Supplementary Material for Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates

Scott Norton [1], Jorge Vaquero-Garcia [1,2], Nicholas F. Lahens [1],
Gregory R Grant [1], and Yoseph Barash [1,2,]

1 Datasets

Four datasets informed the analyses presented in this work. The primary dataset, identified as "Hogenesch", was drawn from Zhang et al. 2014, who tracked gene expression in 12 different tissues, including cerebellum, liver, and muscle, in mice across a 24-hour circadian cycle. RNA was extracted and sequenced at eight timepoints spaced 6 hours apart according to the published procedure, such that each time point was repeated twice.

RT-PCR for a selection of 50 alternative splicing events predicted from these data were validated experimentally using RT-PCR. The procedure is described in the Materials and Methods section (RT-PCR validations) of Vaquero et al. 2016, with the list of events provided as Supplementary File 2 in the same paper.

A second dataset, identified as "MGP", was used to supplement the findings from Hogenesch and evaluate the hyperparameters for MAJIQ weights computation. MGP consists of six biological replicates of each of six tissues, including lung, liver, and hippocampus. RNA was extracted from each animal and sequenced according to the procedure published in Keane et al. 2011.

MGP liver and hippocampus were used to inform simulated RNA-seq experiments. These datasets were created using the procedure described in Section 5.

Samples from the GTEx consortium were used in tandem with Hogenesch and MGP to evaluate the MAJIQ weights hyperparameters. Specifically, heart SRR608096, SRR659637, and SRR808808, and adipose SRR1102631, were acquired and mapped.

Reads from all datasets were mapped using STAR 2.5.3a with a minimum splice junction overhang size of 8 positions. Hogenesch, MGP, and the simulated datasets used the mm10 transcriptome assembly from ENSEMBL, whereas GTEx used the hg19 assembly. For SUPPA, which was only run on the mouse-derived datasets, the RefSeq annotation was used in accordance with the authors' recommendations.

2 Synthetic perturbation procedure

To create an outlier sample by synthetically perturbing a real sample we use three parameters to control the sample's perturbation: $\theta \in [0, 1]$ determined the fraction of LSVs randomly selected to be perturbed, $\delta \in [0, 1]$ controlled the shift of $E[\Psi]$, and $\gamma \geq 0$ set the relative coverage level in the perturbed sample. Specifically, we use the following procedure:

1. For each $l \in$ LSVs, sample the read rates $\mu_{l,j}$ for each junction $1 \leq j \leq J$, where J is the number of junctions in l. The read rates are a proxy for expression and are described in the Materials and Methods section (MAJIQ quantifier) of Vaquero et al. 2016.
2. Set $\theta \in [0, 1]$, $\delta \in [0, 1]$, and $\gamma > 0$.
3. Randomly sample $L \subset$ LSVs with $|L| = \theta|\text{LSVs}|$.
4. For $l \in L$ with per-junction read rates $\mu_{l,j}, j = 1, \ldots, J$:
   (a) Estimate $E[\Psi_{l,j}]$ for each junction.

(b) Sample $\varepsilon \sim U(0,1)$ and let

$$\sigma = \begin{cases} -1, & \varepsilon < E[\Psi_{l,1}], \\ 1, & \text{else.} \end{cases}$$

(c) Set $E[\Psi^*_{l,1}] = \min(\max(E[\Psi_{l,j}] + \sigma\delta, 0), 1)$.

(d) For $2 \leq j \leq J$, set

$$E[\Psi^*_{l,j}] = E[\Psi_{l,j}] + \frac{E[\Psi_{l,1}] - E[\Psi^*_{l,1}]}{J-1}.$$

(e) For $1 \leq j \leq J$, set

$$\mu^*_{l,j} = E[\Psi^*_{l,j}] \frac{\mu_{l,j}}{\sum_{k=1}^{J} \mu_{l,k}}.$$

5. For $l \in$ LSVs, set $\mu^*_{l,j} = \gamma\mu_{l,j}$.

3 Tissue swap procedure

The swaps described in the main paper and portrayed in Figure 5a of the same used **Cer**ebellum, **Liv**er, and **Mus**cle from Hogenesch. As a control, we quantified $\Delta\Psi$ or differential transcript expression levels between Cer_CT28, Cer_CT34, Cer_CT40 and Liv_CT28, Liv_CT34, Liv_CT40. To show reproducibility of detected differences, we quantified differential inclusion or expression levels between Cer_CT46, Cer_CT52, Cer_CT58 and Liv_CT46, Liv_CT52, Liv_CT58. The swaps were quantified in triplicate as follows:

Trial 1: Mus_CT28, Cer_CT34, Cer_CT40 vs Liv_CT28, Liv_CT34, Liv_CT40

Trial 2: Cer_CT28, Mus_CT34, Cer_CT40 vs Liv_CT28, Liv_CT34, Liv_CT40

Trial 3: Cer_CT28, Cer_CT34, Mus_CT40 vs Liv_CT28, Liv_CT34, Liv_CT40

To demonstrate consistency of these results on a different dataset, we executed a similar procedure using MGP and portray the reproducibility ratio for each method in Figure Supplementary 2d. Quantification of differential events between matched sets of three Lung and three Livers served as a control, with the remaining three of each tissue serving to inform reproducibility. In each swap trial, a lung was swapped with a matched hippocampus.

4 Differential splicing quantification tools

MAJIQ 1.0.6 was executed using default parameters against the ENSEMBL annotation for the respective species, and the quantifier was also executed with default parameters for the quantifiability filter and weights tolerance except where noted. MAJIQ-nw (no weights) was executed with `--weights None None`; MAJIQ-gw (global weights) was executed with `--weights Auto Auto --weights_local 0`; and MAJIQ-lw (local weights) was executed with `--weights Auto Auto --weights_local 0.05`. rMATS-turbo version 0.1, SUPPA commit `56b5427`, and DEXSeq 1.18.4 were executed with their respective default parameters. Additional details and specifics are provided in the script release package.

5 Synthetic data generation

Simulated RNA-Seq data was generated using the BEERS simulator [1]. For all analyses described below, gene models from release 75 of the ENSEMBL mm10 annotation were used, and sequence information from the mm10 build of the mouse genome. Empirical gene expression counts and PSI values were inferred separately from five hippocampus and six liver samples. Of the 41,133 annotated genes expressed in the empirical data, 3,055 were chosen at random to reflect the empirical PSI values for their associated transcripts. Let S be this set of genes. These genes will reflect any real differential splicing that exists between the two tissues. The remaining genes were simulated to reflect non-differential spicing as follows: A gene with n

transcripts is assigned the empirical PSI values of a randomly chosen gene in a randomly chosen sample where the gene is chosen randomly from the subset of genes in S which have n splice forms. These PSI values were then assigned to this gene in all samples. To introduce inter-sample variability in the PSI values for all such genes (genes not in S), a small random number uniformly between 0 and 0.025 was generated and added/subtracted from the PSI values for n pairs, so as to maintain that that the PSI values for all transcripts in a gene will still sum to 1 in each sample. This table of PSI values was then transformed into counts for the BEERS feature quantification files to generate simulated data with the same expression distributions as the real samples, assuming the generation of exactly the same number of simulated reads as in the real samples. Simulated data was generated with BEERS under two difference scenarios. The first dataset reflects ideal data with no polymorphisms or errors, and uniform coverage across the length of each transcript. In the second dataset, polymorphisms and errors were introduced according to the following parameters: substitution frequency = 0.001, indel frequency = 0.0001, error rate = 0.005. This dataset also has non-uniform coverage mimicking the 3' bias typical of polyA selection; which was inferred empirically from [2]. Both datasets included 5% intron signal.

6 Technical replicates data generation

Technical replicates were created as descrpibed in the main text by randomly partitioning each original FASTQ file to similarly sized subsets. To simplify notation, we will say, for example, that Cer_CT28 was split into Cer_CT28_1 and Cer_CT28_2. We computed the number N of significant events detected by each method in two runs designed as between groups of these "technical" replicates: Cer_CT28_1, Cer_CT34_1, Cer_CT40_1 vs Cer_CT28_2, Cer_CT34_2, Cer_CT40_2 (Cerebellum), and Liv_CT28_1, Liv_CT34_1, Liv_CT40_1 vs Liv_CT28_2, Liv_CT34_2, Liv_CT40_2 (Liver).

Supplementary 1

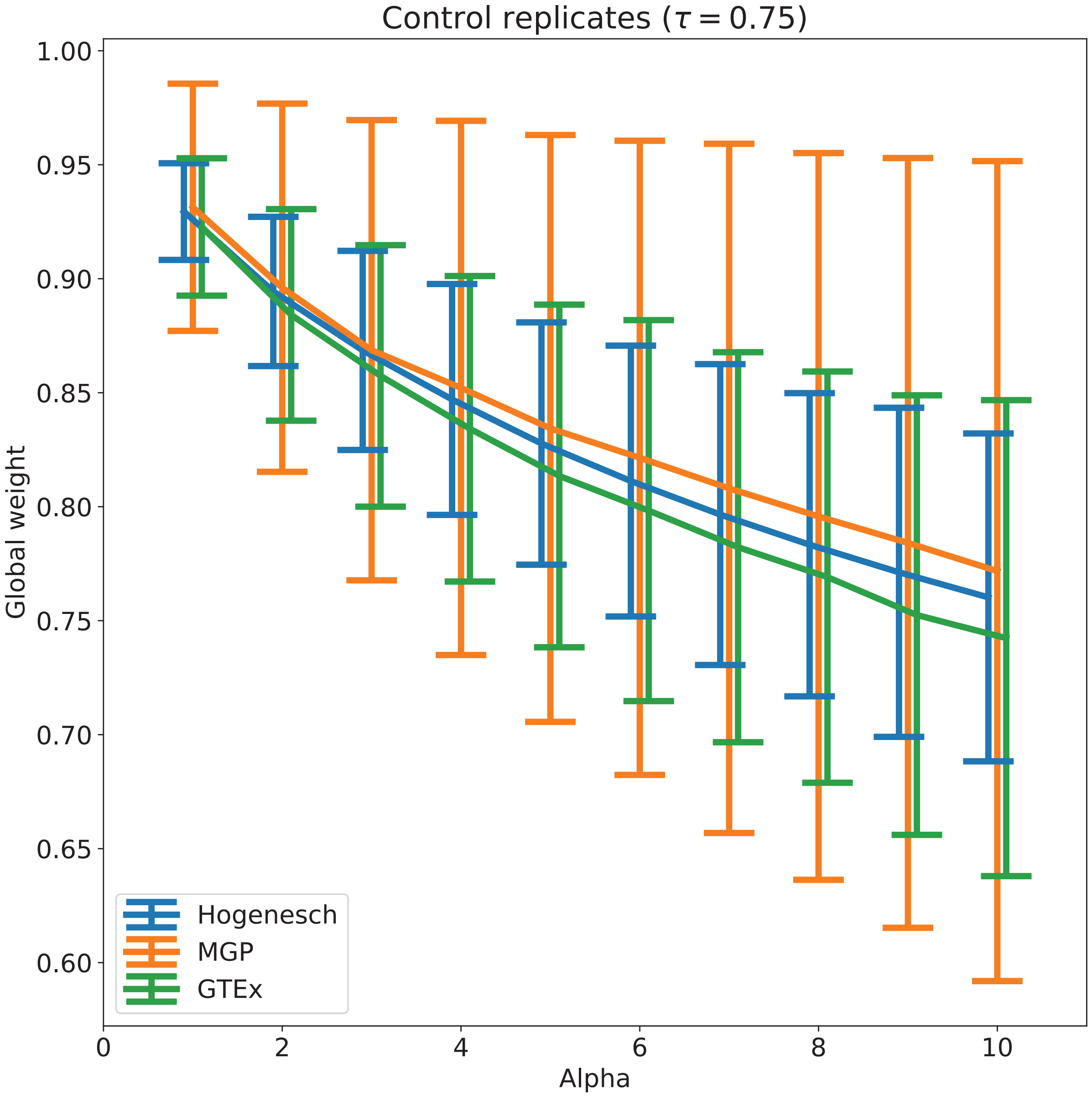

Figure Supplementary 1a: The threshold parameter τ over $\varphi_{t,i} = \max_j d_1(\Psi^t_{i,j}, \hat{\Psi}^T_{i,j})$ as defined in the main text is set to 0.75. Here, the effect of varying the hyper parameter α (Alpha) on the resulting global weight ρ_t is monitored. The effect on ρ_t is tested for three sets of triplicates from the same condition: Hogenesch cerebellum (blue), MGP lung (orange), and GTEx heart (green). Lines represent the mean global weight (ρ) assigned to each dataset; error bars represent ±1 standard deviation. The default MAJIQ is set to $\alpha = 5$. We note that for MAJIQ-lw ρ_t is only used for initially defining a weight average to derive $\overline{P}$ as described in the main text. Also the introduction of an outlier pushes the remaining true replicates to get a value of $\rho_t = 1$ on a wide range of α values (see next figure). Finally, for the mouse biological replicates from MGP we observed significantly higher variability than the one observed between human individuals in GTEx. This variability is also reflected in lower reproducibility ratios for all algorithms for MGP compared to the Hogenesch dataset (see matching figures).

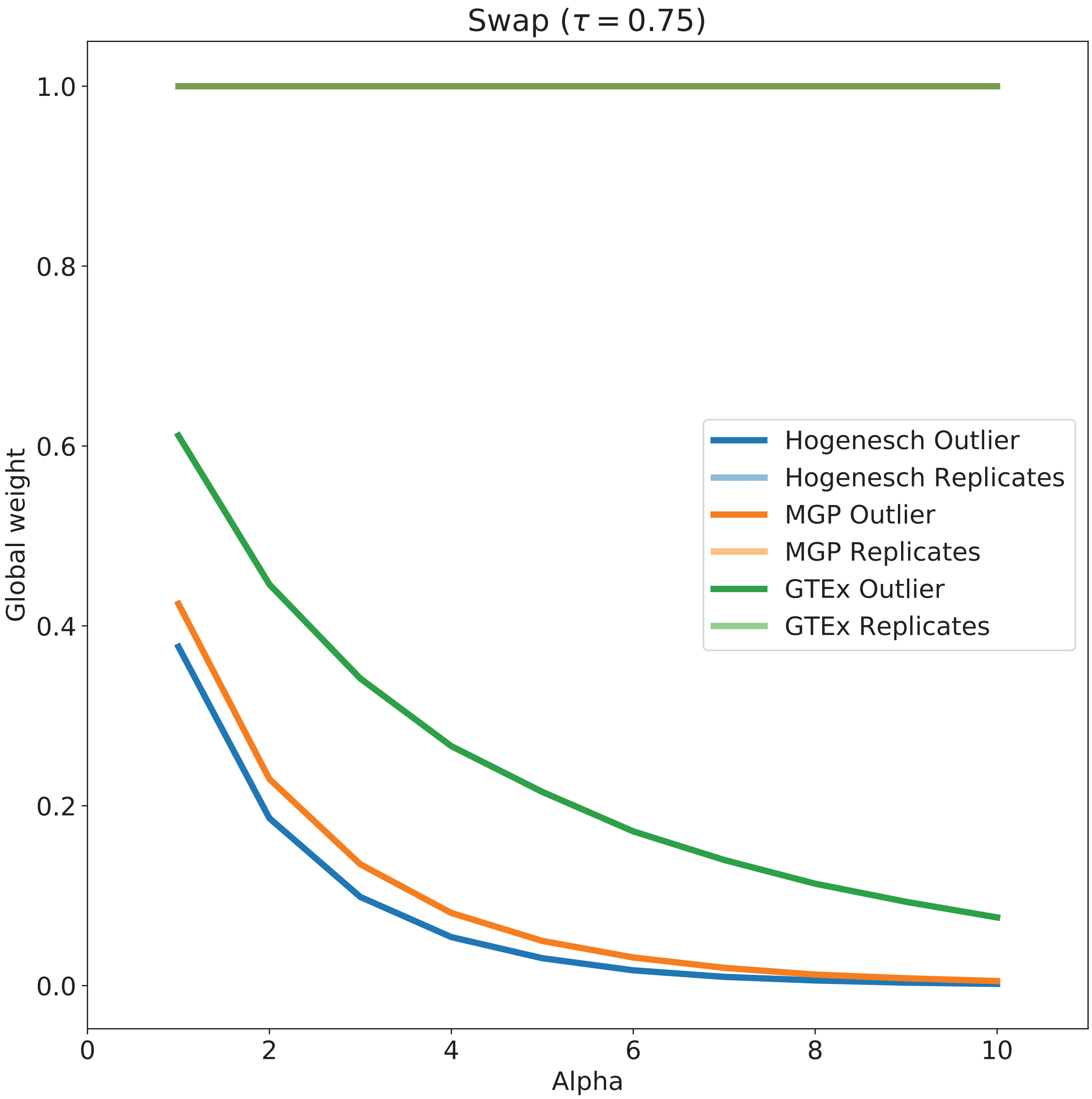

Figure Supplementary 1b: The same methods are applied except now each dataset includes one outlier generated by including in the triplicate a sample from a different tissue in the same dataset. For Hogenesch (blue), the outlier is a muscle sample; for MGP (orange), the outlier is a hippocampus; and for GTEx (green), the outlier is adrenal. Solid lines (bottom) represent the weight on the outlier; The line at the (top) represent the median weight on the remaining two replicates in each group (similar to all three datasets). Note that the introduction of an outlier pushes the remaining true replicates to get a value of $\rho_t = 1$ on a wide range of α values.

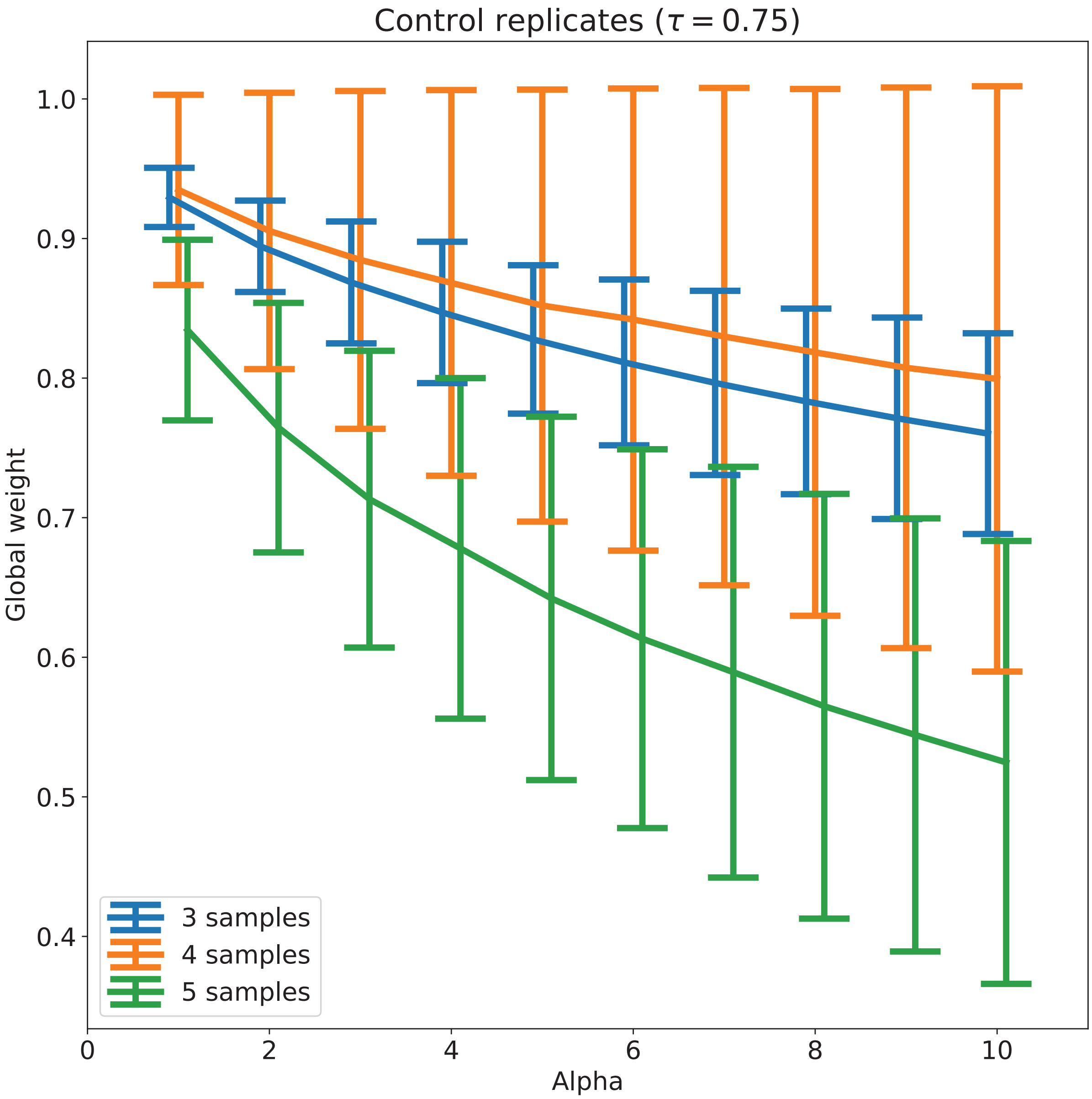

Figure Supplementary 1c: The same methods as in Supplementary Figure 1a are once again applied here, but instead of comparing across datasets, this figure assesses the additional effect of including more replicates (3 (blue), 4 (orange), or 5 (green)) in Hogenesch cerebellum. Again, we note that for MAJIQ-lw ρ_t is only used for initially defining a weight average to derive $\overline{P}$ as described in the main text. Also the introduction of an outlier pushes the remaining true replicates to get a value of $\rho_t = 1$ on a wide range of α values (see next figure)

Supplementary 2

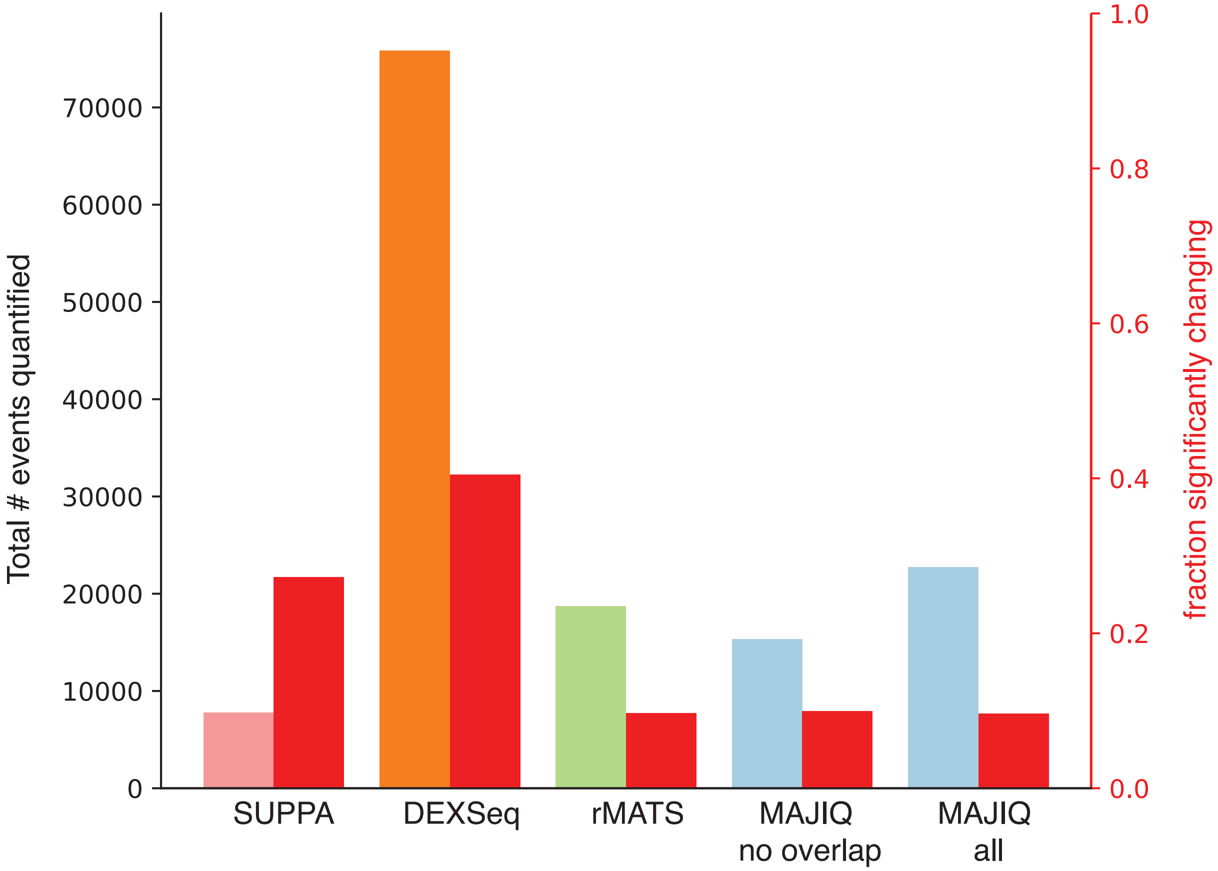

Figure Supplementary 2a: Number of events detected (left y-axis) and fraction reported as differentially spliced (right y-axis, in red) for each method. Data is the same as in the "control" set from Figure Supplementary 2b, which matches main Figure 5a when including intron retention (IR) events (see below). MAJIQ is analyzed when overlapping LSVs are retained (MAJIQ all) and when these are removed (MAJIQ no overlap) (see main text).

.

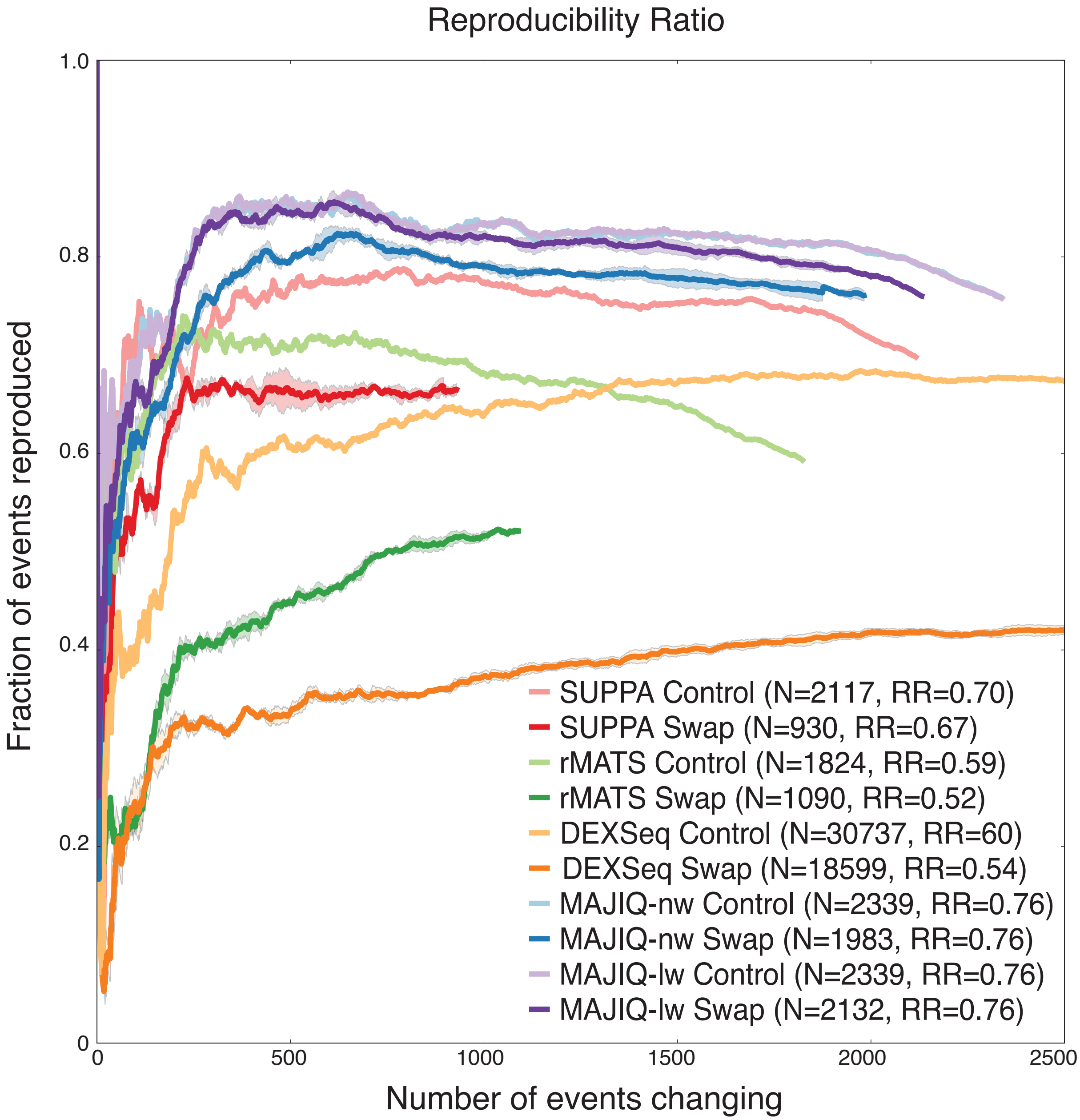

Figure Supplementary 2b: Same as Figure 5a in the main text, but with IR events included.

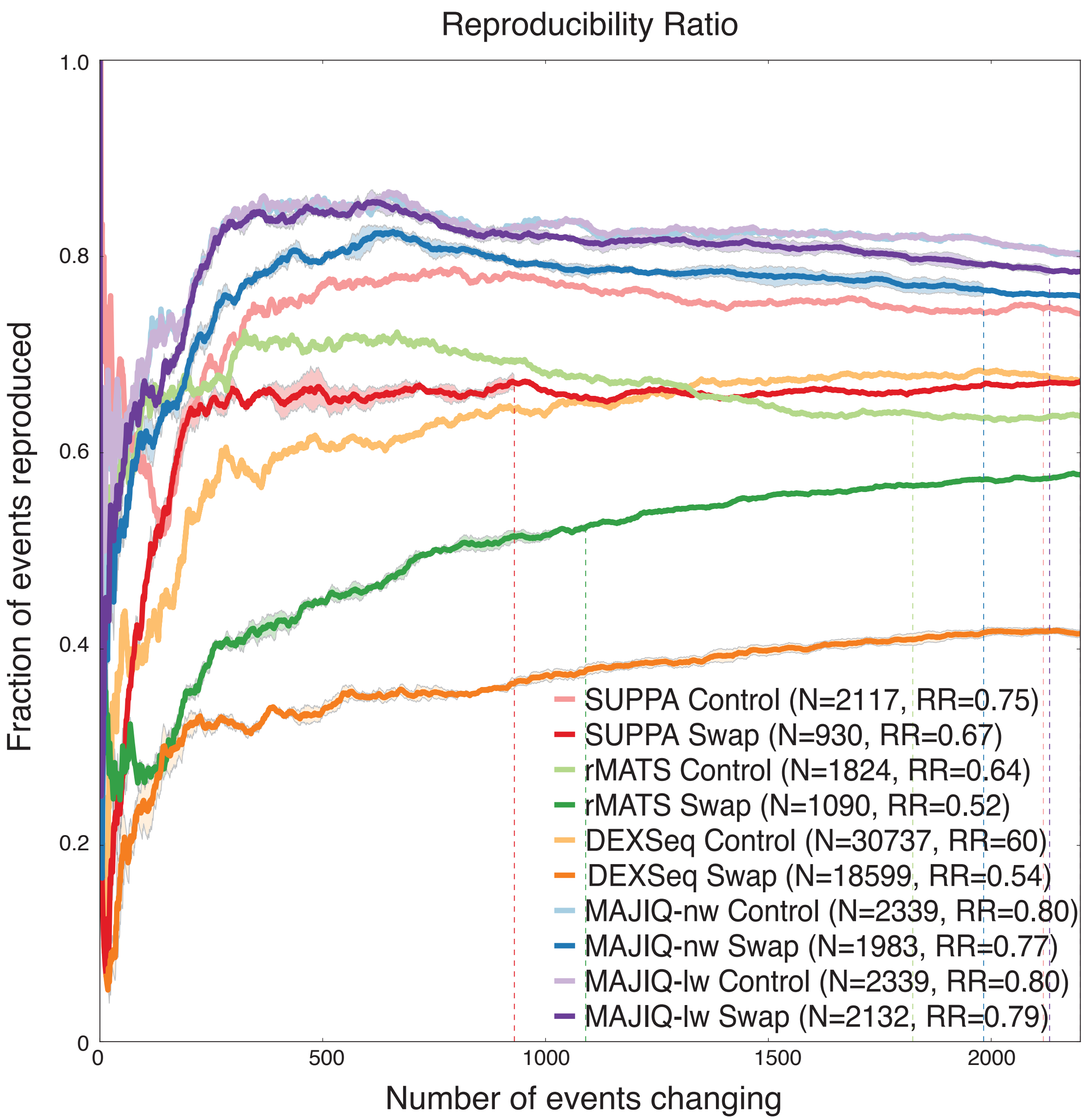

Figure Supplementary 2c: Same as Figure Supplementary 2b, but without using a threshold of p-val < 0.05. Vertical line and N in the legend denote where the p-val $= 0.05$ cutoff occurs. For MAJIQ the events are ordered by $E[|\Delta\Psi|]$ and the N vertical dashed line corresponds to $(P(|\Delta\Psi| > 0.2) > 0.95$.

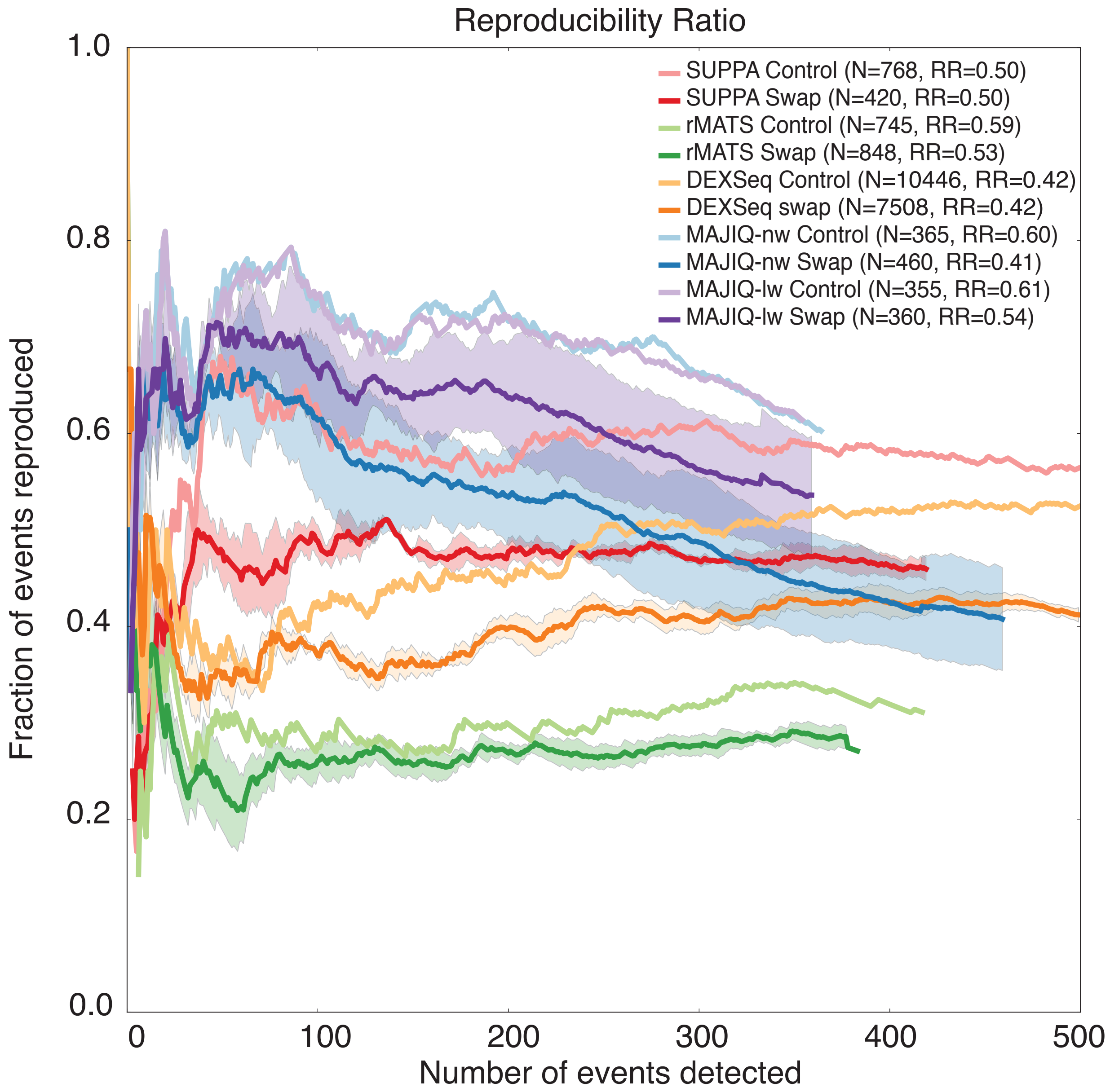

Figure Supplementary 2d: Same as Figure Supplementary 2b, but using biological replicates from the MGP dataset. Control: Three lung are compared against three liver samples, and reproducibility is computed in a second lung-vs-liver execution. Swap: One lung is replaced with a hippocampus in Set 1. The complete RR graph for DEXSeq is given in Figure Supplementary 2f.

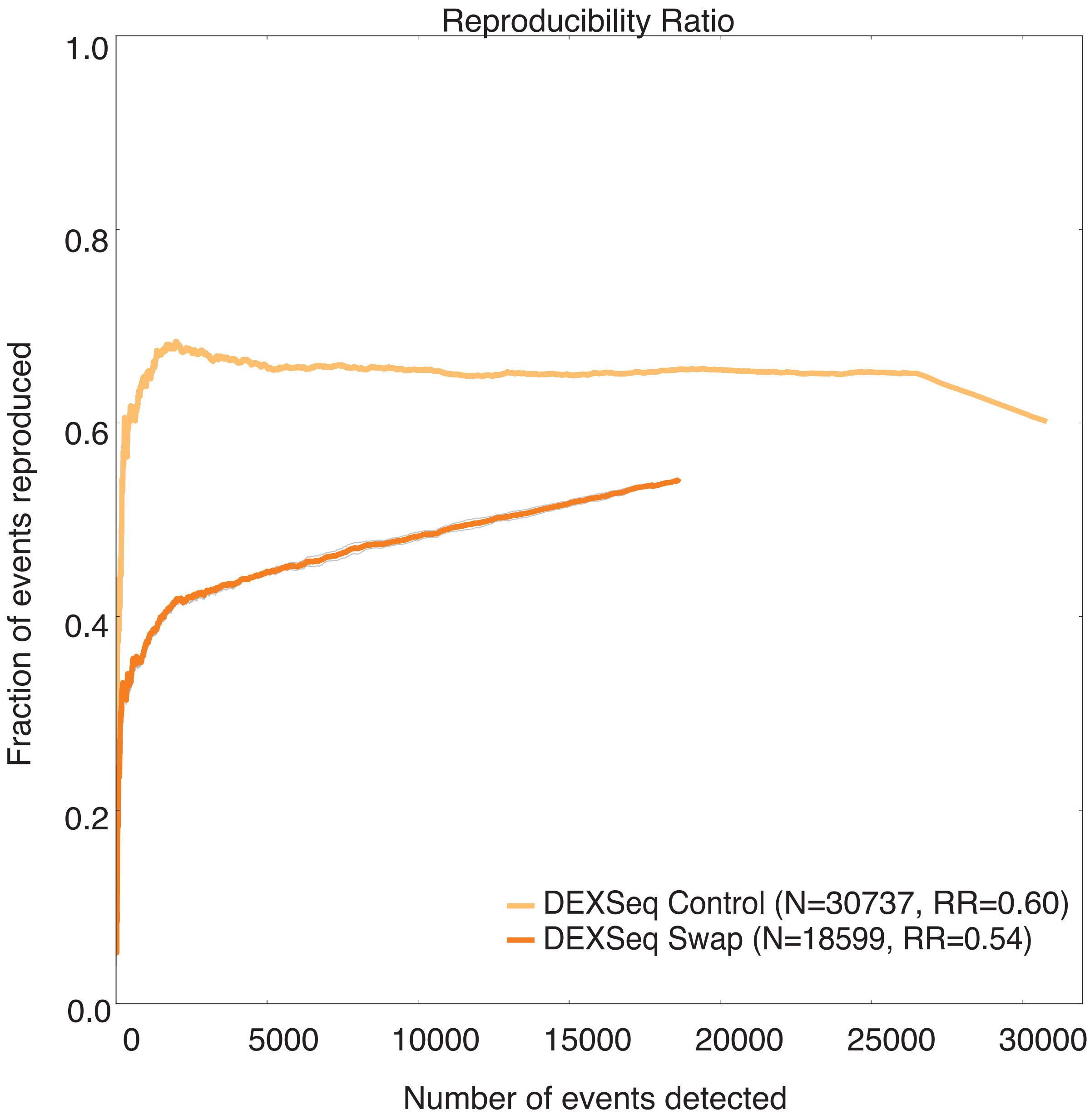

Figure Supplementary 2e: The full reproducibility plot for DEXSeq from Figure Supplementary 2b.

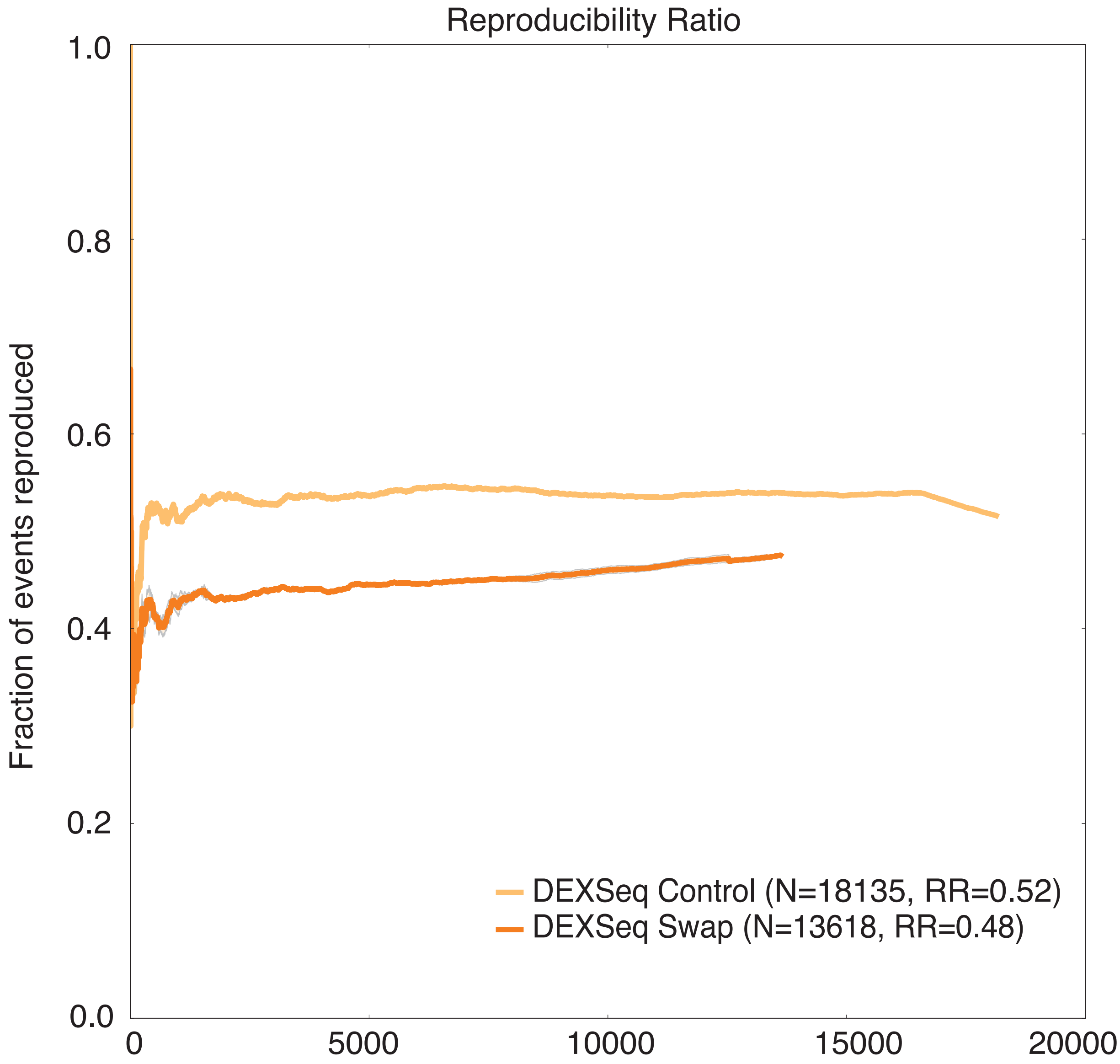

Figure Supplementary 2f: The full reproducibility plot for DEXSeq from Figure Supplementary 2d.

Supplementary 3

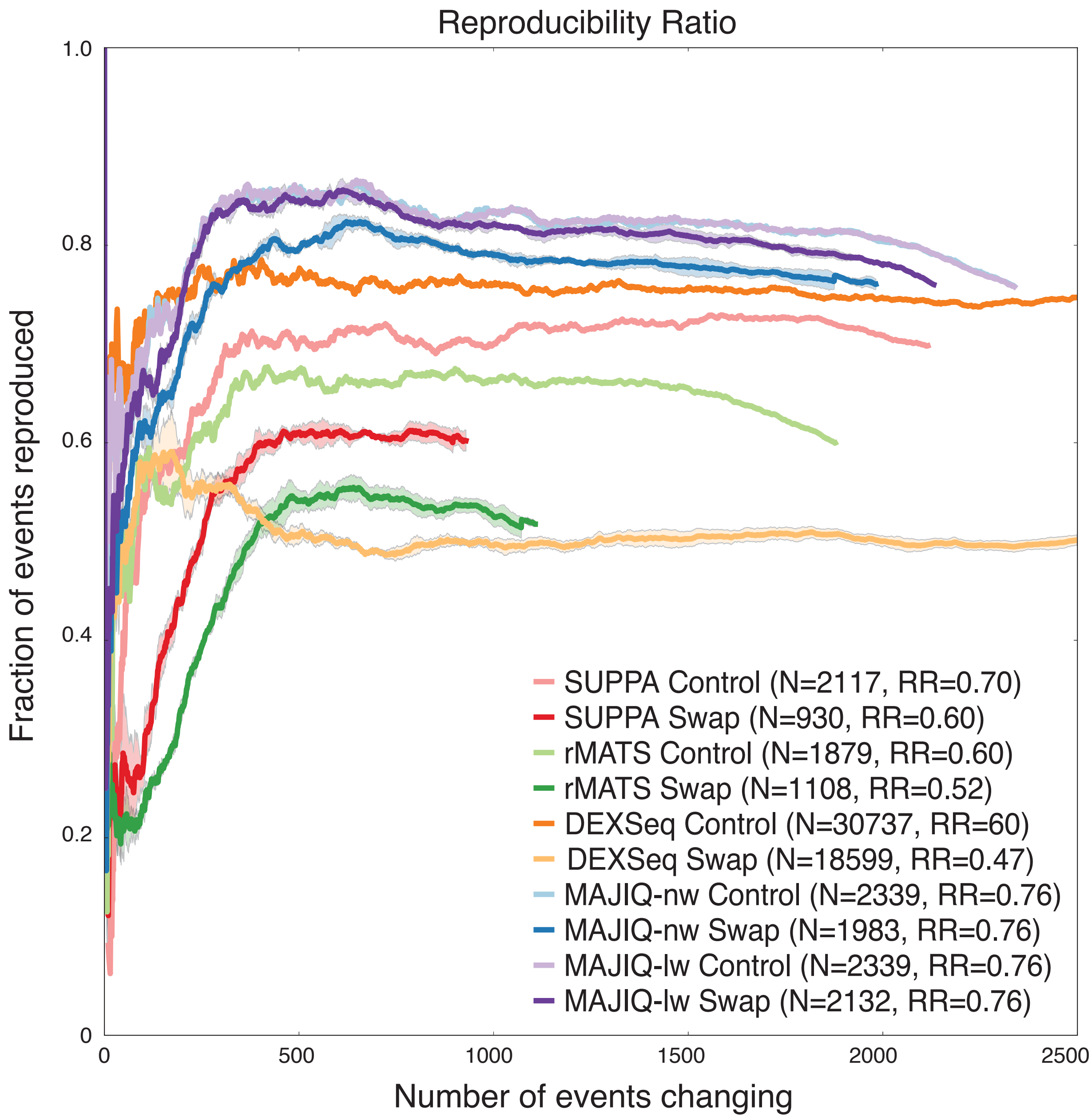

Figure Supplementary 3: Reproducibility ratios plots as in Figure Supplementary 2b but using ranking based on delta PSI instead of p-value for rMATS and SUPPA and using $\log_2$ fold for DEXSeq.

Supplementary 4

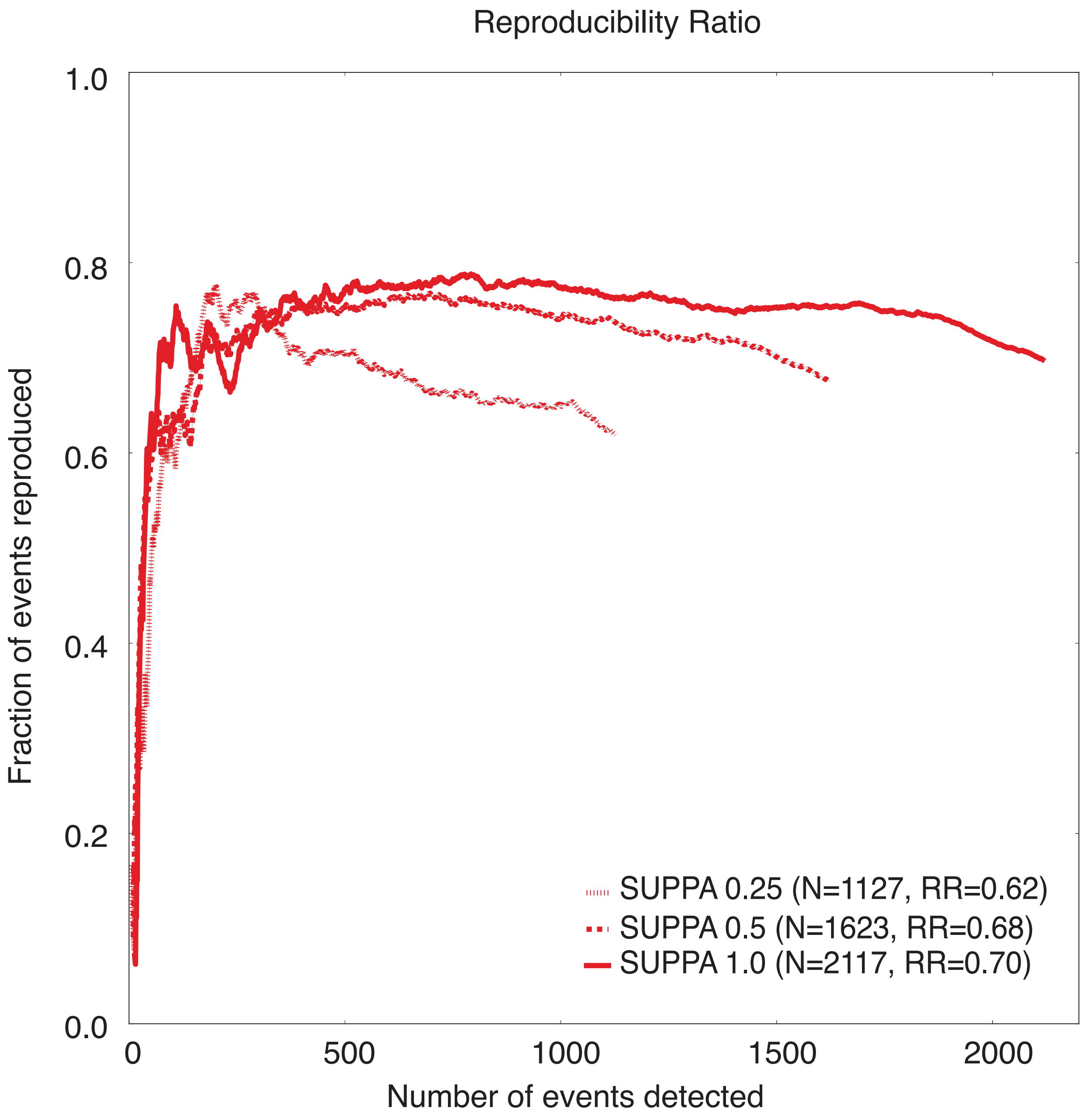

Figure Supplementary 4a: SUPPA reproducibility plots on Hogenesch cerebellum vs liver, with IR, with original files and with files subsampled to 25% and 50% of its original read count.

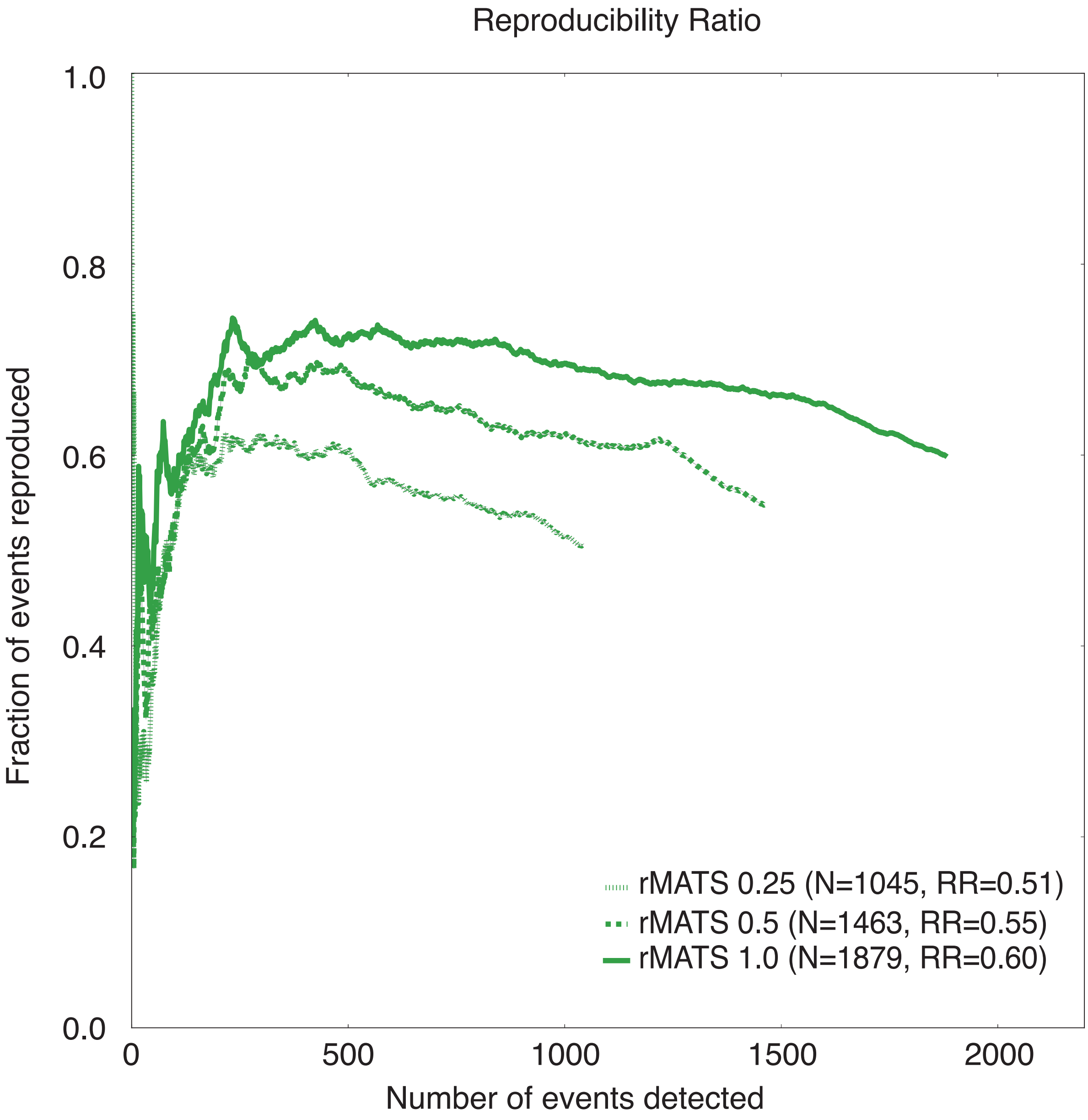

Figure Supplementary 4b: Same as previous figure for rMATS.

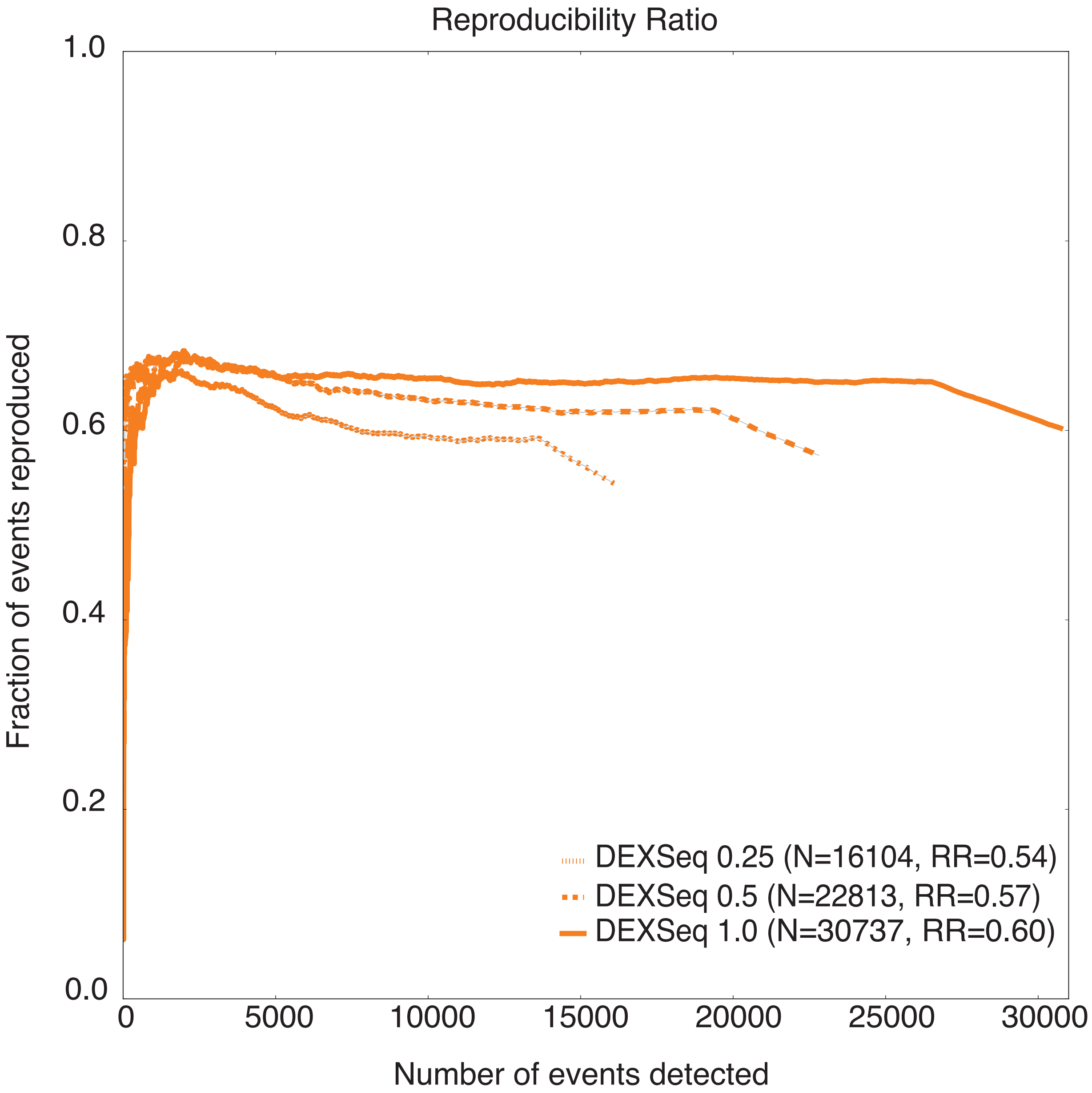

Figure Supplementary 4c: Same as previous figure for DEXSeq.

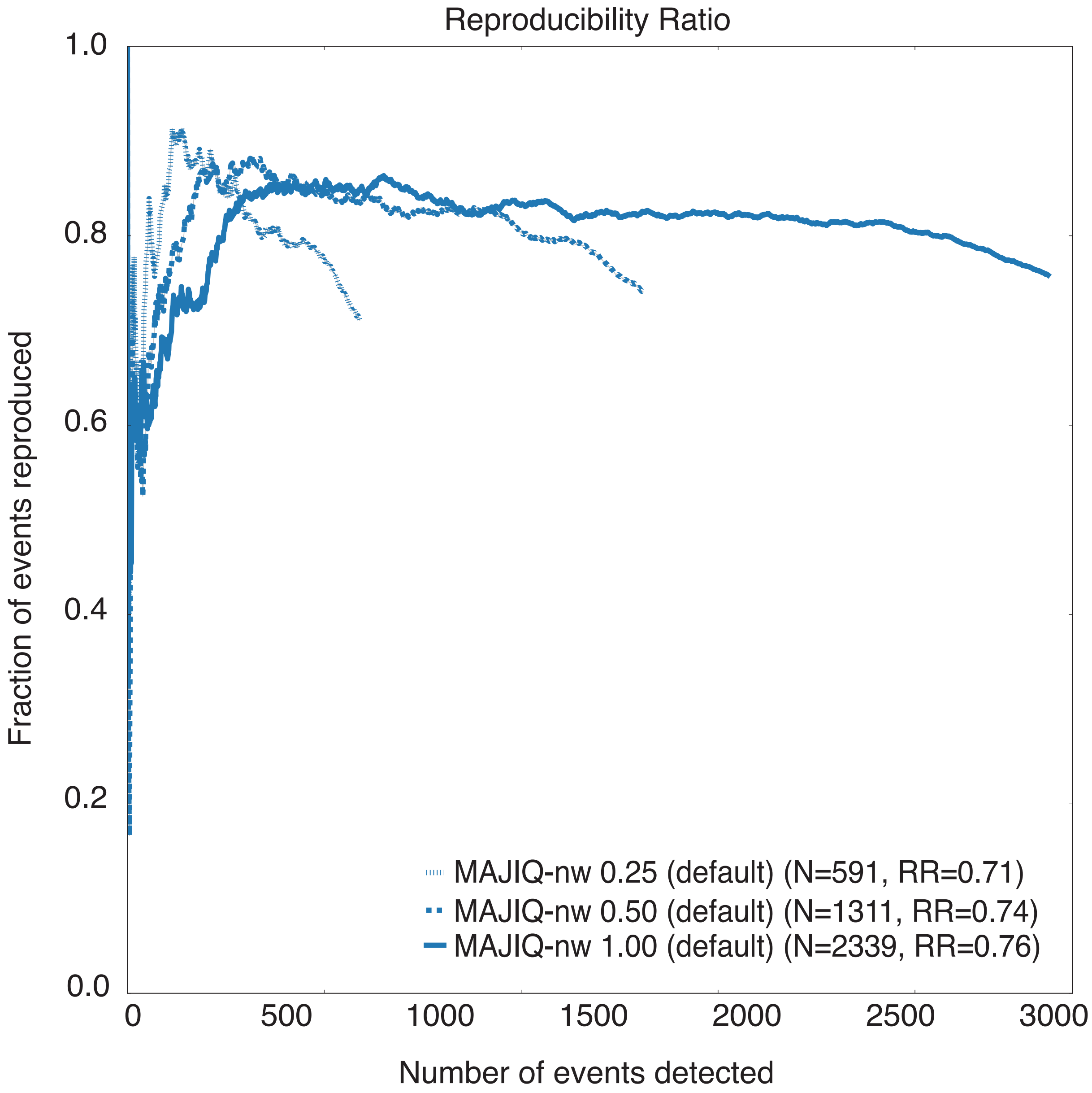

Figure Supplementary 4d: Same as previous figure for MAJIQ without weights (MAJIQ-nw). Default settings were used with respect to the quantifiability filter

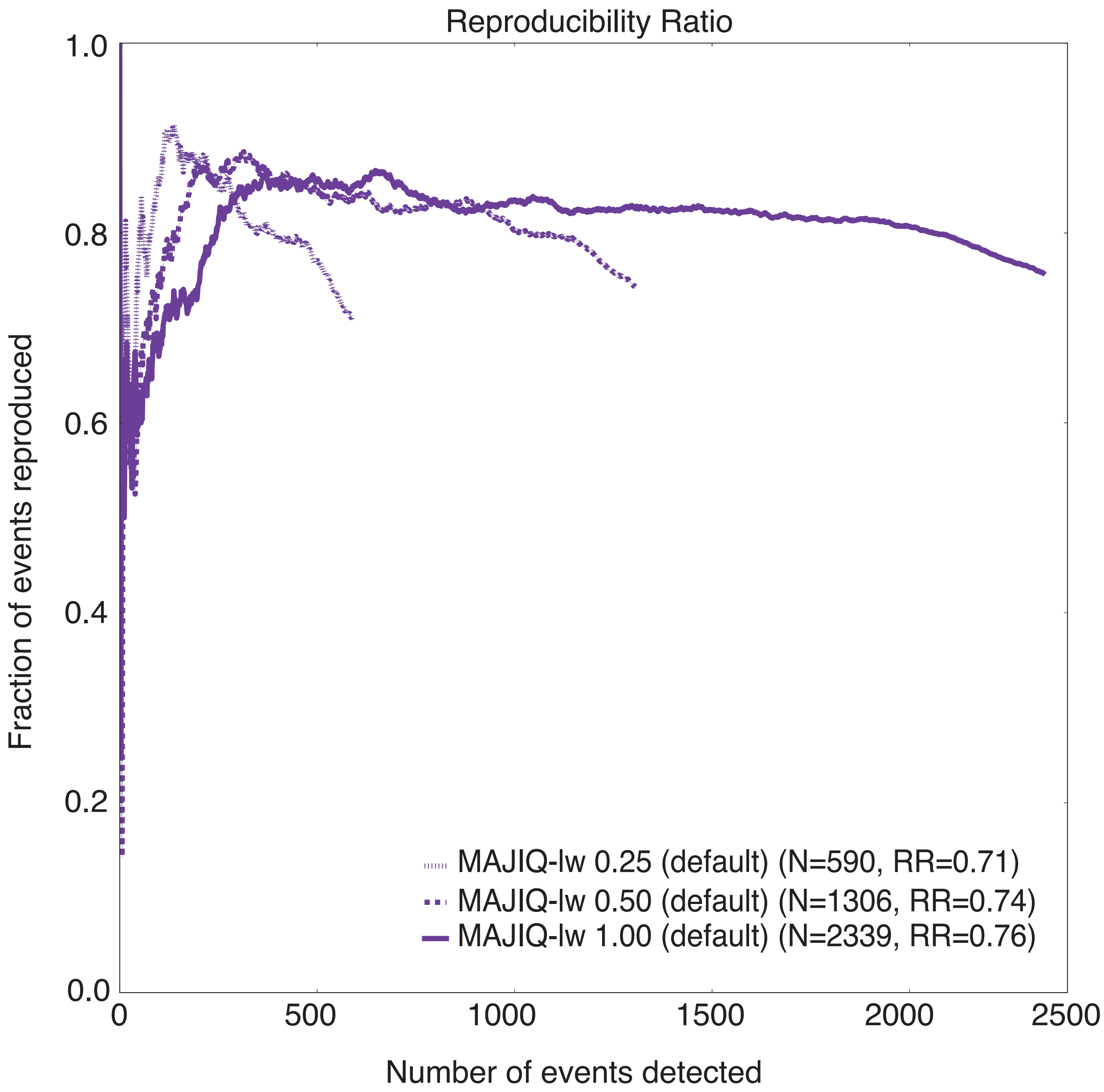

Figure Supplementary 4e: Same as previous figure for MAJIQ with local weights (MAJIQ-lw). Default settings were used with respect to the quantifiability filter

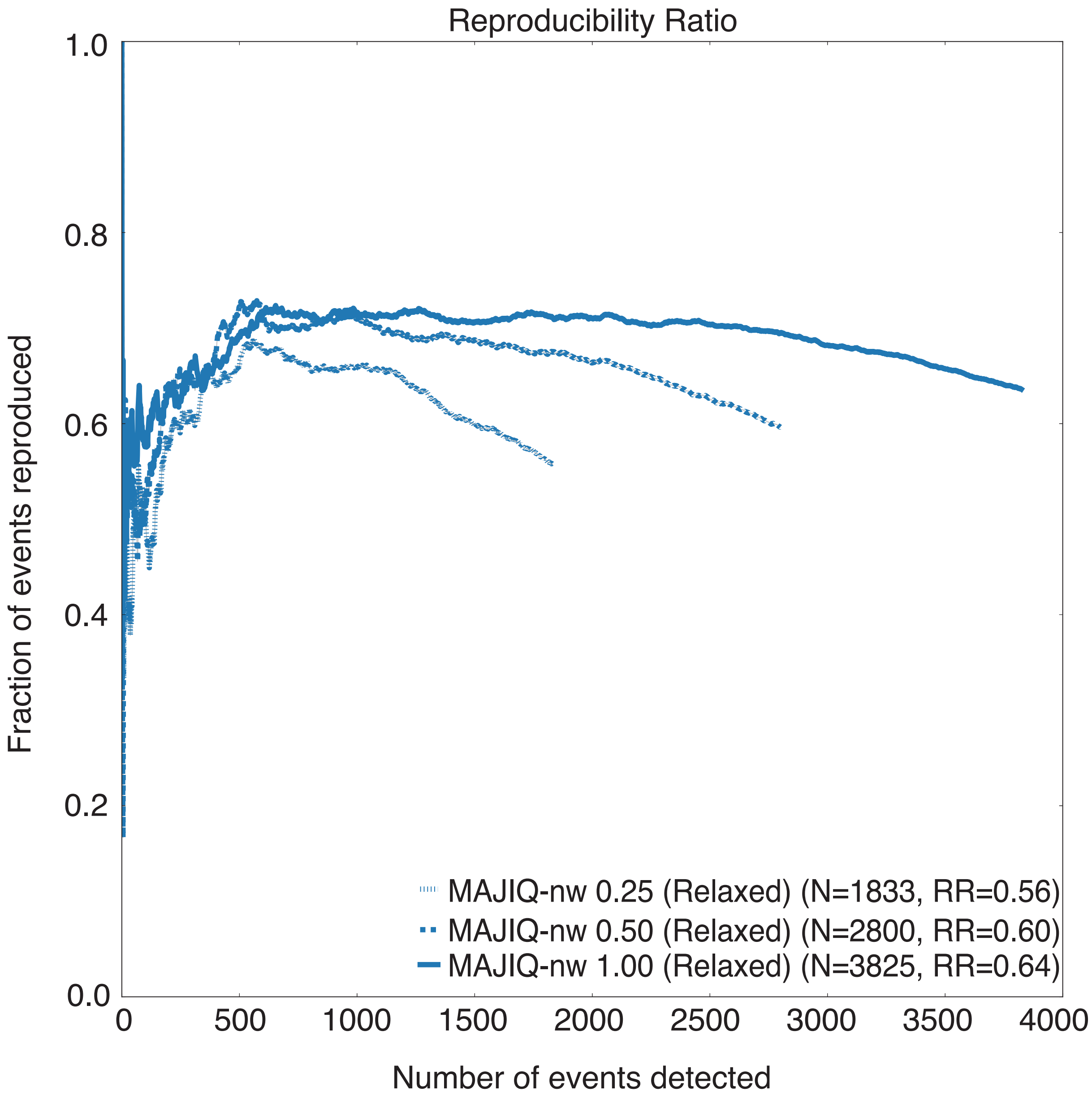

Figure Supplementary 4f: By default, MAJIQ will only attempt to quantify an LSV if it is supported by a minimum of 10 reads with a 3-or-more nucleotide position overhang on each end of the splice junction in at least half the input samples. While this helps to control against false discovery, it comes at the cost of detection power on low-depth experiments. The user can counter this by relaxing the quantifiability filter to allow quantification of LSVs supported by fewer reads, or fewer samples. To demonstrate this, we ran the same MAJIQ execution as in Figure Supplementary 4d, but with the filter relaxed to permit LSVs supported by as few as 2 reads across 2 positions in at least one input sample.

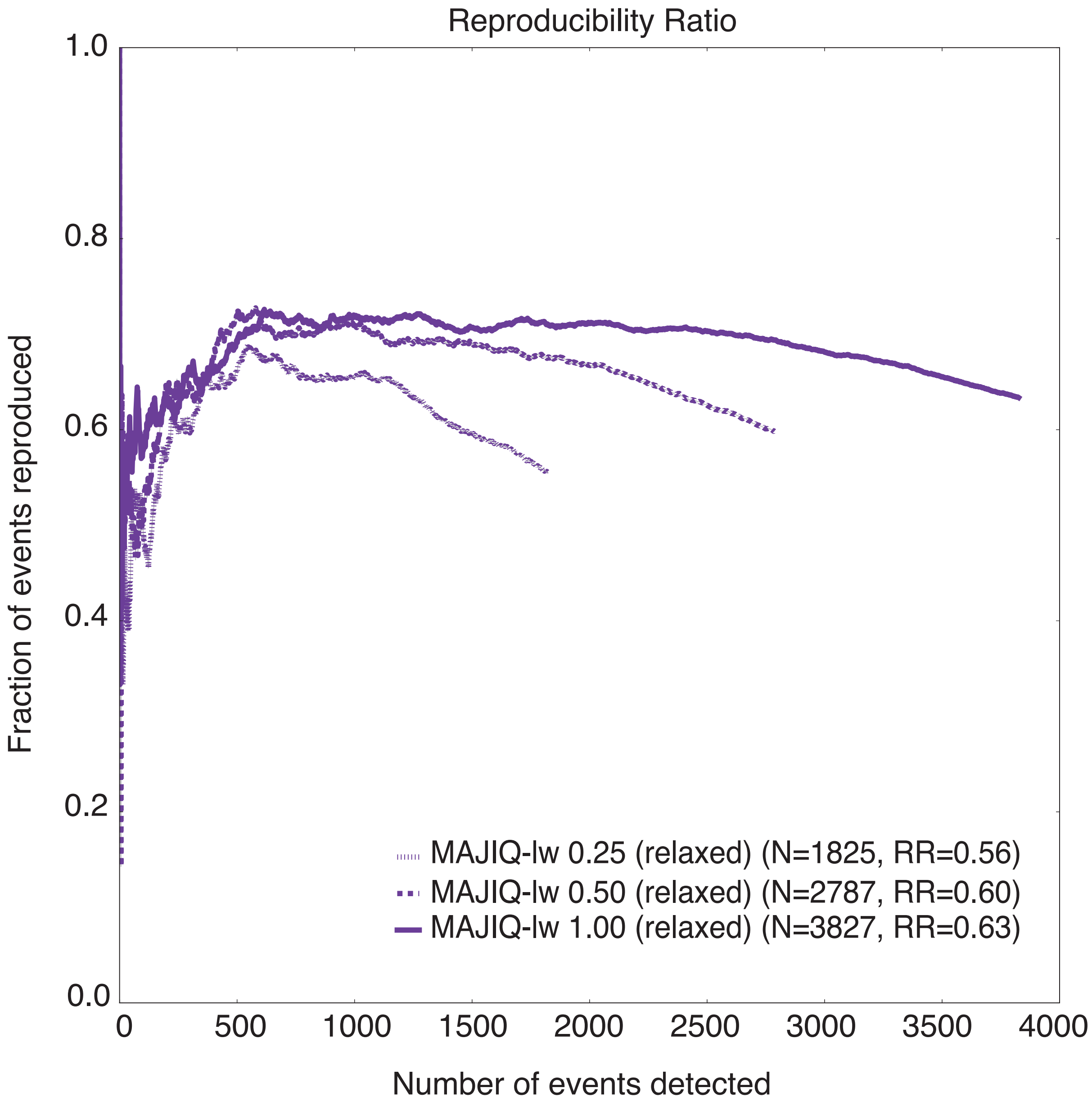

Figure Supplementary 4g: The same relaxed quantifiability filter used to generate the previous figure, is applied to MAJIQ with local weights.

Supplementary 5

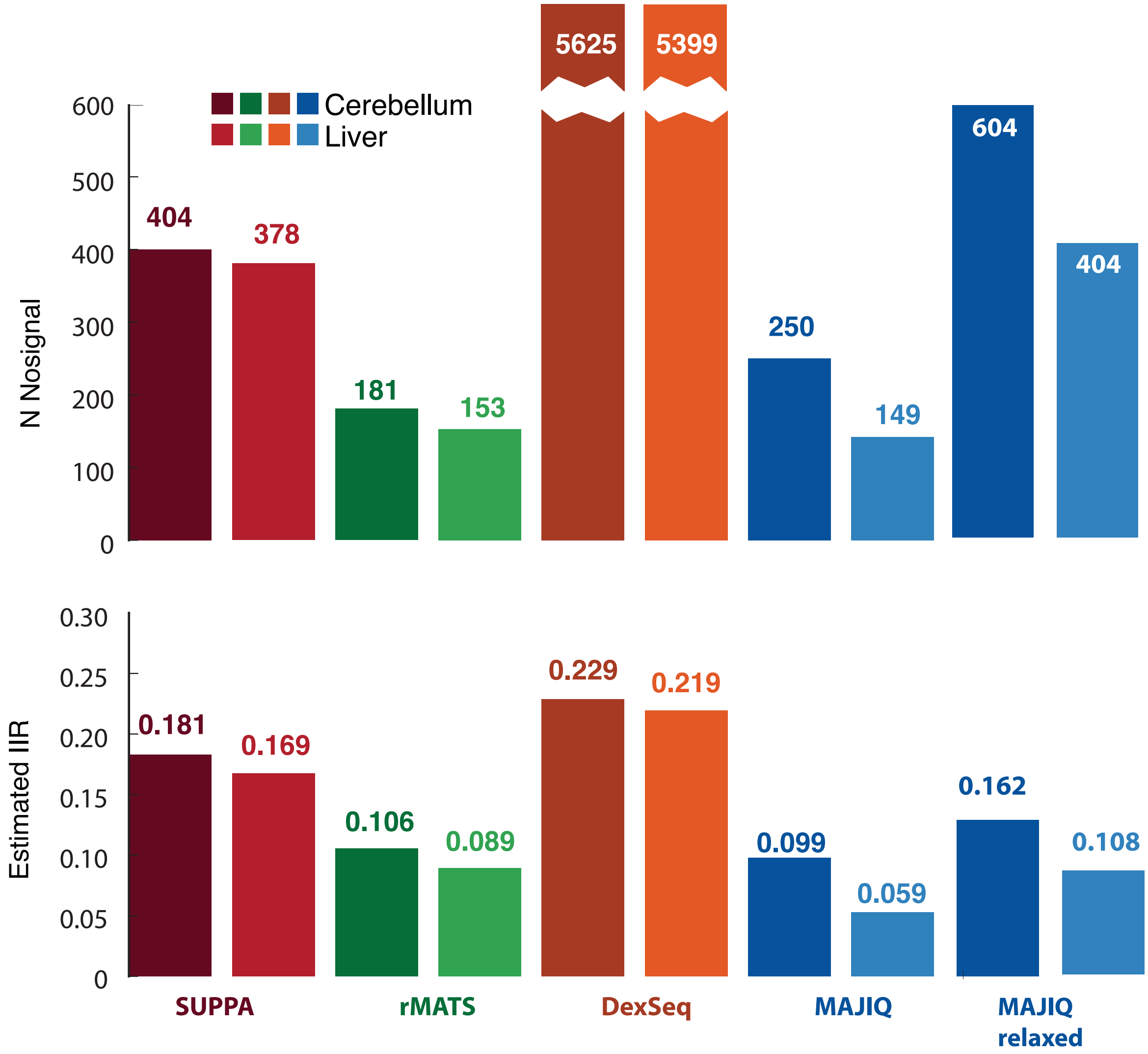

Figure Supplementary 5a: Same as main Figures 5b,c but when IR events are included. We also included the quantifications for MAJIQ with a relaxed quantifiability filter as described in Figure Supplementary 4f.

Supplementary 6

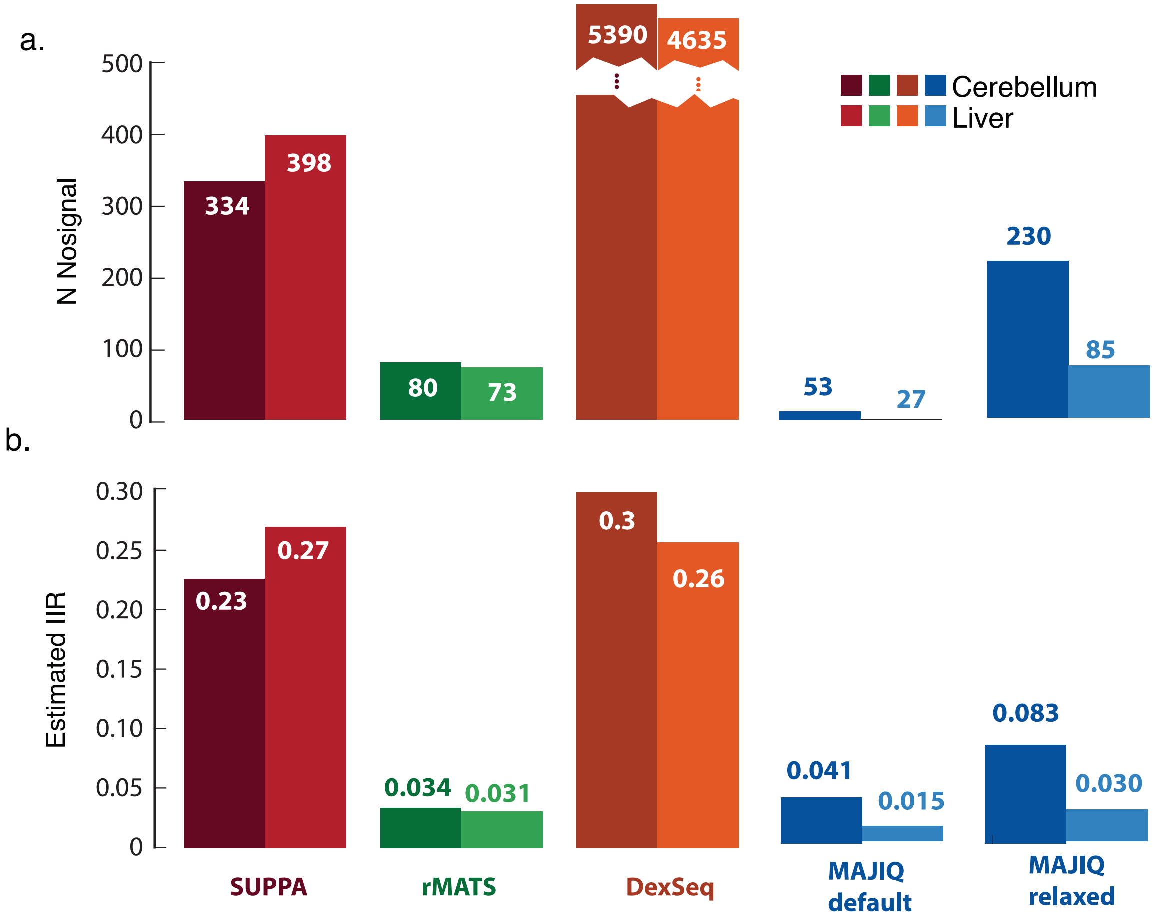

Figure Supplementary 6a: Same as Figure Supplementary 5a but using technical replicates instead of biological replicates. In the main text, these are described as "putative false positives" (PFP). Left to right: SUPPA, rMATS, DEXSeq, MAJIQ with the default quantifiability filter, MAJIQ with the relaxed quantifiability filter.

Figure Supplementary 6b: We estimated the inter-to-intra ratio (IIR) for each method by dividing the N reported in Figure Supplementary 6a with the N calculated from a between-tissues quantification (Cer_CT28_1, Cer_CT34_1, Cer_CT40_1 vs Liv_CT28_1, Liv_CT34_1, Liv_CT40_1).

Supplementary 7

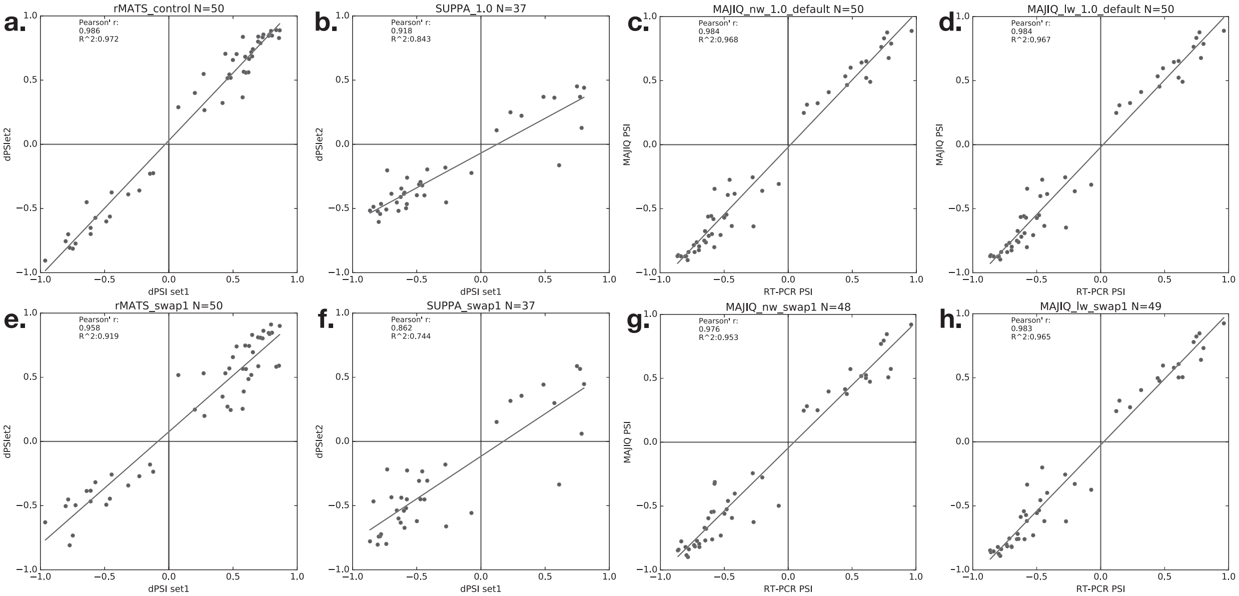

Figure Supplementary 7: RT-PCR scatterplots supporting Figure 5d. Top to bottom: rMATS (a-d), SUPPA (e-h), MAJIQ-nw (i-l), MAJIQ-lw (m-p). Left to right: Control (a,e,i,m) and three swaps (b-d,f-h,j-l,n-p).

Supplementary 8

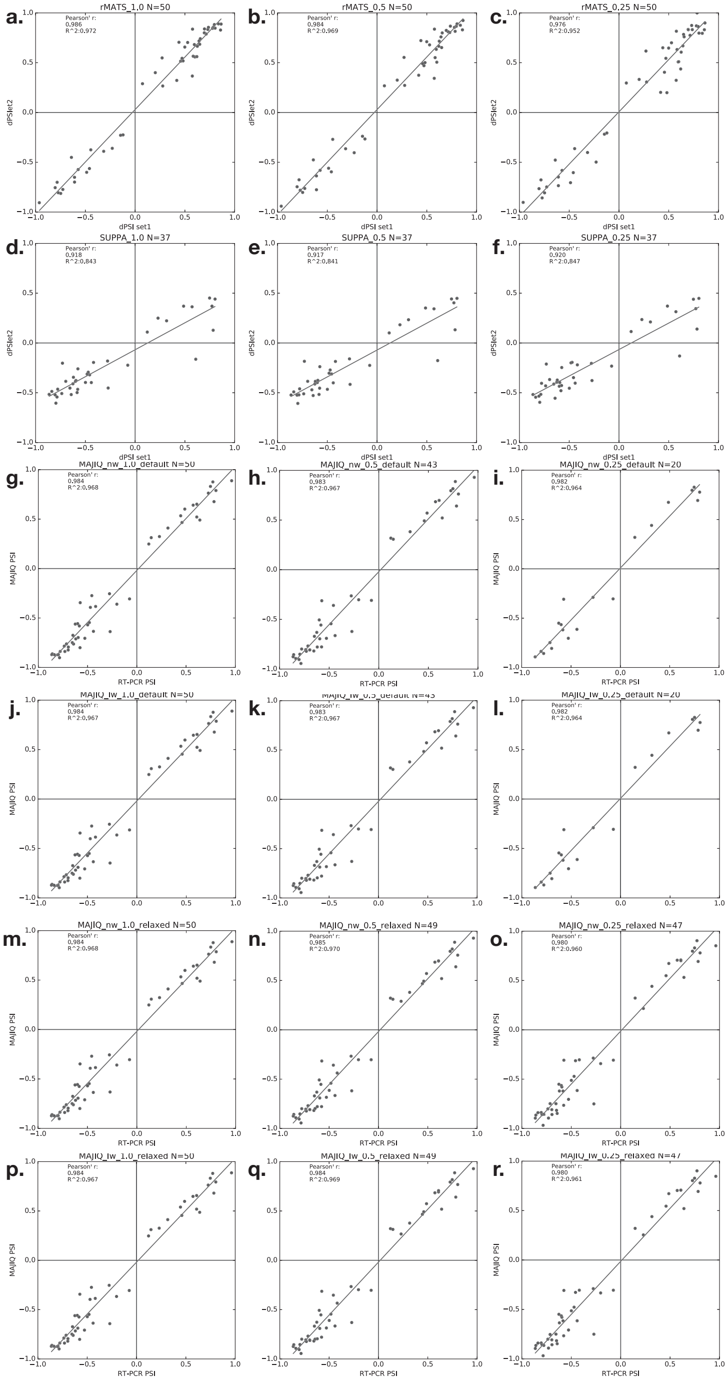

Figure Supplementary 8a: RT-PCR scatterplots supporting Figure 5f. Top to bottom: rMATS (a-c), SUPPA (d-f), MAJIQ-nw default filter (g-i), MAJIQ-lw default filter (j-l), MAJIQ-nw relaxed filter (m-o), MAJIQ-lw relaxed filter (p-r). Left to right: 100% (a,d,g,j,m,p), 50% (b,e,h,k,n,q), 25% (c,f,i,l,o,r) of reads from the original experiments.

References

[1] Gregory R. Grant, Michael H. Farkas, Angel D. Pizarro, Nicholas F. Lahens, Jonathan Schug, Brian P. Brunk, Christian J. Stoeckert, John B. Hogenesch, and Eric A. Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

[2] Nicholas F. Lahens, Ibrahim H. Kavakli, Ray Zhang, Katharina Hayer, Michael B. Black, Hannah Dueck, Angel Pizarro, Junhyong Kim, Rafael Irizarry, Russell S. Thomas, Gregory R. Grant, and John B. Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biology*, 15(6):R86, June 2014.