

# Supporting Information

Li et al. 10.1073/pnas.1113972108

## SI Text

**1. Linear Modeling of RNA-Seq Data.** Linear modeling of paired-end RNA-Seq data has been discussed in *Methods* of the main paper. The main points include (i) the definition of paired-end bins to summarize the key information in RNA-Seq data for isoform discovery, (ii) the enumeration of all possible isoforms from defined subexons, (iii) the modeling of conditional probabilities of observing reads in different bins given an isoform, and (iv) the construction of a linear model to estimate isoform proportions from observed bin counts.

Below, we present more details about modeling the fragment length distribution in  $\mathbf{F}$  and construction of linear model for single-end data.

**1.1. The fragment length distribution.** Modeling the cDNA fragment length distribution is a key part in constructing the design matrix  $\mathbf{F}$  of the linear model. Truncated Exponential is a reasonable candidate for the distribution, based on a Poisson point process assumption on a fragment's 3' end with the 5' end fixed and a size selection step in RNA-Seq protocols. Another widely used candidate in existing RNA-Seq tools is Normal distribution (1). To evaluate the two distributions, we compared them with empirical distributions of cDNA fragment lengths in paired-end RNA-Seq data. However, actual fragment lengths are unknown in genes exhibiting alternative splicing events, thus posing difficulties in obtaining the empirical distributions. To tackle this problem, a conservative solution is to calculate an empirical length distribution of cDNA fragments with both ends in the same subexon where no alternative splicing occurs. The good side of this solution is that the fragment lengths used in the calculation are actual, but the downside is that some long fragments across exons are not considered. Another solution is to calculate an empirical length distribution based on cDNA fragments in genes with no alternative splicing events in the UCSC *Drosophila melanogaster* (September 2010) annotation (16). This solution has the advantage of observing all sorts of fragment lengths, but its disadvantage is that wrong fragment lengths may be used if the annotation is incomplete. We employed both solutions to calculate the empirical distributions from dataset 1, and plotted them against either truncated Exponential or Normal distribution in Q-Q plots (Fig. S1). Parameters in the truncated Exponential and Normal distributions are chosen in such a way that both distributions have the same mean and variance as in the empirical distribution. Q-Q plots in Fig. S1 show that both truncated Exponential and Normal distributions are reasonable approximations of the fragment length distribution.

**1.2. Linear modeling of single-end RNA-Seq data.** For single-end RNA-Seq data, we can derive a similar linear model to the one used for paired-end data (Eq. 3 in the main paper). First, we enumerate possible isoforms in the same way as for the paired-end data, and categorize reads into single-end bins, defined as two-dimensional vectors indicating subexon indices of the reads. For example, single-end bin  $(i, j)$  contains reads whose 5' and 3' ends are in subexon  $i$  and  $j$ , respectively. A single-end bin count is defined as the number of reads in that bin. Bin counts of every gene are normalized as bin proportions, denoted by  $\mathbf{b}$ . Second, we construct a linear model to estimate isoform proportions  $\mathbf{p}$  from observed single-end bin proportions, with a design matrix  $\mathbf{F}$  as the conditional probabilities of observing reads in different single-end bins given an isoform. The modeling and calculation of the conditional probabilities for single-end data are similar to those

for paired-end data in the main paper. We consider a single-end bin as equivalent to a combination of multiple paired-end bins. For example, in a two-subexon gene, reads in single-end bin (1,1) correspond to paired-end reads in bins (1,1,1,1), (1,1,1,2), and (1,1,2,2). So the conditional probability of observing reads in single-end bin (1,1) given an isoform equals to the sum of conditional probabilities of observing reads in each of the three paired-end bins given the same isoform. In general, we calculate the conditional probability of observing reads in single-end bin  $j$  given isoform  $k$  as  $\sum_{r \in S_j} F'_{rk}$ , where  $S_j$  is the set of paired-end bins corresponding to the single-end bin  $j$ , and  $F'_{rk}$  is the conditional probability for paired-end data whose calculation has been described in details in the main paper. Last, we write a linear model in the same formula as in Eq. 3 of the main paper.

For combined paired-end and single-end data, we can simply construct a linear model by concatenating the observation vectors and combining the design matrices by rows in the linear models for paired-end and single-end data, respectively. Hence, the linear model used in SLIDE (sparse linear modeling of RNA-Seq data for isoform discovery and abundance estimation) can accommodate for different types of RNA-Seq data: paired-end, single-end, or both.

**1.3. Identifiability and preselection procedures.** To avoid the unidentifiability issue due to collinearity in the linear model (Eq. 2 in the main paper), we applied a preselection procedure: Only isoforms whose all subexon junctions have been observed are selected as candidates; for genes with more than two subexons, single-subexon isoforms are excluded from the candidates because of their rare existence. With this procedure, the number of parameters for an  $n$ -subexon gene can be reduced from  $2^n - 1$  to a significantly smaller number.

About the observations, there are frequently false zero counts of junction-end bins. We define junction-end bins as bins that include paired-end reads with at least one end across exon junctions [e.g., junction-end bins (1,1,1,2) and (1,2,2,2) include paired-end reads with one end covering the junction between subexons 1 and 2, whereas bin (1,1,2,2) is not a junction-end bin]. When bin (1,1,2,2) has positive counts, the expected counts of bins (1,1,1,2) and (1,2,2,2) should be positive, too; however, due to the difficulty of mapping junction reads, junction-end bins (1,1,1,2) and (1,2,2,2) are often observed with false zero counts. Thus, we exclude false zero junction-end bin proportions from the observations.

As an illustration of the effects of such preselection procedures, we calculate the numbers of genes with unidentifiability issues in their linear models (i.e.,  $\text{rank}(\mathbf{F}) < K$  in Eq. 1 of the main paper) before and after the preselection procedures for every group of  $n$ -subexon genes ( $n = 3, \dots, 10$ ). The numbers are summarized in Table S1, which shows that the preselection procedures have effectively overcome the unidentifiability issue for genes with three subexons and alleviated the problem for a few genes with more subexons. However, the percentage of genes with unidentifiability issues remains high after the preselection procedure for genes with more than three subexons; we see that the sparse estimation in SLIDE is still necessary.

**1.4.  $L_1$  vs.  $L_0$  regularization.** In sparse estimation,  $L_1$  penalty in Lasso is linear and ensures convexity of the objective function (Eq. 4 in the main paper). It also has the convexity property in logistic and Poisson regressions. Lasso does variable selection and shrinkage, thus permitting isoform discovery in SLIDE.  $L_0$

penalty is also a possible choice for sparse estimation. It was reported that  $L_0$  penalty can lead to a sparser model when the number of variables (e.g., the number of isoform candidates) is far larger than the number of relevant variables (e.g., the number of existing isoforms), whereas  $L_1$  penalty in Lasso has to use a large  $\lambda$  to screen out spurious variables and causes biases in retained variables (2, 3). However,  $L_0$  regularization is computationally disadvantageous because it makes the optimization problem nonconvex, and it has been shown that  $L_1$  is a good surrogate for  $L_0$  in many cases. In computational biology,  $L_1$  regularization is shown to be a good approach for high-dimensional and potentially sparse data (4). In our case, SLIDE does isoform discovery and abundance estimation in two steps, so the biased estimates of isoforms in the discovery step would not affect the subsequent abundance estimation step as long as true isoform estimates are not shrunk to zeros by Lasso. This is different from IsoLasso and NSMAP, which use one-step sparse estimation for simultaneous isoform discovery and abundance estimation (5, 6). Moreover, in our estimation, the existence of  $n_k$  (the number of exons in the  $k$ th isoform) in the penalty term would reduce the difference between  $L_1$  and  $L_0$  regularization. Therefore,  $L_1$  regularization is a reasonable choice for our sparse estimation.

### 1.5. Selection of the regularization parameter in sparse estimation.

The selection of the regularization parameter  $\lambda$  (Eq. 4 in the main paper) is by a stability criterion that aims to return the most stable results over different runs of estimation (7). Because genes of the same number of subexons have similar  $\dim(\mathbf{p})$  and  $\dim(\mathbf{b})$  in Eq. 4 of the main paper, we decided to group genes by their numbers of subexons  $n$  and select an optimal  $\lambda^{(n)}$  for each group from 16 candidate values  $(\lambda_i)_{i=1}^{16}$  (see Table 3 of the main paper). This grouping is particularly advantageous for selecting  $\lambda$  for lowly expressed genes, whose signal-to-noise ratios are low. Highly expressed gene signals can counteract the noise in the lowly expressed genes of the same group.

Suppose that there are  $m^{(n)}$  genes with  $n$  subexons,  $n = 3, \dots, 10$ . The selection procedures following the stability criterion are as follows.

1. For the  $r$ th gene with  $n$  subexons,  $r = 1, \dots, m^{(n)}$ , use  $\lambda = \lambda_i$ ,  $i = 1, \dots, 16$  in Eq. 4 to estimate  $\hat{\mathbf{p}}$  for 50 runs. In each run, use randomly sampled one half of the reads in the gene as input into SLIDE. Define  $q_{irk}$  as the proportion of runs in which  $\hat{p}_k > 0$ . Define  $\bar{q}_{ir} = \frac{\sum_{k=1}^K q_{irk}}{\sum_{k=1}^K I(\hat{p}_k > 0 \text{ in some runs})}$ .
2. Calculate the average of  $\bar{q}_{ir}$  over the  $m^{(n)}$  genes as  $\bar{q}_i = \frac{1}{m^{(n)}} \sum_{r=1}^{m^{(n)}} \bar{q}_{ir}$ .
3. Choose  $\lambda^{(n)}$  as  $\lambda_{i^*}$ , where  $i^* = \operatorname{argmax}_i \bar{q}_i$ .

The selected  $\lambda^{(n)}$  for datasets 1–4 and the simulation data are in Table 3 of the main paper.

**2. More Simulation Studies. 2.1. Simulation studies with different read coverages.** To study the isoform discovery accuracy of SLIDE in lowly expressed genes, we did a simulation study with three different read coverages: (i) 10 reads per kilobase of an annotated gene, (ii) 50 reads per kilobase of an annotated gene, and (iii) 100 reads per kilobase of an annotated gene. The simulated reads are paired-end with 37-bp length in each end. Precision and recall rates of SLIDE using the simulated data are summarized in Fig. S2, which shows that SLIDE has improved isoform discovery accuracy as the read coverage increases, as we expected. The improvement is significant when the read coverage increases from 10 reads per kilobase to 50 reads per kilobase, and the improvement becomes less significant when the read coverage increases further to 100 reads per kilobase. Given that many paired-end RNA-Seq data have more than 10 million reads, 10 reads per kilobase would correspond to less than 1 RPKM (number of

reads per kilobase per million of mapped reads) in those data. We note that a gene with such low read coverage and multiple exons is not likely to have all its exon junctions covered by reads, thus posing great difficulties on isoform discovery. As illustrated by this simulation study, SLIDE is robust to changes in gene expression levels when read coverage is beyond a certain threshold, and SLIDE has higher precision and recall rates and lower estimation variance as read coverage increases. When gene expression is too low (e.g., 10 reads per kilobase), some exons or exon junctions would not be observed and the dimensionality of observations in the core linear model would be reduced, thus resulting in incorrect estimation results by SLIDE. At the read coverage of 10 reads per kilobase, we have tried other likelihoods (multinomial and Poisson) to model the responses (i.e., bin counts) in the linear model of SLIDE, but the precision and recall rates are similarly low (see subsection 2.2). [Please note that our Poisson regression has a similar objective function as the maximum-likelihood approach used in NAMAP (6) has in the optimization, except for differences in the design matrix and penalty term.] This missing data problem associated with lowly expressed genes is not unique to SLIDE, because to accurately recover missing reads from observed data remains a big challenge for current RNA-Seq isoform discovery and quantification methods. Because of data noise and biases introduced at many experimental steps of the current RNA-Seq protocol, it would be difficult to recover missing exons or junctions by statistical models.

### 2.2. Simulation studies with different likelihoods in the core linear model.

To explore the effects of using different likelihoods in the generalized linear model of SLIDE (Eq 3 in the main paper), we tried three different likelihoods: Normal (the default), multinomial (logistic regression) and Poisson in the sparse estimation with simulated data. Reads were simulated under two read coverages, 10 and 100 reads per kilobase. Simulation settings are the same as described in the main paper. The results in Fig S3 illustrate that in general, the three different likelihoods do not give very different results in both read coverages. Looking more closely, we find that using Normal likelihood at read coverage 10 reads/kb gives slightly higher precision and recall rates for genes with 3–4 subexons, and using Logistic regression at read coverage 100 reads/kb gives lower precision rates for genes with 3–5 exons. In our SLIDE model, it is naturally to assume that the expected bin counts are linear in isoform quantities and to use an identity link function (Normal likelihood). These exploration results confirm that Normal likelihood is a reasonable choice.

### 2.3. Effects of isoform similarity and missing annotations on isoform discovery.

Similarity between different isoforms of the same gene would pose difficulties on isoform discovery. There are some cases where the isoform deconvolution is not identifiable because of the similarity between true isoforms (8, 9). For example, when some isoforms are fragments of others in the true isoform set, there would usually be more than one possible set of isoforms that can explain the observed exon expression levels and exon junctions.

In situations that annotations have missing but truly expressed isoforms, there are two different cases. First, when missing isoforms have exons not included in annotated isoforms, although SLIDE is not designed to recover missing exons from data, it can solve this issue by using de novo exons assembled by other softwares [e.g., Cufflinks (1), Scripture (10)]. Second, when all the exons in missing isoforms are included in annotated isoforms, SLIDE can discover the missing isoforms with high accuracy, especially if every missing isoform has more than one unique splice junctions. In the difficult case where some missing isoforms are fragments of annotated isoforms and the isoform deconvolution is not identifiable, SLIDE would discover a set of longest isoforms with the highest probability among all the possible sets

of isoforms. For example, we suppose that a three-exon gene has exon RPKMs 10, 20, and 10, respectively, and junction reads are observed between exons 1–2, and exons 2–3. In terms of the isoform deconvolution, there would be two possible sets of isoforms: (i) isoform (1,2,3) with RPKM 10 and isoform (2) with RPKM 10; or (ii) isoform (1,2) with RPKM 10 and isoform (2,3) with RPKM 10. In this case, SLIDE would favor the latter (set ii), which has a smaller penalty term. We design SLIDE to favor longer isoforms in the sparse estimation, by weighting each isoform abundance estimate with the inverse of its number of exons. This is based on our observations that most annotated isoforms contain many instead of few exons. In real data study, there are commonly observed 5' and 3' end biases in RNA-Seq data, that is, in our example above, even if the true isoform is (1,2,3), RNA-Seq read coverage in exons 1 and 3 is very likely to be lower than the read coverage in exon 2. To counteract the end biases in real RNA-Seq data, we allow SLIDE to favor isoforms with more exons in the sparse estimation. Therefore, SLIDE would find the longest isoform containing all the three exons unless the read coverage difference between exons 1 and 3 and exon 2 is significantly high.

We did simulation studies in the following three cases to illustrate the performance of SLIDE when annotations have missing isoforms but contain all the exons. In gene *RhoL*, there are four exons with lengths 379, 172, 286, and 204, respectively. The only annotated isoform is (1,2,3,4) that contains all four exons. In each of the following cases, we did 50 simulation runs with 500 paired-end 37-bp reads simulated in each run.

Case 1. Suppose that isoform (1,3,4) is missing in the annotation and its expression level is the same as that of isoform (1,2,3,4).

We note that (1,3,4) contains a novel junction between exons 1 and 3 that is not in the annotated isoform (1,2,3,4). For all 50 runs, SLIDE correctly discovered both isoforms.

Case 2. Suppose that isoform (2,3,4) is missing in the annotation and its expression level is the same as that of isoform (1,2,3,4).

We note that (2,3,4) is a fragment of the annotated isoform (1,2,3,4). For 49 out of the 50 runs, SLIDE correctly discovered both isoforms.

Case 3. Suppose that both isoforms (1,3,4) and (2,3,4) are missing in the annotation and both of their expression levels are the same as that of isoform (1,2,3,4). SLIDE correctly discovered all three isoforms in 18 runs. It missed isoform (2,3,4) in 18 runs, missed isoform (1,3,4) in 8 runs, and missed both in 6 runs.

From the results, we can see that it is more difficult to discover missing isoforms that are fragments of annotated isoforms, because SLIDE has to tackle end biases in real data. Nevertheless, these simulation results show that SLIDE has satisfactory performance in cases where annotations have missing isoforms.

### 3. More About mRNA Isoform Discovery on modENCODE Data. 3.1.

**More about the comparison between SLIDE and Cufflinks.** In the main paper, we carried out three comparisons between SLIDE and Cufflinks from different perspectives. First, we compare the two methods in their default settings, where SLIDE uses genes and exons from UCSC annotations and Cufflinks uses its de novo assembled genes and exons (Fig. 2B in the main paper). In the evaluation step, we compare discovered isoforms by each method with isoforms in UCSC annotations. We call a discovered isoform and an annotated isoform matched if they have the same number of exons and all of their exons overlap. Thus, our evaluation scheme is not sensitive to exon boundaries as long as de novo assembled exons are in the same loci as annotated exons. We agree that this comparison is not fair for Cufflinks, but the results still reveal two main problems of the Cufflinks results: (i) Cufflinks splits a gene into multiple parts when few junction reads are observed between certain exons; (ii) Cufflinks merges two

genes on opposite strands if they overlap because the read strand information is not properly considered. SLIDE does not have those two problems because it uses annotated gene boundaries that are mostly accurate. In the second comparison, we applied both methods to de novo assembled genes and exons by Cufflinks (i.e., to compare the isoform assembly performance of SLIDE and Cufflinks given the same set of genes and exons) (Fig. 2C in the main paper). The comparison results show that SLIDE and Cufflinks have similar precision and recall rates, which are, however, much lower than the precision and recall rates SLIDE had when using annotated genes and exons. We were concerned that the precision and recall rates in the second comparison might have been dominated by the de novo gene boundaries and exon loci that are different from the annotation. Therefore, we performed a third comparison between SLIDE and Cufflinks with only the genes whose de novo assembled exons agree with annotated exons in their loci. We found that the precision and recall rates of SLIDE are higher than those of Cufflinks. Therefore, we concluded that the isoform discovery performance of SLIDE is better than, or at least comparable to, that of Cufflinks.

The comparison results in the main paper are based on dataset 1 (Table 1 in the main paper). We did the same set of comparisons on datasets 2–4 (Table 1 in the main paper), and the results are summarized in Figs. S4 and S5 (results on dataset 1 are in Fig. 2B and C of the main paper). From Figs. S4 and S5, we observe that the comparison results on datasets 2–4 are consistent with the results on dataset 1.

**3.2. Comparison between SLIDE and IsoLasso/NSMAP.** Here, we compare SLIDE with two other isoform discovery methods with lasso-type sparse estimation: IsoLasso (5) and NSMAP (6).

SLIDE is different from IsoLasso (5) in three aspects. (i) IsoLasso enumerates isoforms based on a connectivity graph used by Scripture (10). This deterministic approach finds the longest paths indicated by connected paired-end reads and would not consider isoforms with alternative starts/ends (i.e., one isoform is a fragment of the other) as isoform candidates. (ii) IsoLasso uses a binary design matrix to relate reads to isoforms. It does not fully capture the quantitative relationship between read counts and isoform abundance. In contrast, our design matrix uses conditional probabilities to relate read counts to isoform abundance, and is flexible in terms of incorporating different types of biological information into the modeling (e.g., using GC content to adjust nonuniform read coverage). (iii) IsoLasso performs isoform discovery and abundance estimation simultaneously with Lasso, a sparse estimation method. However, the penalization term in Lasso would introduce biases to the abundance estimates. To fix this issue, SLIDE uses a two-step approach that first discovers isoforms by sparse estimation and subsequently estimates the abundance of the discovered isoforms by nonnegative least squares that gives less biased estimates than Lasso does. (iv) Unlike IsoLasso, SLIDE favors isoforms with more exons in its sparse estimation. This is because we observe that RNA-Seq data noise often leads the linear model to fit with multiple isoforms each with a small number of exons, contradicting with annotations. To counteract such data noise, we give less penalty to isoforms with more exons in the sparse estimation. In addition, IsoLasso builds isoform candidates from de novo exons directly assembled by mapped reads, without taking annotated gene and exon information into account.

We conducted three numerical comparisons between SLIDE and IsoLasso on isoform discovery based on the same data used in the comparison between SLIDE and Cufflinks. In the first comparison, we evaluated both methods in their default settings, where SLIDE builds isoforms from exons in UCSC annotations and IsoLasso finds isoforms from its de novo assembled exons. The discovered isoforms by either method are evaluated by UCSC annotations, where a discovered isoform is called to match



an annotated isoform if they have the same number of exons and all of their exons overlap. Thus, this evaluation scheme is not sensitive to exon boundaries as long as a discovered isoform has exons in the same loci as exons of an annotated isoform. Precision and recall rates are calculated as described in the main paper. The comparison results in Fig. S6A show that SLIDE has better precision and recall rates than IsoLasso does. These results are similar to the first comparison results between SLIDE and Cufflinks in the main paper. The main reason is that both IsoLasso and Cufflinks find isoforms for de novo assembled genes, whose boundaries are sensitive to RNA-Seq data noise, especially to biases of junction read counts. These results suggest the importance of scrutinizing de novo assembled genes and exons with available annotations before performing isoform discovery; however, because SLIDE and IsoLasso do not start from the same set of genes and exons, this comparison is not a fair evaluation of their isoform assembly performance. Thus, we performed a second comparison based on isoforms discovered by either method from de novo genes and exons assembled by IsoLasso. Fig. S6B shows that SLIDE still has better precision rates than IsoLasso has for most genes. To further exclude the effects of disagreement between annotated and de novo assembled exons, we carried out a third comparison of the two methods using only the de novo assembled exons that agree with the annotation. We found that SLIDE has an average precision rate 0.85 and recall rate 0.91, whereas IsoLasso has an average precision rate 0.79 and recall rate 0.91. This again shows that SLIDE has higher precision than IsoLasso has in isoform assembly from the same set of de novo exons.

NSMAP is a Bayesian model-based method that estimates the abundance of isoform candidates as MAP (maximum a posteriori) estimates (6). It is an extension of the maximum-likelihood abundance estimation method “statistical inferences for isoform expression in RNA-Seq” (SIIER) (11), in the sense of expanding parameters of interest from the abundance of annotated isoforms to that of all isoform candidates. NSMAP uses a Laplace prior to introduce sparseness and then discovers isoforms based on MAP estimates of the abundance of isoform candidates. NSMAP is similar to IsoLasso in four aspects. (i) Both methods use deterministic approaches to construct isoform candidates. NSMAP uses the minimal set of isoforms that can explain all junction reads, and it would miss isoforms with alternative starts/ends (i.e., one isoform is a fragment of the other) in its isoform candidate set. (ii) NSMAP is equivalent to IsoLasso in the optimization step. In the original Lasso paper, it was suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace priors (3). Unlike SLIDE, both NSMAP and IsoLasso do not favor isoforms with more exons in their sparse estimation, but they only keep the longest isoforms in their candidate sets. (iii) Both methods perform isoform discovery and abundance estimation simultaneously with sparse estimation. (iv) Both methods build isoforms from de novo assembled genes and exons. NSMAP constructs genes and exons de novo from read alignment output of Tophat. Therefore, the differences between SLIDE and NSMAP in methodology would be similar to those between SLIDE and IsoLasso. We tried to conduct numerical comparison between SLIDE and NSMAP. However, a code bug in the NSMAP package (Version 0.1.0) prohibited us from using it, and our attempts at contacting the authors were not successful.

**4. More About mRNA Isoform Abundance Estimation on modENCODE Data.** To evaluate the isoform abundance estimation accuracy of SLIDE without knowing the ground truth of isoform quantities in datasets 1–4, we compare SLIDE to two widely used methods: statistical inferences for isoform expression in RNA-Seq (SIIER) (11) and Cufflinks (1). All three methods are used to estimate the isoform proportions of 317 chr3R genes with multiple isoforms in

the UCSC annotation, and the total number of isoforms is 798. On dataset 1, the SLIDE and SIIER estimates have a correlation  $R = 0.75$ , and there are 25 genes with significantly inconsistent estimates between the two methods; i.e.,  $\hat{p}_{\text{SLIDE}} < 0.1$  and  $\hat{p}_{\text{SIIER}} > 0.5$  or  $\hat{p}_{\text{SLIDE}} > 0.5$  and  $\hat{p}_{\text{SIIER}} < 0.1$ . By detailed manual inspection, we find that among the 25 genes there are 20 genes whose SLIDE estimates agree better with the paired-end bin counts. For example, gene *CG9801* has five subexons with RPKMs 8.25, 5.57, 0, 3.92, and 3.16, respectively, and observed junctions between subexons 1–2, 2–4, and 4–5 in dataset 1. There are three isoforms (1,2), (1,2,5), and (1,2,4,5) of *CG9801* in the annotation. SLIDE estimates their proportions as 0.76, 0, and 0.24, respectively, whereas SIIER’s estimates are 0, 0.55, and 0.45, respectively. Because there is no observed junction between subexons 2 and 5 and the expression levels of subexons 1 and 2 are higher than those of subexons 4 and 5, the SLIDE estimates seem more consistent with the data. The rest of the 25 genes include one gene whose SIIER estimates agree better with the paired-end bin counts, and four genes with ambiguous bin counts that cannot differentiate the two sets of estimates. An example of the ambiguous cases is gene *D1* with five subexons. In dataset 1, the RPKMs of the five subexons are 327.79, 326.01, 16.6, 6.23, and 0, respectively, and there are observed junction reads between subexons 1–2, 2–3, and 3–4. SLIDE estimates the proportions of annotated isoforms (1,2,3,4) and (1,2,3) as 0.02 and 0.98, respectively, whereas SIIER returns estimates 1 and 0, respectively. Based on the annotation, we would expect subexons 1, 2, and 3 to have similar expression levels; however, the observed expression in subexon 3 is significantly lower than that of subexons 1 and 2. So the data seriously contradicts with the annotation. Hence, both SLIDE and SIIER cannot reasonably fit the data based on the annotation. After removing those 25 genes, we have a correlation  $R = 0.88$  between the SLIDE and SIIER estimates.

In the comparison between SLIDE and Cufflinks, the correlation between their estimates on the proportions of the 798 isoforms is  $R = 0.67$  on dataset 1. Similarly, we find 35 genes with significantly inconsistent estimates between SLIDE and Cufflinks,  $\hat{p}_{\text{SLIDE}} < 0.1$  and  $\hat{p}_{\text{Cufflinks}} > 0.5$  or  $\hat{p}_{\text{SLIDE}} > 0.5$  and  $\hat{p}_{\text{Cufflinks}} < 0.1$ . Again by detailed manual inspection, we observe that 30 of them have SLIDE estimates that agree better with the paired-end bin counts, 3 have Cufflinks estimates that agree with the bin counts, and 2 have ambiguous bin counts such that both estimates are reasonable. After removing those 35 genes, we have a correlation  $R = 0.85$  between the SLIDE and Cufflinks estimates.

**5. More About the Exploration of Read/Fragment Length Effect.** In the exploration of whether different read lengths would affect the isoform discovery results of SLIDE, we applied SLIDE to datasets 2 and 3, which are from the same Kc167 sample, with similar sequencing depth, but of read lengths 37 and 76 bp, respectively. Surprisingly, the precision and recall rates on the 37-bp data are higher than those on the 76-bp data. In the search for a possible explanation, we observed that the cDNA fragments in single-exon genes have different fragment length distributions in the two datasets:  $N(166, 26^2)$  and  $N(127, 13^2)$  for the 37-bp and 76-bp data, respectively.

To explore whether the read length or the fragment length has larger effects on the isoform discovery, we did a simulation study with two different read lengths (37 and 76 bp) and three different fragment length ranges (50–100 bp, 100–150 bp, and 150–200 bp). In each of the 50 simulation runs, 500 paired-end RNA-Seq reads are generated in each setting for each read length and each fragment length range. We applied SLIDE to the simulated data and summarized the precision and recall rates of each setting in Fig. S7. The figure illustrates that the increase in fragment lengths from 50–100 bp to 100–150 bp significantly improves the precision and recall rates of isoform discovery. Changes in frag-

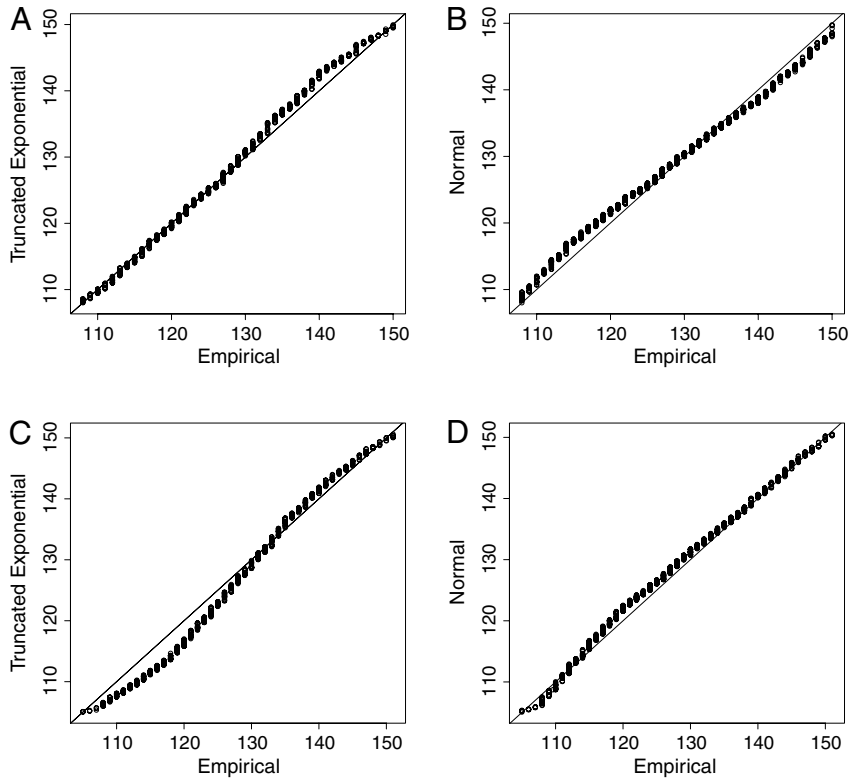
ment lengths from 100–150 bp to 150–200 bp also improve the precision and recall rates by increasing their means to some extent and decreasing the width of their confidence intervals. Compared to the fragment length changes, read length changes from 37 bp to 76 bp have smaller effects on the isoform discovery results.

**6. Read Coverage vs. GC Content.** It has been reported by several groups that read coverage has a strong correlation with GC content in high-throughput DNA sequencing data (2, 12). As high-throughput sequencing technologies (e.g., DNA sequencing, RNA-Seq, ChIP-Seq, etc.) have similar characteristics in the sequencing step, many researchers believe that a strong correlation between read coverage and GC content also exists in RNA-Seq data (13, 14). However, unlike DNA sequencing data, RNA-Seq read coverage varies in different transcribed regions and is mainly determined by expression levels and alternative splicing patterns of the regions (15). It would be difficult to compare read coverage across subexons, which may occur in different transcripts and thus have different expression levels. To check the validity of using GC content correction in our SLIDE model, we study the relationship between RNA-Seq read coverage and GC content within subexons, using RNA-Seq reads on chr3R in dataset 1 (see Table 1 of the main paper). We use three different window sizes: 10 bp, 30 bp, and 50 bp. For every subexon, we calculate the correlation coefficient of its windowed average read coverage vs. GC content. Then, we calculate the percentage of subexons giving positive correlations among all the subexons with more than  $n$  windows ( $n = 3, 10, \dots, 100$ ), and find that the percentage increases as  $n$  increases. This trend is observed with all the three window sizes. The percentages for 10-bp windows are summarized in Table S2. A histogram of the correlations in subexons with more than 100 windows is in Fig. S8. Because we expect that correlations calculated in subexons with more windows can better represent the relationship between read coverage and GC content, we conclude that there is a positive correlation between read coverage and GC content.

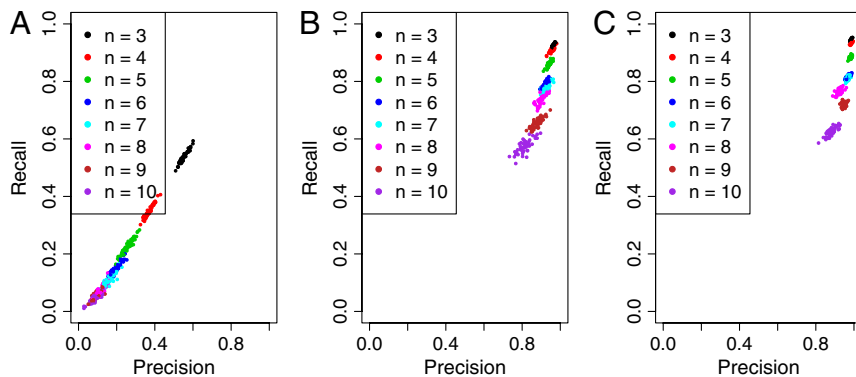
## 7. Some Other Details in the Analysis.

- In the simulation study of the main paper, we simulated reads from the 1,972 genes of 3–10 subexons (defined in the main paper) on chr3R from *D. melanogaster* annotation (September 2010) of UCSC Genome Browser (16). For each gene, reads are generated from the annotated isoforms, whose proportions  $p_k$  are randomly sampled from  $\{0, 0.1, \dots, 0.9, 1\}$  subject to the constraint that  $\sum_k p_k = 1$ . For every gene, we simulate 500 reads in each run, with 50 runs in total.
- In the sparse estimation, we selected an optimal  $\lambda$  for each group of genes with  $n$  subexons ( $n = 3, \dots, 10$ ) by a stability criterion (7). However, there are a small number of genes where zero isoforms were identified under the selected  $\lambda$ . For those genes, we reselected a gene-specific  $\lambda$ . In more details, we replace the previous  $\lambda$  by  $\lambda^* = \max(\lambda - 0.1, \lambda/2)$  until non-trivial results were obtained.
- For isoform discovery, SLIDE uses sparse linear model estimation to find isoforms. We note that the linear model for paired-end data (Eq. 3 in the main paper) is identifiable; i.e.  $\mathbf{F}^T \mathbf{F}$  is invertible (8), for a few genes with 3–10 subexons of *D. melanogaster* (Table S1). In those cases, we additionally attempted to use nonnegative least squares (NNLS), whose estimation results should be less biased than those of  $L_1$  penalized estimation. However, compared to the penalized estimation results in the main paper, we found that the NNLS results include a lot of short isoforms as truncated fragments of isoforms in the UCSC annotation. In the example of gene *jumu*, whose three subexons have RPKMs 20.86, 42.62, and 25.97, respectively, there are observed junctions between subexons 1–2 and 2–3. NNLS discovered isoforms (1,2), (2,3), and (1,2,3) for *jumu*, whereas SLIDE only found the longest isoform (1,2,3), which agrees with the annotation. The possible reason of NNLS finding short isoforms is that RNA-Seq data have unexpected read coverage variation among exons in the same transcript (12–14).

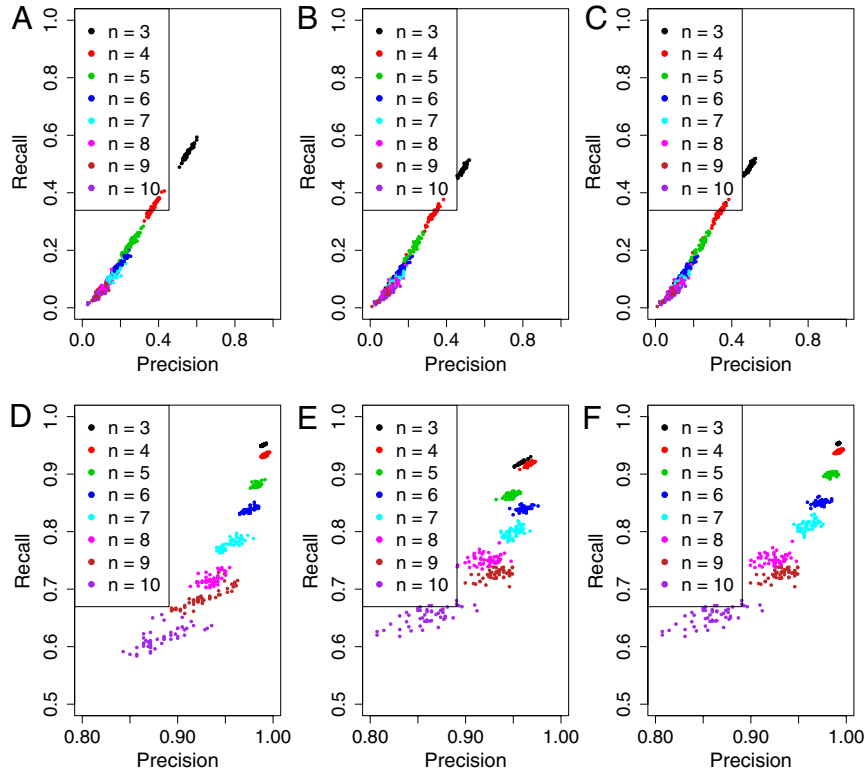
1. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28:511–515.
2. Hiller D (2010) Alternative splicing analysis using RNA-seq data. PhD dissertation (Department of Statistics, Stanford University, Palo Alto, CA).
3. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58:267–288.
4. Dahinden C, Parmigiani G, Emerick MC, Bhlmann P (2007) Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics* 8:1–11.
5. Li W, Feng J, Jiang T (2011) IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. *15th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2011), Lecture Notes on Computer Science* (Springer, Berlin/Heidelberg), 6577:168–188.
6. Xia Z, Wen J, Chang C, Zhou X (2011) NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics* 12: 162.
7. Meinshausen N, Bühlmann P (2010) Stability selection. *J R Stat Soc Series B Stat Methodol* 72:417–473.
8. Hiller D, Jiang H, Xu W, Wong WH (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics* 25:3056–3059.
9. Zheng S, Chen L (2009) A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res* 37:e75.
10. Guttman M et al. (2010) Ab initio reconstruction of cell typespecific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 28:503–510.
11. Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25:1026–1032.
12. Dohm JC, Lottaz C, Borodina T, Himmelbauer H, (2010) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36: e105.
13. Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 38: e131.
14. Li J, Jiang H, Wong WH (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* 11: R50.
15. Srivastava S, Chen L (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* 38: e170.
16. Fujita PA (2011) The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* 39(Suppl 1):D876–D882.



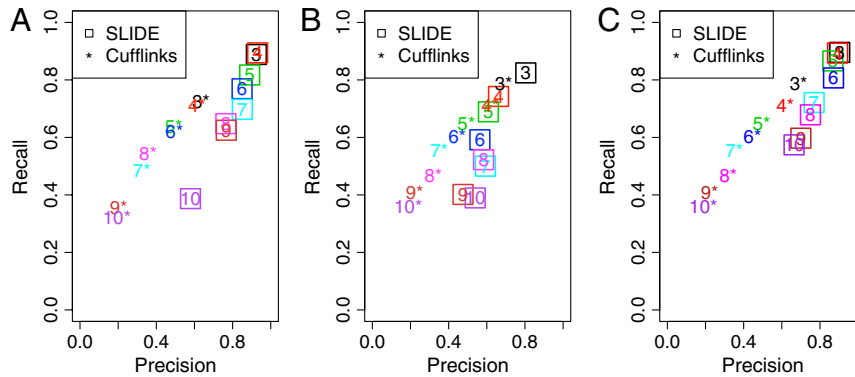
**Fig. 51.** Q-Q plots of modeled vs. empirical fragment length distribution on dataset 1. Note that only the fragment lengths between the 5% and 95% percentiles of the empirical distribution are used to construct the Q-Q plots, because extremely long or short fragments may be results of mapping errors. (A) Q-Q plot of truncated Exponential distribution vs. empirical length distribution of cDNA fragments within single-exon genes. (B) Q-Q plot of Normal distribution vs. empirical length distribution of cDNA fragments within single-exon genes. (C) Q-Q plot of truncated Exponential distribution vs. empirical length distribution of cDNA fragments within single-isoform genes. (D) Q-Q plot of Normal distribution vs. empirical length distribution of cDNA fragments within single-isoform genes.



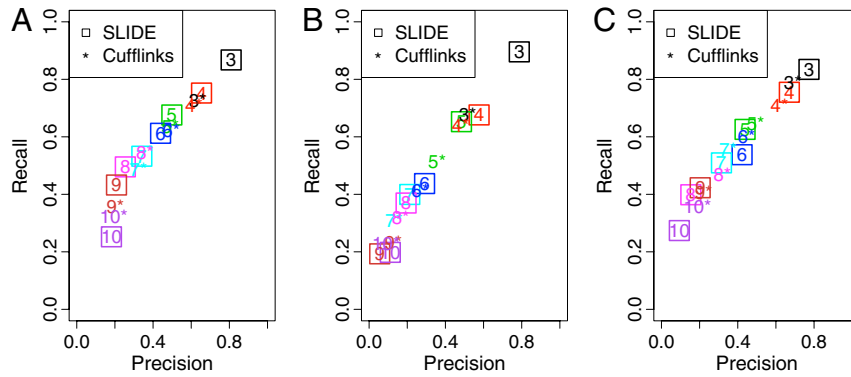
**Fig. 52.** Precision and recall rates of SLIDE on simulated data with different read coverages. (A) Read coverage is 10 reads per kilobase. (B) Read coverage is 50 reads per kilobase. (C) Read coverage is 100 reads per kilobase.



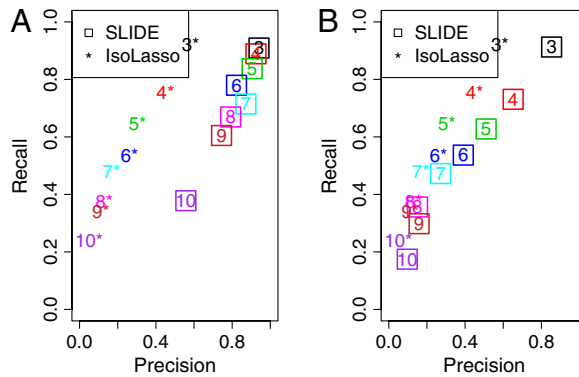
**Fig. S3.** Precision and recall rates of SLIDE using different likelihoods in simulation with two different read coverages. (A) Normal likelihood, (B) Poisson likelihood, and (C) multinomial likelihood (logistic regression) with read coverage 10 reads/kb. (D) Normal likelihood, (E) Poisson likelihood, and (F) multinomial likelihood (logistic regression) with read coverage 100 reads/kb.



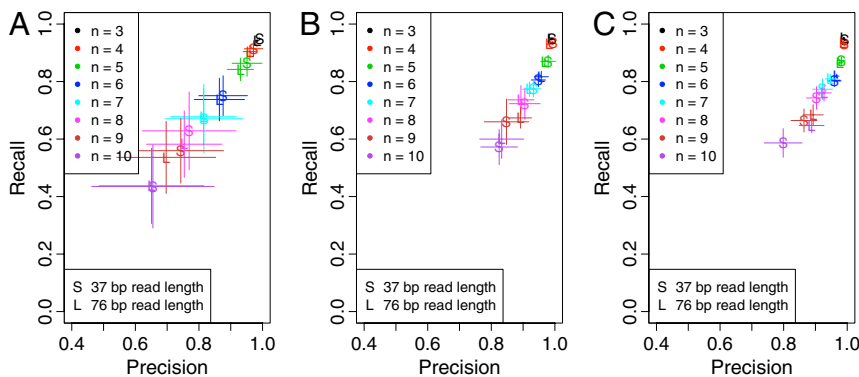
**Fig. S4.** Comparison of isoform discovery results by SLIDE (using genes and exons from the UCSC annotation) and Cufflinks. (A) Precision and recall rates of SLIDE and Cufflinks on dataset 2. The numbers in the figure are the group indices of genes (i.e., numbers of subexons). The squares and stars represent SLIDE and Cufflinks results, respectively. (B) Precision and recall rates of SLIDE and Cufflinks on dataset 3. The numbers, squares, and stars have the same meaning as in A. (C) Precision and recall rates of SLIDE and Cufflinks on dataset 4. The numbers, squares, and stars have the same meaning as in A.



**Fig. 55.** Comparison of isoform discovery results by SLIDE (using de novo genes and exons assembled by Cufflinks) and Cufflinks. (A) Precision and recall rates of SLIDE and Cufflinks on dataset 2. The numbers in the figure are the group indices of genes (i.e., numbers of subexons). The squares and stars represent SLIDE and Cufflinks results, respectively. (B) Precision and recall rates of SLIDE and Cufflinks on dataset 3. The numbers, squares, and stars have the same meaning as in A. (C) Precision and recall rates of SLIDE and Cufflinks on dataset 4. The numbers, squares, and stars have the same meaning as in A.



**Fig. 56.** Comparison of isoform discovery results by SLIDE and IsoLasso. (A) Precision and recall rates of SLIDE (using annotated genes/exons) and IsoLasso on dataset 1. The numbers in the figure are the group indices of genes (i.e., numbers of subexons). The squares and stars represent SLIDE and Cufflinks results, respectively. (B) Precision and recall rates of SLIDE (using IsoLasso assembled genes/exons) and IsoLasso on dataset 1. The numbers, squares, and stars have the same meaning as in A.



**Fig. 57.** Simulation study of read/fragment length effects on isoform discovery. (A) Precision and recall rates of SLIDE on simulated paired-end RNA-Seq data with fragment lengths in the range of 50–100 bp and two different read lengths (37 bp vs. 76 bp). 95% confidence intervals of precision and recall rates are shown as error bars parallel to the x and y axes, respectively. (B) Precision and recall rates of SLIDE on simulated paired-end RNA-Seq data with fragment lengths in the range of 100–150 bp and two different read lengths (37 bp vs. 76 bp). The confidence intervals are shown in the same way as in A. (C) Precision and recall rates of SLIDE on simulated paired-end RNA-Seq data with fragment lengths in the range of 150–200 bp and two different read lengths (37 bp vs. 76 bp). The confidence intervals are shown in the same way as in A.



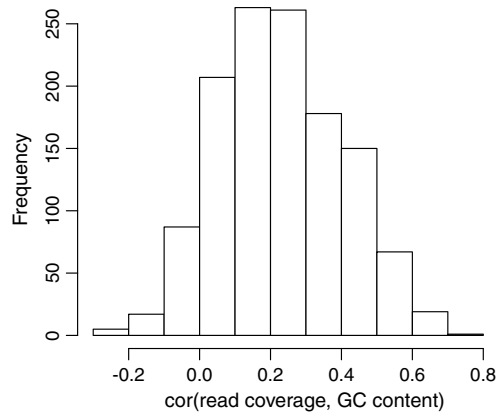


Fig. S8. A histogram of correlations between windowed read coverage and GC content in subexons with more than 100 windows of 10-bp size.

**Table S1. Number of genes with unidentifiability issues before and after preselection procedures**

No. of subexons	Total no. of genes	No. of genes with unidentifiability issue before the preselection procedures	No. of genes with unidentifiability issue after the preselection procedures
3	295	204 (69.2%)	1 (0.3%)
4	237	228 (96.2%)	198 (83.5%)
5	165	165 (100%)	155 (93.9%)
6	142	142 (100%)	137 (96.5%)
7	82	82 (100%)	80 (97.6%)
8	72	72 (100%)	70 (97.2%)
9	56	56 (100%)	55 (98.2%)
10	35	35 (100%)	35 (100%)

**Table S2. Percentages of subexons (>*n* 10-bp windows) with positive correlation (R) between read coverage and GC content**

<i>n</i>	3	10	20	40	60	80	100
Percentage	77.3%	78.6%	83.0%	87.9%	90.1%	91.0%	91.3%
Mean(R)	0.171	0.174	0.193	0.219	0.227	0.232	0.229