Cell Systems, Volume *9*

Supplemental Information

Detecting, Categorizing, and Correcting Coverage Anomalies of RNA-Seq Quantification

Cong Ma and Carl Kingsford

FDR	Salmon	SAD-adjusted	percentage reduced
0.01	6088	5854	3.84%
0.05	10132	9907	2.46%
0.1	13555	13316	2.29%

Table S1: **The number of DE transcripts detected at a given FDR threshold.** Related to Figure 1. Among the 30 samples, there should not be any DE transcripts. With SAD-adjusted expression quantification, the number of false positively detected DE transcripts is reduced.

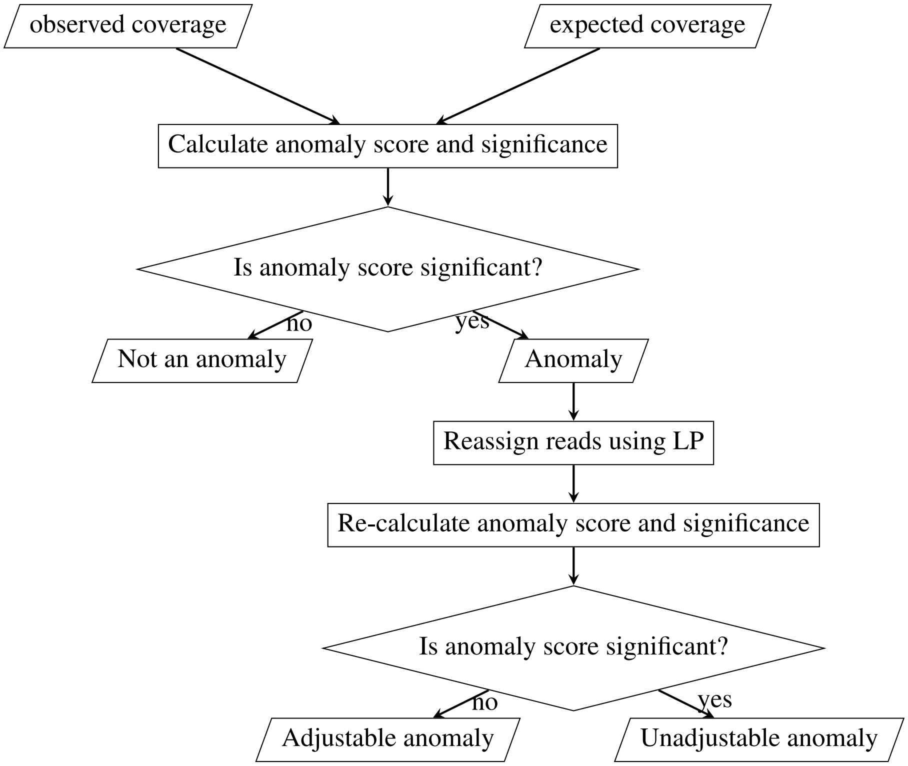
observed coverage
expected coverage
Calculate anomaly score and significance
Is anomaly score significant?
no
yes
Not an anomaly
Anomaly
Reassign reads using LP
Re-calculate anomaly score and significance
Is anomaly score significant?
no
yes
Adjustable anomaly
Unadjustable anomaly

Figure S1: **Diagram of SAD.** Related to Figure 1 and Figure 2. SAD detects anomalies by calculating an anomaly score and the significance of its value. To further distinguish the potential cause of the anomalies, it reassigns the reads across isoforms and checks whether the anomaly score becomes insignificant after reassignment. The anomalies whose anomaly scores become insignificant are categorized as adjustable anomalies and considered to be caused by quantification algorithm mistake. The anomalies whose anomaly scores remain significant are categorized as unadjustable anomalies and considered to be caused by external reasons. When the expected coverages are accurate, the external reason is likely the incompleteness of the reference transcriptome.

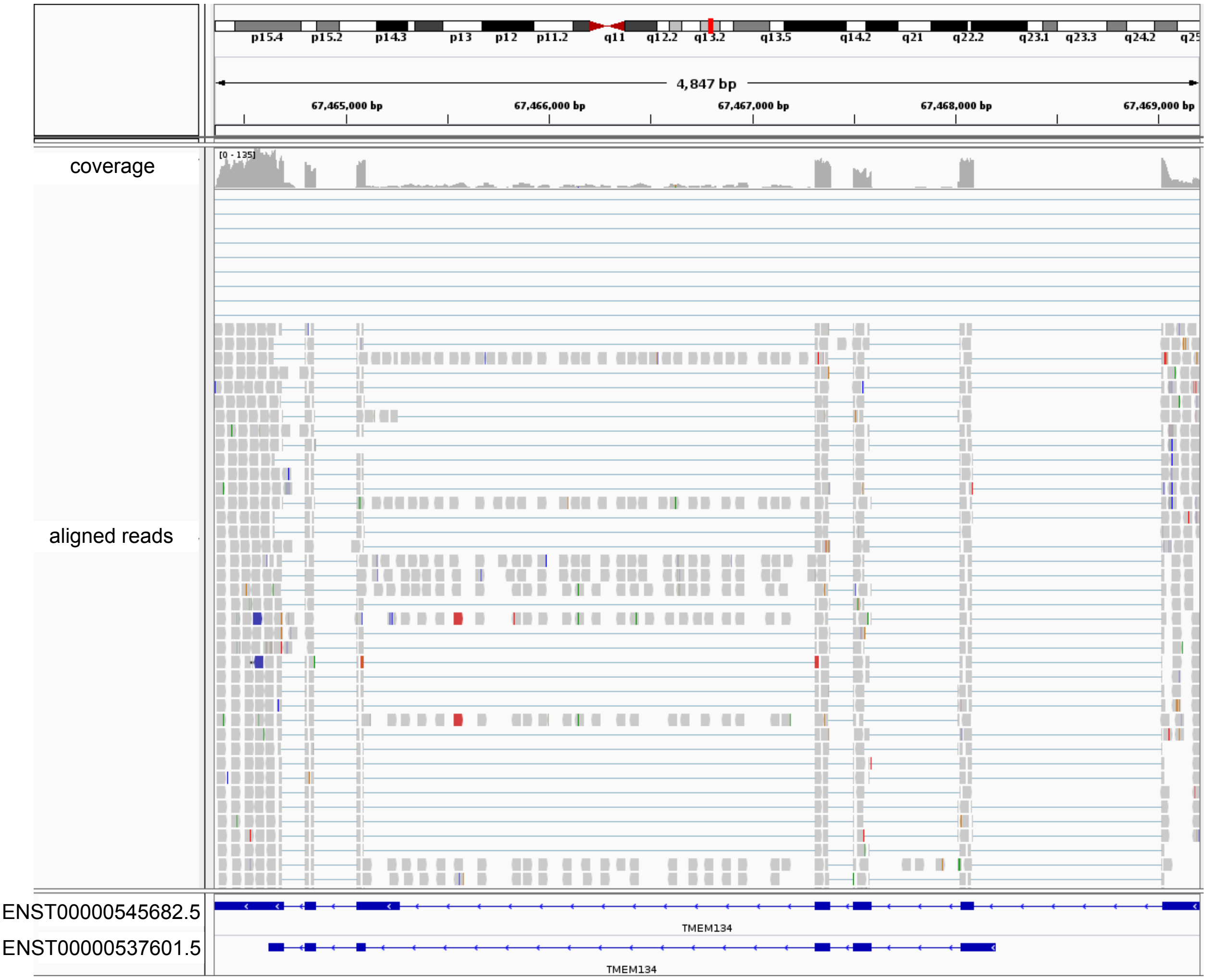

Figure S2: **IGV visualization of alignments on *TMEM134* of the kidney sample.** Related to Figure 1. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000545682.5; the intron-exon structure of transcript ENST00000537601.5. SAD identifies the region after the first splicing junction of transcript ENST00000545682.5 as an under-expressed region. The anomaly is adjustable by re-shuffling reads with transcript ENST00000537601.5. Before SAD adjustment, expression of ENST00000545682.5 is 1.4 times that of ENST00000537601.5. After SAD adjustment expression of ENST00000545682.5 is 9 times that of ENST00000537601.5. The first splicing junction (right-most junction) of ENST00000545682.5 is highly expressed, which is more consistent with a larger abundance ratio.

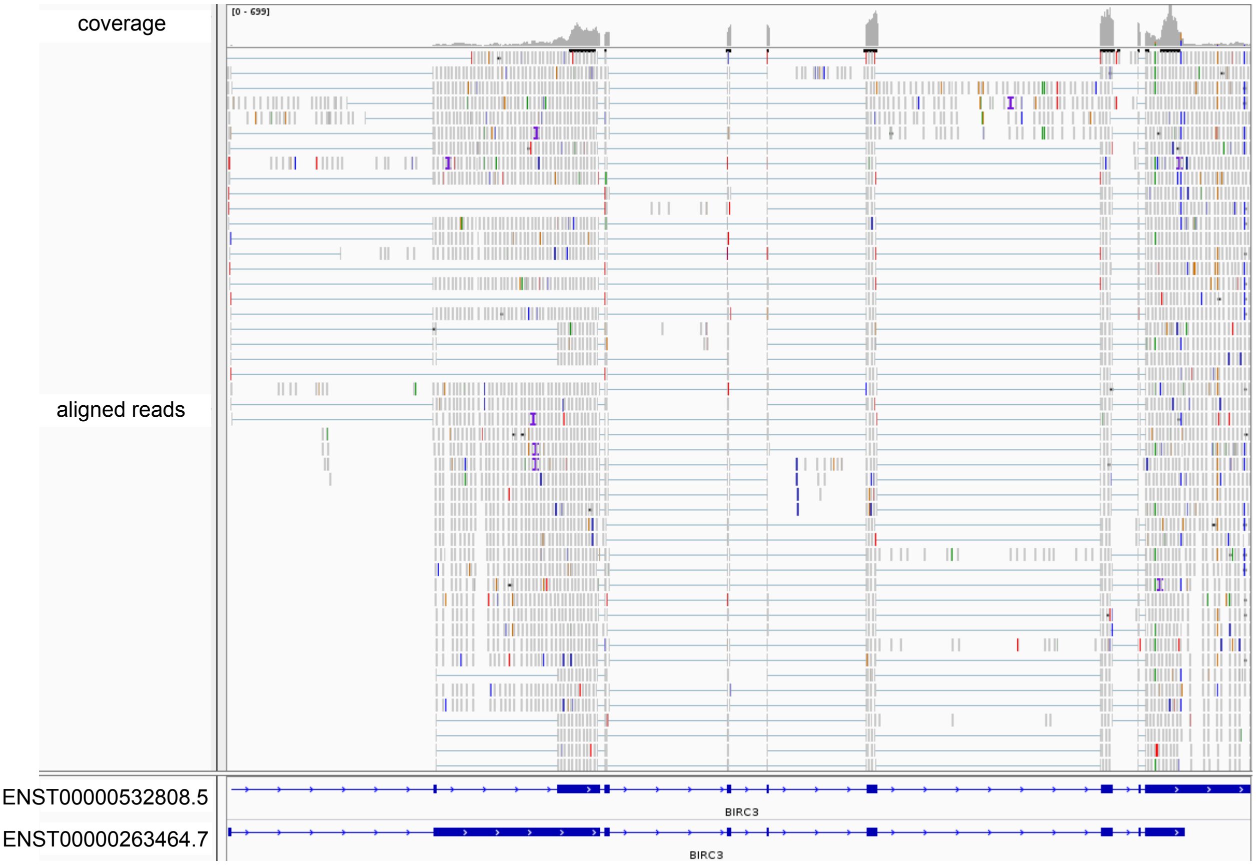

Figure S3: **IGV visualization of alignments on *BIRC3* of a GEUVADIS sample.** Related to Figure 1. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000532808.5; the intron-exon structure of transcript ENST00000263464.7. SAD identifies the 3' region of ENST00000532808.5 as an under-expressed region. The anomaly is adjustable by re-shuffling reads with transcript ENST00000263464.7. Before SAD adjustment, the two isoforms has similar expression. After SAD adjustment expression of ENST00000263464.7 is 3 times that of ENST00000532808.5. According to the expected coverages, the 3' sides of both transcripts are expected to have higher coverage than 5' sides. That the 3' end of ENST00000532808.5 has low coverage suggests that ENST00000532808.5 may be of low abundance.

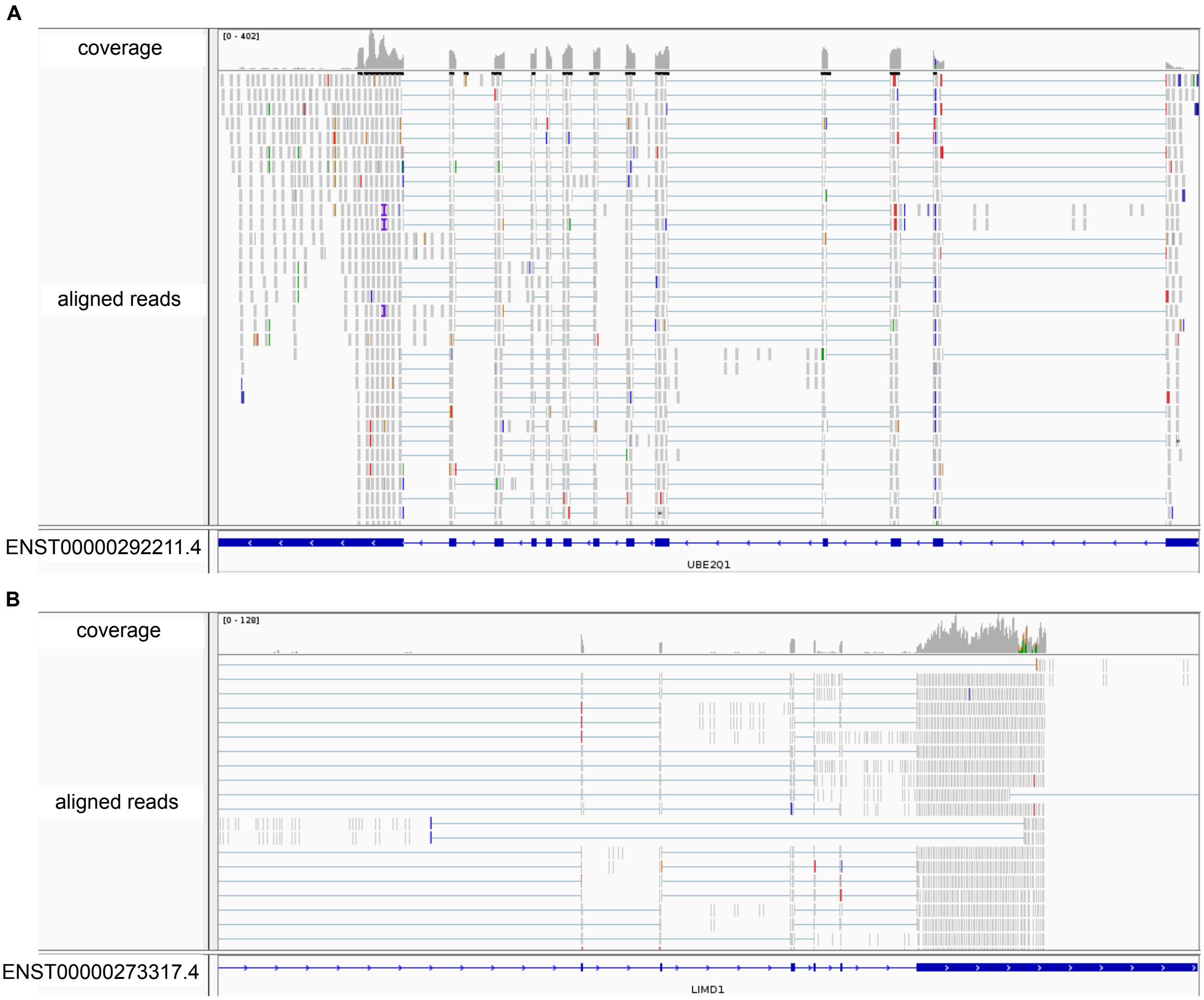

Figure S4: **IGV visualization for the unadjustable anomaly examples.** Related Figure 2. (A). Gene *UBE2Q1* of the heart sample. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000292211.4. There is an under-expressed region of the transcript at 3' end (left-most region). The anomaly is unadjustable. (B) Gene *LIMD1* of the heart sample. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000273317.4. There is an under-expressed region of the transcript at 3' end (right-most region). The anomaly is unadjustable.

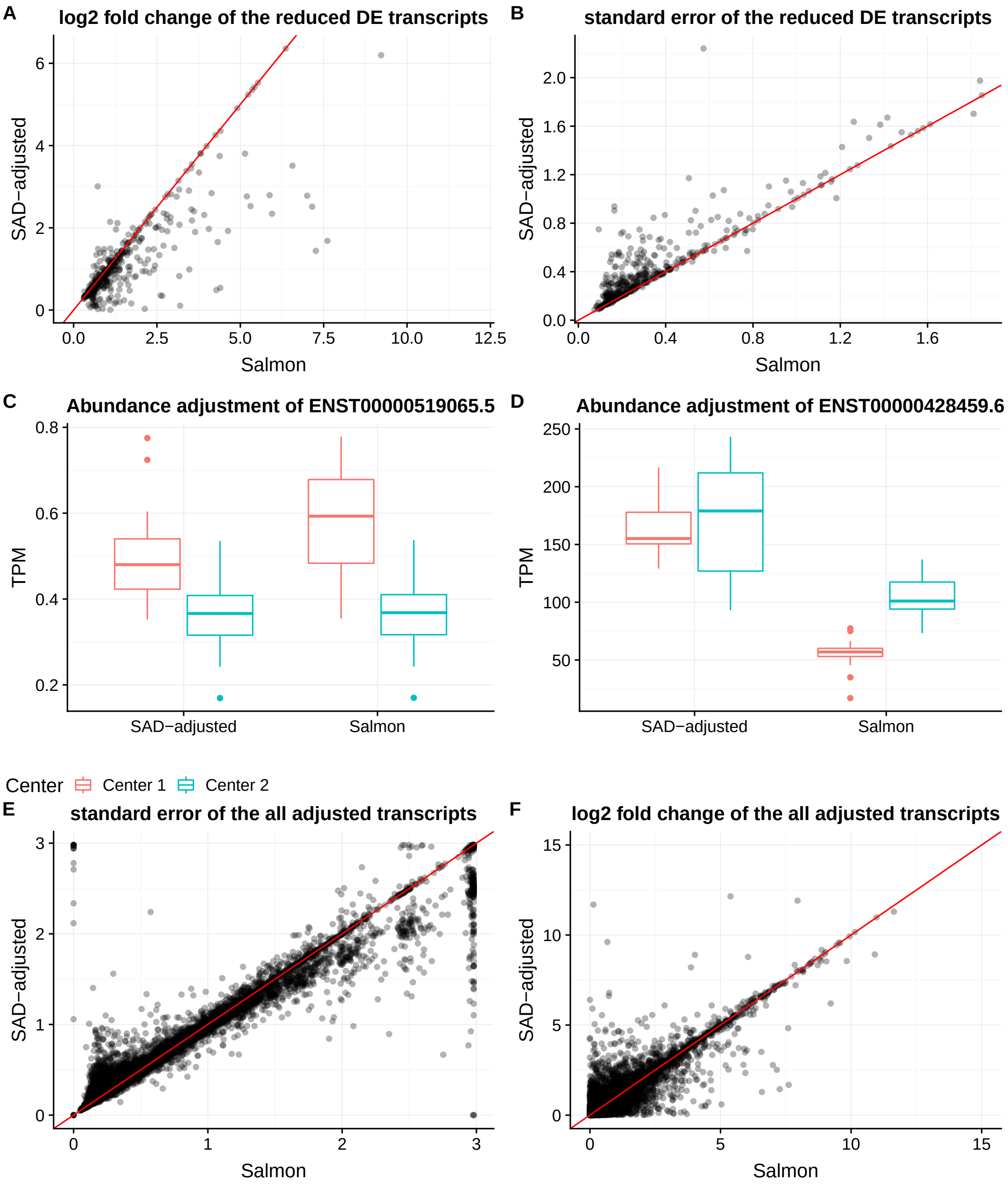

Figure S5: **Changes in statistics of DE detection by using SAD-adjusted quantification for adjustable anomalies.** Related to Figure 1. (A) Absolute $\log_2$-fold change between the two sequencing centers for the transcripts labeled as DE under Salmon but not under SAD-adjusted quantification. The $\log_2$-fold change is often reduced by SAD-adjustment for these transcripts. (B) Standard error of the $\log_2$-fold change between sequencing centers for the transcripts that are labeled as DE under Salmon but not under SAD-adjusted quantification. SAD adjustment may increase the variance of expression. These two panels show the two reasons why potential false positive DE calls are reduced by SAD: increased variance and decreased fold change. (C,D) Two examples of transcript that is detected as DE by Salmon but not detected by SAD-adjusted quantification. Each box indicates the range of estimated expression across RNA-seq samples corresponding to each sequencing center. (E) Standard error of the $\log_2$-fold change for all transcripts under Salmon and SAD-adjusted quantification. (F) Absolute $\log_2$-fold change between the two sequencing centers for all transcripts.

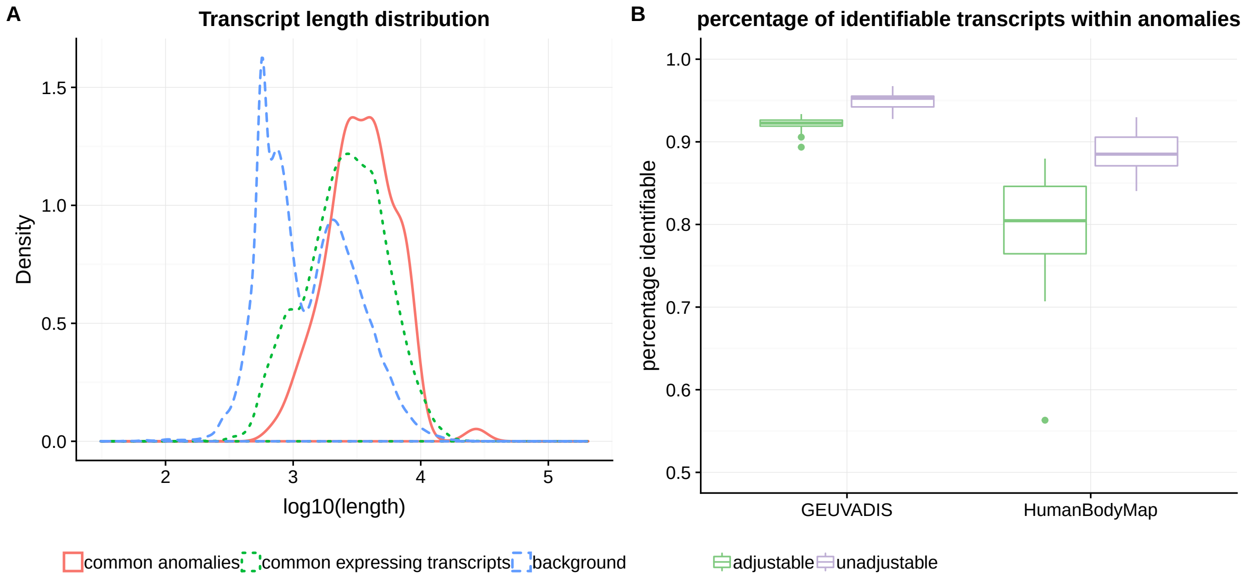

Figure S6: **Length distribution of unadjustable anomalies and identifiability status.** Related to Figure 1 and Figure 2. (A) Density curve of length distribution of the common unadjustable anomalies and commonly expressed transcripts across all 46 samples. The length distribution of common unadjustable anomalies generally follows that of the commonly expressed transcripts. Some transcripts are not commonly expressed. When including these transcripts into the background, the length distribution contains a large proportion of transcripts with a shorter length than the commonly expressed ones and the common unadjustable anomalies. (B) Percentage of unadjustable and adjustable anomalies that are identifiable in the quantification model. Each box indicates the range of percentages across the transcripts in the indicated dataset and the indicated anomaly category. Transcripts with identifiable expression indicate the optima is unique and the anomaly is not a result due to the intrinsic uncertainty of quantification optimization objective. Identifiability is determined by eXpress, which uses a different quantification model from Salmon but still reflect the degree of identifiability.

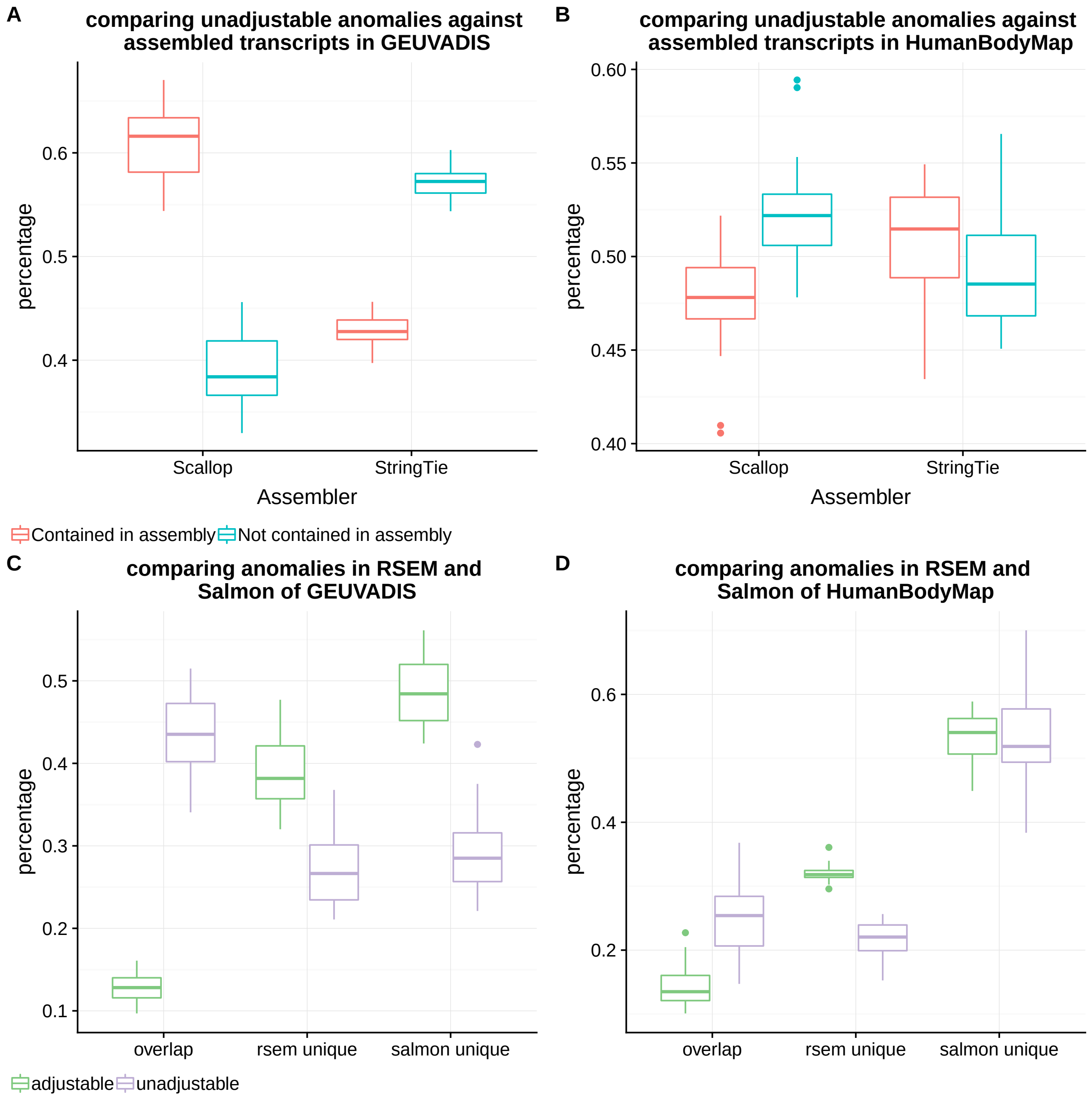

Figure S7: **Comparing Salmon anomalies with transcriptome assembly and RSEM anomalies.** Related to Figure 1 and Figure 2. (A–B) Proportion of the unadjustable-anomaly-containing genes that can (cannot) be detected by transcriptome assemblers. Each box indicates the range of percentages across samples in the corresponding dataset. (A) For the GEUVADIS dataset, about 40% of the genes do not have corresponding unannotated isoforms predicted by Scallop, and about 60% of the genes do not have unannotated isoforms predicted by StringTie. (B) For the Human Body Map dataset, about 53% of unadjustable-anomaly-containing genes cannot be detected by Scallop, and about 50% of them cannot be detected by StringTie. The lower percentage of detection from StringTie in GEUVADIS dataset may be an effect of using the “Guided by reference” option. (C–D) Overlapping of unadjustable anomalies predicted based on Salmon and RSEM on (C) GEUVADIS dataset and (D) Human Body Map dataset. Each box indicates the range of percentages across samples in the corresponding dataset. The denominator of the percentage calculation is the number of transcripts that are detected as unadjustable (or adjustable) anomalies under either Salmon or RSEM quantification.

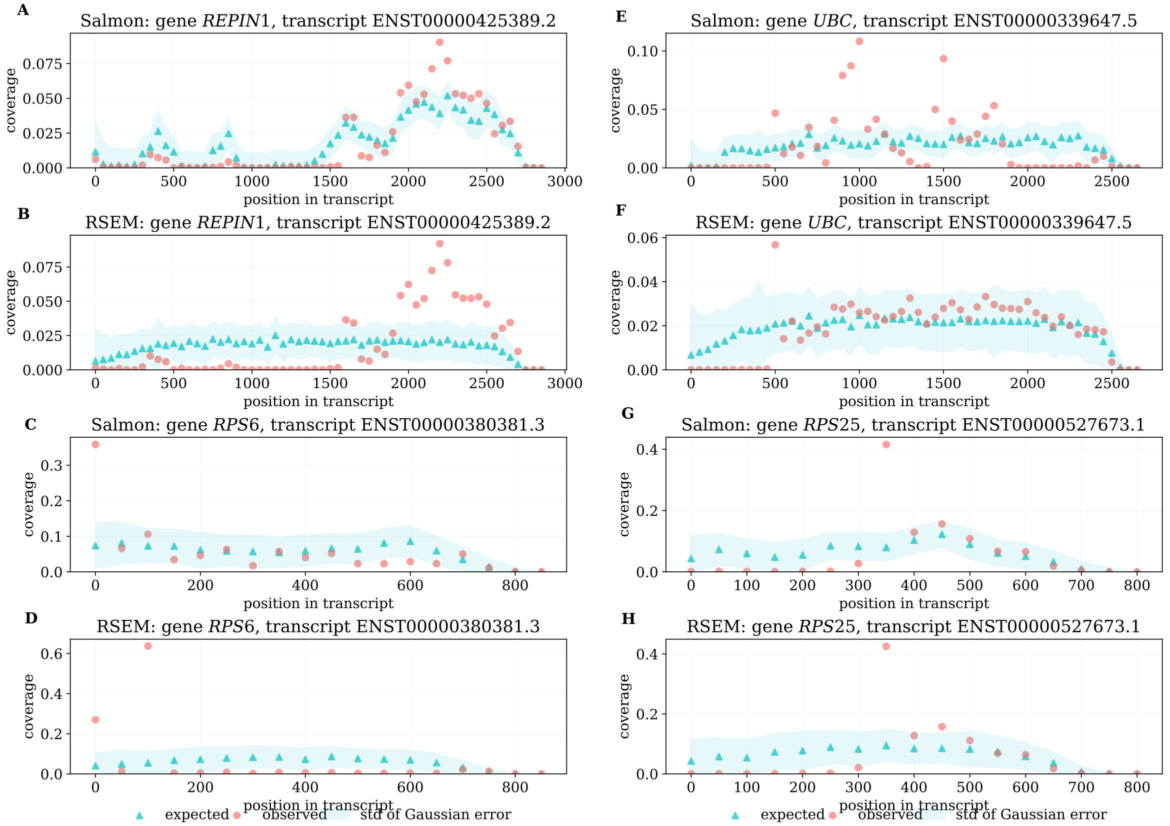

Figure S8: **Differences between Salmon and RSEM unadjustable anomalies.** Related to Figure 2. All examples are from one GEUVADIS sample (accession ERR188265). Red and blue points are the observed and expected coverage distribution separately, and the blue shade is the standard deviation of the expected distribution estimation. (A–B) Expected and observed coverage distribution for transcript ENST00000425389.2 under (A) Salmon and (B) RSEM. The expected distribution refers to the estimated expected distribution by the quantifier subtracted by the mean of Gaussian error. Each point is a 50 bp bin along the transcript. The transcript is identified to be unadjustable anomaly under only RSEM. The observed distributions under both quantifiers are similar. However, the expected distribution derived from Salmon bias correction is closer to the observed distribution than the one derived from RSEM bias correction. The difference in the estimated expected distribution causes the transcript to be detected as an unadjustable anomaly under RSEM but not Salmon. (C–D) Expected and observed coverage distribution for transcript ENST00000380381.3 under (C) Salmon and (D) RSEM quantification. The transcript is identified to be unadjustable anomaly under only RSEM. The observed coverage distributions has large difference between the two quantifiers around position 150. The large difference in observed distribution persists after read re-assignment and causes the transcript to be detected as an unadjustable anomaly under RSEM but not Salmon. (E–F) Expected and observed coverage distribution for transcript ENST00000339647.5 under (E) Salmon and (F) RSEM. The transcript is identified to be unadjustable anomaly under only Salmon. The observed coverage distributions has large difference between the two quantifiers. The difference between observed and expected coverages under Salmon persists after read re-assignment. (G–H) Expected and observed coverage distribution for transcript ENST00000527673.1 under (G) Salmon and (H) RSEM. The transcript is identified to be unadjustable anomaly under only Salmon. Both the observed and the expected coverage distribution under the two quantifiers are similar. However, RSEM has a relatively larger variance of Gaussian error in the expected distribution estimation and leads to a larger p-value. The different variance of Gaussian error in expected distribution causes the transcript to be detected as an unadjustable anomaly only under Salmon but not RSEM.

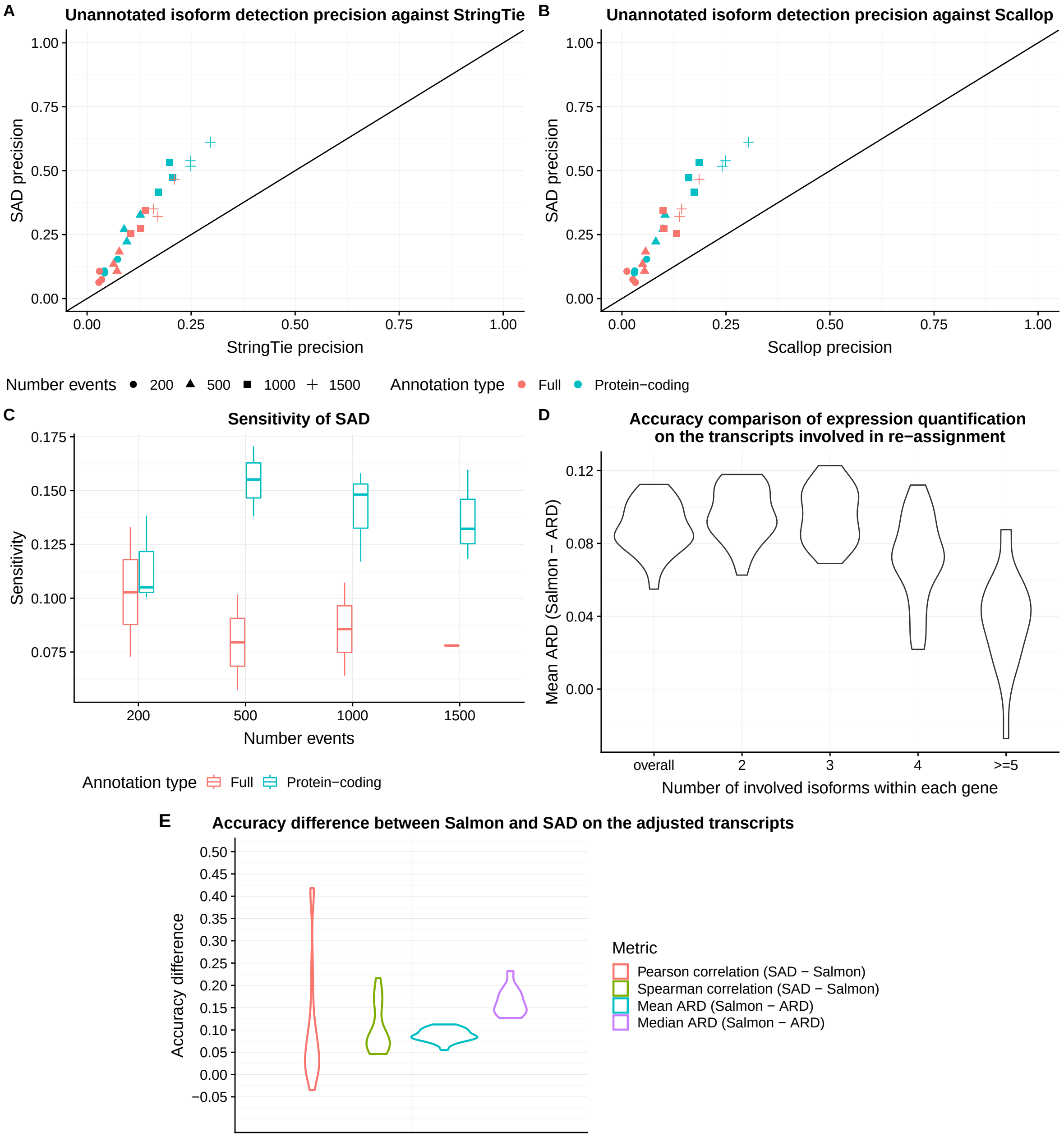

Figure S9: **Simulation performance of anomaly detection and correction.** Related to Figure 1 and Figure 2. (A) Precision of unannotated isoform detection using SAD unadjustable anomalies and StringTie assembly. Point color and shape refers to different simulation settings. The simulated unannotated isoforms do not contain unannotated splicing junctions, but only contain unannotated starting / ending sites, or unannotated combinations of known splicing junctions. (B) Precision of unannotated isoform detection using SAD unadjustable anomaly and Scallop. (C) Sensitivity of the unadjustable anomalies of SAD. Most of the simulated unannotated isoforms do not affect the coverage significantly enough to be detected by SAD. The boxes and the violins in the next two panels indicate the ranges of y-axis values across simulated datasets. (D) Quantification accuracy improvement of SAD compared to original Salmon. Each violin refers to a subset of transcripts where the corresponding genes contain a certain number of isoforms in the adjustment according to the x-axis. “Overall” in the x-axis is the overall mean ARD improvement of all adjusted isoforms without distinguishing the number of isoforms involved. (E) Overall quantification accuracy improvement of SAD compared original Salmon under four metrics. Positive accuracy differences indicate that SAD-adjusted quantification is an improvement under the metric.

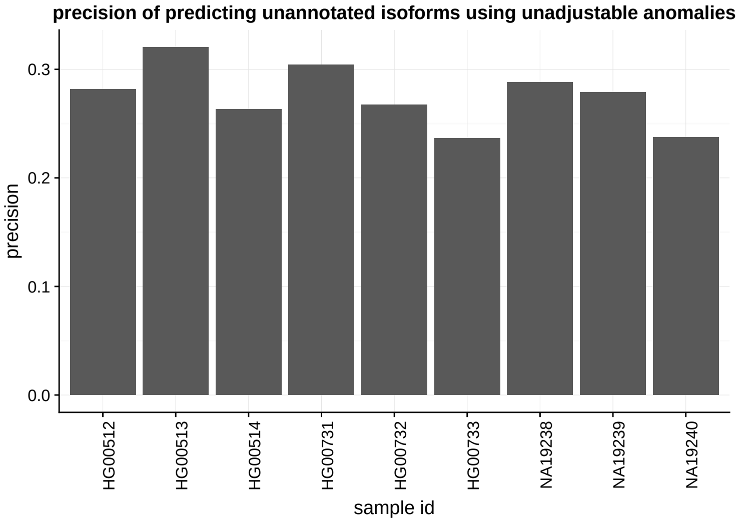

Figure S10: **Validating unadjustable anomaly prediction using full-length transcript sequencing.** Related to Figure 2. Whether an unadjustable anomaly is caused by an unannotated isoform can be validated by PacBio full-length transcript sequencing. When a sequencing reads contain a large proportion of the predicted over-expressed region and exclude a large proportion of the predicted under-expressed region, the unadjustable anomaly is considered to be supported by the long reads and correctly predicted. Y-axis shows the percentage of unadjustable anomalies that have long reads supports, that is, the prediction of unadjustable anomaly prediction. The precision is around 23% – 32% for all 9 samples from 1000 Genome project.

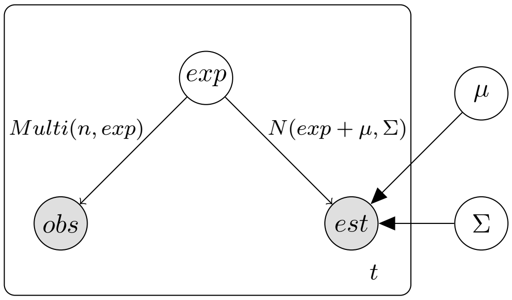

Figure S11: **The probability model of the expected distribution, the observed distribution, and the estimator of the expected distribution. Related to STAR Method Section "Probabilistic model for coverage distribution".** exp is the expected coverage, obs is the observed coverage, est is the estimation for the expected coverage. Here, exp is a hidden variable, while obs and est are observed. obs follows a multinomial distribution parameterized by the number of reads n and the expected coverage exp. est follows a Gaussian distribution with mean shift μ and covariance matrix Σ. We assume that the estimation errors of the expected coverage have the same pattern for all transcripts, and therefore μ and Σ are shared among all transcripts.