Biostatistics (2012), **0**, 0, *pp.* 1–28
doi:10.1093/biostatistics/manuscript-acf-v5

Differential expression analysis of RNA-seq data at base-pair resolution

Alyssa C. Frazee[1], Sarven Sabunciyan[2], Kasper D. Hansen[1], Rafael A. Irizarry[1*], Jeffrey T. Leek[1*]

1. Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA

2. Stanley Division of Developmental Neurovirology, The Johns Hopkins School of Medicine, 615 North Wolfe Street, Baltimore, MD 21205, USA

rafa@jhu.edu, jtleek@gmail.com

Summary

Since the invention of microarrays, measuring genome-wide gene expression has become a common experiment performed by molecular biologists and clinicians. Detecting differentially expressed genes is arguably the most common application of this technology. RNA-sequencing (RNA-seq) is a more flexible technology for measuring genome-wide expression that is rapidly replacing microarrays as costs become comparable. The raw data from RNA-seq experiments are billions of short sequences, obtained from individual RNA transcripts, referred to as *reads*. Current statistical methods for differential expression analysis based on RNA-seq data fall into two broad classes based on how they summarize the information in the reads: (1) methods that count the number of reads within the boundaries of genes previously published in databases and (2) methods that attempt to reconstruct full length RNA transcripts. Both methods have limitations. The first cannot discover differential expression outside of previously known genes, which negates one of the novel aspects of the technology. While the second approach does possess discovery capabilities, the existing implementation grossly underestimates the uncertainty introduced during the summary step and thus cannot reliably detect differential expression. Here we propose a statistical pipeline that preserves the discovery capability of the second approach while achieving similar stability to the first approach. We do this by measuring the number of reads overlapping each individual base-pair, then grouping consecutive base-pairs with common differential expression patterns into differentially expressed regions (DERs). Novel regions and regions that overlap known genes are then labeled for downstream use. We refer to our approach as DER Finder. We compare our approach to leading competitors from both current classes of differential expression methods and highlight the strengths and weaknesses of each. A software implementation DER Finder is available on github.

Key words: RNA sequencing, differential expression, bioinformatics, genomics, false discovery rate

1. Introduction

Microarrays revolutionized the way we measure gene expression by providing, for the first time, genome-wide transcript-level measurements, where *transcript* is used here to refer to the molecule associated with gene expression. However, assigning only one measurement to each known gene has greatly over-simplified the biological process in two ways. The first is that we have not yet discovered or annotated all regions of the genome capable of expressing biologically functional transcripts. Second, experimentally, most genes produce not one but several transcripts through the process of *alternative splicing* (Mortazavi *and others* 2008, Trapnell *and others* 2010, Katz *and others* 2010). In principle, RNA-sequencing (RNA-seq) provides measurements of transcript expression from which we can obtain a more complete picture of reality. While microarrays rely on hybridization to predefined probes by explicitly sequencing transcripts, RNA-seq is potentially capable of measuring expression in regions not previously annotated (Guttman *and others* 2009, Cabili *and others* 2011, Guttman *and others* 2010, Clark *and others* 2011) and to measure multiple transcripts for individual genes (Trapnell *and others* 2010, Graveley *and others* 2010, Mortazavi *and others* 2008). This flexibility, coupled with rapidly declining sequencing costs, has led to explosive growth in the use of RNA-seq technology (Stein *and others* 2010).

The most common goal among investigators using either microarrays or RNA-seq is detecting differential expression, for example: discovering transcripts showing different average expression levels across two populations. A major difference between the two technologies is that in microarrays, measurement units are fixed in advance: only the abundances of the specific RNA sequences that correspond to probes on the microarrays are measured. With this approach, differential expression is relatively straightforward to quantify: measurements from the same probe are compared across samples. In contrast, RNA-seq reads out short sequences of molecules produced by shearing and reading RNA transcripts (the measurements produced are referred to as *reads*). Unlike with a microarray, across-sample comparisons are not straightforward as measurement units are not defined in advance. Therefore, reads must be summarized into units of expression before differential expression analysis can be performed. Different summarization approaches can lead to very different statistical inference.

Here we group the most popular differential expression analysis approaches into two categories based on the ways that the reads are summarized. We refer to these two categories as (1) *annotate-then-identify* and (2) *assemble-then-identify.* The first category counts the number of reads that fall within previously identified boundaries of known genes. The second class seeks to assemble full transcripts directly from the reads. In either case, differential expression analysis is then performed on the resulting measurements at the gene or transcript level.

In Section 2 we describe the limitations with the existing approaches and propose a new intermediate class of differential expression methods which we refer to as *identify-then-annotate*. This class of methods calculates the number of reads that cover each genomic location or base-pair (coverage), then forms a statistic measuring differential expression for each base-pair. Consecutive locations showing a common differential expression signal are grouped into differentially expressed regions (DERs), that are then assigned a measure of statistical significance. The identified regions may then be *annotated* by calculating similarity metrics between identified regions and features appearing in reference annotation databases. Regions of differential expression that cannot be easily annotated are reported separately for easy identification of novel transcriptional events.

In Section 3, we propose a specific implementation of the identify-then-annotate class of methods, which we call Differentially Expressed Region Finder (DER Finder). And in Section 4, using an example dataset, we show that identify-then-annotate models provide a good compromise between current RNA-seq analysis methods. Identify-then-annotate methods more flexibly deal with variability or errors in reference annotation than annotate-then-identify models. At the same time, identify-then-annotate models do not incur the added variability and ambiguity of full transcriptome assembly. We show that these advantages allow identify-then-annotate models to: (1) identify novel transcribed regions even outside reference annotation, (2) identify differential expression signals obscured by transcript assembly, and (3) allow for easily summarized and visualized differential expression results. We have implemented our approach in the Bioconductor package *derFinder*, which is be freely available for download on github (https://github.com/alyssafrazee/derfinder).

2. Previous work and motivation

To understand the complications arising with RNA-seq data analysis, we need to understand the process of gene expression, so we provide a quick review (for more, refer to Alberts (2008) or Twyman (2003)). The genomic regions we refer to as genes are typically subdivided into protein-coding units (*exons*) and non-protein coding units (*introns*). Transcription is the first step in expression, during which a single-stranded DNA copy of the gene is created. The introns are then removed (*spliced out*), leaving a messenger RNA (mRNA) molecule composed solely of exons, which are then translated into proteins (the building blocks of our body). During the splicing step, exons can also be removed, resulting in different mRNA versions being produced. These versions are referred to as *isoforms*.

RNA-sequencing generates millions or billions of short sequences of from individual mRNA molecules. Analyzing these data requires several steps. First, each read must be matched to the position it originates from in the genome in a process called alignment. Then, the number of reads aligned to specific regions must be summarized into quantitative measurements. The measurements are then normalized for the total number of reads measured for a particular sample and statistical models are applied to the summarized units. Oshlack *and others* (2010) describe this RNA-seq data analysis process in much more detail. Based on the summarization step, current statistical methods for the analysis of RNA-seq data can be grouped into two major classes.

The first class of methods, which we call annotate-then-identify methods, summarize the reads by counting the number that fall within pre-specified exons or genes. The exon and gene specifications, collectively called the *annotation*, are obtained from databases of previously identified genomic features. Once the reads have been summarized at the exon or gene level, the statistical problem is very similar to statistical analysis of microarray data, with some deviations because the raw measurements take the form of counts. Note that the results from this step can be naturally summarized into matrices such as those produced by microarray experiments, where rows are genes or exons and columns are samples. The most mature statistical methods for RNA-seq data analysis fall into the category of annotate-then-identify methods. The most widely used methods in this category are EdgeR (Robinson *and others* 2010, McCarthy *and others* 2012) and

DESeq (Anders and Huber 2010).

The advantages of the annotate-then-identify approach are that it provides a straightforward and interpretable analysis and that mature, tested statistical methodology is available once raw read counts have been summarized into a gene-level matrix. The first disadvantage of annotate-then-identify methods is that they rely heavily on the accuracy of annotation databases of gene and exon boundaries, but current annotation may be unreliable or hard to interpret (Klimke *and others* 2011). For example, in some regions of the human genome, a large number of highly similar exons with complicated overlap structure all correspond to the same region (Figure 1). This overlap may be biologically real due to complexities of transcription, or may be due to misspecification of exon boundaries in the database. In either case, it requires the analyst to make a decision about how the exons in a region should be summarized, and this decision can have a major impact on downstream statistical results. A second disadvantage is that annotate-then-identify methods do not allow for discovery outside of previously defined exons or genes.

Another downside is the number of reads that must be discarded

The second class of methods, which we call assemble-then-identify methods, attempt to assemble the full sequences of the mRNA molecules that produced the short reads. The advantage of this category of statistical methods is that they rely less heavily on annotation databases of exon or gene boundaries. Another advantage is that assemble-then-identify methods aim to fully quantify all the potential isoforms of mRNA molecules emanating from each gene. One disadvantage of assemble-then-identify methods is that the short length of typical sequencing reads leads to inevitable ambiguity when attempting to quantify the abundances of individual mRNA molecules (Figure 2). This ambiguity also leads to varying and structured covariances between transcript measurements within genes, which complicates statistical analysis. A second disadvantage is that there is increased computational time associated with attempting to assemble full transcripts as compared to the more direct annotate-then-identify approach. The most widely used algorithm in this category is Cufflinks/Cuffdiff (Trapnell *and others* 2010).

Here we propose an intermediate class of methods which we call identify-then-annotate. These methods first summarize the reads by counting the number of reads with alignments overlapping each individual base-pair in the genome. Then we form a base-pair-by-base-pair statistic to identify base-pairs that are differentially expressed between groups. Consecutive base pairs showing a common differential expression signature are then grouped into differentially expressed regions (DERs). The unit of statistical analysis is then the DER, which can be evaluated for statistical significance using permutation or bootstrap approaches. DERs can then be compared to previous databases of exons and genes to identify: (1) regions of differential expression corresponding to known exons or genes and (2) novel regions of differential expression. The rnaSeqMap Bioconductor package (Leśniewska and Okoniewski 2011) is the only existing implementation of an identify-then-annotate algorithm.

The advantages of the proposed identify-then-annotate model are that: (1) it allows for detection of differential expression in regions outside of known exons or genes, (2) it allows for direct evaluation of differential expression of known genes and exons, and (3) it does not incur the added ambiguity and computational cost of assembly from short reads. The primary disadvantage is that the proposed approach does not allow for direct quantification of alternative transcription. However, regions of potential alternative transcription can be easily identified where a subset of exons for a gene overlaps DERs but another subset does not. In that sense, identify-then-annotate methods could be coupled with assemble-then-annotate methods to focus the computational effort of assembly on regions already showing basic differential expression signatures.

3. DER Finder

rnaSeqMap seeks to facilitate the computation of basic identify-then-annotate models by providing a database backend to manage base-pair resolution coverage. The software suite also provides basic functions for identifying regions with coverage above a fixed genome-wide threshold. However, it lacks statistical sophistication and does not, for example, attach uncertainty assessment to its findings outside of what can be obtained from EdgeR and DESeq. Our approach generalizes the idea behind rnaSeqMap by allowing a more flexible statistical model to obtain the base-pair level summaries, implementing an HMM for DER identification that relaxes the need for an arbitrary genome-wide threshold, and providing a measure of statistical significance for candidate DERs. Although our approach is modular, each key step is motivated by statistical reasoning.

3.1 *Base-pair resolution estimates*

The first step in DER Finder is to quantify the evidence for differential expression at the base-pair-level. Since RNA-seq produces reads from mRNA transcripts, rather than directly from the genome, reads must be aligned using a strategy that accounts for reads that span multiple exons, called *junction reads*. In identify-then-annotate approaches like DER Finder, these junction reads are treated identically to reads that map directly to the genome when computing coverage. Tophat (Trapnell *and others* 2009) is an example of an aligner that appropriately handles junction reads. Since shorter reads are less likely to originate from a sequence containing a splice junction, another alignment approach might be to first divide the RNA-seq reads into shorter sub-reads, termed *readlets*, and align each readlet to the genome using an aligner like Bowtie (Langmead *and others* 2009). Whatever alignment strategy is used, the result is ultimately a large matrix with rows corresponding to base-pairs and columns corresponding to samples. We refer to this matrix as the *coverage matrix*.

To account for biological variability and possible confounders, we fit a linear regression model to each row of the coverage matrix. Specifically, we let

$$g(Y_{i,j}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^{K} \gamma_k(l_j)W_{il} + \varepsilon_{ij} \tag{3.1}$$

where $Y_{i,j}$ is coverage for sample i at location l_j, g is a Box-Cox style transformation that makes the linear assumption acceptable, $\alpha(l_j)$ represents the baseline gene expression level at location l_j, X_i is the covariate of interest for sample i, $\beta(l_j)$ is the parameter of interest that quantifies differential expression at location l_j, W_{ik} $(k = 1, \ldots, K)$ are possible confounders that may include sample-specific GC effect (Hansen *and others* 2012, Risso *and others* 2011), sex or other demographic variables, or processing data, $\gamma_k(l_j)$ represent the effects associated with the confounders, and ε_{ij} represents residual measurement error.

Our goal is to identify contiguous regions A where $\beta(l_j) > 0$ or $\beta(l_j) < 0$ for all $l_j \in A$. Instead of modeling $\beta(l_j)$ as a function (for example, with wavelet models or splines), we adopt a modular approach in which we first estimate $\beta(l_j)$ for each location l_j and then divide the estimates into regions in a separate step. To do this, we apply the methods developed by Smyth *and others* (2004).

3.2 *Finding candidate DERs*

In this section, we refer to the base-pair resolution test statistic (resulting from testing the null hypothesis that $\beta(l_j) = 0$) as $s(l_j)$. (For ease of notation, we omit the j subscript in the discussion that follows). For most experiments, we expect the function $s(l)$ to be a step function that is mostly 0, since most of the genome is not differentially expressed. We do not expect $s(l)$ to be smooth because gene expression usually has a clear-cut start and end location. Hidden Markov Models (HMMs) are a natural way of modeling s, and we describe the specifics of our implementation here.

Assumes just 2 groups

We assume there is an underlying Markov process along the genome $D(l)$ with three hidden states: $D(l) = 0$ if $\alpha(l) = \beta(l) = 0$, $D(l) = 1$ if $\alpha(l) \neq 0$ and $\beta(l) = 0$, and $D(l) = 2$ if $\beta(l) \neq 0$. State $D(l) = 0$ corresponds to regions producing practically no gene expression. This state will be the most common, as most base-pairs will not be covered by any reads because abundant gene expression is confined to a relatively small fraction of the genome. State $D(l) = 1$ corresponds to regions for which gene expression is observed but does not differ between populations. We are interested in finding regions in the differentially expressed state, $D(l) = 2$.

We assume that $D(l)$ is a stationary first-order Markov chain with invariant hidden state probabilities $\pi_d = \Pr(D(l) = d)$. We treat the transition matrix as fixed. The transition probabilities can be roughly estimated based on the relative frequencies of base-pairs covered or not covered by genes, along with a prior estimate of the number of differentially expressed genes. As defaults, we set the retain state probabilities as very high with low transition probabilities between states, due to the sparsity of genes in the genome.

Conditional on the hidden state of each base-pair l, we then assume $s(l)$ follows a normal distribution. Specifically, $s(l)|D(l) = d \sim N(\mu_d, \sigma_d^2)$. When $D(l) = 0$ there is little expression observed for base-pair l so we model the distribution as $N(0, \delta)$, where δ is an arbitrary, very small positive number, to restrict values to very close to zero. We estimate π_0 empirically by calculating the fraction of base-pairs where the average coverage is less than a threshold c.

The model parameters for states $D(l) = 1$ and $D(l) = 2$ (μ_1, μ_2, σ_1^2, and σ_2^2) can be estimated using a standard two-groups mixture model, first proposed for the analysis of differential expression in microarray experiments (Efron 2008). We assume that the statistics $s(l)$ from these two states are drawn from a mixture $f(s) = f_1(s)\pi_1^* + f_2(s)\pi_2^*$, where $\pi_1^* + \pi_2^* = 1$. (Estimates for π_1^* and π_2^* are scaled by the estimate of π_0 to obtain estimates for the overall state probabilities, π_1 and π_2, such that $\pi_0 + \pi_1 + \pi_2 = 1$.) Each mixture component is again assumed to be normal and can be estimated using the empirical null distribution. We can then directly estimate the most likely path of unobserved states $D(l)$ based on the observed statistics $s(l)$ using standard estimation techniques for HMMs. Details on specific parameter estimation techniques implemented in *derFinder* are available in supplementary material (section 1).

3.3 *Statistical significance*

The Hidden Markov Model identifies regions with predicted latent state $D(l) = 2$ as potentially differentially expressed regions (candidate DERs). To assign statistical significance to these DERs, we consider the size of the individual statistics within each region, since regions with very large test statistics are more likely to be truly differentially expressed. We can assign a p-value to each potential DER using a permutation or bootstrap procedure. We apply an approach similar to Jaffe *and others* (2012): first, we calculate the average base-pair level test statistic within each potential DER r: $\bar{s}_r = \sum_{l \in DERr} s(l)$. In the simple case-control scenario with no confounders, we can assign p-values to DERs with the following procedure:

1. Permute the values of the covariate of interest (X_i) for all samples.

2. Re-calculate the base-pair level statistics using equation 3.1. Denote these null statistics by $s^0(l)$.

3. Re-run the HMM on the $s^0(l)$s to identify a set of null DERs, indexed by ρ and denoted by DER^0_ρ.

4. Calculate the average base-pair level statistic within each null DER $\bar{s}^0_\rho = \sum_{l \in DER^0 \rho} s^0(l)$

Steps 1-4 are repeated B times, and the empirical p-value for region r is $p_r = \frac{1}{\sum_{b=1}^{B} P_b} \sum_{b=1}^{B} \sum_{\rho=1}^{P_b} 1(\bar{s}^0_\rho > \bar{s}_r)$, where P_b is the number of null DERs for permutation b. This quantity is the percent of null DERs with average statistic as or more extreme than the observed statistic for candidate DER r calculated on the observed data. Standard false discovery rate calculations can then be applied to adjust these p-values for multiple testing.

In the case where confounders or additional covariates are included in model 3.1, a straightforward bootstrap extension of this permutation approach can be derived.

3.4 *Methods for annotating output*

Output from the HMM pipeline is a table of regions, giving each region's chromosome, start position, end position, predicted state D, and if $D = 2$, an adjusted p-value. To bring biological meaning to the table of regions, each region is flagged if it may indicate an event of interest (Table 1). If the event involves a specific gene or exon, that information is included in the flag. The flag "Unknown Event of Interest" means that at least one, but not all, of the exons in a given gene are differentially expressed: this phenomenon could indicate a number of genomic events, including possible alternative splicing. Further exploration of these regions is possible using assemble-then-annotate methods to evaluate potential alternative or differential splicing events. Due to variance in read coverage across the genome, we observed some regions shorter than the length of an individual read. These small regions are particularly detrimental in the annotation and labeling step. We therefore choose to disregard regions shorter than the read length. Regions flanking very short transitions between states are merged.

4. RESULTS

We implemented our proposed pipeline in a series of R functions which we refer to as *derFinder* (available for download on github, https://github.com/alyssafrazee/derfinder). The exuberance over RNA-seq technology led many to publish studies without biological replicates (Hansen *and others* 2011). Our method is designed for differential expression with biological replicates because we anticipate the standard experimental designs will eventually converge on sensible ones. However, finding public assessment datasets proved difficult. We therefore designed an analysis comparing human males and females to assess the competing methods: the Y chromosome was tested for differential expression between sexes.

The data consisted of unpaired, 101-bp RNA-seq reads from 15 control samples (9 male, 6 female) of postmortem brain tissue. The new pipeline was compared to the Tophat-Cufflinks-

Cuffdiff pipeline, EdgeR, and DESeq. EdgeR and DESeq analyses were run at the exon level, using exon-by-sample count tables created based on the Tophat alignment file with RSamtools (Morgan and Pagès) and GenomicRanges (Aboyoun *and others*). It has been pointed out that reads aligning to overlapping genomic features introduce extra complexity into this analysis pipeline because they have the potential to be "double-counted" (Obenchain and Morgan 2012), so we discarded reads aligning to overlapping exons. Note that this is distinct from discarding reads that align to two exons via, e.g., a spliced alignment. Default parameters and library size adjustments were used in EdgeR and DESeq. For Tophat-Cufflinks-Cuffdiff, default parameters were used. Detailed commands used are available upon request from the corresponding authors. The DER Finder model was fit using sex as the covariate of interest and median of nonzero coverage values for each sample as a known confounder. All p-values (from all pipelines) were adjusted for multiple testing by controlling the false discovery rate and thus using the q-value (Storey and Tibshirani 2003) as a measure of statistical significance.

Two sets of results were obtained: one analysis compared males to females, and the other compared a randomly selected set of five of the males to the other four males. We expect virtually all genomic features of the Y chromosome (barring pseudogenes and other irregularities) to be differentially expressed between males and females, since females do not have a Y chromosome, and no genomic features to be differentially expressed between control males.

4.1 *DER Finder results*

DER Finder identified 633 Y-chromosome regions as differentially expressed ($q < 0.05$) between males and females. Thirty of these regions were classified as underexpressed in males, which we know to be some sort of artifact, but the other 603 were identified as overexpressed in males as expected. Additionally, we found 399 novel differentially transcribed regions ($q < 0.05$). These novel transcribed regions ranged in length from 15 to 1450 base pairs. These regions may indicate noise from the method, but they also may point to regions that should be examined further, either because they have interesting mapability characteristics or because they might truly be expressed and not yet annotated. The 633 differentially expressed regions pointed to 172 differentially expressed exons, using the criteria outlined in Table 1. These 172 exons came from 17 different genes, which means we found those 17 genes to be differentially expressed or indicate an event of interest. In comparing males to each other, we did not identify any differential expression on the Y chromosome: the minimum q-value for the regions found to be differentially expressed in the HMM step was 0.84.

4.2 *Tophat-Cufflinks-Cuffdiff results*

Of 808 assembled transcripts tested for differential expression on the Y chromosme between males and females, the Tophat-Cufflinks-Cuffdiff pipeline found no differentially expressed transcripts. The minimum q-value for these assembled transcripts was 0.45. While 736 of these transcripts showed nonzero abundance in males and zero abundance in females, these differences were not found to be statistically significant using the Cuffdiff methodology. In the comparison of normal males, none of the 818 assembled transcripts were called differentially expressed: the minimum q-value was 0.63.

4.3 *EdgeR and DESeq results*

Both of these methods tested 280 exons on the Y chromosome for differential expression between males and females. The other annotated exons on the Y chromosome did not have any reads mapping to them. Of these 280 exons, EdgeR classified 138 and DESeq classified 133 as differentially expressed between males and females ($q < 0.05$). 129 exons were found by both EdgeR and DESeq. When comparing the males to each other, neither method found any exons to be differentially expressed: all q-values were 1.

4.4 *Comparison of results across methods*

MA plots (Dudoit *and others* 2002) were used to examine results from each method (Figure 3). This type of plot shows the relationship between each unit's average expression (denoted with M) and the magnitude of differential expression it exhibits (denoted with A), where the unit is a region for DER Finder, an exon for EdgeR and DESeq, or a transcript for Cufflinks. The MA plots reveal that DER Finder, EdgeR, and DESeq all produce reasonable results, but the findings from Cufflinks are somewhat problematic. While there does seem to be more overexpression of transcripts in males in the male/female differential expression analysis done by Cufflinks, we observe several extreme fold changes in the opposite direction, and the male-to-male comparison also produced these extreme fold changes. These problems do not exist in the other methods, whose MA plots illustrate high fold changes found between males and females and very little change found between males, as expected.

To assess validity of statistical tests and multiple testing p-value adjustments, p-value histograms for each method were created (Figure 4). P-values were assigned to each region assigned latent state $D = 2$ by the HMM step in DER Finder, to each transcript in Cufflinks, and to each exon in EdgeR and DESeq. The observed distributions were shaped as expected in the results from DER Finder, EdgeR, and DESeq: in the comparison between sexes, many low p-values were observed, corresponding to the fact that most of the Y chromosome should be differentially expressed. However, based on the p-value histogram generated from the Cufflinks transcripts, the analysis of differential expression between sexes did not produce a very substantial number of small p-values. Instead, it produced a cluster of p-values between 0.2 and 0.4, which is an unexpected finding given the nature of Y chromosome expression differences between males and females.

To compare the final results produced by the four methods we evaluated the tables of differentially expressed regions between sexes and between males produced by each method. As described in the DER Finder methodology, regions shorter than the read length were ignored. We combined results from both the male versus male comparison (negative results) to the male versus female comparison (positive results) and ordered these by the absolute value of the test statistic. An algorithm ranking all positive results ahead of the negative ones is preferred. Figure 5 shows at each percentile of the differential expression statistic the percent of regions that are positive. This is analogous to finding what percentage of the findings were truly positive findings at different significance cutoffs, assuming all tests in the sex comparison should be positives and tests in the male comparison should be negatives. We find that EdgeR, DESeq, and DER Finder perform comparably: the top 20% of regions all came from comparisons between sexes. Cufflinks does much worse: only about 60% of the top 20% of their top transcripts came from the male-to-female comparison.

Most of the above results demonstrate that DER Finder performs comparably to EdgeR and DESeq. However, DER Finder has the added advantage of being agnostic to annotation, which gives it an additional benefit over the annotate-then-identify methods. To illustrate these benefits, consider the case where the location of the exon may be incorrectly annotated (Figure 6). In the example shown, the exon's location appears misspecified in the annotation given the observed alignments. As such, the exon's expression is underestimated, since many reads that should belong to this exon fall outside the annotated region. This causes EdgeR and DESeq not to call this Y-chromosome exon differentially expressed ($q = 1$) between males and females, which is incorrect. DER finder reports the shown differentially expressed region as overlapping 49.5% of an annotated exon with a q-value < 0.001. Another advantage of DER Finder over EdgeR and DESeq is its ability to find regions of interest that fall outside of annotated exons (Figure 7). Supplementary figures (supplementary material, section 2) illustrate these two phenomena ocurring extensively in the dataset, demonstrating tha the sensitivity of identify-annotate methods like DER Finder to these types of differential expression is essential in facilitating further biological analyses, such as transcript discovery and changes to existing annotation.

5. Discussion

We have identified major challenges faced by existing RNA-seq differential expression analysis approaches. We propose DER Finder as a specific implementation of a new class of methods. The new class deals with the identified challenges by (a) not relying on existing annotation when calling differential expression and (b) avoiding the immensely difficult problem of full transcript assembly by putting differential expression into a more straightforward framework. The proposed method outperforms Cufflinks, the current leading assemble-then-identify pipeline, and performs comparably to EdgeR and DESeq, popular annotate-then-identify methods, while having the added advantages of producing sensible results even in the presence of incorrect annotation and having transcript discovery capability. An identify-then-annotate method like DER Finder is an inportant step in development of a new way to analyze RNA-seq data.

6. Software

All software and code used in this analysis is available upon request from the corresponding authors.

7. Supplementary Material

Supplementary material has been submitted as separate document.

Acknowledgements

We acknowledge helpful discussions with Geo Pertea and Steven Salzberg. *Conflict of Interest*: None declared.

References

Aboyoun, P., Pages, H. and Lawrence, M. *GenomicRanges: Representation and manipulation of genomic intervals*. R package version 1.6.7.

Alberts, B. (2008). *Molecular Biology of the Cell: Reference edition*, Molecular Biology of the Cell. Taylor & Francis Group.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**, R106.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J. L. (2011, Sep). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25**(18), 1915–1927.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A. *and others*. (2011). The reality of pervasive transcription. *PLoS Biology* **9**(7), e1000625.

Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica* **12**(1), 111–140.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**(1), 1–22.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W. *and others*. (2010). The developmental transcriptome of Drosophila melanogaster. *Nature* **471**(7339), 473–479.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N. *and others*. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235), 223–227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S. *and others*. (2010, May). Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. *Nature Biotechnology* **28**(5), 503–510.

Hansen, K.D., Irizarry, R.A. and Zhijin, Wu. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**(2), 204–216.

Hansen, K.D., Wu, Z., Irizarry, R.A. and Leek, J.T. (2011). Sequencing technology does not eliminate biological variability. *Nature Biotechnology* **29**(7), 572–573.

Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P. and Irizarry, R.A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology* **41**(1), 200–209.

Katz, Y., Wang, E.T., Airoldi, E.M. and Burge, C.B. (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature Methods* **7**(12), 1009–1015.

Klimke, W., O'Donovan, C., White, O., Brister, J.R., Clark, K., Fedorov, B., Mizrachi, I., Pruitt, K.D. and Tatusova, T. (2011). Solving the problem: Genome annotation standards before the data deluge. *Standards in Genomic Sciences* **5**(1), 168.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L. *and others*. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**(3), R25.

Leśniewska, A. and Okoniewski, M.J. (2011). rnaseqmap: a bioconductor package for rna sequencing data exploration. *BMC Bioinformatics* **12**(1), 200.

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 Research* **40**(10), 4288–4297.

Morgan, Martin and Pagès, Hervé. *Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import*. R package version 1.6.3.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**(7), 621–628.

Obenchain, V. and Morgan, M. (2012). user2012!: High-throughput sequence analysis with R and Bioconductor. `http://www.bioconductor.org/help/course-materials/2012/useR2012/Bioconductor-tutorial.pdf`.

Oshlack, A., Robinson, M.D. and Young, M.D. (2010). From RNA-seq reads to differential expression results. *Genome Biology* **11**(12), 220.

Risso, D., Schwartz, K., Sherlock, G. and Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinformatics* **12**(1), 480.

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**(1), 139–140.

Smyth, G.K. *and others*. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1), 3.

Stein, L.D. *and others*. (2010). The case for cloud computing in genome informatics. *Genome Biology* **11**(5), 207.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**(16), 9440–9445.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9), 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010, May). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**(5), 511–515.

Twyman, Richard. (2003). Gene expression. `http://genome.wellcome.ac.uk/doc_WTD020757.html`.

Figures and Tables

Result	**Flag**
Region of state $D = 2$ overlaps more than 80% of an annotated exon	Differentially Expressed Exon
There exists a set of regions of state $D = 2$ with differentially expressed exon flags such that all exons in a given gene are flagged by the set	Differentially Expressed Gene
There exists a set of regions of state $D = 2$ with differentially expressed exon flags such that at least one, but not all, of the exons in a given gene are flagged by the set	Unknown Event of Interest
Region of state $D = 1$ does not overlap any annotated exons	Novel Transcribed Region
Region of state $D = 2$ does not overlap any annotated exons	Novel Differentially Transcribed Region

Table 1. Genomic events indicated by HMM results

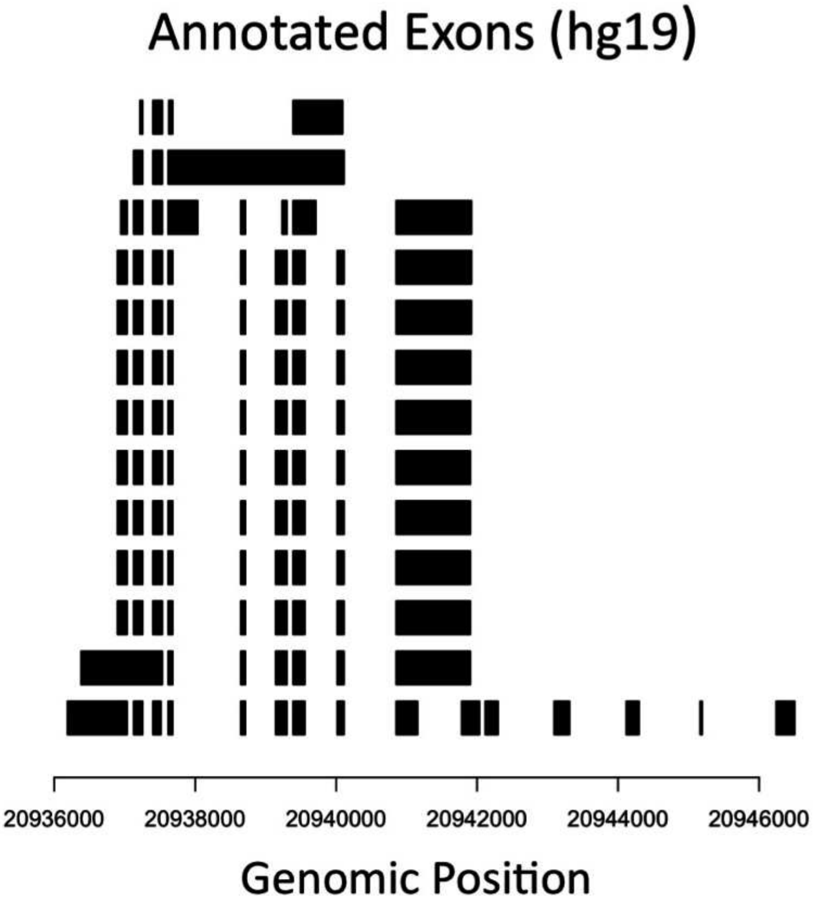

Fig. 1. Locations of all the annotated exons in a 10kb region of the human genome. The region in the middle where 10 exons are annotated at essentially the same location causes problems in annotate-identify analysis pipelines, as there is no clear way to determine which of the overlapping exons generated each read.

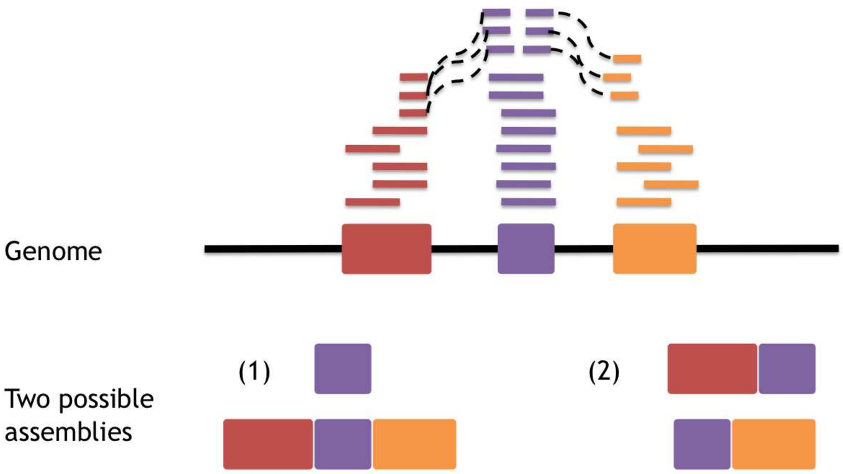

Fig. 2. Illustration of the ambiguity that can occur when assembling transcripts based on RNA-seq data. This hypothetical gene contains three exons: red, purple, and orange. The purple exon appears to be expressed at higher levels than either the red or orange exons. From this information, we can guess this gene has multiple transcripts, but it is not clear which set of transcripts is correct: for example, both set (1) and set (2), shown at the bottom of the figure, are possible.

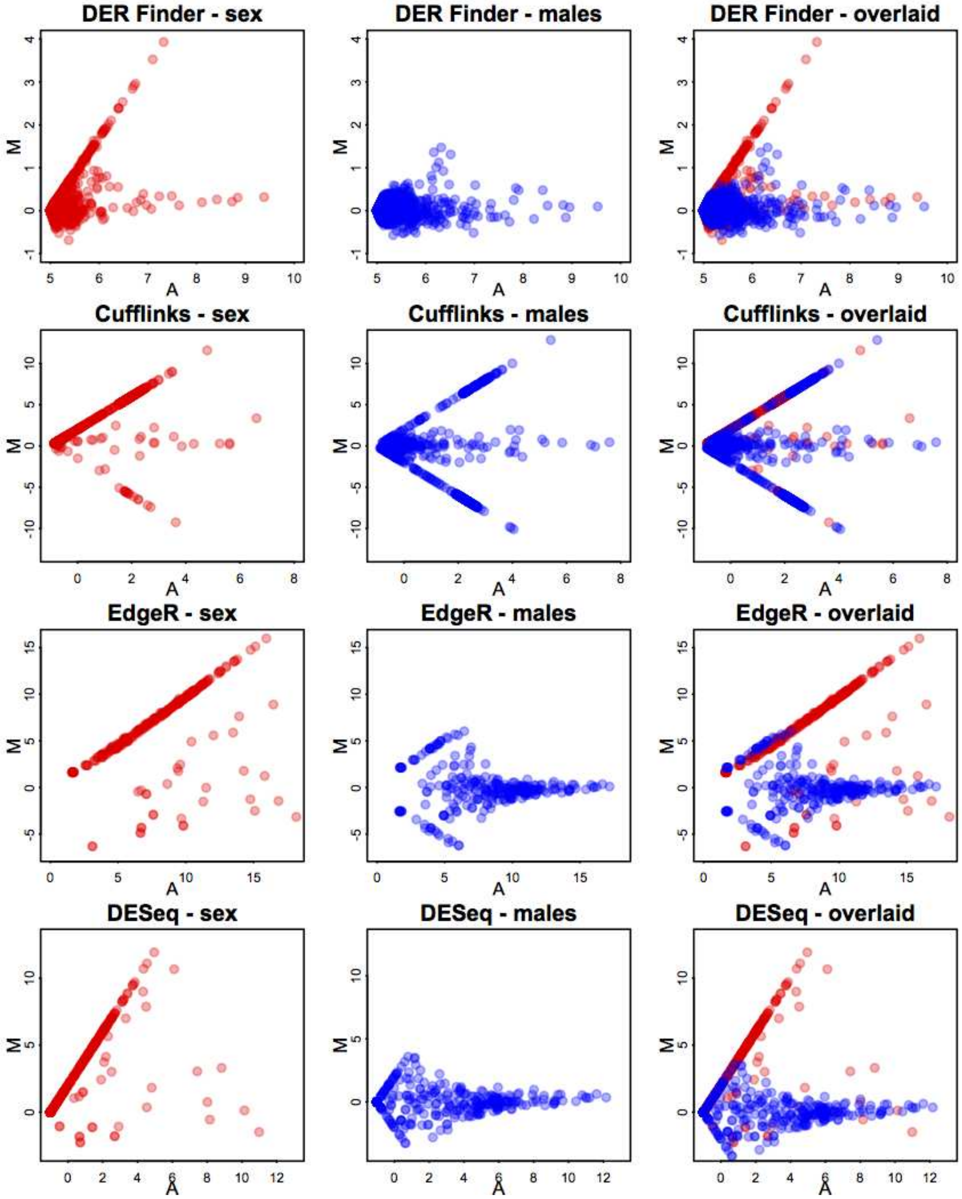

Fig. 3. MA plots for Y chromosome regions, transcripts or exons, for each method and for both male vs. female (red) and male vs. male (blue) comparisons. On each plot, the x-axis represents the average log (base 2) abundance for each unit (region for DER Finder, transcript for Cufflinks, exon for EdgeR and DESeq), and the y-axis represents the log (base 2) fold change between males and females (red points) or the two groups of males (blue points). We expect to see the red, positively-sloped diagonal on all plots: this represents genomic regions expressed in males but not in females. In DER Finder, EdgeR, and DESeq, this diagonal corresponds with differential expression detected, however, no differential expression was detected in Cufflinks even though the red diagonal exists as expected.

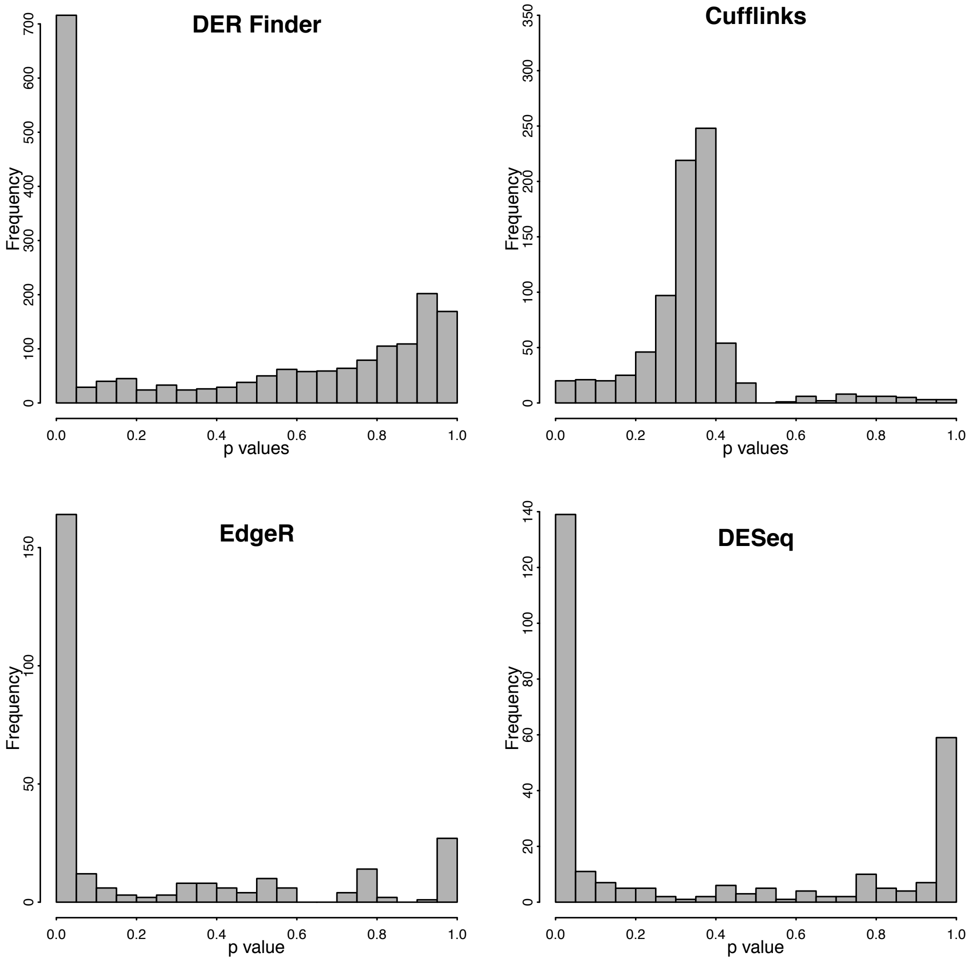

Fig. 4. P-value histograms for tests of differential expression on the Y chromosome between males and females. For all methods except Cufflinks, substantial differential expression is evident in the comparisons between sexes, as expected. The Cufflinks p-value distribution is quite unusual and indicates that using p-values adjusted for multiple testing to assess significance may be problematic.

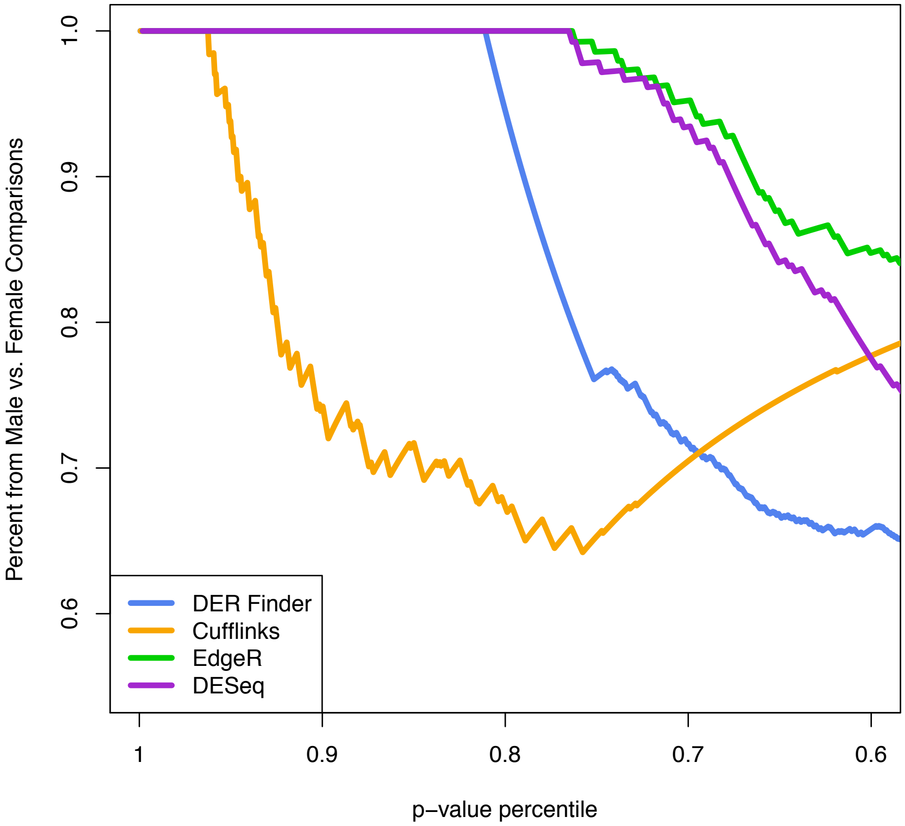

Fig. 5. Percentage of significantly differentially expressed regions/transcripts/exons originating from male-to-female comparisons, using various percentiles of the p-value distribution as a significance cutoff. We find that most highly significant results are true positives, i.e., results with low p-values and high test statistics stem from comparing males to females, for DER Finder, EdgeR, and DESeq, while Cufflinks exhibits problems in this area.

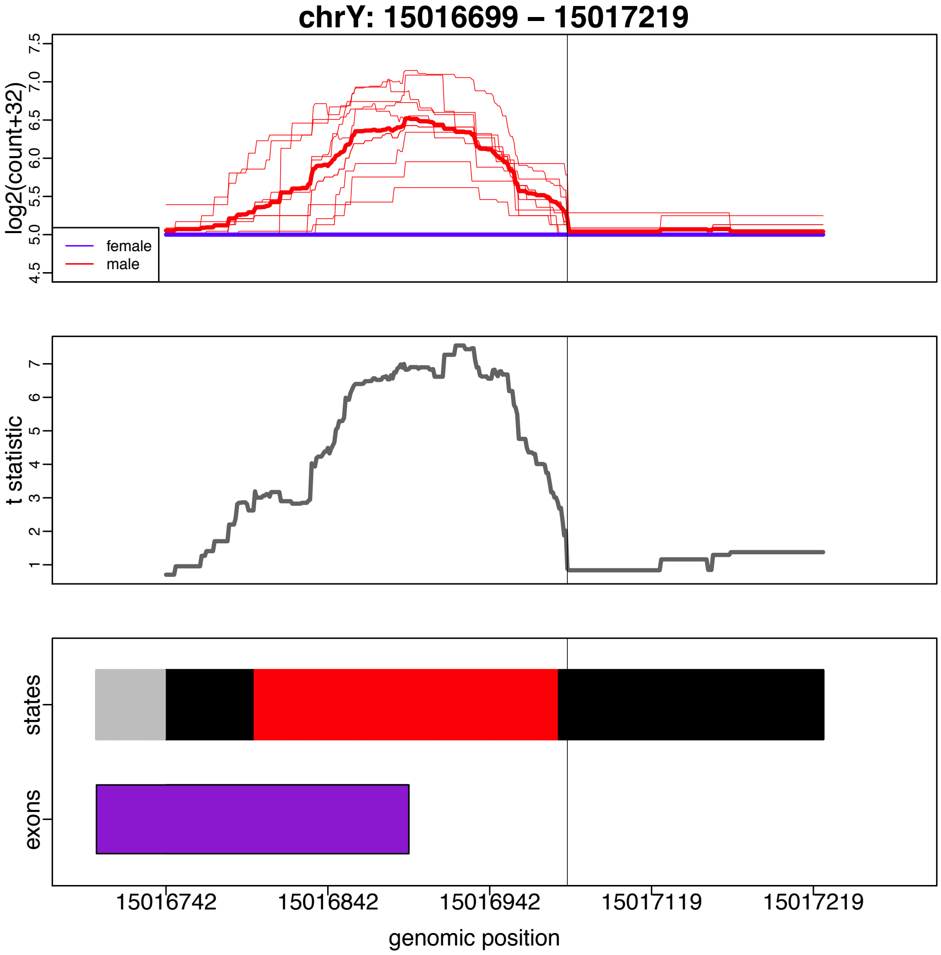

Fig. 6. Example of an exon (from gene *DDX3Y*) whose location appears to be mis-annotated, leading EdgeR and DESeq to underestimate the exon's abundance and therefore incorrectly call this exon not differentially expressed. Top panel: base-pair resolution coverage (on log2 scale). Middle panel: t statistic from linear model fit by DER Finder. Bottom panel: exon location and state calls from DER Finder: gray = not expressed, black = equally expressed, red = overexpressed in men.

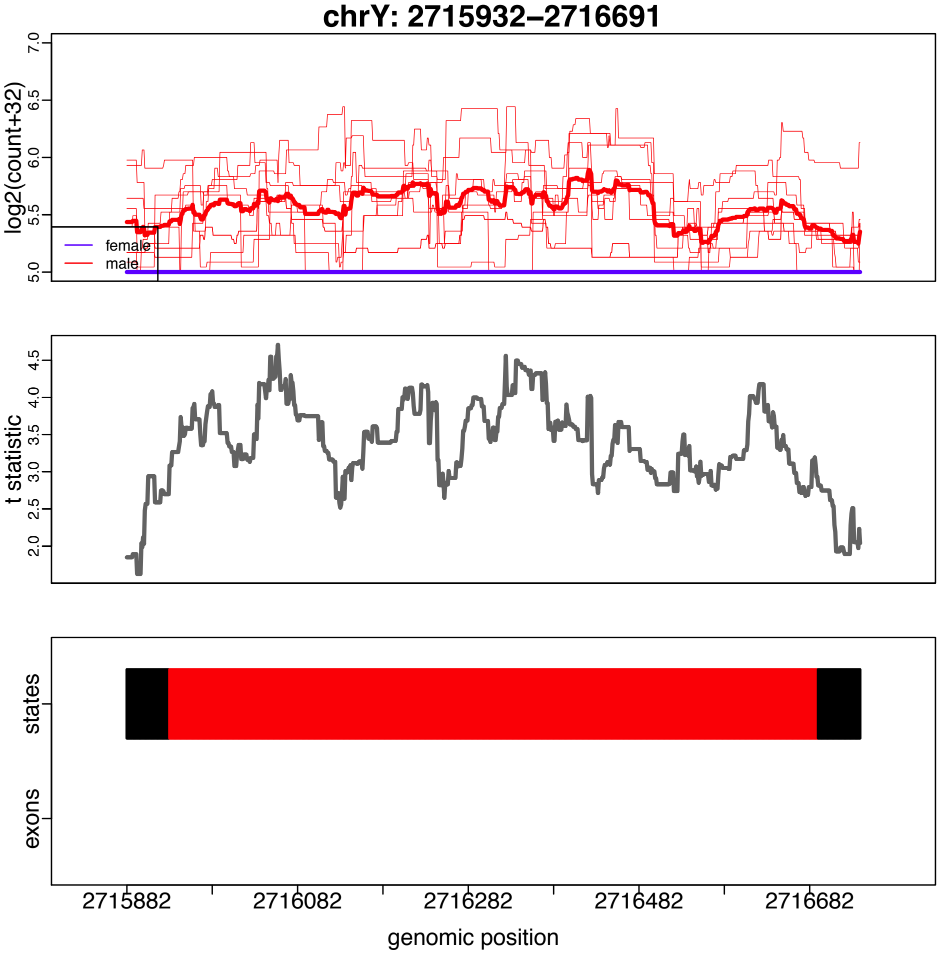

Fig. 7. Example of a differentially expressed region ($q = 0.03$) falling outside of an annotated exon, which can be found by DER Finder but not by EdgeR or DESeq. Top panel: base-pair resolution coverage (on log2 scale). Middle panel: t statistic from linear model fit by DER Finder. Bottom panel: exon location (no exon in this region) and state calls from DER Finder: black = equally expressed, red = overexpressed in men.