Supplementary Methods and Results

This document describes R implementation details of `derfinder`, the single base-level approach, and results from applying the single base-level approach to the *BrainSpan* data set. It also includes the simulation results when performing the statistical tests using `edgeR`-robust [1] or `DESeq2` [2] instead of `limma` [3].

1 Supplementary Results

1.1 R implementation

The `derfinder` package can be used for different types of analyses such as DER finding (single base-level and ER-level approaches) as well as creating a feature counts matrix. The overall relationship between these functions is shown in section *Flow charts* subsection *DER analysis flow chart* of the *derfinder users guide* vignette available at `www.bioconductor.org/packages/derfinder`.

For the single base-level approach, the main function is `analyzeChr()` which makes it easier for users to run this type of analysis. This function is a wrapper for other functions available in `derfinder`, as can be seen section *Flow charts* subsection *analyzeChr() flow chart* of the *derfinder users guide* vignette. It splits the data, calculates the F-statistics, identifies the null regions, and annotates them.

The expressed regions (ERs) approach is described in section *Flow charts* subsection *regionMatrix() flow chart* of the *derfinder users guide* vignette. This type of analysis requires fewer functions, as the user only needs to load the data and then identify the ERs with the `regionMatrix()` function. The *regionMatrix() flow chart* shows which other functions are internally used by `regionMatrix()` that filter the coverage by using a mean cutoff, identify the regions, and produce the region-level count matrix. The function `railMatrix()` is optimized for identifying ERs from BigWig files, specially those created with `Rail-RNA` (DOI: 10.1101/019067).

1.2 Differential expression in the developing human brain via expressed region-level analysis

Figure S1 complements Figure 5 with the results of performing principal component analysis of ERs found in the *BrainSpan* data set given the known annotated elements they overlap with. The results are consistent regardless of the type of ERs under study.

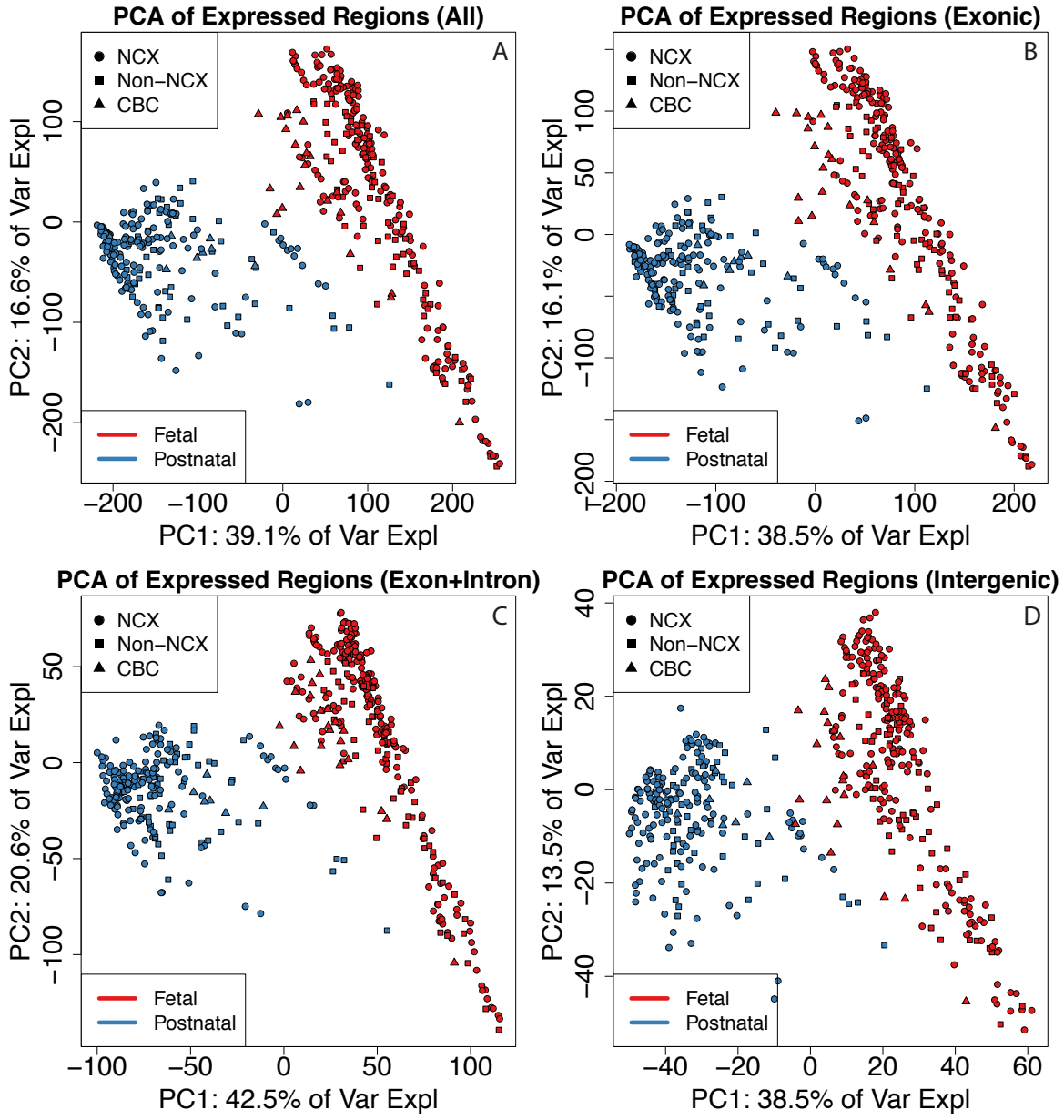

Figure S1: Principal components analysis reveals clusters of samples in the BrainSpan data set. First two principal components (PCs) with samples colored by sample type (F: Fetal or P: Postnatal) and shape given by brain region using all ERs (top left), strictly exonic ERs (top right), ERs overlapping exons and introns (bottom left) and strictly intergenic ERs (bottom right).

1.3 Single base-level statistical test

A single base-level resolution analysis in `derfinder` starts with read alignment and coverage calculation as done in the ER-level approach. 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 [4], demographic variables, and batch effects [5].

Once an F-statistic is calculated at each base, we identify differentially expressed regions (DERs) using a "bump hunting" approach [6]. First we find candidate DERs by identifying regions of the genome where the base-level F-statistics pass a genome-wide threshold (Figure S2 with *BrainSpan* data set, see Supplementary Section 2.1). 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. Alternatively, the region-level p-values can be adjusted for multiple testing using standard false discovery rate techniques [7, 8].

1.4 Differential expression in the developing human brain via single base-level analysis

At the single base-level, we identified 113,691 genome-wide significant DERs (FWER $< 5\%$) with the same statistical models used with the ER-level analysis described in the main text. These resulting single base-level DERs largely distinguished the fetal and postnatal samples representing the first principal component and 49.4% of the variance of the mean coverage levels within the DERs (Figure S3). The most significant DERs map to genes previously implicated in development, and contained many of the DERs we previously identified in the frontal cortex in 36 independent subjects [9]. For example, 59% 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 S3, left panel).

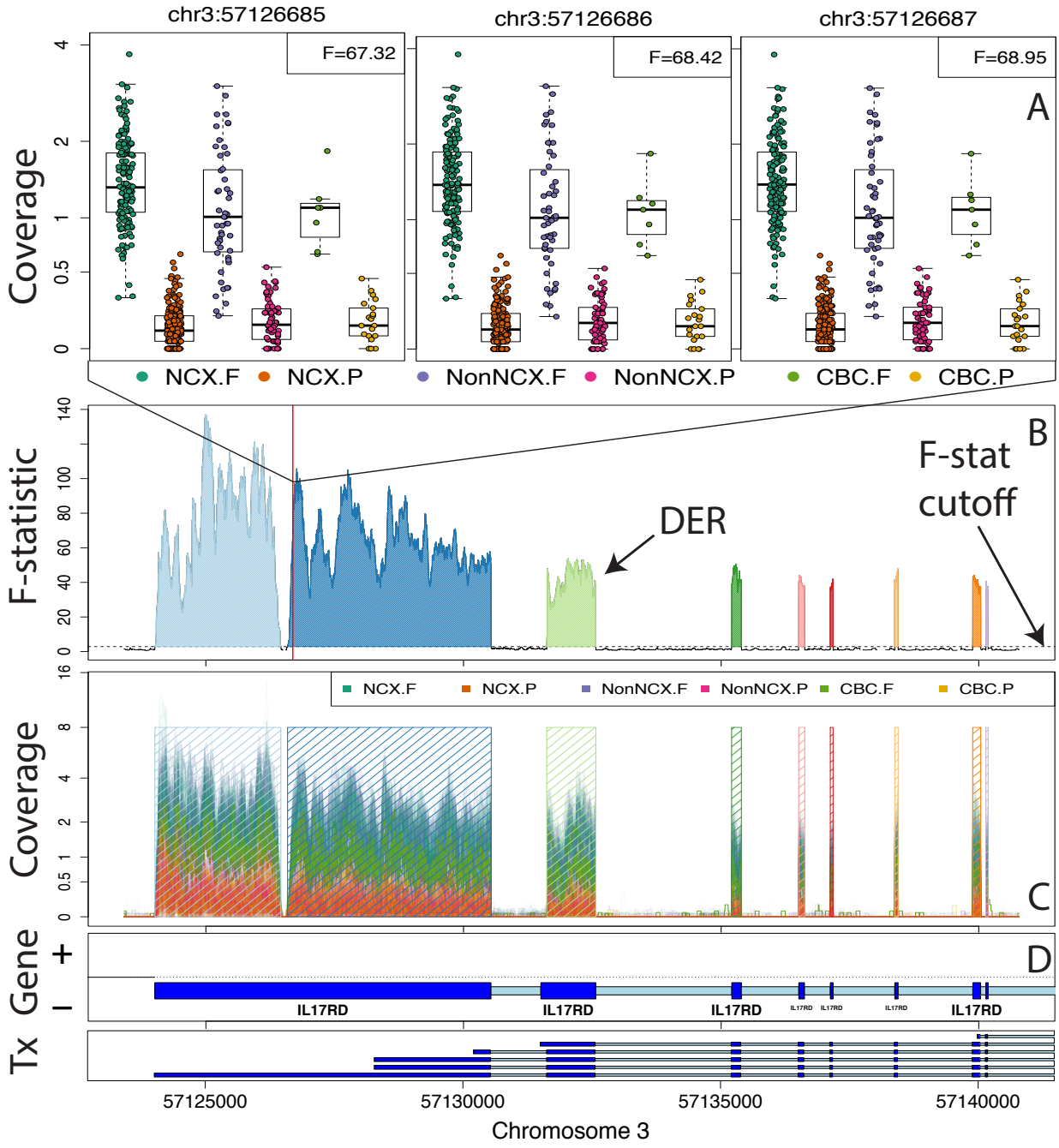

Figure S2: Finding DERs on chromosome 3 with *BrainSpan* data set using six groups: Neocortical regions (NCX: DFC, VFC, MFC, OFC, M1C, S1C, IPC, A1C, STC, ITC, V1C), Non-neocortical regions (NonNCX: HIP, AMY, STR, MD), and cerebellum (CBC) split by whether the sample is from a fetal (F) or postnatal (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.

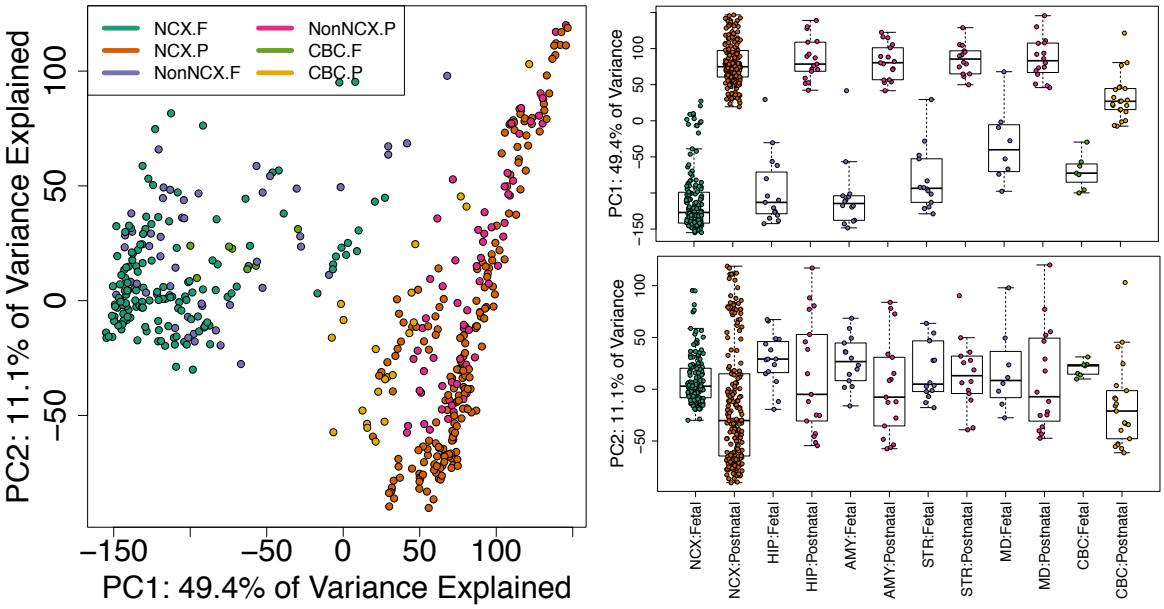

Figure S3: Principal components analysis reveals clusters of samples in the BrainSpan data set. (Left) First two principal components (PCs) with samples colored by sample type (F: Fetal or P: Postnatal) and shape given by brain region. (Right) Boxplots for PCs 1 and 2 by brain region (NCX: neocortex, 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 (68.1%) of single base-level DERs overlap exclusively exonic sequence using Ensembl database v75, we find that a fraction (22.2%) 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 [9]. When the single base-level DERs are stratified by brain region and developmental period with the highest expression levels (Table S1), 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 [10]. 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, background expression, or previously undetected artifacts.

Table S1: 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. Region assignment is prioritized by exon > intron > intergenic.

Group		Exonic	Intergenic	Intronic	Total
NCX	Fetal	15583	1946	1196	18725
	Postnatal	2750	882	415	4047
HIP	Fetal	12511	889	523	13923
	Postnatal	1021	237	144	1402
AMY	Fetal	14705	1178	727	16610
	Postnatal	1193	229	167	1589
STR	Fetal	6952	1706	1199	9857
	Postnatal	4734	1060	905	6699
MD	Fetal	4671	890	431	5992
	Postnatal	2922	425	348	3695
CBC	Fetal	9984	1815	1118	12917
	Postnatal	11382	2932	3921	18235

1.5 Exploratory analysis of the cutoff used for the expressed regions-level analysis in the developing human brain

The cutoff used in the expressed regions-level `derfinder` analysis impacts how many ERs are found (Figure S4A), their length in base pairs (width, Figure S4B). It can also affect the percent of the known annotation that at least overlaps one ER (Figure S4C) and conversely the percent of ERs that overlap at least one known exon (Figure S4D). Figure S4 shows the effect of the cutoff used with the *BrainSpan* data set for a range of cutoffs from 0.025 to 0.5 in increments of 0.025. Note that this data set was already normalized to a library size of 1 million reads. We recommend choosing a cutoff in the elbow of these curves. In Section 3.6 we present the results for cutoffs 0.1 and 0.25 which are at the beginning and the end of the elbow, respectively.

1.6 Simulation analysis

1.6.1 Simulation results with `DESeq2` or `edgeR-robust`

Table S2 shows the empirical power, false positive rate (FPR) and false discovery rate (FDR) for the different analysis pipelines that result in a count matrix which we analyzed with `DESeq2` [2]

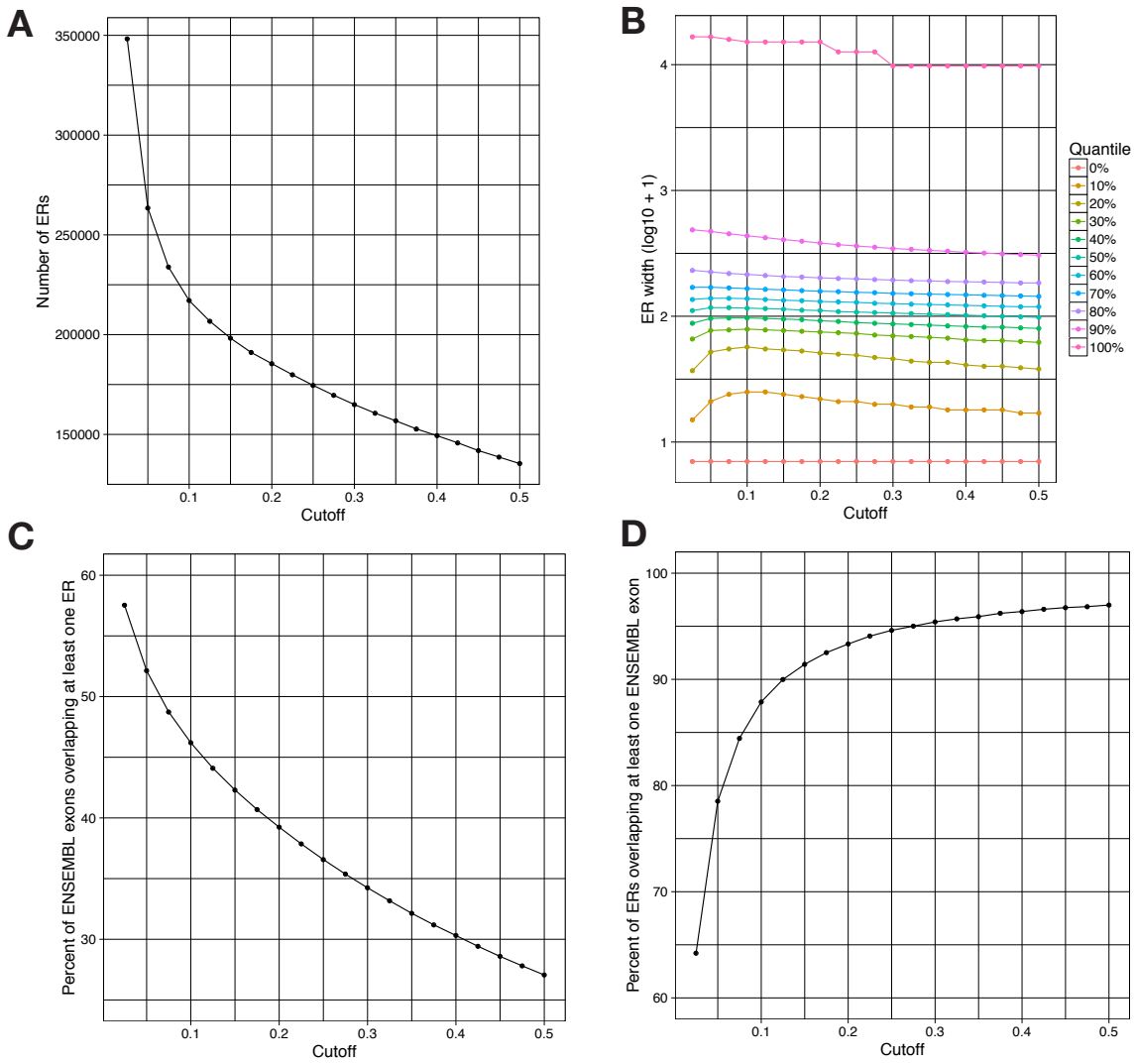

Figure S4: Exploratory analysis of the expressed regions cutoff used for the *BrainSpan* data set. **A** Relationship between number of ERs of at least 6 base-pairs in length against the cutoff used in Figure 2A. **B** Distribution of the width of the ERs for each cutoff summarized by quantiles in 10% increments and $\log_1 0$ transformed. **C** Percent of ENSEMBL v75 exons overlapping at least one ER by cutoff. **D** Percent of ERs overlapping at least one ENSEMBL v75 exon by cutoff.

or `edgeR`-robust [11] while controlling the FDR to 5%. The observed power for `edgeR`-robust is slightly higher than the corresponding results using `DESeq2` [2]. The observed FPR and FDR with `edgeR`-robust are higher than in the `DESeq2` results, with overlapping ranges for the `derfinder` analyses and non-overlapping ones when summarizing the data with `featureCounts` [1].

Table S2: Simulation results for pipelines that used `DESeq2` or `edgeR`-robust for the statistical tests. Minimum and maximum empirical power, false positive rate (FPR) and false discovery rate (FDR) observed from the three simulation replicates for each analysis pipeline that resulted in a count matrix analyzed with `DESeq2` or `edgeR`-robust. `ballgown` analyses were done at either the exon or transcript levels. Pipelines that rely on annotation were run with the full annotation or with 20% of the transcripts missing (8.28% exons missing). fCounts is short for featureCounts.

Power	FPR	FDR	Annotation complete	Aligner	Summary method	Statistical method
(94.3-95.5)	(6.8-9)	(12.2-16.1)		HISAT	derfinder	DESeq2
(94.1-95.1)	(6.5-8.3)	(11.1-15.7)		Rail-RNA	derfinder	DESeq2
(95.6-96.1)	(8.4-10.5)	(13.9-18.9)		HISAT	derfinder	edgeR
(94.7-95.9)	(8.3-10.2)	(12.8-18.5)		Rail-RNA	derfinder	edgeR
(70-78.3)	(2.1-3.6)	(5.6-9.8)	No	HISAT	fCounts	DESeq2
(95.1-95.9)	(2.9-4.8)	(6.3-9.7)	Yes	HISAT	fCounts	DESeq2
(71.2-79.2)	(5-8.1)	(12-19.5)	No	HISAT	fCounts	edgeR
(96.4-97.2)	(6.8-9.9)	(12.7-18.1)	Yes	HISAT	fCounts	edgeR

1.6.2 Timing and computational resources used

Table S3 shows a summary of the computational resources used for the different pipelines used in the simulation as well as the time for running them. In general, the maximum memory per core is low (most are below 3.2 GB) regardless of the analysis step. The exception is alignment with `Rail-RNA` because of how our computing cluster measures memory usage: it artificially increases when processes spawn shared-many sub-processes by counting more than once the memory used by shared objects. Time-wise all analysis steps except for alignment take only 11 minutes at most. Notably, the ER-level approach is much faster with `Rail-RNA` output than with `HISAT` output. This is because `derfinder` can load the data much faster from BigWig files than from BAM alignment files and the `railMatrix` has been optimized for the BigWig files that `Rail-RNA` produces. In this particular simulation, `Rail-RNA` is slower than `HISAT` for aligning reads, but this is expected since `Rail-RNA` is better suited at analyzing larger data sets in the cloud and decreasing false positives when determining new splice junctions. This is reflected on Table 1 and S2 with slightly reduced FPR and FDR when using `Rail-RNA` compared to `HISAT`. The timing results for each computing job are available in the Supplementary Website.

Table S3: Summary of computing resources required for each analysis step for the different simulation pipelines. This table shows the maximum memory (GB) per core, the time in minutes to run the analysis with all jobs running sequentially and the maximum number of cores used in any step of the simulation analysis for the different pipelines. Note that the ERs (HISAT), the feature-level counts and ballgown pipelines rely on HISAT alignments.

Max memory by core (GB)	Time (minutes)	Peak cores	Pipeline	Analysis step
(2.8-3.1)	(2.1-3.2)	10	ER-level (Rail-RNA)	Align prep
(32.8-39.1)	(137.6-218.4)	10	ER-level (Rail-RNA)	Align
(3.2-3.2)	(47.2-72.1)	40	HISAT	Align
(1.4-1.4)	(1.5-1.5)	1	ER-level (Rail-RNA)	Summarize
(0.6-0.6)	(1.3-1.9)	1	ER-level (Rail-RNA)	Statistical tests
(0.8-0.8)	(5-7)	4	ER-level (HISAT)	Summarize
(0.6-0.6)	(1.5-1.9)	1	ER-level (HISAT)	Statistical tests
(2.2-2.2)	(1.6-1.6)	8	Feature counts	Summarize
(0.6-0.6)	(3.7-5.3)	2	Feature counts	Statistical tests
(2.1-2.1)	(8.7-11)	80	StringTie-Ballgown	Summarize
(0.7-0.7)	(0.7-0.8)	2	StringTie-Ballgown	Statistical tests

2 Supplementary Methods

2.1 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 can include a library size normalization factor for raw data and optionally other potential confounders like age, sex, and batch variables. There are different library size normalization factors you can consider using and `derfinder` implements a version in the `sampleDepth` function based on Paulson et. al [12].

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

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.

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

`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 [3] 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.

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

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 S2B). We have previously applied this approach to identify local differentially and variably methylated regions and more long range changes in methylation [6, 13, 14]. 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 S2C).

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) [6]. Alternatively, the empirical p-values can be adjusted to control the false discovery rate (FDR) via `qvalue` [7].

2.2 Data Processing: BrainSpan data

For the single base-level analysis, we used a scaling factor of 1 and chose the F-statistic cutoff T such that $P(F > T) = 10^{-6}$. We used the same alternative model described for the expressed region analysis in the main text. We compared the alternative 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-level 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 S3.

References

[1] Liao, Y., Smyth, G. K., and Shi, W. (April, 2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* **30**(7), 923–930.

[2] Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* **15**(12), 550.

[3] Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology,* **3**(1), 1–25.

[4] 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.

[5] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics,* **11**(10), 733–739.

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

[7] Dabney, A. and Storey, J. D. qvalue: Q-value estimation for false discovery rate control (2014) R package version 1.40.0.

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

[9] Jaffe, A. E., Shin, J., Collado-Torres, L., Leek, J. T., Tao, R., Li, C., Gao, Y., Jia, Y., Maher, B. J., Hyde, T. M., Kleinman, J. E., and Weinberger, D. R. (January, 2015) Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Neuroscience,* **18**(1), 154–161.

[10] Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Z., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., Weinberger, D. R., Mane, S., Hyde, T. M., Huttner, A., Reimers, M., Kleinman, J. E., and Sestan, N. (October, 2011) Spatio-temporal transcriptome of the human brain. *Nature,* **478**(7370), 483–489.

[11] Zhou, X., Lindsay, H., and Robinson, M. D. (June, 2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research,* **42**(11), e91.

[12] Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (September, 2013) Differential abundance analysis for microbial marker-gene surveys. *Nature methods,*.

[13] Jaffe, A. E., Feinberg, A. P., Irizarry, R. A., and Leek, J. T. (January, 2012) Significance analysis and statistical dissection of variably methylated regions. *Biostatistics,* **13**(1), 166–178.

[14] Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., and Feinberg, A. P. (August, 2011) Increased methylation variation in epigenetic domains across cancer types. *Nature genetics,* **43**(8), 768–775.

Last compiled at 00:15:52 (GMT) on 2016/07/29