^{-4}	18	0.9	0.03	10 479
RNU12	6.4	2×10^{-4}	19	0.5	0.47	16 918
RNU5B-1	6.4	2×10^{-4}	20	1.1	0.06	10651

leukaemia (64); FOLR1 has altered expression in lung adenocarcinoma (65,66); OS9 is overexpressed in sarcomas (67); AKR1C1 promotes metastasis in NSCLC (68). Each of these genes is ranked among the top 200 for differential variability, but with the exception of SFTPC, none are ranked within the top 1000 for differential expression.

Beyond the gene set enrichment analysis, Table 6 shows the top 20 genes identified by differential variability analysis overall. Some of these are pseudogenes, but the majority are non-protein coding genes—small nuclear RNAs, small nucleolar RNAs or long noncoding RNAs—with the exception being WDR74, a 60S ribosome assembly factor which has been shown to promote lung cancer growth and metastasis (69). Of the noncoding genes among the top 20, SNORD13 (70), ST8SIA6-AS1 (71,72) and Y RNA (73) have been shown to have altered expression in NSCLC or in lung adenocarcinoma specifically.

DISCUSSION

Identifying differentially variable or distributed genes from 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 NB 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. However, GAMLSS also outperformed MDSeq, and gave nearly identical performance to the hierarchical model on larger sample sizes. Which method is preferred for differential dispersion analysis may depend on the study design. In particular, the hierarchical model allows for testing at a minimum LFC, which GAMLSS does not.

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, false discovery curves (Figure 1C) show that, for a given number of discoveries, combining separate tests generally resulted in fewer false discoveries. This may at least in part reflect a general 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 rarely discussed in the literature, but affects any method that uses posterior sampling, such as BASiCS, when applied to data with a large number of parameters such as genome-wide expression data. It is evident in the false discovery curves for both the differential dispersion and differential distribution tests from the hierarchical model, for which there are sometimes several hundred top-ranked

genes which cannot be separated (Figure 1B and C). 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 NB 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 (46) suggests that such an approach is worth exploring. This type of approach would also greatly increase computational speed, which is another limitation of posterior inference-based methods. On the lung adenocarcinoma data, for example (23 416 genes after filtering and 43 samples per group), the hierarchical model takes around three hours to run on a 2.7 GHz, four core processor with 16GB RAM, compared to around 12 min for GAMLSS and 4 min for MDSeq. While our view is that processing time is not a critical consideration for a method that will typically be run only a very small number of times per experiment, a decrease from several hours to several minutes or less would be a vast improvement. Another way of avoiding the issue of returning many equally-ranked genes is to incorporate information on the magnitude of changes in mean or dispersion into gene rankings, as was done for the analysis of the TCGA data. The issue may also be mitigated somewhat by performing tests at a minimum LFC, as is evident from Supplementary File 1, Figure S5.

Small sample sizes—from 2 to 50 samples per group—were used for the methods comparisons in this study. Extreme small sample sizes were included in order to test the limits of differential variability detection, and while the results show that the hierarchical model has advantages over methods that treat each gene independently for very small sample sizes, this is not enough to 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 is best reserved for situations where larger numbers of samples are available.

Differential variability and distribution in cancer

Expression variability has been shown to be a trait that is under genetic control (74), as well as being influenced by chromatin structure and promoter architecture (18,75). Changes in variability define different pluripotent cell states and developmental stages (4,5,76), and cell-to-cell expression variability has been found to be correlated with divergence in transcriptional responses to immune stimuli between species (77), suggesting a link between variability at the single-cell level and species-level evolution of transcriptional regulation. Given these findings, a role for changes in expression variability in cancer—whether cause, effect, or both—should not be surprising.

While there is strong evidence in the existing literature that differentially variable genes are important in cancer, this work provides the first clear demonstration of the inverse: that differential variability can be used to identify cancer-related genes. Importantly, these results also show that differential expression and differential variability identify distinct sets of cancer-related genes. Differential variability 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.

The results for lung adenocarcinoma in Table 6 provide a concrete demonstration of the ability of differential variability analysis to identify potential cancer-related genes that may be missed by differential expression analysis. Among these genes are several that have previously been shown to have altered expression patterns in NSCLC or in lung adenocarcinoma specifically, and which would not have been identified based on differential expression alone for the dataset considered here. Of particular interest is the long noncoding RNA ST8SIA6-AS1. This gene has been associated with multiple cancers, including lung adenocarcinoma (71,72), but whereas it has previously been found to be overexpressed in cancer, differential expression analysis in this study found no change in mean expression levels between normal and paired tumour samples from TCGA.

As well as identifying different genes, gene set enrichment analysis showed that differential expression and differential variability identify functionally distinct sets of genes. GO terms commonly enriched among differentially expressed genes are often related to cell structure and migration (for example extracellular structure organisation), signalling (e.g. G protein-coupled receptor activity) or immune system functions (e.g. antigen binding). In contrast, GO terms relating to basic cellular processes such as transcription and translation are frequently enriched among genes with differences 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. Top-ranked genes among the most highly enriched GO categories for lung adenocarcinoma exemplify this pattern, including genes coding for histone components and with roles in protein degradation, endoplasmic reticulum–Golgi transport, and regulation of transcription (Table 5). Several of these genes have previously been linked with lung adenocarcinoma or other cancers, and, importantly, most were not ranked highly enough by differential expression analysis to be identified as potentially cancer-related from the TCGA data used here.

Functions relating to transcription and translation have previously been identified among low-variability genes (5,76,78). 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 variability 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 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 variability between normal and tumour tissues. Decreased variability 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 (79,80), 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 (5,76) and at the tissue level (18,78) 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 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.

One potential way of distinguishing between changes in variability within a single tissue or population of cells and apparent changes in variability caused by tissue heterogeneity is to assess the shape of gene expression distributions. In a tumour predominantly composed of a mixture of two cell types, for example, a gene that is differentially expressed between these cell types will have a bimodal distribution, whereas a gene whose expression variability is uniformly increased compared to normal tissue will still have a unimodal distribution, but with a wider spread. There is a similar distinction to be made at the individual level, between a uniform change in variability among a group of individuals, and an increase in variability arising from changes in mean expression between sub-groups. The methods used in this study cannot distinguish between these two situations, but there are methods specifically designed to identify changes in gene expression distributions arising from mixtures of cell types (81,82).

Assessment of some of the top-ranked genes identified from differential dispersion analysis for lung adenocarcinoma gives some clue as to possible sources of the detected changes in variability. For example, changes in expression of FOLR1, found to have increased variability in lung adenocarcinoma in this study (Table 5), have previously been found to depend on factors such as smoking history, tumour differentiation and stage, and mutations in other genes (65,66). It is possible, therefore, that the increased expression variability observed here could be a result of changes in mean expression levels among a subset of tumour samples. CAND1 had increased variability in this study, along with increased mean expression in tumours (Table 5). Increased expression of CAND1 has previously been found in NSCLC (62). CAND1 is a regulator of cullin-RING ligases (CRLs), and conflicting reports on CAND1 as a promoter or inhibitor of CRLs have led to suggestions that there is an optimal level of CAND1 required for appropriate protein degradation driven by CRLs (83,84). It seems plausible, then, that changes in the expression of CAND1 in either direction could be associated with cancer phenotypes. Yet another possibility is suggested by the abundance of histone genes among the differentially dispersed genes: several of these genes are located in a cluster on chromosome 6p22.1, which as been identified as a region frequently amplified in NSCLC (85). Increased variability in the expression of these genes could, therefore, be equally plausibly a result of chromosomal instability or a cause of disruption of normal patterns of transcription. It seems likely that differentially variable genes in cancer could variously be artifacts of tumour heterogeneity, or drivers or consequences of cellular dysfunction.

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 variability alone (17). Differential distribution analysis may therefore be preferable to separate tests of differential expression and variability both in terms of identifying potential cancer-related genes, and for studies into diagnostic or prognostic prediction.

DATA AVAILABILITY

R code used to generate synthetic data, introduce changes in mean and dispersion into real RNA-seq data, and process the GTEx and TCGA data is given in Supplementary File 7. Code used for analyses is given in Supplementary Files 2 and 5, and at https://github.com/aedanr/DiffDist (DOI: 10.5281/zenodo.4544153).

SUPPLEMENTARY DATA

Supplementary data are available at NARGAB online.

ACKNOWLEDGEMENTS

This research is supported by a NSW Health PhD Scholarship and an Australian Government Research Training Program Scholarship (both AGKR). The authors would like to thank Thomas Lysaght and George Mundackal for their contributions to optimising the MCMC code, and two anonymous reviewers, whose comments and suggestions greatly improved the paper. 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

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Cheung,V.G., Conlin,L.K., Weber,T.M., Arcaro,M., Jen,K.-Y., Morley,M. and Spielman,R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.
2. Komurov,K. and Ram,P.T. (2010) Patterns of human gene expression variance show strong associations with signaling network hierarchy. *BMC Syst. Biol.*, **4**, 154.
3. Mar,J.C., Matigian,N.A., Mackay-Sim,A., Mellick,G.D., Sue,C.M., Silburn,P.A., McGrath,J.J., Quackenbush,J. and Wells,C.A. (2011) Variance of gene expression identifies altered network constraints in neurological disease. *PLoS Genet.*, **7**, e1002207.
4. Kalmar,T., Lim,C., Hayward,P., Muñoz-Descalzo,S., Nichols,J., Garcia-Ojalvo,J. and Arias,A.M. (2009) Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLOS Biol.*, **7**, e1000149.
5. Hasegawa,Y., Taylor,D., Ovchinnikov,D.A., Wolvetang,E.J., Torrenté,L.D. and Mar,J.C. (2015) Variability of gene expression identifies transcriptional regulators of early human embryonic development. *PLOS Genet.*, **11**, e1005428.
6. Bahar,R., Hartmann,C.H., Rodriguez,K.A., Denny,A.D., Busuttil,R.A., Dollé,M.E.T., Calder,R.B., Chisholm,G.B., Pollock,B.H., Klein,C.A. *et al.* (2006) Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, **441**, 1011–1014.
7. Somel,M., Khaitovich,P., Bahn,S., Pääbo,S. and Lachmann,M. (2006) Gene expression becomes heterogeneous with age. *Curr. Biol.*, **16**, R359–R360.
8. Li,Z., Wright,F.A. and Royland,J. (2009) Age-dependent variability in gene expression in male Fischer 344 rat retina. *Toxicol. Sci.*, **107**, 281–292.
9. Viñuela,A., Brown,A.A., Buil,A., Tsai,P.-C., Davies,M.N., Bell,J.T., Dermitzakis,E.T., Spector,T.D. and Small,K.S. (2018) Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Hum. Mol. Genet.*, **27**, 732–741.
10. Zhang,F., Shugart,Y.Y., Yue,W., Cheng,Z., Wang,G., Zhou,Z., Jin,C., Yuan,J., Liu,S. and Xu,Y. (2015) Increased variability of genomic transcription in schizophrenia. *Sci. Rep.-UK*, **5**, 17995.
11. Yu,K., Ganesan,K., Tan,L.K., Laban,M., Wu,J., Zhao,X.D., Li,H., Leung,C.H.W., Zhu,Y., Wei,C.L. *et al.* (2008) A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet.*, **4**, e1000129.
12. Gorlov,I.P., Byun,J., Zhao,H., Logothetis,C.J. and Gorlova,O.Y. (2012) Beyond comparing means: the usefulness of analyzing interindividual variation in gene expression for identifying genes associated with cancer development. *J. Bioinf. Comput. Biol.*, **10**, 1241013.
13. Corrada Bravo,H., Pihur,V., McCall,M., Irizarry,R.A. and Leek,J.T. (2012) Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics*, **13**, 272.
14. Ecker,S., Pancaldi,V., Rico,D. and Valencia,A. (2015) Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. *Genome Med.*, **7**, 8.
15. Gorlov,I.P., Yang,J.-Y., Byun,J., Logothetis,C., Gorlova,O.Y., Do,K.-A. and Amos,C. (2014) How to get the most from microarray data: advice from reverse genomics. *BMC Genomics*, **15**, 223.
16. Dinalankara,W. and Corrada Bravo,H. (2015) Gene expression signatures based on variability can robustly predict tumor progression and prognosis. *Cancer Informatics*, **2015**, 71–81.
17. Strbenac,D., Mann,G.J., Yang,J.Y.H. and Ormerod,J.T. (2016) Differential distribution improves gene selection stability and has competitive classification performance for patient survival. *Nucleic Acids Res.*, **44**, e119.
18. Alemu,E.Y., Carl,J.W., Corrada Bravo,H. and Hannenhalli,S. (2014) Determinants of expression variability. *Nucleic Acids Res.*, **42**, 3503–3514.
19. Prieto,C., Rivas,M.J., Sánchez,J.M., López-Fidalgo,J. and De Las Rivas,J. (2006) Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics*, **22**, 1103–1110.
20. Ho,J.W.K., Stefani,M., Remedios,C.G. and and Charleston,M.A. (2008) Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, **24**, i390–i398.
21. Bar,H.Y., Booth,J.G. and Wells,M.T. (2012) A mixture-model approach for parallel testing for unequal variances. *Stat. Appl. Genet. Mol.*, **11**, 8.
22. Ouyang,W., An,Q., Zhao,J. and Qin,H. (2016) Integrating mean and variance heterogeneities to identify differentially expressed genes. *BMC Bioinformatics*, **17**, 497.
23. Ran,D. and Daye,Z.J. (2017) Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Res.*, **45**, e127.
24. Rigby,R.A. and Stasinopoulos,D.M. (2005) Generalized additive models for location, scale and shape. *J. R. Stat. Soc. C Appl.*, **54**, 507–554.
25. de Jong,T.V., Moshkin,Y.M. and Guryev,V. (2019) Gene expression variability: the other dimension in transcriptome analysis. *Physiol. Genomics*, **51**, 145–158.
26. Vallejos,C.A., Richardson,S. and Marioni,J.C. (2016) Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.*, **17**, 70.
27. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Met.*, **57**, 289–300.
28. Newton,M.A., Noueiry,A., Sarkar,D. and Ahlquist,P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.
29. Ventrucci,M., Scott,E.M. and Cocchi,D. (2011) Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation. *Biostatistics*, **12**, 51–67.
30. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N., Foster,B. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet*, **45**, 580–585.
31. Collado-Torres,L., Nellore,A., Kammers,K., Ellis,S.E., Taub,M.A., Hansen,K.D., Jaffe,A.E., Langmead,B. and Leek,J.T. (2017) Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, **35**, 319.
32. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
33. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.

34. Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Tamborero,D., Schroeder,M.P., Jene-Sanz,A., Santos,A. and Lopez-Bigas,N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.
35. Kanehisa,M., Sato,Y., Furumichi,M., Morishima,K. and Tanabe,M. (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
36. Rappaport,N., Twik,M., Plaschkes,I., Nudel,R., Iny Stein,T., Levitt,J., Gershoni,M., Morrey,C.P., Safran,M. and Lancet,D. (2017) MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.*, **45**, D877–D887.
37. R Core Team (2017) In: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
38. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Corrada Bravo,H., Davis,S., Gatto,L., Girke,T. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.
39. Soneson,C. (2014) compcodeR—an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*, **30**, 2517–2518.
40. Pickrell,J.K., Marioni,J.C., Pai,A.A., Degner,J.F., Engelhardt,B.E., Nkadori,E., Veyrieras,J.-B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.
41. Cheung,V.G., Nayak,R.R., Wang,I.X., Elwyn,S., Cousins,S.M., Morley,M. and Spielman,R.S. (2010) Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.*, **8**, e1000480.
42. 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.
43. 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.
44. Lun,A. T.L., Chen,Y. and Smyth,G.K. (2016) It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. *Methods Mol. Biol.*, **1418**, 391–416.
45. 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.
46. 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.
47. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
48. Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
49. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
50. Phipson,B. and Oshlack,A. (2014) DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol.*, **15**, 465.
51. Phipson,B., Maksimovic,J. and Oshlack,A. (2016) missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*, **32**, 286–288.
52. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
53. Yu,G., Li,F., Qin,Y., Bo,X., Wu,Y. and Wang,S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.
54. Supek,F., Bošnjak,M., Škunca,N. and Šmuc,T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
55. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
56. 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.
57. Yu,D., Huber,W. and Vitek,O. (2013) Shrinkage estimation of dispersion in negative binomial models for RNA-seq experiments with small sample size. *Bioinformatics*, **29**, 1275–1282.
58. van de Wiel,M.A., Leday,G.G.R., Pardo,L., Rue,H., van der Vaart,A.W. and van Wieringen,W.N. (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, **14**, 113–128.
59. Zheng,J., Yang,X., Harrell,J.M., Ryzhikov,S., Shim,E.H., Lykke-Andersen,K., Wei,N., Sun,H., Kobayashi,R. and Zhang,H. (2002) CAND1 binds to unneddylated CUL1 and regulates the formation of SCF ubiquitin E3 ligase complex. *Mol. Cell*, **10**, 1519–1526.
60. Pointud,J.-C., Mengus,G., Brancorsini,S., Monaco,L., Parvinen,M., Sassone-Corsi,P. and Davidson,I. (2003) The intracellular localisation of TAF7L, a paralogue of transcription factor TFIID subunit TAF7, is developmentally regulated during male germ-cell differentiation. *J. Cell Sci.*, **116**, 1847–1858.
61. Zhou,H., Kaplan,T., Li,Y., Grubisic,I., Zhang,Z., Wang,P.J., Eisen,M.B. and Tjian,R. (2013) Dual functions of TAF7L in adipocyte differentiation. *eLife*, **2**, e00170.
62. Kang,M., Li,Y., Zhao,Y., He,S. and Shi,J. (2018) miR-33a inhibits cell proliferation and invasion by targeting CAND1 in lung cancer. *Clin. Transl. Oncol.*, **20**, 457–466.
63. Oh,H.R., An,C.H., Yoo,N.J. and Lee,S.H. (2015) Frameshift mutations of TAF7L gene, a core component for transcription by RNA polymerase II, in colorectal cancers. *Pathol. Oncol. Res*, **21**, 849–850.
64. Yazarloo,F., Shirkoohi,R., Mobasheri,M.B., Emami,A. and Modarressi,M.H. (2013) Expression analysis of four testis-specific genes AURKC, OIP5, PIWIL2 and TAF7L in acute myeloid leukemia: a gender-dependent expression pattern. *Med. Oncol.*, **30**, 368.
65. Iwakiri,S., Sonobe,M., Nagai,S., Hirata,T., Wada,H. and Miyahara,R. (2008) Expression status of folate receptor alpha is significantly correlated with prognosis in non-small-cell lung cancers. *Ann. Surg. Oncol.*, **15**, 889–99.
66. Nunez,M.I., Behrens,C., Woods,D.M., Lin,H., Suraokar,M., Kadara,H., Hofstetter,W., Kalhor,N., Lee,J.J., Franklin,W. *et al.* (2012) High expression of folate receptor alpha in lung cancer correlates with adenocarcinoma histology and EGFR mutation. *J. Thorac. Oncol.*, **7**, 833–840.
67. Su,Y.A., Hutter,C.M., Trent,J.M. and Meltzer,P.S. (1996) Complete sequence analysis of a gene (OS-9) ubiquitously expressed in human tissues and amplified in sarcomas. *Mol. Carcinogen.*, **15**, 270–275.
68. Zhu,H., Chang,L.-L., Yan,F.-J., Hu,Y., Zeng,C.-M., Zhou,T.-Y., Yuan,T., Ying,M.-D., Cao,J., He,Q.-J. *et al.* (2018) AKR1C1 activates STAT3 to promote the metastasis of non-small cell lung cancer. *Theranostics*, **8**, 676–692.
69. Li,Y., Chen,F., Shen,W., Li,B., Xiang,R., Qu,L., Zhang,C., Li,G., Xie,H., Katanaev,V.L. *et al.* (2020) WDR74 induces nuclear beta-catenin accumulation and activates Wnt-responsive genes to promote lung cancer growth and metastasis. *Cancer Lett.*, **471**, 103–115.
70. Liao,J., Yu,L., Mei,Y., Guarnera,M., Shen,J., Li,R., Liu,Z. and Jiang,F. (2010) Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. *Mol. Cancer*, **9**, 198.
71. Cao,Q., Yang,W., Ji,X. and Wang,W. (2020) Long non-coding RNA ST8SIA6-AS1 promotes lung adenocarcinoma progression through sponging miR-125a-3p. *Front. Genet.*, **11**, 597795.
72. Luo,M.-L., Li,J., Shen,L., Chu,J., Guo,Q., Liang,G., Wu,W., Chen,J., Chen,R. and Song,E. (2020) The role of APAL/ST8SIA6-AS1 lncRNA in PLK1 activation and mitotic catastrophe of tumor cells. *J. Natl. Cancer Inst.*, **112**, 356–368.
73. Christov,C.P., Trivier,E. and Krude,T. (2008) Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Brit. J. Cancer*, **98**, 981–988.
74. Ansel,J., Bottin,H., Rodriguez-Beltran,C., Damon,C., Nagarajan,M., Fehrmann,S., François,J. and Yvert,G. (2008) Cell-to-cell stochastic variation in gene expression is a complex genetic trait. *PLOS Genet.*, **4**, e1000049.

75. Tirosh,I. and Barkai,N. (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome Res.*, **18**, 1084–1091.
76. Kolodziejczyk,A.A., Kim,J.K.., Tsang,J.C.H., Ilicic,T., Henriksson,J., Natarajan,K.N., Tuck,A.C., Gao,X., Bühler,M., Liu,P. *et al.* (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, **17**, 471–485.
77. Hagai,T., Chen,X., Miragaia,R.J., Rostom,R., Gomes,T., Kunowska,N., Henriksson,J., Park,J.-E., Proserpio,V., Donati,G. *et al.* (2018) Gene expression variability across cells and species shapes innate immunity. *Nature*, **563**, 197–202.
78. Bashkeel,N., Perkins,T.J., Kærn,M. and Lee,J.M. (2019) Human gene expression variability and its dependence on methylation and aging. *BMC Genomics*, **20**, 941.
79. Dong,D., Shao,X., Deng,N. and Zhang,Z. (2011) Gene expression variations are predictive for stochastic noise. *Nucleic Acids Res.*, **39**, 403–413.
80. Ecker,S., Pancaldi,V., Valencia,A., Beck,S. and Paul,D.S. (2018) Epigenetic and transcriptional variability shape phenotypic plasticity. *Bioessays*, **40**, 1700148.
81. Korthauer,K.D., Chu,L.-F., Newton,M.A., Li,Y., Thomson,J., Stewart,R. and Kendziorski,C. (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**, 222.
82. Zhang,Y., Wan,C., Wang,P., Chang,W., Huo,Y., Chen,J., Ma,Q., Cao,S. and Zhang,C. (2019) M3S: a comprehensive model selection for multi-modal single-cell RNA sequencing data. *BMC Bioinformatics*, **20**, 672.
83. Wu,S., Zhu,W., Nhan,T., Toth,J.I., Petroski,M.D. and Wolf,D.A. (2013) CAND1 controls in vivo dynamics of the cullin 1-RING ubiquitin ligase repertoire. *Nat. Commun.*, **4**, 1642.
84. Straube,R., Shah,M., Flockerzi,D. and Wolf,D.A. (2017) Trade-off and flexibility in the dynamic regulation of the cullin-RING ubiquitin ligase repertoire. *PLoS Comput. Biol.*, **13**, e1005869.
85. Lo,F.-Y., Chang,J.-W., Chang,I.-S., Chen,Y.-J., Hsu,H.-S., Huang,S.-F.K., Tsai,F.-Y., Jiang,S.S., Kanteti,R. *et al.* (2012) The database of chromosome imbalance regions and genes resided in lung cancer from Asian and Caucasian identified by array-comparative genomic hybridization. *BMC Cancer*, **12**, 235.
86. The Cancer Genome Atlas Research Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
87. The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**, 43–49.
88. The Cancer Genome Atlas Research Network (2014) Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, **159**, 676–690.
89. The Cancer Genome Atlas Research Network (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.
90. The Cancer Genome Atlas Research Network (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, **169**, 1327–1341.
91. The Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.
92. The Cancer Genome Atlas Research Network (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**, 1011–1025.
93. The Cancer Genome Atlas Research Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.