Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants

JAE-HYUNG LEE,[1] JASON K. ANG, and XINSHU XIAO[2]

Department of Integrative Biology and Physiology, Bioinformatics Interdepartmental Program, and the Molecular Biology Institute, University of California, Los Angeles, California 90095, USA

ABSTRACT

RNA-sequencing (RNA-Seq) technologies hold enormous promise for novel discoveries in genomics and transcriptomics. In the past year, a surge of reports has analyzed RNA-Seq data to gain a global view of the RNA editome. Opposing results have been presented, giving rise to extensive debate surrounding one of the first such studies in which a daunting list of all 12 types of RNA–DNA differences (RDDs) were identified. Although a consensus is forming that some of the initial "paradigm-shifting" results of this study may be questionable, recent reports on this topic differed in terms of the number and relative abundance of each type of RDD. Many outstanding issues exist, most importantly, the choice of bioinformatic approaches. Here we discuss the critical data analysis and experimental design issues of such studies to enable improved systematic investigation of the largely unexplored frontier of single-nucleotide variants in RNA.

Keywords: RNA editing; RNA–DNA difference; RNA-Seq; read mapping; mapping error

INTRODUCTION

Recently, high-throughput RNA-sequencing (RNA-Seq) data were analyzed in multiple studies to identify RNA–DNA mismatches in mammalian mRNAs (Li et al. 2011; Bahn et al. 2012; Chen et al. 2012; Danecek et al. 2012; Gu et al. 2012; Kleinman et al. 2012; Park et al. 2012; Peng et al. 2012; Ramaswami et al. 2012, 2013; Dillman et al. 2013; Lagarrigue et al. 2013). One of the first studies reported a striking list of all 12 types of RNA–DNA differences (RDDs) in human cells (Li et al. 2011), while others presented opposing results (Schrider et al. 2011; Bahn et al. 2012; Danecek et al. 2012; Gu et al. 2012; Park et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). There is still extensive debate in the scientific community as to whether a large number of RDDs exist, especially those that cannot be explained by known types of RNA editing. According to recent analyses of published results, various artifacts or technical errors may account for many apparent RDDs (Schrider et al. 2011; Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012). Along this line, a number of studies have suggested the necessity of applying extensive filtering on RDD results to remove false positives obtained by nominal mapping methods (Danecek et al. 2012; Gu et al. 2012; Park et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). Here we highlight a different perspective, supported by analytical comparisons, that upfront stringent mapping of the reads can bypass much of the need for post-filtering and be advantageous in quantifying editing levels. In addition, we discuss several other critical issues in the design and analysis of RNA-Seq experiments for studying RNA editing and provide guidance for this direction of research. Notably, the issues and discussions here are generally applicable to studies of any single-nucleotide variants (SNVs), such as genetic variants, expressed in the RNA. Through rigorous experimental design and bioinformatic analysis, we stand poised to make rapid progress toward deciphering the RNA editome in a variety of tissues, organisms, diseases, and environmental conditions.

TWO STRATEGIES FOR RDD PREDICTION USING RNA-Seq

In the pursuit of SNVs in RNA-Seq data, read mapping is crucial. This is because mismapped reads often contain apparent mismatches to the DNA, thus giving rise to false-positive RDDs. As alluded to above, two types of mapping strategies may be adopted. The first strategy applies stringent mapping criteria to reduce potential bias or artifacts due to the presence of sequence variants or homologous regions in the genome (Bahn et al. 2012) (Strategy 1) (Fig. 1A). In contrast, the second strategy, which is more often used, applies nominal mapping approaches, followed by a series of

[1]**Present address:** Department of Maxillofacial Biomedical Engineering, School of Dentistry, Kyung Hee University, Seoul 130-701, Korea

[2]**Corresponding author**

E-mail gxxiao@ucla.edu

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.037903.112.

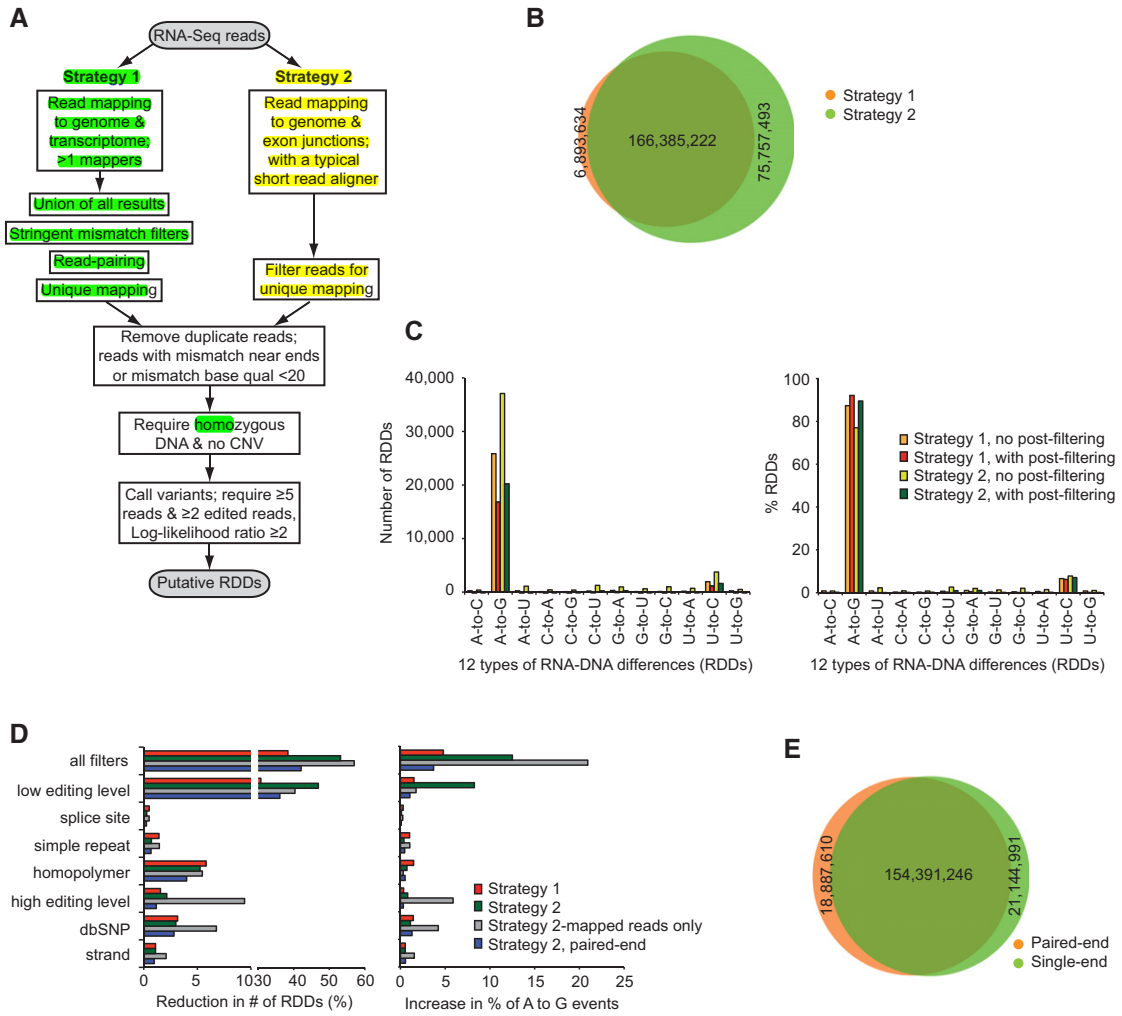

FIGURE 1. Performance comparison of alternative read mapping strategies. (*A*) Flowchart of read mapping and RDD analysis using RNA-Seq data with two alternative strategies for read mapping. Strategy 1 was adapted from a previous study (Bahn et al. 2012) and includes usage of multiple aligners (Bowtie and Blat were used for all results shown in this work), stringent filters on allowed number of mismatches per read, and a read-pairing procedure. A typical short read aligner and its default mismatch filter were used in Strategy 2 (specifically, BWA was used as described previously) (Ramaswami et al. 2012). A few basic filters were applied to both strategies following read mapping, such as duplicate read removal and other quality control measures. Statistical significance of the identified RDDs (log-likelihood ratio) was evaluated using a previously published method (Bahn et al. 2012). (CNV) Copy number variation. (*B*) Using paired-end poly(A^+) RNA-Seq data obtained previously (Peng et al. 2012) (same below), the number of final mapped reads common to both strategies or exclusive to one strategy is shown. (*C*) Number of RDDs and % of A-to-G events in the RNA-Seq data in *B* identified by the two mapping strategies prior to and following application of seven artifact filters of RDDs (see *D* and text). Since the RNA-Seq libraries were nonstrand specific, the types of RDDs were determined based on RefSeq gene annotation. As a result, the complementary types of RDDs may not be clearly distinguished (e.g., some of the T-to-C events may be genuine A-to-G events in unannotated antisense genes). (*D*) Reduction in the total number of events and increase in % of A-to-G events following seven artifact filters of RDDs defined previously (Peng et al. 2012; Ramaswami et al. 2012). Results are shown, respectively, for all reads mapped by Strategies 1 or 2, reads mapped exclusively by Strategy 2, and reads mapped by Strategy 2 plus a read-pairing procedure. Note that the paired-end mapping of Strategy 2 was implemented by pairing only uniquely mapped reads by BWA, as BWA currently provides limited mismatch information for nonuniquely mapped reads. Definition of filters are as follows: "low editing level" filter: requires three or more edited reads and an editing level of ≥0.1; "splice site" filter: removes intronic RDDs within 4 bp of splice junctions; "simple repeat" filter: discards sites in simple repeats; "homopolymer" filter: removes sites in homopolyer runs of ≥5 bp; "high editing level" filter: removes sites with 100% editing levels; "dbSNP" filter: removes RDDs annotated as SNPs (dbSNP 135); "strand" filter: removes sites whose reads exhibited strand bias in distribution ($P < 0.01$, Fisher's Exact test). (*E*) Number of final mapped reads with or without read pairing in Strategy 1.

filters to remove likely false-positive RDDs (Strategy 2) (Danecek et al. 2012; Gu et al. 2012; Peng et al. 2012; Ramaswami et al. 2012).

The stringency of Strategy 1 is mainly imposed by the use of carefully designed filters on mismatches per read in the read-mapping stage (Bahn et al. 2012). Specifically, a dual-filtering scheme on mismatches is applied to require that (1) a read is mapped uniquely with up to n_1 mismatches to the reference (genome and transcriptome); and (2) the read does not map to regions other than the unique position in

(1) if up to n_2 mismatches were allowed ($n_2 > n_1$). The values of n_1 and n_2 were determined using simulated reads with sequencing error profiles derived from the specific RNA-Seq data at hand (see more discussions below). Moreover, another feature of this strategy is the usage of Blat (Kent 2002) in the first step of read mapping, combined with an often-used short read aligner such as Bowtie (Langmead et al. 2009) and BWA (Li and Durbin 2009). Owing to the fundamental algorithmic difference between Blat and most short read aligners, their read mapping results are often different and complementary, especially in handling reads derived from spliced junctions. A third feature of Strategy 1 is the incorporation of a read-pairing step to retain only uniquely paired reads if paired-end data are available (see below).

In the second strategy for RNA-Seq based RDD prediction, one of the short read aligners, such as Bowtie (Langmead et al. 2009) or BWA (Li and Durbin 2009), is used in the same way as adopted in most gene-expression analysis, where a universal mismatch cutoff is applied in parsing the mapping results. RDDs derived by this strategy have been associated with a number of features indicative of false-positive predictions. Consequently, these RDD sites necessitate a series of post-filtering steps to reduce false positives. In analyzing human RNA-Seq data, a total of seven filters have been applied (listed in Fig. 1D), primarily (except the known single nucleotide polymorphism [SNP] filter) to remove such artifacts (Peng et al. 2012; Ramaswami et al. 2012). First, a "strand filter" was applied to remove RDDs covered by reads with a "strand bias." This bias refers to the observation that some reads containing the presumably edited base map more often to one strand of the genome than the other strand (Gu et al. 2012; Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012). It may reflect a difference in sequencing error rate or mapping accuracy between the two strands of a particular genomic locus (Meacham et al. 2011; Pickrell et al. 2012). Another filter removes RDDs with 100% editing, which, based on simulation studies, often arise from mapping errors to homologous regions (Peng et al. 2012). RDD sites with low-editing levels (e.g., <10% and less than three edited reads) are also removed based on the assumption that they may reflect mapping errors in a minority of reads, although many authentic low-level editing sites do exist. In addition, RDD sites close to splice sites (e.g., ≤4 nt from spliced junctions) and those within simple repeats or homopolymers are removed due to the potential inaccuracy of read mapping to such regions. Lastly, RDD sites overlapping known SNPs in public databases were excluded, as they are likely expressed genetic variants.

READ MAPPING STRINGENCY IMPROVES ACCURACY IN RDD PREDICTION

The accuracy of RDD prediction has been evaluated using either laborious experimental validation or simulated reads that may not faithfully reflect all artifacts in the actual data. Here, we propose two additional measures to assess the overall accuracy of a read-mapping approach in predicting RDDs. Since the post-mapping filters described above for Strategy 2 were designed to remove artifacts in RDD prediction, one way to evaluate mapping accuracy is to examine the number of RDDs excluded by these filters. If the filters eliminate only a small number of RDDs, the accuracy of the predicted RDDs is relatively high (at least for the specific artifact targeted by the filter) suggesting a high mapping accuracy. A second measure of RDD accuracy is the "change" in % A-to-G events after artifact filtering. A general assumption is currently accepted in the field that A-to-G mismatches, corresponding to A-to-I editing, should be dominant among all predicted RDDs in mammals, as observed previously (Bahn et al. 2012; Danecek et al. 2012; Gu et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). As a result, there has been a tendency to assume that the higher the % A-to-G is, the more accurate the predicted RDD events are (Peng et al. 2012; Ramaswami et al. 2012). However, for any specific data set, the correct % of A-to-G among all RDDs is not known and it may not be 100% since other types of genuine mismatches (e.g., C-to-U) may exist. Thus, it is not ideal to evaluate the mapping and RDD accuracy solely based on the absolute % A-to-G value. Instead, if the "increase" in % A-to-G is high following artifact filtering, it can be assumed that the initial RDD results are contaminated by relatively high artifacts. Thus, we propose to use the "change" in % A-to-G events after artifact filtering as another way to evaluate accuracy in RDD prediction.

We applied the two mapping strategies to a public RNA-Seq data set (paired-end) derived from human lymphoblastoid cells (Peng et al. 2012). For Strategy 1, we followed the exact procedures and parameters as described in Bahn et al. (2012), where Blat and Bowtie were used for read alignment. For Strategy 2, we used BWA and related parameters to align the reads as described in Ramaswami et al. (2012). Note that similar (only slightly worse) results were obtained if Bowtie were used in Strategy 2 and that the current BWA version cannot be used in Strategy 1 due to the limited mismatch output information for nonuniquely mapped reads. As shown in Figure 1B, the two methods led to a significant overlap in mapped reads, but with Strategy 2 retaining more reads, which is consistent with the mapping stringency in Strategy 1. Both methods generated a large number of RDDs, most of which were A-to-G changes (Fig. 1C). Further application of the seven artifact filters as described above reduced the number of RDDs and increased the proportion of A-to-G sites (Fig. 1C, D). Overall, the RDD results of Strategy 1 demonstrated much fewer artifacts and biases examined by the filters compared with the results of Strategy 2. Notably, using reads that were only mapped by Strategy 2, but not Strategy 1, led to an enhanced level of artifacts removed by the "high editing level" filter (Fig. 1D). The most potent filter for both strategies was the "low editing level" filter, which required at least three reads demonstrating the mismatch in question and an "editing"

level of at least 0.1 (note that we use the term "editing level" in a generic sense here, although RNA editing may not be responsible for some of the RDDs). This filter was first used based on the assumption that false-positive RDDs often have low numbers of "edited" reads (Ramaswami et al. 2012). Although genuine RDDs with low-editing levels should exist, the drastic increase in the % A-to-G events in the results of Strategy 2 (but not Strategy 1) indicates that many sites removed by this filter may be false positives. Based on the above comparison, the filters considered here had remarkably less impact on the results of Strategy 1. In contrast to the reliance on post-filtering by Strategy 2, this stringent mapping method alone is effective in reducing false positive results.

READ MAPPING STRINGENCY CAN BE ENHANCED BY PAIRED-END ANALYSIS

In designing RNA-Seq experiments, an often encountered question is whether to choose the single- or paired-end sequencing mode. Paired-end sequencing is often preferred in whole-genome sequencing or novel transcript isoform discovery, since the pairing information can improve genome or transcript assembly. However, the higher cost of paired-end sequencing often prohibits its usage in applications where only gene or transcript expression levels are sought after. For the purpose of identifying and quantifying SNVs in RNA, there has been no comparison of the impacts of single- and paired-end analyses. For simplicity, some studies (such as the above Strategy 2) may decide to analyze paired-end data in the single-end mode by ignoring the pairing information.

Here, we show that RDD analysis may benefit significantly from the added layer of stringency if read mapping were carried out to filter for reads that pair properly. It is important to note that pairing of reads does not necessarily reduce the number of mapped reads considerably (Fig. 1E). Paired-end mapping can rescue some reads that were mapped uniquely in pairs, but nonuniquely as singletons. In contrast, single-end mapping may retain reads in scenarios where only one read was mapped uniquely, but its partner was unmapped or nonuniquely mapped, or when both reads were mapped uniquely but did not pair properly (e.g., mapped to the same genomic strand). Such mapping results are likely wrong since the location of one read cannot be confirmed by a correct topology relative to its paired partner. Mapping Strategy 1 already capitalizes on the power of read pairing. Remarkably, incorporating a simple read-pairing procedure into Strategy 2 can considerably improve RDD predictions, as shown in Figure 1D. Note that due to the inadequate information provided by BWA for mismatches in nonuniquely mapped reads, read pairing here was conducted for uniquely mapped singleton reads only. Thus, reads resulted from Strategy 2 paired end represent a subset of those resulted from the original Strategy 2. These results strongly suggest that paired-end sequencing and data analysis should be adopted when RNA nucleotide variants are examined via RNA-Seq.

READ MAPPING STRINGENCY IMPROVES ACCURACY IN ESTIMATED RNA EDITING LEVELS

The effectiveness of RNA-Seq in RNA-editing studies should be evaluated from two perspectives: (1) accuracy of predicted RDD/editing sites, and (2) accuracy of predicted levels of RNA editing (i.e., quantitative ratios between RNA variants). Most studies (except Bahn et al. 2012) only focused on the first aspect. However, the second aspect is also essential, in that the extent of RNA editing readily exerts biological consequences. Importantly, alteration of editing levels during development or disease could have profound functional impacts (Paz et al. 2007; Gallo and Galardi 2008; Rula et al. 2008; Wahlstedt et al. 2009; Silberberg et al. 2012). In addition, the same issue exists in the analysis of allele-specific expression, where the allelic ratios of genetic variants in the expressed RNA need to be estimated accurately. Thus, estimating the quantitative ratios between RNA variants represents an important question in the application of RNA-Seq.

Evidently read mapping accuracy can directly influence the accuracy of estimated editing levels. Mapping bias favoring either of the alternative nucleotides in the RNA will lead to a quantification bias. Thus, stringent mapping without a systematic bias is necessary to ensure quantification accuracy. On the global scale, mapping bias can be evaluated by examining RNA-Seq allelic ratios of known heterozygous genomic SNVs in the sample (Fig. 2A). If no or little mapping bias exists, the overall allelic ratios of all expressed heterozygous SNVs should not be significantly different from a 1:1 ratio (assuming allele-specific expression only exists in a minority of SNVs). As shown in Figure 2A, a insignificant allelic bias was observed only for the mapping results based on

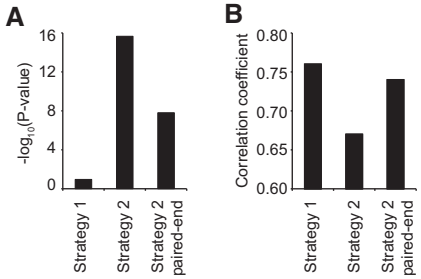

FIGURE 2. Evaluation of the bias and accuracy of the estimated editing levels via RNA-Seq data. (*A*) *P* values of a binomial test to determine whether the allelic ratio (number of reads containing the reference allele/total number of reads) deviated from 0.5 for known heterozygous SNVs in the genome of the sample under study (Peng et al. 2012). A lower *P* value (i.e., higher –log 10(P)) suggests a higher mapping bias favoring one of the alleles. (*B*) Pearson correlation coefficient between experimentally calculated editing levels (based on clonal sequencing) and those estimated from RNA-Seq data using different mapping strategies. RNA-Seq data and experimental validation results were obtained from a previous study (Bahn et al. 2012).

Strategy 1. The accuracy of editing levels can also be examined directly through traditional Sanger sequencing of RT-PCR clones (Li et al. 2009; Bahn et al. 2012). The results shown in Figure 2B confirm that stringent mapping approaches (Strategy 1 and Strategy 2-paired-end) resulted in better correlation with experimentally confirmed editing levels, with Strategy 1 providing the best correlation.

The accuracy of estimated RNA-editing levels closely depend on the accuracy of read mapping. Post-mapping artifact filters can only eliminate an assumed false-positive RDD site. Since such filters cannot remove or rescue individual mismapped reads, true-positive RDD sites are still potentially associated with read mapping bias and, consequently, compromised quantification accuracy. For genetic variants this problem is especially prominent, since analysis of allele-specific expression greatly depends on read counts for quantification. Thus, stringent mapping approaches are advantageous in quantitative analyses. Nevertheless, these approaches often identify a relatively smaller number of RDDs than those more lenient mapping strategies, although the difference may not be pronounced (Fig. 1C), especially when sequencing depth is high.

RECOMMENDATIONS FOR ANALYSIS OF RNA NUCLEOTIDE VARIANTS VIA RNA-Seq

Based on the above observations, we advocate for paired-end sequencing and enhanced stringency in the read-mapping step when analyzing nucleotide variants in RNA-Seq data, including RNA-editing and allele-specific expression analyses. The dual-filtering scheme on mismatches in the above mapping Strategy 1 is an effective way to reduce mapping errors resulting from the existence of highly similar genomic regions. It is also effective in eliminating potential mapping bias for reads carrying the alternative nucleotides whose inherent difference from the reference genome sequence may render a disadvantage. In addition, combined usage of multiple types of aligners with fundamental algorithmic differences appears to be advantageous, such as combining Blat with one or more of the often-used short read aligners (the choice of the specific short read aligner may not be very critical, although differences do exist). Such stringency is necessary to reduce false-positive RDDs given the current state-of-the-art of sequencing technologies. However, the false-negative rate may be increased due to this stringency, which remains to be explored.

For a specific RNA-Seq data set, the choice of read aligners, mapping parameters, and mismatch filtering parameters need to be evaluated individually. We recommend two basic approaches to this end. First, simulated reads should be generated with the same read length, similar insert size distribution (for paired-end reads), similar read distribution across the genome, similar base quality scores, and sequencing error profiles as the actual data set (Bahn et al. 2012). The simulated reads should include alternative alleles of known SNPs (e.g., from dbSNP) with equal probability. Subsequently, different variables related to mapping (aligners, parameters, and mismatch filters) can be evaluated, and the best approach can be chosen to reach unbiased allelic ratios of the simulated SNPs. In the second approach, the RNA allelic ratios of known SNPs are examined after the actual RNA-Seq data has been mapped using the chosen method. Insignificant bias from an expected 1:1 allelic ratio on average should be reached with the optimal mapping method. The above procedures are necessary to ensure relatively accurate read mapping and quantification of RNA allelic ratios in RNA-Seq data.

DESIGN OF RNA-Seq EXPERIMENTS FOR RNA EDITING STUDIES

In addition to the paired-end or single-end sequencing mode, there are a number of other variables to be considered in the design of RNA-Seq experiments (Table 1). Sequencing depth is one of the most critical considerations. The number of predicted editing sites depends heavily on sequencing depth (Fig. 3), which is expected since deeper sequencing enables higher genome-wide read coverage. In addition, accuracy of the estimated editing levels was also shown to be higher for sites with higher read coverage (Bahn et al. 2012). Interestingly, the percentage of A-to-G events among all identified RDDs also increases with sequencing depth until it reaches a plateau at about 90 million (i.e., 45 million pairs) total mapped reads for the human lymphoblastoid cells (Fig. 3). This observation indicates again that the absolute % A-to-G value alone should not be used as a measure to evaluate the accuracy of an analysis method without taking into account sequencing depth (and possibly other variables discussed below).

TABLE 1. Recommended variables for consideration in the design of RNA-Seq experiments for identifying RNA-editing events

Variables	Rationale and consideration
Sequencing depth	Number of RDDs and % of A-to-G events increase with sequencing depth; accuracy of estimated editing levels increases with read coverage of putative RDDs.
Biological replicates	Recommended in order to ensure high total coverage of candidate RDD sites after removal of duplicate reads.
Paired or single-end sequencing	Paired-end sequencing and read pairing during data analysis can significantly improve RDD accuracy.
Quality of sequencing library	High fidelity enzymes for RT and PCR should be adopted. Rate of duplicate reads should be evaluated and minimized. Base quality of reads should be inspected and optimized by sequencing chemistry.
Type of sequencing library	Strand-specific libraries are advantageous for pinpointing specific types of RDDs.

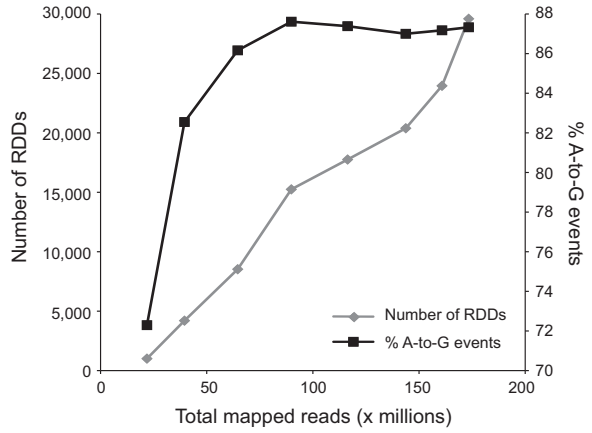

FIGURE 3. Number of RDDs and % of A-to-G events identified in RNA-Seq closely depend on the sequencing depth. Paired-end poly (A^+) RNA-Seq data obtained previously (Peng et al. 2012) were down-sampled to varying depths as shown. RDDs were identified using Strategy 1 as illustrated in Figure 1A (without applying artifact filters). The *x*-axis corresponds to the total number of mapped reads counted in singletons.

Another question in experimental design is whether to achieve the desirable sequencing depth using one sample or multiple biological replicates. One problem in current RNA-Seq data is the existence of duplicate reads, i.e., reads that map to an identical genomic locus. Such reads often arise from reverse transcription, PCR, or RNA fragmentation bias during library preparation (Pepke et al. 2009; Roberts et al. 2011; Sendler et al. 2011). It is crucial in RNA-editing studies that only one copy of such reads is retained to ensure accuracy in the quantification of editing levels (as applied for both Strategies 1 and 2) (Fig. 1A). The removal of duplicate reads should be carried out for data from each biological replicate for which an independent sequencing library was generated. In this case, very deep sequencing of the same library is not advantageous, as the maximum read coverage of a locus after removal of duplicate reads is equal to the length of the read in the single-end sequencing mode. If multiple biological replicates are obtained, a common practice is to remove duplicate reads within each replicate and, subsequently, combine all replicates for RNA-editing analysis to maximize read coverage and statistical power (on the condition that the biological replicates were confirmed to be largely consistent). Thus, it is highly desirable to obtain biological replicates for which library preparation and sequencing procedures are independent. Nevertheless, this concern may be irrelevant if statistical models are available to distinguish artifactual duplicates from authentic ones resulting from high gene expression and deep coverage of one library.

The quality of the RNA-Seq library and sequencing run are critical to the accuracy of predicted editing events. Sequence mutations introduced during library generation, e.g., in the step of reverse transcription or PCR, may be predicted as bona fide RDDs. To this end, use of high-fidelity enzymes that are less error prone is desirable. In addition, careful examination of sequencing error occurrence and analysis via statistical models (e.g., Bahn et al. 2012) may help to reduce false positives in the predicted RDDs. Beside library quality, the type of library also affects the scope of identifiable editing events. For example, most standard RNA-Seq libraries do not preserve the strand information of the original RNA. In this case, complementary types of RDDs (e.g., A-to-G vs. T-to-C) cannot be distinguished unless the RDD site is located within a known gene, and reads covering the RDD are assumed to originate from this gene. Thus, RDDs in unannotated regions or in regions with bidirectional transcription (sense and antisense genes) cannot be accurately categorized into a specific type. To this end, strand-specific libraries (e.g., Parkhomchuk et al. 2009; Levin et al. 2010; Vivancos et al. 2010; Zhong et al. 2011) are advantageous, since the type of RDDs can be inferred directly from the reads independent of gene annotations. Indeed, our analysis of data from a strand-specific library derived from U87MG cells (data not shown) suggested that ~7% of all identified RDDs are antisense to annotated genes and 40% reside in intergenic regions based on RefSeq annotation (as of June 29, 2012).

The number of reported RDDs (and % A-to-G events) differs greatly among the recent RNA-Seq-based studies in various human samples (Li et al. 2011; Bahn et al. 2012; Chen et al. 2012; Gu et al. 2012; Peng et al. 2012; Ramaswami et al. 2012). This apparent discrepancy has puzzled many, since the correct editome profile seems to be unclear. We stress here that many factors beside bioinformatic methods, as summarized in Table 1, may affect the final profile of the identified editome. Importantly, RNA-editing profiles may differ depending on the species (Eisenberg et al. 2005), tissue type (Song et al. 2004), disease (Maas et al. 2006), or environmental conditions (Garrett and Rosenthal 2012) under study, all of which are interesting aspects that remain to be fully explored. As a result, these variables should always be considered in the design and interpretation of RDD analysis. In addition, performance of bioinformatic pipelines for RNA-editing analysis can only be compared using identical RNA-Seq data sets.

DO UNKNOWN TYPES OF RDDs REALLY EXIST?

To date, most recent studies focused on the false-positive issue in RDD identification motivated by the striking list of reported RDDs in a previous report (Li et al. 2011). A consensus is arising that A-to-I and C-to-U editing should account for most of the observed RDD events and a large number of the other types of RDDs (noncanonical RDDs) are likely false positives (Schrider et al. 2011; Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012). However, it remains to be determined whether all noncanonical RDDs are false positives, or whether a small fraction of them are genuine events resulting from unknown mecha-

nisms. A small number of studies provided experimental confirmation of some of the noncanonical events (Bahn et al. 2012; Chen et al. 2012; Peng et al. 2012). It is possible that paralogous regions may also complicate experimental validation (Piskol et al. 2013). However, even under extreme scrutiny, some of the validated events cannot be attributed to the existence of paralogous regions. Thus far, only a small number of noncanonical RDDs were experimentally tested. This issue needs to be examined using effective experimental approaches on a much larger scale.

CONCLUSIONS

Due to a recent debate surrounding the discovery of RDDs in the literature, many are questioning the usefulness of RNA-Seq in RNA editing studies. Nevertheless, RNA-Seq has great potential to provide an unbiased profile of the editome of any species. Accurate bioinformatic methods and careful experimental design are key to the success of this application. With improved methods and sequencing technologies, RNA-Seq will continue to enable exciting discoveries in the field of RNA editing.

ACKNOWLEDGMENTS

The work was supported in part by NIH grant R01HG006264, Alfred P. Sloan Foundation Research Fellowship and Research Grant No. 5-FY10-486 from the March of Dimes Foundation to X.X., and an American Heart Association Postdoctoral Fellowship to J.H.L.

REFERENCES

Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22:** 142–150.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148:** 1293–1307.

Danecek P, Nellaker C, McIntyre RE, Buendia-Buendia JE, Bumpstead S, Ponting CP, Flint J, Durbin R, Keane TM, Adams DJ. 2012. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol* **13:** r26.

Dillman AA, Hauser DN, Gibbs JR, Nalls MA, McCoy MK, Rudenko IN, Galter D, Cookson MR. 2013. mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nat Neurosci* **16:** 499–506.

Eisenberg E, Nemzer S, Kinar Y, Sorek R, Rechavi G, Levanon EY. 2005. Is abundant A-to-I RNA editing primate-specific? *Trends Genet* **21:** 77–81.

Gallo A, Galardi S. 2008. A-to-I RNA editing and cancer: From pathology to basic science. *RNA Biol* **5:** 135–139.

Garrett S, Rosenthal JJ. 2012. RNA editing underlies temperature adaptation in K^+ channels from polar octopuses. *Science* **335:** 848–851.

Gu T, Buaas FW, Simons AK, Ackert-Bicknell CL, Braun RE, Hibbs MA. 2012. Canonical A-to-I and C-to-U RNA editing is enriched at 3′UTRs and microRNA target sites in multiple mouse tissues. *PLoS One* **7:** e33720.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12:** 656–664.

Kleinman CL, Majewski J. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335:** 1302; author reply 1302.

Kleinman CL, Adoue V, Majewski J. 2012. RNA editing of protein sequences: A rare event in human transcriptomes. *RNA* **18:** 1586–1596.

Lagarrigue S, Hormozdiari F, Martin LJ, Lecerf F, Hasin Y, Rau C, Hagopian R, Xiao Y, Yan J, Drake TA, et al. 2013. Limited RNA editing in exons of mouse liver and adipose tissue. *Genetics* doi: 10.1534/genetics.112.149054.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7:** 709–715.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210–1213.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333:** 53–58.

Lin W, Piskol R, Tan MH, Li JB. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335:** 1302; author reply 1302.

Maas S, Kawahara Y, Tamburro KM, Nishikura K. 2006. A-to-I RNA editing and human disease. *RNA Biol* **3:** 1–9.

Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12:** 451.

Park E, Williams B, Wold BJ, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res* **22:** 1626–1633.

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123.

Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, Barbash ZS, Adamsky K, Safran M, Hirschberg A, et al. 2007. Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res* **17:** 1586–1595.

Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, et al. 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30:** 253–260.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6:** S22–S32.

Pickrell JK, Gilad Y, Pritchard JK. 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335:** 1302; author reply 1302.

Piskol R, Peng Z, Wang J, Li JB. 2013. Lack of evidence for existence of noncanonical RNA editing. *Nat Biotechnol* **31:** 19–20.

Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. 2012. Accurate identification of human *Alu* and non-*Alu* RNA editing sites. *Nat Methods* **9:** 579–581.

Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. 2013. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods* **10:** 128–132.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12:** R22.

Rula EY, Lagrange AH, Jacobs MM, Hu N, Macdonald RL, Emeson RB. 2008. Developmental modulation of $GABA_A$ receptor function by RNA editing. *J Neurosci* **28:** 6196–6201.

Schrider DR, Gout JF, Hahn MW. 2011. Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One* **6:** e25842.

Sendler E, Johnson GD, Krawetz SA. 2011. Local and global factors affecting RNA sequencing analysis. *Anal Biochem* **419:** 317–322.

Silberberg G, Lundin D, Navon R, Ohman M. 2012. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Hum Mol Genet* **21:** 311–321.

Song W, Liu Z, Tan J, Nomura Y, Dong K. 2004. RNA editing generates tissue-specific sodium channels with distinct gating properties. *J Biol Chem* **279:** 32554–32561.

Vivancos AP, Guell M, Dohm JC, Serrano L, Himmelbauer H. 2010. Strand-specific deep sequencing of the transcriptome. *Genome Res* **20:** 989–999.

Wahlstedt H, Daniel C, Enstero M, Ohman M. 2009. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res* **19:** 978–986.

Zhong S, Joung JG, Zheng Y, Chen YR, Liu B, Shao Y, Xiang JZ, Fei Z, Giovannoni JJ. 2011. High-throughput Illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb Protoc* **2011:** 940–949.

Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants

Jae-Hyung Lee, Jason K. Ang and Xinshu Xiao

RNA 2013 19: 725-732 originally published online April 18, 2013
Access the most recent version at doi:10.1261/rna.037903.112