`derfinder`: Software for annotation-agnostic RNA-seq differential expression analysis

Authors: Leonardo Collado-Torres[1,2], Alyssa C. Frazee[1], Michael I. Love[3], Rafael A. Irizarry[3], Andrew E. Jaffe[1,2,4*], Jeffrey T. Leek[1,4*]

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205 USA

2. Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205, USA

3. Department of Biostatistics, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115 USA

4. Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21025, USA

* Corresponding authors: AEJ: andrew.jaffe@libd.org ; JTL: jtleek@gmail.com

Abstract

Background Differential expression analysis of RNA sequencing (RNA-seq) data typically relies on reconstructing transcripts or counting reads that overlap known gene structures. Previously we introduced an intermediate approach called differentially expressed region (DER) finder that seeks to identify contiguous regions of the genome showing differential expression signal at single base resolution that does not rely on existing annotation or potentially inaccurate transcript discovery. However, there were computational challenges involved with performing base-resolution analyses in large numbers of samples at genome scale.

Results Here we describe a new version of the `derfinder` software that allows for: (1) genome-scale analyses in a large number of samples, (2) flexible statistical modeling, including multi-group and time course analyses, and (3) a new, computationally efficient approach to re-analysis at base resolution called expressed-region analysis. We also introduce functionality for annotating and plotting base-resolution data to identify artifacts and confirm results. We apply this approach

to public RNA-seq data from the developing human brain to illustrate the types of analyses and results possible using `derfinder`.

Conclusions Single-base and expressed-region RNA-sequencing analysis provides compromise between full transcript reconstruction and gene-level analysis. `derfinder` is software designed to identify, visualize, and interpret differentially expressed regions. The package is available from `Bioconductor` at `www.bioconductor.org/packages/release/bioc/html/derfinder.html`.

Key words: RNA sequencing, differential expression analysis, coverage, gene annotation, gene expression

1 Introduction

The increased flexibility of RNA sequencing (RNA-seq) has made it possible to characterize the transcriptomes of a diverse range of experimental systems, including human tissues (Farrell et al. 2014; GTEx Consortium 2013), cell lines (ENCODE Project Consortium et al. 2012; Lappalainen et al. 2013) and model organisms (Daines et al. 2011; Dillman et al. 2013). The goal of many experiments involves identifying differential expression with respect to disease, development, or treatment. In experiments using RNA-seq, messenger RNA (mRNA) is sequenced to generate short "reads" (36-200+ base pairs). These reads are aligned to genome, and this alignment information is used to quantify the transcriptional activity of both annotated (existing in databases like Ensembl) and novel transcripts and genes.

The ability to quantitatively measure expression levels in regions not previously annotated in gene databases, particularly in tissues or cell types that are difficult to ascertain, is one key advantage of RNA-seq over hybridization-based assays like microarray technologies. As complicated transcript structures are difficult to completely characterize using short read sequencing technologies (Steijger et al. 2013), the most mature statistical methods used for RNA-seq analysis rely on existing annotation for defining regions of interest - such as genes or exons - and counting reads that overlap those regions (Anders, Pyl, and Huber 2014). These counts are then used as measures of gene expression abundance for downstream differential expression analysis (Anders and Huber 2010; Robinson, McCarthy, and Smyth 2010). Unfortunately, the gene annotation may be incorrect or incomplete, which can affect downstream modeling of the number of reads that cross these defined

features.

We proposed an alternative approach for finding differentially expressed regions (DERs) that first identifies regions that show differential expression signal and then annotates these regions using previously annotated genomic features (Frazee et al. 2014a). The approach uses coverage-level data (i.e. the number of reads aligned to each base in the genome) to identify differential expression signal at each individual base and our previous implementation merged adjacent bases with similar signal into candidate regions using a Hidden Markov Model (HMM). Using Y-chromosome and simulated data, we showed that this approach maintains power when compared to the feature-level approaches while allowing for discovery of novel transcriptional events, without incurring the potential inaccuracies of transcript assembly. However, the original software was inefficient for handling large RNA-seq data sets across the entire genome.

The largest limitation to RNA-seq is arguably handling the vast amounts of data generated in a single experiment, as there are tens to hundreds of millions of reads per sample. RNA-seq experiments can therefore generate terabytes of data across billions of measurements in a single experiment involving dozens of samples. The original implementation was statistically rigorous, but it was difficult to apply the approach genome-wide. We seek to extend this general approach allowing the base-level differential expression analysis on hundreds of samples at the genome-scale.

Here we describe an extended framework and corresponding R/Bioconductor software package called `derfinder` that performs differential expression analysis at single-base and expressed-region levels across the entire genome, all within the `R` statistical environment (R Core Team 2014). We first introduce the new analysis framework at the single-base and expressed-region levels using publicly-available RNA-seq data measured in the developing human brain (BrainSpan 2011) as motivation, which extends our recently published work limited to the frontal cortex (Jaffe et al. 2014). We subsequently apply `derfinder` to this *BrainSpan* data and showcase the types of results possible using `derfinder`. Next using data simulated from differentially expressed transcripts, we demonstrate that approach has high sensitivity and specificity. Finally, we close with computational considerations for performing `derfinder` on RNA-seq data. Our method offers a powerful approach for leveraging biological insight from new and existing RNA-seq data sets.

2 Results

We present the results of implementing `derfinder` at single-base and expressed-region levels on the publicly-available *BrainSpan* (BrainSpan 2011) data set, as well as simulated data. We explain different analyses and visualizations on the resulting DERs, and describe the differential expression in the developing human brain. We demonstrate that the `derfinder` approach controls the family-wise error rate while retaining sufficient power to detect differential expression at base resolution. Finally, we describe the computational considerations for using `derfinder`.

2.1 Overview of the `derfinder` approaches

The idea behind `derfinder` is to identify regions of the genome that are differentially expressed with respect to a qualitative or quantitative outcome. Conceptually this approach is a middle ground between gene counting based on previously annotated gene regions and full assembly of transcripts (Frazee et al. 2014a). The approach is to map reads to the genome, calculate coverage at base resolution, and identify regions that show differential coverage profiles between conditions.

The first step is to map reads using a splicing aware alignment tool such as *Tophat2* (Kim et al. 2013). Then we calculate the number of reads in each sample that cover each genomic base. The result is a genome-length coverage vector for each sample. At this stage `derfinder` permits two separate strategies for identifying regions that show differential coverage between conditions: single-base and expressed-region analyses.

Single-base analysis

A single-base resolution analysis in `derfinder` first filters out bases that show low levels of expression across all samples. This typically reduces the number of bases that must be analyzed by up to 90%, reducing both CPU and memory usage (Supplementary Section 1.2). Next, a standard differential expression analysis is performed at each base by comparing nested null and alternative linear models using an F-statistic. The statistical models may include adjustments for confounders such as library size (Mortazavi et al. 2008), demographic variables, and batch effects (Leek et al. 2010).

Once an F-statistic is calculated at each base, we identify differentially expressed regions (DERs) using a "bump hunting" approach (Jaffe et al. 2012a). First we find candidate DERs by identifying

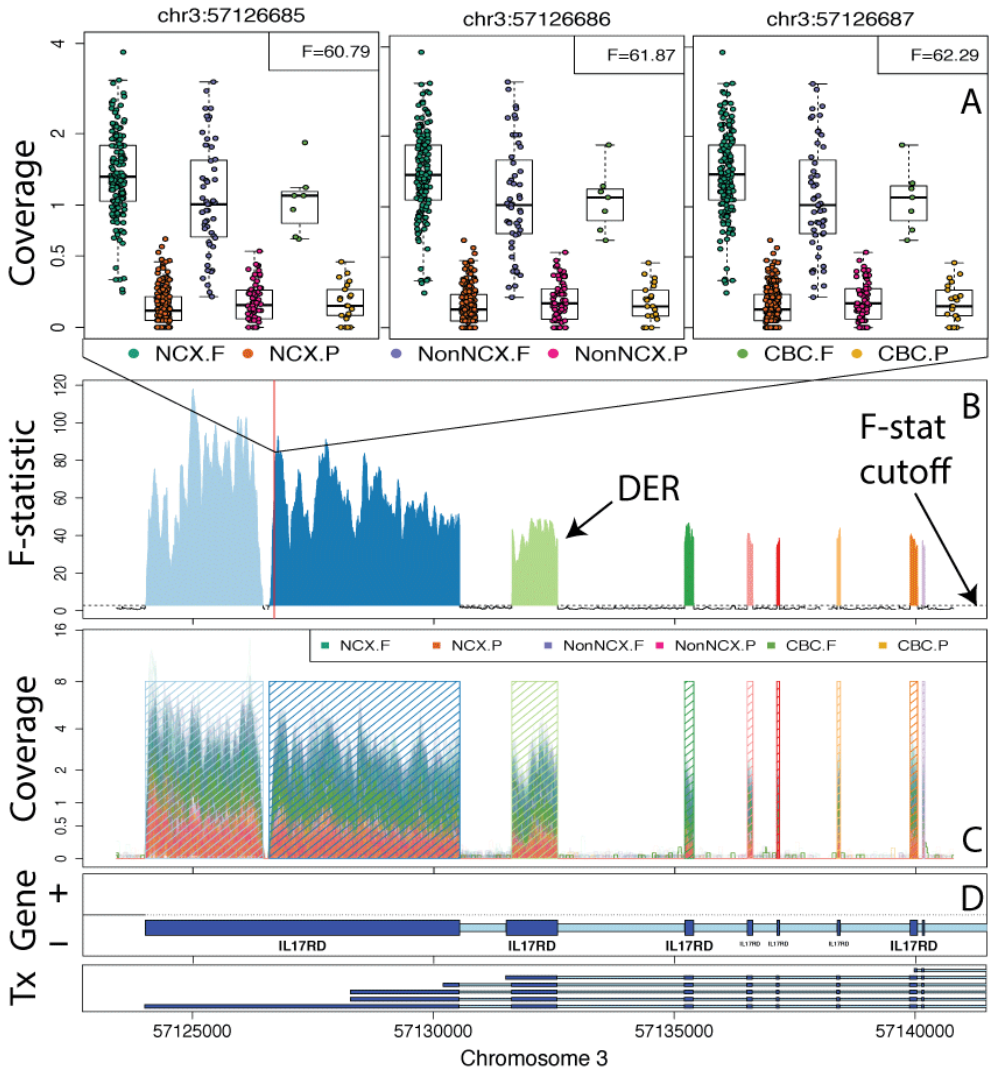

Figure 1: Finding DERs on chromosome 3 with *BrainSpan* data set (see Methods) using six groups: Neocortical regions (NCX: DFC, VFC, MFC, OFC, M1C, S1C, IPC, A1C, STC, ITC, V1C), Non-neocortical regions (NonNCX: HIP, AMY, STR, MD), and cerebellum (CBC) split by whether the sample is from a fetal (F) or postnatal (P) subject. **A** Boxplots for three specific bases. **B** F-statistics curve with regions passing the F-stat cutoff marked as candidate DERs. **C** Raw coverage curves superimposed with the candidate DERs. **D** Known exons (dark blue) and introns (light blue) by strand. The third DER matches the shorter version of the second exon shown in the *Tx* track.

regions of the genome where the base-level F-statistics pass a genome-wide threshold (Figure 1 with *BrainSpan* data set, see Methods). We then calculate a summary statistic for each candidate region based on the length of the region and the size of the statistics within the region. To evaluate the statistical significance of these candidate regions, we permute the sample labels and recompute candidate regions and summary statistics. The result is a region-level p-value, which can be adjusted to control the family-wise error rate (FWER). Alternatively, the region-level p-values

can be adjusted for multiple testing using standard false discovery rate techniques (Dabney and Storey 2014; Storey and Tibshirani 2003).

Expressed-region analysis

In our original work describing `derfinder`, we introduced a general single-base resolution framework for differential expression analysis (Frazee et al. 2014a). However, performing region-level analyses can potentially be a more flexible and computationally attractive solution. This type of analysis starts with read alignment and coverage calculation like the single-base level analysis. In the expressed-region approach, rather than calculating a test statistic at each base, we instead identify contiguous regions in the genome where the average coverage across samples passes a permissive threshold (Leśniewska and Okoniewski 2011) as shown in Figure 2 with the *BrainSpan* data set (see Methods). Similar to the gene counting approach, we summarize these expressed regions by counting the number of reads (including fractions of reads) that overlap the region. We then analyze the resulting coverage matrix using statistical models that have been developed for gene counts such as `limma` (Smyth 2005), `voom` (Law et al. 2014), `edgeR` (Robinson, McCarthy, and Smyth 2010), or `DESeq` (Anders and Huber 2010). The difference between an expressed-region analysis and standard gene counting approaches is that the expressed regions are annotation-agnostic, defined entirely using the observed data.

While both the single-base and expressed-region level analyses are annotation-agnostic, there are several key differences between the two approaches. The single-base approach directly identifies DERs without first explicitly defining what constitutes expressed sequence beyond very liberal filtering, but permits only a single biological question posed via a single pair of full and nested linear models per application. Searching for differential expression in a subset of data or for a different effect of interest requires rerunning the entire `derfinder` approach, including potentially resource-intensive permutations. As the expressed-region approach first identifies expressed regions, secondary analyses on subsets of the data, or sensitivity analyses including additional covariates, are simple. However, the region-level analysis loses spatial resolution, particularly when exonic and intronic sequence are contained in the same expressed region, but only the exon is differentially expressed. We revisit considerations for implementation of these complementary approaches in the Discussion section.

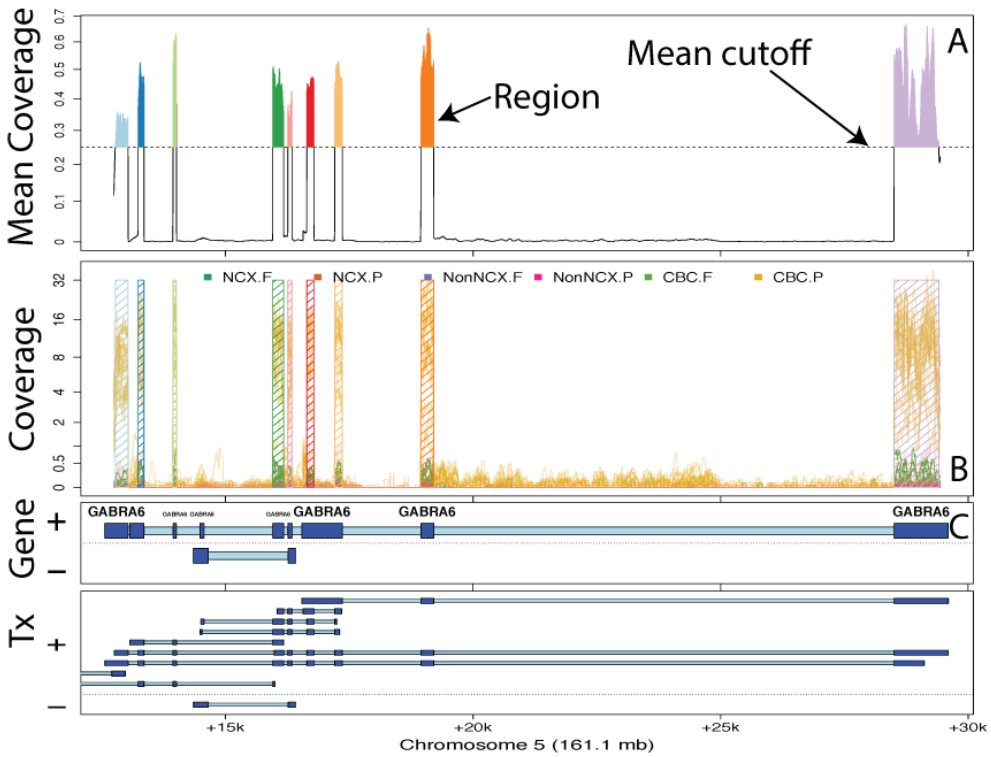

Figure 2: Finding regions via expressed-region approach on chromosome 5 with *BrainSpan* data set (see Methods). **A** Mean coverage with segments passing the mean cutoff (0.25) marked as regions. **B** Raw coverage curves superimposed with the candidate regions. Coverage curves are colored by brain region and developmental stage (NCX: Neocortex: Non-NCX: Non-neocortex, CBC: cerebellum, F: fetal, P: postnatal). **C** Known exons (dark blue) and introns (light blue) by strand for genes and subsequent transcripts in the locus.)

2.2 Visualizing and analyzing DERs

For both types of approaches, `derfinder` produces a set of DERs (as a `GRanges` object (Lawrence et al. 2013), format details in Supplementary Website) with several summary statistics per region. For example, the mean and overall sum of the F-statistics is stored for single-base analyses, and the mean and total base-level coverage are saved for region-level analyses. The DERs can be grouped into larger regions by distance (that can contain many nearby DERs), which can be useful to identify artifacts such as coverage dips (Figure 3). The DERs can be annotated to their nearest gene or known feature using `bumphunter` (Jaffe et al. 2012a). Using this information, visualizing the location of the DERs can be easily made with different software tools such as `ggbio` (Yin, Cook, and Lawrence 2012). With the results from `derfinder`, it is straightforward to make "MA" plots comparing pairs of sample groups post hoc within the DERs. It is important to note that in the case of the single-base level approach, MA and p-value plots used to identify potential artifacts are

biased towards regions that have differential expression signal due to the nature of the analysis.

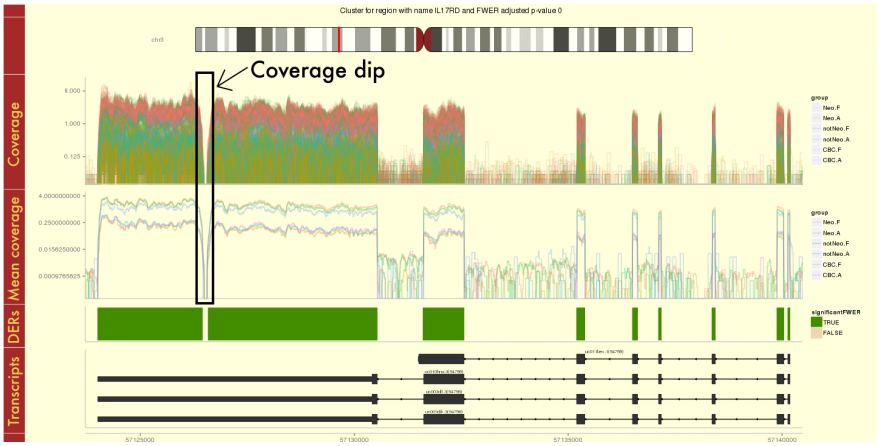

Figure 3: Example of a coverage dip from *BrainSpan* single-base analysis from cluster number 16 in terms of overall signal. `ggbio` plot with tracks showing: ideogram, base level coverage by sample, mean base level coverage by sample group, and known transcripts. Other examples are shown in the Supplementary Section 1.3 and the Supplementary Website.

The sets of DERs from both approaches can also be compared to each other using the number, width, and/or proportion overlapping (Supplementary Website). With both approaches, candidate DERs can further be compared to known gene annotation tables to identify candidate novel transcription events (regions overlapping known intergenic parts of the genome), run-off transcription, and other events. The regions can be exported to CSV files or other biologist-friendly file formats for followup and downstream analyses.

2.3 Differential expression in the developing human brain

We wanted to detect regions that were differentially expressed across the lifespan in human brains. To achieve this, we applied both the single-base and expressed-region `derfinder` approaches to the *BrainSpan* RNA-seq coverage data (see Methods), a publicly available data set consisting of 487 samples on 40 unique individuals across the lifespan across 16 brain regions (BrainSpan 2011). At the single-base level, we identified 115,658 genome-wide significant DERs (at family-wise error rate, FWER $< 5\%$) where expression levels were associated with developmental stage (fetal versus postnatal) and/or brain region (see Methods section 4.5.1). These resulting single-base level DERs largely distinguished the fetal and postnatal samples representing the first principal component and 49.1% of the variance of the mean coverage levels within the DERs (Figure 4). The most significant DERs map to genes previously implicated in development (see Supplementary Website),

and contained many of the DERs we previously identified in the frontal cortex in 36 independent subjects (Jaffe et al. 2014). For example, 59.2% of our previously published 50,650 developmental DERs (and 72.6% in the 10,000 most significant) in the frontal cortex overlapped these DERs identified in the *BrainSpan* data set. The potential lack of overlap may be explained by unmodeled artifacts as there appear to be clusters in the principal components calculated on the base resolution data (Figure 4, left panel).

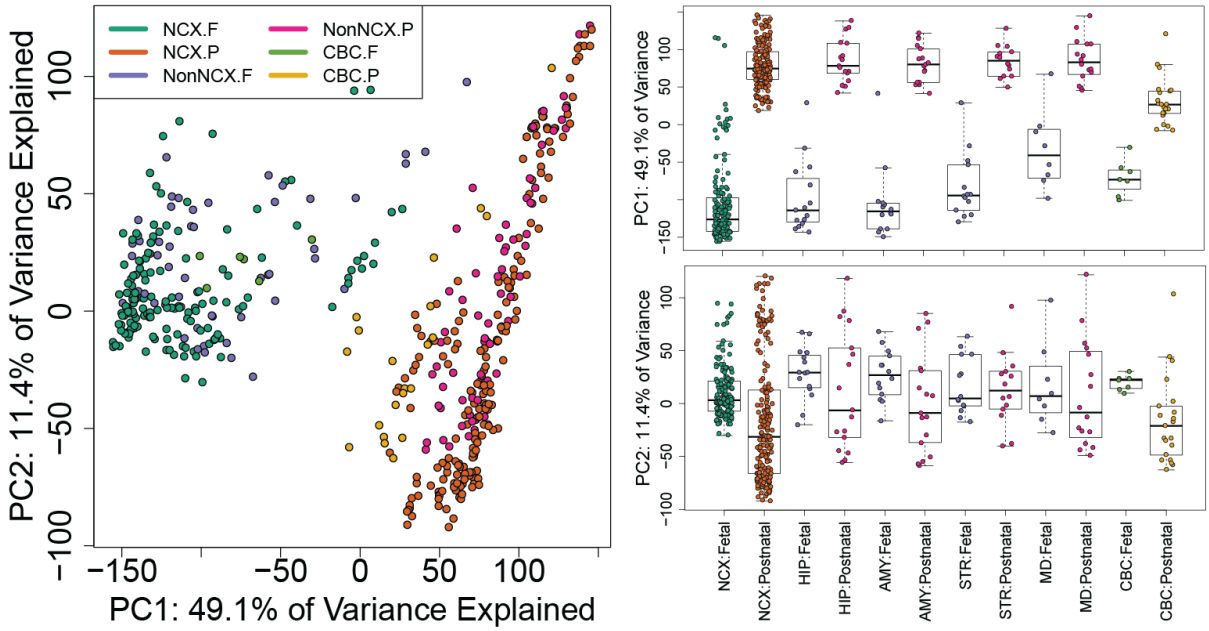

Figure 4: (Left) First two principal components (PCs) with samples colored by brain region and sample type (F: Fetal or P: Postnatal). (Right) Boxplots for PCs 1 and 2 by brain region (NCX: neocortex, HIP: hippocampus, AMY: amygdala, STR: striatum, MD: thalamus, CBC: cerebellum) and sample type with non-neocortex brain decomposed into its specific regions.

While the majority (67.4%) of single-base level DERs overlap exclusively exonic sequence in the latest Ensembl database (v75), we find that a fraction (22.5%) of the single-base level DERs map to sequence previously annotated as non-exonic (e.g. solely intronic or intergenic). The proportion of exonic sequence is higher than our previous analyses in the frontal cortex (Jaffe et al. 2014). When the single-base DERs are stratified by brain region and developmental period with the highest expression levels (Table 1), we find the highest degree of unannotated regulation in the cerebellum, the brain region with the largest degree of region-specific genes in a previous analyses (Kang et al. 2011). The majority of DERs, regardless of their annotation, are most highly expressed in fetal life,

particularly within the neocortex, hippocampus, and amygdala. Non-exonic expression might be due to incomplete transcript annotation in reference databases, backround expression, or previously undetected artifacts.

Table 1: Classification of single-base level DERs in the *BrainSpan* project. For each statistically significant DER, we identified the developmental period and region with the highest average expression levels, stratified by annotation relative to the Ensembl gene database. NCX: neocortex, HIP: hippocampus, AMY: amygdala, STR: striatum, MD: thalamus, CBC: cerebellum.

Group		Exonic	Intergenic	Intronic	Total
NCX	Fetal	14035	1813	1115	16963
	Postnatal	2869	911	418	4198
HIP	Fetal	13036	934	553	14523
	Postnatal	1047	248	149	1444
AMY	Fetal	15364	1240	762	17366
	Postnatal	1234	232	162	1628
STR	Fetal	7327	1800	1261	10388
	Postnatal	4840	1112	916	6868
MD	Fetal	4865	928	452	6245
	Postnatal	3049	437	356	3842
CBC	Fetal	10298	1901	1169	13368
	Postnatal	11661	3086	4078	18825

We then assessed differential expression using the complementary expressed-region level approach with the same statistical models, and first identified 174,639 contiguously expressed regions across the 487 samples (with mean across-sample normalized coverage > 0.25) constituting 34.52 megabases of expressed sequence. The majority (80.3%) of these expressed regions were strictly exonic while only a small subset (5.4%) were strictly non-exonic by Ensembl annotation. Using the same statistical models as the single-base level analysis (see Methods section 4.5.1), we found that 128,345 (73.5%) were differentially expressed by brain region and/or developmental stage at the region-level. These differentially expressed regions overlapped a total of 17,458 Ensembl genes (12,979 with gene symbols), representing a large portion of the known transcriptome. Of the significant expressed-region level DERs, 93,622 (72.9%) overlapped at least 1 significant single-base level DER (previously described). Lack of overlap results from almost half (45.8%) of single-base level DERs having an average coverage lower than the expression level determining expressed regions (0.25). For example, there was high expression only in the samples from a few brain regions, or only one development period. Decreasing the cutoff that defines the expressed regions from 0.25 to 0.1

results in a larger number of regions (217,178) that have a higher proportion of non-exonic sequence (12.2%), suggesting that the choice of this expression cutoff requires some initial exploratory data analysis.

Lastly, we highlight the utility of the expressed-region level analysis (using the original 0.25 cutoff) to identify regions differentially expressed within subsets of the data, for example across brain regions within a single developmental period. We identified that 1,176 expressed regions were differentially expressed comparing striatum versus hippocampus samples in the fetal developmental stage. These DERs mapped to 302 unique genes. Genes more highly expressed in the striatum include *ARPP-21*, previously shown to localize in the basal ganglia (Ouimet, Hemmings, and Greengard 1989), and dopamine receptor genes *DRD1* and *DRD2* (Cachope and Cheer 2014). Genes more highly expressed in the hippocampus in fetal life were strongly enriched for neuro-developmental genes including *FZD7* (Melchior et al. 2008), *ZBTB18* (Tatard et al. 2010), and *NEUROD1* (Poulin, Turgeon, and Drouin 1997). The expressed-region level analysis therefore permits subgroup analysis without the need to rerun the full `derfinder` single-base level pipeline.

2.4 Simulation results

We next performed a simulation to assess the ability of `derfinder` to identify true differential expression signal at both the single-base and expressed-sequence levels. We simulated data for 60 genes from chromosome 22 using the `polyester` framework (Frazee et al. 2014b). We used a 3 group design with 10 samples per group, used 40x read depth for the control/reference group, and then induced differential expression by inserting fold changes of 2x and $\frac{1}{2}$x for high and low expression, respectively, in certain transcripts. From the 60 selected genes, 24 genes have a single transcript, and we set half of them (12) to be differentially expressed. The remaining 36 genes have two transcripts - 12 genes have both transcripts differentially expressed, 12 have a single transcript differentially expressed, and the remaining 12 have neither transcript differentially expressed. Then, for each strand we divided the exons belonging to these transcripts into non-overlapping segments to determine the sensitivity and specificity of the derfinder approach. This resulted in 280 non-overlapping strand-specific exonic segments, where 169 were generated to be differentially expressed.

The single-base analysis resulted in 469 candidate DERs, of which 126 were significant when controlling the family-wise error rate (FWER adjusted p-value < 0.05). Table 2 (Left) shows

Table 2: Comparison between the 280 strand-specific exonic segments and the 126 and 125 significant single-base (Left) and expressed-region level DERs (Right), respectively.

		Single-base level				Expressed-region level		
		Overlaps significant DER				Overlaps significant DER		
		Yes	No	Sum		Yes	No	Sum
Diff. expressed	Yes	132	37	169	Yes	144	25	169
	No	0	111	111	No	0	111	111
	Sum	132	148	280	Sum	144	136	280

whether a strand-specific exonic segment overlaps a significant DER which is not strand specific. We did not identify any false positives, suggesting that the FWER was accurately controlled in this simulation. Conversely, the approach identified 37 false negatives, resulting in an empirical power of 78.1%. We note that 28 (75.7%) of the false negatives were from two-transcript genes where only one transcript was set to be differentially expressed. In most of these cases, a lower F-statistic cutoff would have likely identified these regions as differentially expressed.

Using the expressed-region level approach we identified a total of 249 expressed regions, of which 125 were significant (at Bonferroni adjusted p-value < 0.05). Table 2 (Right) shows the results from cross-tabulating differential expression between the exonic segments and these DERs. Like in the single-base level analysis, the approach did not detect any false positive regions, empirical power was 85.2%, and 18 (72%) of these false negatives were from two-transcript genes with only one set to be differentially expressed. These results further suggest that the single-base level analysis is more conservative than the expressed-region level analysis.

2.5 Computational considerations

We have demonstrated the ability of our `derfinder` software to identify annotation-agnostic regions of differential expression in real and simulated data. However, calculating coverage, summary statistics, and permutation statistics at base resolution can be extremely computationally intensive. We have made efforts to address these computational concerns by: (1) parallelizing computations where possible, (2) storing intermediate data files to manage memory load, and (3) implementing the expressed region approach to minimize the number of base-resolution statistics that must be calculated (Supplementary Section 1.1).

We use `IRanges`' efficient run-length encoding (Rle) infrastructure (Lawrence et al. 2013) to

minimize the required memory whenever possible. However, the F-statistics are more efficiently calculated outside of `Rle` objects, for which our implementation relies on efficient sparse matrix operations via the `Matrix` package (Bates and Maechler 2014). To mitigate the time consumed transforming the data for each permutation, we save subsets of data to temporary files speeding up the computation.

Several steps of the analysis can be parallelized by splitting the data into contiguous regions on chromosomes which the software takes advantage of, thus reducing the wall clock computation time (Supplementary Website). `derfinder` is flexible enough to run with different parallel computing implementations available in R. However, reducing the overall wall clock time does require a high performance computing environment. This becomes more necessary as the richness and size of the data set increase.

We present some data on computational time and memory use for the single-base level analysis with the *BrainSpan* data set as a benchmark. This analysis resulted in more than 170 megabases (5.5% of the genome) passing the initial data filter and approximately 171 x 10^9 F-statistics calculated after all 1000 permutations. At its peak `derfinder` used 510 cores to analyze all chromosomes simultaneously. In this analysis, chromosome 1 took 7 days to complete with 86 GB of memory used at the peak. Analyses of smaller data sets took 140 and 58 minutes to complete, using at its peak 96 and 48 cores respectively (Supplementary Website).

The expressed-region level analysis can be performed much faster and with lower number of cores than the single-base analysis. For example, the *BrainSpan* data set took 2.7 hours to load the data and 7 hours to determine the expressed regions with 10 and 5 cores respectively. That is, it took a total time of 9.7 hours to go from BigWig files to resulting expressed regions. Further timing and memory use results appear in Supplementary Section 1.6 and the Supplementary Website.

3 Conclusions

Here we introduced the `derfinder` statistical software for performing genome-scale annotation-agnostic RNA-seq differential expression analysis. This approach utilizes coverage-level information to identify differentially expression regions (DERs) at the single-base or expressed-region levels, and then generates useful summary statistics, visualizations and reports to further inspect and validate

candidate regions. Feature-level coverage can be easily computed, allowing for the implementation of complementary statistical analyses, such as differential gene and exon analysis (Supplementary Section 1.5). The reduced dependence on the transcriptome annotation permits the discovery of novel regulated transcriptional activity, such as the expression of sequences previously annotated as intronic or intergenic, which we highlight in publicly available RNA-seq data and our previous `derfinder` application (Jaffe et al. 2014). Furthermore, the structure of DERs across a given gene can permit the direct identification of differentially expressed transcripts (e.g. Figure 1), providing useful information for biologists running validation experiments.

The software pipeline, starting with BAM or BigWig files, and ending with lists of DERs, reports, and visualizations, runs at comparable speeds to existing RNA-seq analysis software. Given the appropriate computing resources, `derfinder` can scale to analyze studies with several hundred samples. Further work might be needed to scale to sample sizes in the thousands, which we foresee becoming a reality in the near future. For such large studies, it will be important to correct for batch effects and potentially expand `derfinder`'s statistical model for base-level covariates. The flexibility in defining the statistical model is one key advantage over other RNA-seq pipelines like `Cuffdiff` (Trapnell et al. 2013). Flexible models allow for a wide variety of biological questions that can be interrogated in the data, while `Cuffdiff` only allows for two groups comparisons. In summary, `derfinder` provides a unique statistical method for RNA-seq data to identify regions in the genome likely associated with a particular trait or disease, that can be integrated with complementary statistical approaches and used to prioritize regions of interest in a wide range of molecular systems.

4 Methods

4.1 Overview of R Implementation

We chose to implement `derfinder` entirely in the `R` statistical environment (R Core Team 2014). Our software includes upstream pre-processing of BAM and/or BigWig files into base-resolution coverage. At this stage the user can choose to summarize the base resolution coverage into feature level counts and apply popular feature-level RNA-seq differential expression analysis tools like `DESeq` (Anders and Huber 2010), `edgeR` (Robinson, McCarthy, and Smyth 2010), and `voom` (Law

et al. 2014). Since gene-counting approaches frequently filter out reads that overlap multiple gene features, using a base resolution counting approach before counting within features may allow more of the reads to be considered (Supplementary Section 1.5).

`derfinder` can be used to identify regions of differential expression agnostic to existing annotation. This can be done with either the single-base or the expressed-region level approaches, described in detail in the following subsections. The resulting regions can then be visualized to identify novel regions and filter out potential artifacts.

After differential expression analysis, `derfinder` can plot DERs using base-resolution coverage data by accessing the raw reads within differentially expressed regions for posthoc analysis like clustering and sensitivity analyses. We have also created a lightweight annotation function for quickly annotating DERs based on existing transcriptome annotation, including the UCSC knownGene hg19, Ensembl p12, and Gencode v19 databases.

Vignettes with detailed instructions and examples are available through the Bioconductor pages for `derfinder` (Collado-Torres et al. 2014) and `derfinderPlot` (Collado-Torres, Jaffe, and Leek 2014).

4.2 Single-base level `derfinder`

The single-base level approach implemented in `derfinder` requires two models. The alternative model (1) contains an intercept, the primary covariate of interest, and optionally adjustment variables. The primary variable can be as simple as a case-control variable or a more complicated model including smoothing functions (e.g. splines) over time. The adjustment variables include a library size normalization factor for raw data and optionally other potential confounders like age, sex, and batch variables.

$$y_{ij} = \alpha_i + \sum_{p=1}^{n} \beta_{ip} X_{jp} + \sum_{q=1}^{m} \gamma_{iq} Z_{jq} + \epsilon_{ij} \quad (1)$$

In both models y_{ij} is the scaled $\log_2$ base-level coverage for genomic position i and sample j. That is, $y_{ij} = \log_2 \left(\text{coverage}_{ij} + \text{scaling factor}\right)$. The model is completed by the n group effects $\beta_{\mathbf{i}}$, m adjustment variable effects $\gamma_{\mathbf{i}}$ and potentially correlated measurement error ϵ. The null model (2) is nested within model (1) and contains only the intercept and adjustment variables.

$$y_{ij} = \alpha_i + \sum_{q=1}^{m} \gamma_{iq} Z_{jq} + \epsilon_{ij} \quad (2)$$

`derfinder` uses a fixed design matrix, testing the same hypothesis at every base. This permits fast vectorized differential expression analysis. At each base we compute a moderated F-statistic (Smyth 2005) of the form in equation (3), where RSS0_i and RSS1_i are the residual sum of squares of the null and alternative models for base i. Furthermore, df_0 and df_1 are the degrees of freedom for the null (2) and alternative (1) models respectively, n is the number of samples, and an offset can be used for smaller experiments to shrink large F-statistics that may be driven by few biological replicates that cluster tightly.

$$F_i = \frac{(\text{RSS0}_i - \text{RSS1}_i)/(\text{df}_1 - \text{df}_0)}{\text{offset} + (\text{RSS1}_i/(n - \text{df}_1))} \quad (3)$$

We then perform "bump hunting" adapted to `Rle` objects in order to identify candidate DERs, R_k. Candidate DERs are defined as contiguous sets of bases where $F_i > T$ for a fixed threshold T. We then calculate an "area" statistic for each candidate DER which is the sum of the F-statistics above the threshold within the region: $S_k = \sum_{j \in R_k} F_j$ (Figure 1B). We have previously applied this approach to identify local differentially and variably methylated regions and more long range changes in methylation (Hansen et al. 2011; Jaffe et al. 2012a,b). One key difference compared to previous implementations in DNA methylation data is that we do not explicitly smooth the F-statistics, allowing for precise discovery of intron-exon boundaries in the data (Figure 1C).

Permutation analysis generates statistical significance for each of these candidate DERs by permuting the sample labels, re-calculating the F-statistics, identifying null candidate regions and region-level statistics in this permuted data set, and then calculating empirical p-values and/or directly estimating the family-wise error rate (FWER) (Jaffe et al. 2012a). Alternatively, the empirical p-values can be adjusted to control the false discovery rate (FDR) via `qvalue` (Dabney and Storey 2014).

4.3 Expressed-region level analysis

In the expressed-region approach, we compute the mean coverage for all base pairs from all the samples and filter out those below a user specified cutoff. Contiguous bases passing this filtering step are then considered a candidate region (Figure 2A). Then for each sample, we sum the base level coverage for each such region in order to create an expression matrix with one row per region and one column per sample. This matrix can then be used with feature-level RNA-seq differential expression analysis tools.

4.4 Annotation and "Genomic State" Objects

We have implemented a "genomic state" framework to efficiently annotate and summarize resulting regions, which assigns each base in the genome to exactly one state: exonic, intronic, or intergenic, based on any existing or user-defined annotation (e.g. UCSC, Ensembl, Gencode). At each base, we prioritize exon > intron > unannotated across all annotated transcripts.

Overlapping exons of different lengths belonging to different transcripts are reduced into a single "exonic" region, while retaining merged transcript annotations. We have a second implementation that further defines promoters and divides exonic regions into coding and untranslated regions (UTRs) which may be useful for the user to more specifically annotate regions - this implementation prioritizes coding exon > UTR > promoter > intron > unannotated.

4.5 Data Processing for Results in Main Manuscript

4.5.1 BrainSpan data

BigWig files for all 487 samples across 16 brain regions were downloaded from the *BrainSpan* website (BrainSpan 2011). Based on exploratory analyses, coverage was assumed to be reads per million in this data set, and we set the coverage filter to 0.25 for both the single-base and region-level `derfinder` approaches.

In the single base analysis, library size was not included in the models (as the coverage was already adjusted for this factor), a scaling factor of 1 was used, and the F-statistic cutoff T was chosen such that $P(F > T) = 10^{-6}$. We sought to identify differences in expression across brain region (neocortical regions: DFC, VFC, MFC, OFC, M1C, S1C, IPC, A1C, STC, ITC, V1C and

non-neocortical regions: HIP, AMY, STR, MD, and CBC) and developmental stage (fetal versus postnatal). We therefore fit the following region-by-stage interaction alternative model, which included main effects for fetal versus postnatal (binary) and categorical brain region variable (15 levels, relative to A1C), and interaction terms for each brain region and developmental stage. This resulted in a total of 32 terms in the model (intercept; 16 main effects, 15 interaction terms).

$$y_{ij} = \alpha_i + \beta_i Fetal_j + \sum_{q=1}^{m} \gamma_{iq} Region_{jq} + \sum_{q=1}^{m} \zeta_{iq} Fetal_j * Region_{jq} + \epsilon_{ij} \quad (4)$$

We compared the above model to an intercept-only model, and identified DERs using the single-base level analysis. We then calculated the mean coverage for each significant single-base DERs in each sample, resulting in a mean coverage matrix (DERs by samples), and we performed principal component analysis (PCA) on this $\log_2$-transformed matrix (after adding an offset of 1) , which were subsequently plotted in Figure 4.

For the expressed-region analysis, we used the same alternative model as in (4) where y_{ij} is the log(mean base-level coverage +1). We then compared it against the corresponding intercept-only model using the `lmFit` function from `limma` (Smyth 2005). The p-values for the expressed-region level DERs were adjusted via the Bonferroni method and those with adjusted p-values less than 0.05 were determined to be significant.

4.5.2 Simulated data

We simulated 100 bp paired-end reads (250bp fragments, sd = 25) with `polyester` (Frazee, Jaffe, and Leek 2014) for 3 groups with 10 samples each from human chromosome 22 at 40x coverage. 12 single transcript genes were set to be differentially expressed, 24 two transcript genes were differentially expressed (half only one transcript, half both transcripts), and 24 genes were set to background expression levels (half single, half two transcript genes). For each set of differentially expressed genes, 4 were assigned to each group with 2 genes having lower $\left(\frac{1}{2}\text{x}\right)$ and 2 higher (2x) expression levels in that group relative to the other two groups in order to balance the differential expression signal. The number of reads was generated with the `NB` function with size equal to one third of the mean. Reads were aligned to the genome using `TopHat` v2.0.13 using known transcripts of UCSC knownGene in the UCSC hg19 Illumina iGenomes distribution to guide the alignment,

as well as the known fragment information (the "`--transcriptome-index`", "`--mate-inner-dis`", and "`--mate-std-dev`" arguments in the software).

The single-base analysis approach was performed using models that adjusted for library size and tested for differences between the 3 groups. We used a data filter cutoff of 0 as we didn't expect reads to align to regions outside of the simulated genes. While a few reads aligned to other chromosomes, none of them resulted in DERs. Finally, we used a scaling factor of 32, a F-statistic cutoff T corresponding to $P(F > T) = 10^{-3}$, and 100 permutations.

For the expressed-region analysis, a mean coverage cutoff of 5 was used to determine the expressed regions. Then, differential expression was determined by fitting the same models as in the single-base analysis using the `lmFit` function from `limma` (Smyth 2005). Like with the *BrainSpan* data set, p-values for the expressed-region level DERs were adjusted via the Bonferroni method with a 0.05 cutoff for determining statistical significance.

Supplementary Materials

Supplementary files 1 through 3 contain the identified candidate single-base level DERs in CSV format (gzip compressed). The Supplementary Methods and Results describes in more detail the R implementation, coverage artifacts, benefits from base-level counting, and resources used by `derfinder`. The flexibility of the software is showcased by analyses of the effects of drug abuse on the human hippocampus (Zhou et al. 2011) and time course expression measurements in blood (Chen et al. 2012) from publicly available data sets.

The code used for all the analyses described in this paper is available at the Supplementary Website: `http://lcolladotor.github.io/derSoftware`. Several HTML reports are available in this site including

1. Brief results overview for each experiment.
2. Detailed description of fields from the CSV supplementary files 1, 2, and 3.
3. Timing and memory results from the different analyses.
4. Detailed information about the simulation including all code for generating the data and evaluating the results.

5. Using the effects on drug abuse data set, we compare single-base level `derfinder` results versus previously published results (Zhou et al. 2011) (Supplementary Section 1.4).

The `derfinder` (Collado-Torres et al. 2014) vignettes detail how to use the software and its infrastructure. In particular, the advanced vignette has visualizations showing the relationship between the different functions.

Funding

J.T.L. was supported by NIH Grant 1R01GM105705, L.C.T. was supported by Consejo Nacional de Ciencia y Tecnología México 351535.

Competing Interests

The authors declare that they have no competing interests.

References

Anders, Simon and Wolfgang Huber (2010). "Differential expression analysis for sequence count data". eng. In: *Genome biology* 11.10. PMID: 20979621 PMCID: PMC3218662, R106. ISSN: 1465-6914. DOI: 10.1186/gb-2010-11-10-r106.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (2014). "HTSeq–A Python framework to work with high-throughput sequencing data". In: *bioRxiv*.

Bates, Douglas and Martin Maechler (2014). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.1-4. URL: http://CRAN.R-project.org/package=Matrix.

BrainSpan (2011). *Atlas of the Developing Human Brain*. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. URL: http://developinghumanbrain.org.

Cachope, Roger and Joseph F. Cheer (2014). "Local control of striatal dopamine release". eng. In: *Frontiers in Behavioral Neuroscience* 8. PMID: 24904339, p. 188. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2014.00188.

Chen, Rui et al. (2012). "Personal omics profiling reveals dynamic molecular and medical phenotypes". eng. In: *Cell* 148.6. PMID: 22424236 PMCID: PMC3341616, pp. 1293–1307. ISSN: 1097-4172. DOI: 10.1016/j.cell.2012.02.009.

Collado-Torres, Leonardo, Andrew E. Jaffe, and Jeffrey T. Leek (2014). *derfinderPlot: Plotting functions for derfinder*. R package version 1.0.3. URL: http://www.bioconductor.org/packages/release/bioc/html/derfinderPlot.html.

Collado-Torres, Leonardo et al. (2014). *derfinder: Annotation-agnostic differential expression analysis of RNA-seq data at base-pair resolution*. R package version 1.0.10. URL: http://www.bioconductor.org/packages/release/bioc/html/derfinder.html.

Dabney, Alan and John D. Storey (2014). *qvalue: Q-value estimation for false discovery rate control*. R package version 1.40.0. URL: http://www.bioconductor.org/packages/release/bioc/html/qvalue.html.

Daines, Bryce et al. (2011). "The Drosophila melanogaster transcriptome by paired-end RNA sequencing". eng. In: *Genome research* 21.2. PMID: 21177959 PMCID: PMC3032934, pp. 315–324. ISSN: 1549-5469. DOI: 10.1101/gr.107854.110.

Dillman, Allissa A et al. (2013). "mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex". eng. In: *Nature neuroscience* 16.4. PMID: 23416452 PMCID: PMC3609882, pp. 499–506. ISSN: 1546-1726. DOI: 10.1038/nn.3332.

ENCODE Project Consortium et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". eng. In: *Nature* 489.7414. PMID: 22955616 PMCID: PMC3439153, pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247.

Farrell, Catherine M et al. (2014). "Current status and new features of the Consensus Coding Sequence database". eng. In: *Nucleic acids research* 42.Database issue. PMID: 24217909 PMCID: PMC3965069, pp. D865–872. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1059.

Frazee, Alyssa C., Andrew E. Jaffe, and Jeffrey T. Leek (2014). *polyester: Simulate RNA-seq reads*. R package version 1.1.0.

Frazee, Alyssa C et al. (2014a). "Differential expression analysis of RNA-seq data at single-base resolution". ENG. In: *Biostatistics (Oxford, England)*. PMID: 24398039. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxt053.

Frazee, Alyssa C et al. (2014b). "Polyester: simulating RNA-seq datasets with differential transcript expression". In: *bioRxiv*. DOI: 10.1101/006015.

GTEx Consortium (2013). "The Genotype-Tissue Expression (GTEx) project". eng. In: *Nature genetics* 45.6. PMID: 23715323, pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

Hansen, Kasper Daniel et al. (2011). "Increased methylation variation in epigenetic domains across cancer types". eng. In: *Nature genetics* 43.8. PMID: 21706001 PMCID: PMC3145050, pp. 768–775. ISSN: 1546-1718. DOI: 10.1038/ng.865.

Jaffe, A. E. et al. (2014). "Developmental regulation of human cortex transcription and its clinical relevance at single base resolution". In: *Nat. Neurosci.*

Jaffe, Andrew E et al. (2012a). "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies". eng. In: *International journal of epidemiology* 41.1. PMID: 22422453 PMCID: PMC3304533, pp. 200–209. ISSN: 1464-3685. DOI: 10.1093/ije/dyr238.

Jaffe, Andrew E et al. (2012b). "Significance analysis and statistical dissection of variably methylated regions". eng. In: *Biostatistics (Oxford, England)* 13.1. PMID: 21685414 PMCID: PMC3276267, pp. 166–178. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxr013.

Kang, Hyo Jung et al. (2011). "Spatio-temporal transcriptome of the human brain". eng. In: *Nature* 478.7370. PMID: 22031440, pp. 483–489. ISSN: 1476-4687. DOI: 10.1038/nature10523.

Kim, Daehwan et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". ENG. In: *Genome biology* 14.4. PMID: 23618408, R36. ISSN: 1465-6914. DOI: 10.1186/gb-2013-14-4-r36.

Lappalainen, Tuuli et al. (2013). "Transcriptome and genome sequencing uncovers functional variation in humans". eng. In: *Nature* 501.7468. PMID: 24037378 PMCID: PMC3918453, pp. 506–511. ISSN: 1476-4687. DOI: 10.1038/nature12531.

Law, Charity W et al. (2014). "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts". ENG. In: *Genome biology* 15.2. PMID: 24485249, R29. ISSN: 1465-6914. DOI: 10.1186/gb-2014-15-2-r29.

Lawrence, Michael et al. (2013). "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9 (8). DOI: 10.1371/journal.pcbi.1003118. URL: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118.

Leek, Jeffrey T et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10, pp. 733–739.

Leśniewska, Anna and Michał J Okoniewski (2011). "rnaSeqMap: a Bioconductor package for RNA sequencing data exploration". In: *BMC bioinformatics* 12.1, p. 200.

Melchior, Kai et al. (2008). "The WNT receptor FZD7 contributes to self-renewal signaling of human embryonic stem cells". eng. In: *Biological Chemistry* 389.7. PMID: 18681827, pp. 897–903. ISSN: 1431-6730. DOI: 10.1515/BC.2008.108.

Mortazavi, Ali et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature methods* 5.7, pp. 621–628.

Ouimet, C. C., H. C. Hemmings, and P. Greengard (1989). "ARPP-21, a cyclic AMP-regulated phosphoprotein enriched in dopamine-innervated brain regions. II. Immunocytochemical localization in rat brain". eng. In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 9.3. PMID: 2538585, pp. 865–875. ISSN: 0270-6474.

Poulin, G., B. Turgeon, and J. Drouin (1997). "NeuroD1/beta2 contributes to cell-specific transcription of the proopiomelanocortin gene". eng. In: *Molecular and Cellular Biology* 17.11. PMID: 9343431, pp. 6673–6682. ISSN: 0270-7306.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". eng. In: *Bioinformatics (Oxford, England)* 26.1. PMID: 19910308 PMCID: PMC2796818, pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

Smyth, Gordon K (2005). "Limma: linear models for microarray data". In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by R. Gentleman et al. New York: Springer, pp. 397–420.

Steijger, Tamara et al. (2013). "Assessment of transcript reconstruction methods for RNA-seq". eng. In: *Nature methods* 10.12. PMID: 24185837 PMCID: PMC3851240, pp. 1177–1184. ISSN: 1548-7105. DOI: 10.1038/nmeth.2714.

Storey, John D and Robert Tibshirani (2003). "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16, pp. 9440–9445.

Tatard, Valérie M. et al. (2010). "ZNF238 is expressed in postmitotic brain cells and inhibits brain tumor growth". eng. In: *Cancer Research* 70.3. PMID: 20103640, pp. 1236–1246. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-09-2249.

Trapnell, Cole et al. (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq". eng. In: *Nature biotechnology* 31.1. PMID: 23222703 PMCID: PMC3869392, pp. 46–53. ISSN: 1546-1696. DOI: 10.1038/nbt.2450.

Yin, Tengfei, Dianne Cook, and Michael Lawrence (2012). "ggbio: an R package for extending the grammar of graphics for genomic data". In: *Genome Biology* 13.8, R77.

Zhou, Zhifeng et al. (2011). "Substance-specific and shared transcription and epigenetic changes in the human hippocampus chronically exposed to cocaine and alcohol". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.16. PMID: 21464311 PMCID: PMC3081016, pp. 6626–6631. ISSN: 1091-6490. DOI: 10.1073/pnas.1018514108.

Last compiled at 18:46:54 (GMT) on 2015/02/17