High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing

RACHEL M. SIMPSON,[1] ANDREW E. BRUNO,[2] JONATHAN E. BARD,[3] MICHAEL J. BUCK,[4] and LAURIE K. READ[1]

[1]Department of Microbiology and Immunology, University at Buffalo, Jacobs School of Medicine and Biomedical Sciences, Buffalo, New York 14214, USA
[2]Center for Computational Research, University at Buffalo, Buffalo, New York 14203, USA
[3]University at Buffalo Genomics and Bioinformatics Core, Buffalo, New York 14222, USA
[4]Deparment of Biochemistry, University at Buffalo, Jacobs School of Medicine and Biomedical Sciences, Buffalo, New York 14214, USA

ABSTRACT

Uridine insertion/deletion RNA editing in kinetoplastids entails the addition and deletion of uridine residues throughout the length of mitochondrial transcripts to generate translatable mRNAs. This complex process requires the coordinated use of several multiprotein complexes as well as the sequential use of noncoding template RNAs called guide RNAs. The majority of steady-state mitochondrial mRNAs are partially edited and often contain regions of mis-editing, termed junctions, whose role is unclear. Here, we report a novel method for sequencing entire populations of pre-edited partially edited, and fully edited RNAs and analyzing editing characteristics across populations using a new bioinformatics tool, the Trypanosome RNA Editing Alignment Tool (TREAT). Using TREAT, we examined populations of two transcripts, RPS12 and ND7-5′, in wild-type *Trypanosoma brucei*. We provide evidence that the majority of partially edited sequences contain junctions, that intrinsic pause sites arise during the progression of editing, and that the mechanisms that mediate pausing in the generation of canonical fully edited sequences are distinct from those that mediate the ends of junction regions. Furthermore, we identify alternatively edited sequences that constitute plausible alternative open reading frames and identify substantial variability in the 5′ UTRs of both canonical and alternatively edited sequences. This work is the first to use high-throughput sequencing to examine full-length sequences of whole populations of partially edited transcripts. Our method is highly applicable to current questions in the RNA editing field, including defining mechanisms of action for editing factors and identifying potential alternatively edited sequences.

Keywords: trypanosome; kinetoplastid; RNA editing; mitochondria; bioinformatics

INTRODUCTION

Flagellated protozoa of the order Kinetoplastida are early branching eukaryotes, several members of which cause devastating human diseases (Stuart et al. 2008; Bilbe 2015). These organisms share unique biology, including the essential process termed mitochondrial uridine (U) insertion/deletion RNA editing (Aphasizhev and Aphasizheva 2011b; Hashimi et al. 2013; Read et al. 2016). Kinetoplastids are characterized by their unusual mitochondrial DNA structure, the kinetoplast, or kDNA. In *Trypanosoma brucei*, the kDNA consists of dozens of nearly identical ~22-kb maxicircles and thousands of heterogenous ~1 kb minicircles interlocked into a unique structure (Jensen and Englund 2012). Twelve of the 18 protein coding genes encoded in the maxicircles are referred to as cryptogenes because they do not encode functional open reading frames. Prior to translation, these mRNAs must be altered by the specific addition and less frequent deletion of U's by RNA editing (Aphasizhev and Aphasizheva 2011b; Hashimi et al. 2013; Read et al. 2016). The precise insertion and deletion of U residues is guided by small, noncoding guide RNAs (gRNAs), which act as templates. gRNAs are encoded almost exclusively in the minicircle component of kDNA. The enzymes that catalyze U insertion/deletion during kinetoplastid RNA editing are part of a multiprotein complex termed the RNA editing core complex (RECC) or 20S editosome (Rusche et al. 1997, 2001; Cruz-Reyes et al. 2002; Aphasizhev et al. 2003;

Corresponding author: lread@buffalo.edu
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.055160.115.

Ernst et al. 2003; Panigrahi et al. 2003, 2006; Stuart et al. 2005; Trotter et al. 2005; Carnes et al. 2008, 2011; Read et al. 2016). Distinct RECC variants catalyze U insertion and U deletion, although it is not known whether these are stable complexes, or whether proteins specific for insertion or deletion shuttle on and off a common base of proteins (Carnes et al. 2008, 2011). Though RECC is necessary for the editing process, it is not sufficient. Numerous recent studies have demonstrated that a large, possibly dynamic and heterogeneous complex termed MRB1 (mitochondrial RNA binding complex 1) or RESC (RNA editing substrate binding complex) is also essential for RNA editing in kinetoplastids (Fisk et al. 2008; Hashimi et al. 2008, 2009, 2013; Weng et al. 2008; Acestor et al. 2009; Ammerman et al. 2011, 2012, 2013; Kafkova et al. 2012; Aphasizheva et al. 2014; Huang et al. 2015; Madina et al. 2015; Read et al. 2016). Indeed, the MRB1 complex likely acts as the platform for RNA editing, since it contains readily detectable mRNA and gRNA, while purified RECC lacks associated RNA (Weng et al. 2008; Aphasizheva et al. 2014; Madina et al. 2014, 2015). MRB1 appears to be composed of interacting subcomplexes with distinct roles in the editing process, and it transiently associates with other proteins that impact mitochondrial RNA editing and processing (Hashimi et al. 2008, 2009, 2013; Weng et al. 2008; Ammerman et al. 2011, 2012; Kafkova et al. 2012; Aphasizheva et al. 2014; Madina et al. 2014, 2015; Read et al. 2016).

Of the maxicircle-encoded mRNAs that require editing, three are edited only in short regions and are thus termed minimally edited. However, nine mRNAs are edited throughout their lengths, with editing often nearly doubling their sizes, and these mRNAs are termed pan-edited. Pan-editing of a given mRNA requires the sequential use of dozens of gRNAs (Koslowsky et al. 1991, 2014; Maslov and Simpson 1992). In this process, the initiating gRNA anchors to a short never-edited region at the 3′ end of the transcript. Early in the process, this gRNA will contain intramolecular hairpins, and it does not fully align to the pre-edited mRNA (Koslowsky et al. 1991, 2004; Schmid et al. 1995; Leung and Koslowsky 1999, 2001a,b; Reifur and Koslowsky 2008; Reifur et al. 2010). Once the editing guided by a gRNA is complete, that gRNA is complementary to the fully edited mRNA through a combination of Watson–Crick and G-U base-pairing. That gRNA then needs to be at least partially removed because the subsequent gRNA forms an anchor duplex with a portion of the edited region guided by the first gRNA (Maslov and Simpson 1992). The mechanism of gRNA exchange is poorly understood, but may involve mitochondrial helicases (Hernandez et al. 2010; Li et al. 2011; Madina et al. 2014, 2015). This gRNA utilization process proceeds throughout the length of a given mRNA, resulting in the general, although not precise, progression of editing from the 3′ to the 5′ end of the transcript (Sturm and Simpson 1990; Koslowsky et al. 1991; Souza et al. 1992). The sequential utilization of gRNAs results in an exceedingly heterogenous and complex steady-state mitochondrial RNA population comprised of mRNAs that are edited to different extents at their 3′ ends and pre-edited at their 5′ ends (Sturm and Simpson 1990; Koslowsky et al. 1991, 2014; Ammerman et al. 2010). These mRNAs are typically referred to as partially edited.

Further complicating the picture is the existence of regions within partially edited sequences of varying lengths that arise between the 3′ fully edited and 5′ pre-edited regions and which contain edited sequence matching neither the fully edited nor the pre-edited mRNA sequence. These regions of "mis-edited" sequence, which are present in the majority of mRNAs undergoing editing, are termed junction regions (Koslowsky et al. 1991). Three hypotheses as to the role of junction sequences have arisen in the field. First, junctions are hypothesized to be regions of active ongoing editing, suggesting that consistent errors and remodification occur in generating the final functional mRNA sequence (Koslowsky et al. 1991). Second, junctions may also represent regions of mis-editing that are not remodeled to final edited sequence, thereby defining dead-end products (Sturm and Simpson 1990). Finally, a subset of junctions may reflect alternative editing that could lead to the production of alternative proteins (Ochsenreiter and Hajduk 2006; Ochsenreiter et al. 2008a,b). While the latter would not then be mis-editing, for simplicity we use the term "mis-edited" here to refer to all regions of edited sequence that differ from pre-edited or canonical fully edited sequence. The origin of junctions is not known; however, they have been proposed to arise from misalignment of cognate gRNAs or utilization of noncognate or alternative gRNAs. Supporting the essential role of junctions is the fact that depletion of the MRB1 complex protein, TbRGG2, leads to a decrease in the length and prevalence of junctions that coincides with massive RNA editing and growth defects (Ammerman et al. 2010).

While catalysis of U insertion/deletion at a single editing site (ES; see Table 1 for terminology) is relatively well understood, there remain numerous unanswered questions with respect to the 3′ to 5′ progression of editing along a given transcript. For example, are there specific barriers to editing progression? If so, do these barriers arise due to factors such as bottlenecks in gRNA exchange, exchange between insertion and deletion RECCs, or other sequence characteristics? Do these barriers serve as regulatory points? To date, the study of the editing process has been limited by several factors. Primary among these is that existing in vitro editing assays catalyze editing only at a single ES (Seiwert and Stuart 1994) (with one exception of dual site editing [Alatortsev et al. 2008]), and therefore do not allow analysis of site-to-site progression within a gRNA-defined block or the process of gRNA exchange. Second, in the past, minimal tools were available for the large-scale sequencing of partially edited transcripts. Advancement in deep sequencing now allows for high-throughput RNA-seq analysis of the mitochondrial transcriptome. Large-scale analysis of mitochondrial mRNAs in wild-type cells will provide a comprehensive picture of

TABLE 1. Glossary of terms

Term	Definition
Editing site (ES)	Any space between two non-T nucleotides (cDNA) has the potential to be edited at the RNA level and is termed an editing site (ES). ESs are numbered from 3′ to 5′ following the direction of editing.
Editing stop site	Moving 3′ to 5′, the editing stop site is the final (5′ most) ES that matches the canonical fully edited sequence correctly. All ESs 3′ of the editing stop site match the canonical fully edited sequence.
Junction start site	The first ES, moving 3′ to 5′, which does not match the canonical fully edited sequence correctly (can match pre-edited or mis-edited).
Junction end site	The 5′ most ES with any editing action, whether canonical or mis-edited.
Intrinsic pause site (IPS)	An editing stop site at which the total number of sequences sharing this editing stop site is greater than the outlier threshold. Intrinsic pause sites (IPSs) represent ESs at which canonical editing frequently pauses.
Major junction end site (MJES)	An ES that comprises the junction end site in a large number of sequences, greater than the outlier threshold for junction end sites. Major junction end sites (MJESs), thus, represent ESs where all editing action frequently stops.

partially edited mRNA populations that may provide insight into the 3′ to 5′ progression of editing, the role of junction sequences, and the presence of alternatively edited mRNAs that could encode novel proteins. Examination of mRNA populations in transgenic trypanosomes depleted of key editing factors will reveal specific defects in mRNAs that arise upon loss of these factors, providing invaluable insight into the functions of these proteins in the editing process. To address gaps in our understanding of the editing process, we developed the Trypanosome RNA Editing Alignment Tool (TREAT), an open source program that allows analysis of large populations of partially edited transcripts. Here, we describe the capabilities of this tool and provide an overview of its basic usage. TREAT is a multiple sequence alignment-based program that generates a searchable database from large sequencing data sets and makes resulting data accessible through a user-friendly web application. TREAT permits the analysis of large populations of full-length pre-edited, partially edited, and fully edited transcripts. Currently, paired-end Illumina MiSeq analysis can sequence reads of ~550 nucleotides (nt), thus permitting examination of complete populations of shorter edited RNAs such as RPS12, whose fully edited length is 325 nt (Read et al. 1992). This technology will permit direct analysis of all three minimally edited transcripts and six of the pan-edited RNAs, including the 5′ region of ND7. Longer RNAs can be examined in blocks using multiple primer sets. To compare partially edited sequences, TREAT aligns all non-T bases in the cDNA and then replaces the T's, cataloging each potential editing site as pre-edited, fully edited, or mis-edited. Editing characteristics are then determined for each sequence, including the full editing stop site, junction start site, and junction end site. The data output is searchable by multiple parameters that can be viewed in the form of a searchable table of individual sequences and in constrainable histograms, allowing us to compare multiple parameters across populations. For the first time, TREAT allows us to analyze large data sets of the full-length sequences derived from partially edited mRNAs and to thereby identify intrinsic pause sites (IPSs) in editing (Ammerman et al. 2010), examine junction sequences, establish correlations, and group the data in a meaningful way. Here, we present the method of our analysis and a case study performed using TREAT to define editing characteristics of two pan-edited transcripts (RPS12 and ND7-5′) in wild-type procyclic form *T. brucei*.

RESULTS

Algorithm of TREAT and major definitions

The Trypanosome RNA Editing Alignment Tool (TREAT) is a special purpose multiple sequence aligner designed to permit the user to analyze variation in sequences caused by U insertion/deletion RNA editing. In U insertion/deletion RNA editing, U's are added and deleted throughout the sequence, and any space between two non-U nucleotides has the potential to be modified through the addition or deletion of U's. Thus, we define each space between two non-U nucleotides as an editing site (ES) and number them from 3′ to 5′ following the direction of editing (Fig. 1A). This includes both sites that require editing to achieve the fully edited sequence as well as those that would not need to be changed to generate fully edited sequence, but which have the potential to be altered, as seen in junction sequences. TREAT aligns sequences using three bases (A,C,G) and assembles ESs to detect the extent of U (T in cDNA) editing. TREAT requires two user provided template sequences: fully edited and pre-edited. The fully edited template represents a mature edited mRNA (complete canonically edited mRNA). In this study, we use the canonical, fully edited mRNA sequences reported in a series of studies performed in the late 1980's and early 1990's (Benne et al. 1986; Feagin et al. 1988; Bhat et al. 1990; Koslowsky et al. 1990; Read et al. 1992, 1994; Souza et al. 1992, 1993; Corell et al. 1994; http://dna.kdna.ucla.edu/trypanosome/files/tbmaxi.html). The pre-edited template represents the corresponding genomically encoded sequence, which will become edited to the mature mRNA. Optionally, TREAT accepts one or more alternatively edited templates. All template sequences must be identical with respect to the non-edited bases (A,C,G); i.e., they must be the

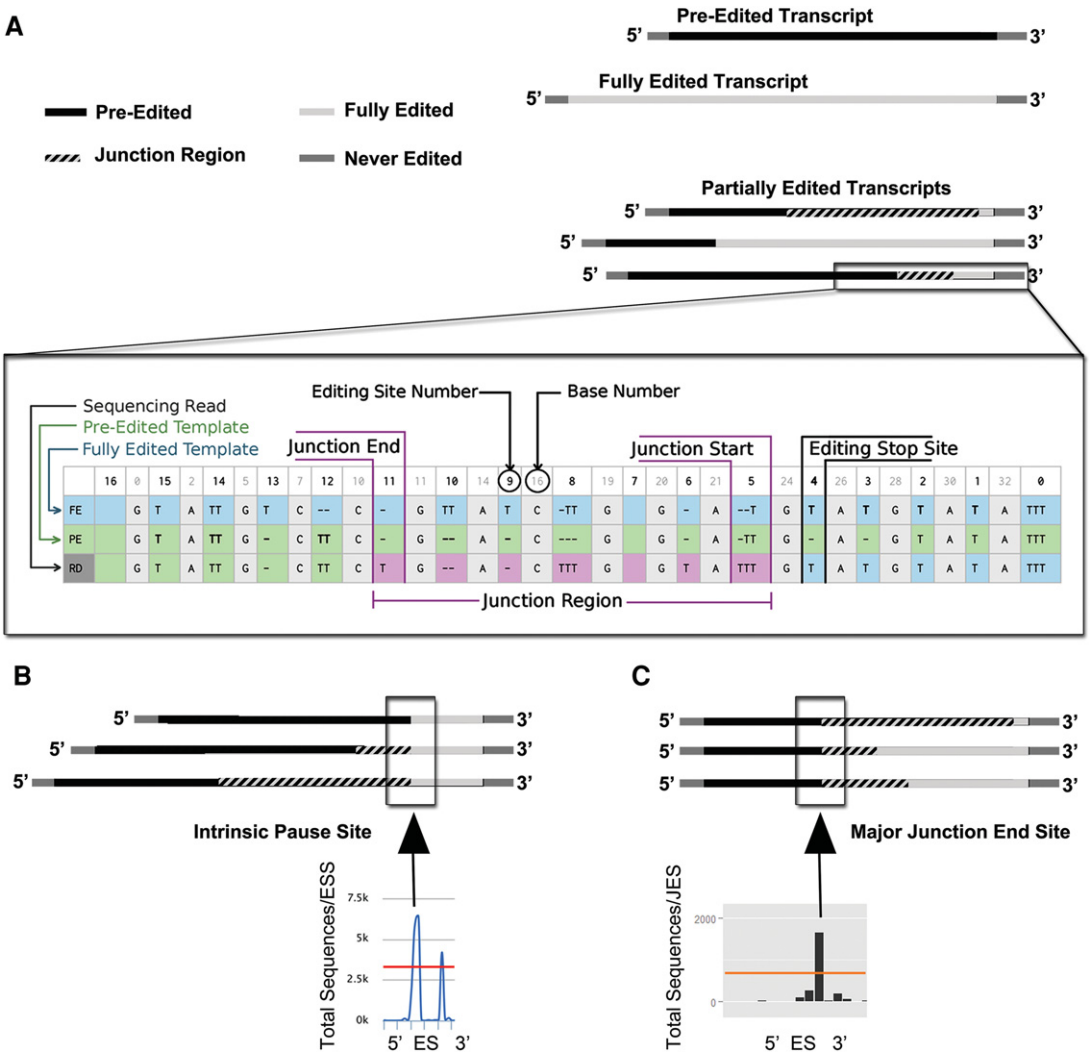

FIGURE 1. Major definitions used in TREAT and subsequent analyses. (*A*) Schematics show pre-edited and fully edited transcripts flanked by small never-edited regions (*upper right*) and a series of partially edited transcripts (*middle right*) demonstrating a long junction (*top*), no junction (*middle*), and short junction (*bottom*). The short junction (not to scale) is expanded, showing the TREAT program alignment of the transcript beginning with ES0. Editing stop site, junction start site, junction region, junction end, all templates, and the editing site and base pair notation systems are all labeled. In this example read, ES0, ES1, and ES2 are never edited, ES3 and ES4 are the fully edited sites, the junction start site is ES5, and the junction end site is ES11. (*B*) Schematics show three transcripts with partially edited sequences that share a common editing stop site (ESS). When the total number of sequences is above the outlier threshold (red line on *lower* graph), this is considered an intrinsic pause site (IPS). (*C*) Schematic shows three partially edited sequences with a common junction end site (JES). When the number of sequences is above the outlier threshold for junction end sites (orange line on graph), this is termed a major junction end site (MJES).

exact same sequence after removing the edited base (T). TREAT preprocesses the template sequences by numbering the editing sites in the 3′ to 5′ direction and recording the number of T's found at each ES.

TREAT processes a single sequence as follows. To determine the extent of canonical editing, TREAT begins at the 3′ end of the sequence and determines which editing sites match the canonical fully edited sequence moving from 3′ to 5′. If an ES does not match the canonical fully edited template, TREAT determines whether the site matches the pre-edited template, an alternative template given by the user, or is noncanonically edited (matching neither the fully edited, pre-edited, nor alternative templates). Based on this, the final ES that matches the canonical fully edited sequence is termed the editing stop site and marks the 5′ boundary of fully edited sequence (Fig. 1A). The junction region for each sequence is defined as beginning at the first ES that does not match fully edited sequence and extending to the 5′ boundary of all editing action in the sequence. The ES that is the start of the junction is termed the junction start site. TREAT then continues in the 3′ to 5′ direction, defining whether each ES matches pre-edited, fully edited, alternatively edited, or mis-edited sequence. Having reached the 5′ end of a sequence, TREAT then analyzes the same sequence in the 5′

to 3′ direction and identifies the first site that does not match the pre-edited sequence. This ES is termed the junction end site and represents the 5′ boundary of all editing action in a given sequence. The junction length is the number of ESs contained within the region between, and inclusive of, the junction start site and junction end site (Fig. 1A). These basic measures are determined for each individual sequence and used for downstream analysis of an entire population of sequences. The web-based user interface of TREAT displays various constrainable histograms to allow easy examination of characteristics across data sets (Supplemental Fig. S1).

When examining a population of partially edited transcripts, we compile the individual sequence characteristics to examine the extent of editing action, moving 3′ to 5′, in an entire population (Fig. 1B,C). We define two terms to accomplish this. First, the intrinsic pause site (IPS), as introduced in Ammerman et al. (2010), is a measure to define ESs at which fully edited sequence stops in a large portion of a population of partially edited transcripts (Fig. 1B). Statistically, we define an IPS as an editing stop site at which the total number of sequences sharing this editing stop site is greater than the outlier threshold (see Materials and Methods). Second, we define major junction end sites (MJESs) as ESs that comprise the junction end in a large number of sequences; these MJESs thus represent ESs where all editing action frequently stops (Fig. 1C). MJESs are defined using the same outlier threshold, but examining the number of sequences containing the same junction end sites.

Identification of IPSs in RPS12 and ND7-5′ mRNAs

Fully edited RPS12 mRNA is 325 nt (Read et al. 1992), and thus the sequences of entire RPS12 transcripts can be obtained in one fragment using Illumina MiSeq 300 cycle paired-end analysis. ND7 mRNA is edited in two domains. The 5′ domain can be edited prior to completion of editing of the 3′ domain (Koslowsky et al. 1990). The edited sequence of the ND7-5′ domain is 234 nt, including the entire 3′ never-edited region, and editing of this entire domain can also be examined within a single fragment. Analysis of RPS12 and ND7-5′ RNAs is also advantageous since limited sequencing of these transcripts has been previously reported (Koslowsky et al. 1991; Ochsenreiter et al. 2008b; Ammerman et al. 2010; Guo et al. 2010; Madina et al. 2014), thereby allowing a comparison of our method to previous studies. Thus, we engaged in a case study of TREAT utilization by examining these two mRNAs in procyclic form cells. To begin, we amplified cDNA populations corresponding to all pre-edited, fully edited, and partially edited mRNAs for each transcript using primers specific to their 5′ and 3′ never-edited regions and containing bar codes to facilitate downstream analysis. Importantly, these PCR reactions were performed within their linear ranges, as determined using qRT-PCR analysis with the same cDNAs and primer sets. High-throughput, paired-end analysis of the resulting cDNA populations was then performed. Using this method, we obtained sequences from three biological replicates each for RPS12 and ND7-5′. The number of sequences obtained was comparable in two of the three replicates, and greater in the remaining replicate. After removal of sequences with non-T mutations, we obtained 251,006 (replicate 1), 55,894 (replicate 2), and 69,944 (replicate 3) normalized reads for RPS12 and 798,405 (replicate 1), 370,926 (replicate 2), and 413,533 (replicate 3) reads for ND7-5′. In this article, we examine major parameters across all three replicates and, given the consistency of these analyses, use the replicate with the greatest coverage, designated replicate 1, as a representative data set for specific examples as noted in the text. In RPS12 replicate 1, 14% (35,451 reads) matched pre-edited, 0.007% (19 reads) matched fully edited, and 86% (218,536 reads) were partially edited. In ND7-5′ replicate 1, 31% (249,326 reads) were pre-edited, no canonical fully edited sequences were isolated, and 69% of sequences (549,079 reads) were partially edited. Given we found so few fully edited sequences, and that gRNAs directing editing in the 5′ UTRs of RPS12 and ND7-5′ were either not found or were in low abundance (Koslowsky et al. 2014), we searched for sequences that encoded the canonical ORFs but allowed for variation in the 5′ UTRs, reasoning that these could represent translatable mRNAs. This search returned 14,385 reads in RPS12 and 46,488 reads in ND7-5′, suggesting that 5′ UTRs can vary in this strain (29–13) compared with the strain used to determine canonical fully edited sequences (EATRO 164). In ND7-5′, a large proportion of these sequences have a single nucleotide difference from the canonical 5′ UTR, while in RPS12 a greater variety of unique 5′ UTRs was observed. Overall, the number of reads obtained through this method marks a substantial improvement in the coverage of partially edited sequences from the previously published efforts that relied on conventional sequencing methods (Koslowsky et al. 1991; Ochsenreiter et al. 2008b; Ammerman et al. 2010; Guo et al. 2010; Madina et al. 2014).

We previously reported evidence, based on limited sequence analysis of RPS12 cDNAs, for the existence of IPSs within partially edited transcripts (Ammerman et al. 2010). In those studies, we observed accumulation of transcripts in the steady-state mitochondrial RNA population having common editing stop sites, thereby suggesting that full, canonical editing, moving in the general 3′ to 5′ direction, tends to stall at these sites. To determine the presence of IPSs in our current data sets, we began by quantifying the abundance of transcripts with editing stop sites at each ES for RPS12 (Fig. 2A) and ND7-5′ (Fig. 2B). If there were no IPSs, we would expect to see approximately equal numbers of sequences with editing stop sites at each ES and thus a normal distribution of pausing. Figure 2 demonstrates that this is not the case for either transcript. Rather, we observe large numbers of sequences whose editing stop sites occur at a relatively small number of ESs for both RPS12 and ND7-5′, and these data were consistent across all three replicates. To define IPSs

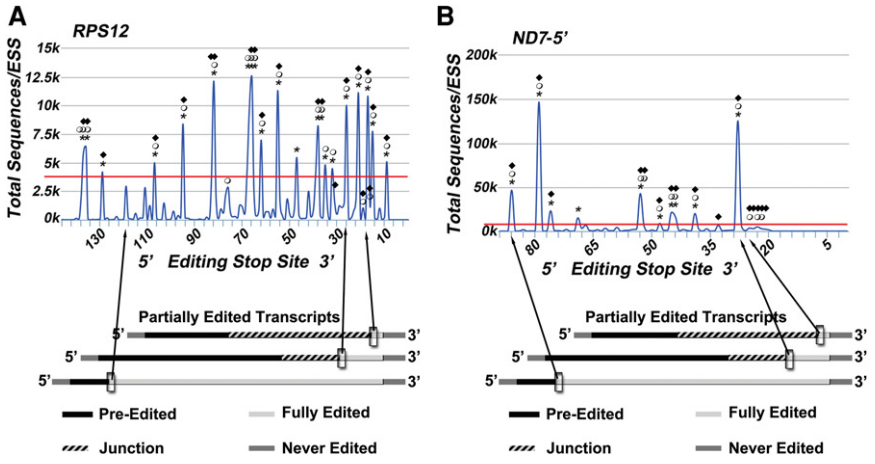

FIGURE 2. Amplitude of pausing across RPS12 and ND7-5′ populations of partially edited transcripts. Graphs show the total number of sequences (normalized) per editing stop site across the RPS12 (*A*) and ND7-5′ (*B*) transcripts in replicate 1. The red lines denote the outlier thresholds for each transcript, and asterisks *above* peaks denote IPSs present in replicate 1. The open circle *above* editing sites denotes sites that are IPSs in replicate 2, and black diamonds represent IPSs in replicate 3. Multiple asterisks, diamonds, and circles in a row denote adjacent editing sites that are all above the outlier threshold and appear as a wider peak in the line graph.

quantitatively, we determined the mean and median numbers of sequences whose editing stop sites occur at each ES, and found that the distribution is not normal, suggesting the presence of outliers in the data (Materials and Methods, Supplemental Fig. S2; Crawley 2011). Given this, we hypothesized that if there are IPSs, their magnitude should be great enough to be considered outliers relative to non-pause sites, which would oscillate in a more normally distributed manner. The formula for determination of outliers (Crawley 2011) was used to determine the threshold above which an ES would be considered an IPS (red lines, Fig. 2). This analysis was performed for each of the three replicates, and the IPSs were remarkably consistent as illustrated in Figure 2. For each replicate, >80% of all IPSs were common with at least one other replicate, and when the significance of the overlap was tested in a pairwise manner using the hypergeometric distribution, the *P*-value for all pairs in both transcripts was highly significant (RPS12: 2.89×10^{-22} for replicate 1 vs. replicate 2 [R1:R2], 5.73×10^{-15} for R2:R3, 3.83×10^{-14} for R1:R3; ND7-5′: 3.41×10^{-6} for R1:R2, 3.70×10^{-10} for R2:R3, 3.80×10^{-8} for R1:R3). Moreover, in all three replicates <25% of all ESs in both RPS12 and ND7-5′ mRNAs constitute IPSs, and the majority of ESs exhibit a minimal level of pausing. For example, in the first RPS12 replicate, 3.7% of ESs (excluding primer regions) were never observed to constitute editing stop sites, and 75% of ESs constitute editing stop sites in less than 1500 sequences (Fig. 2A). In contrast, IPSs were represented by between 3578 and 12,612 sequences at each site. Said another way, RPS12 sequences containing IPSs (Fig. 2A, asterisks) make up 62.3% of the total reads recovered, but the IPSs represent only 15.7% of all ESs in the edited domain. Similarly for ND7-5′ replicate 1, sequences containing IPSs make up 58.3% of all sequences, but the IPSs represent only 13.5% of all ESs (Fig. 2B). Similar percentages were observed for both transcripts in replicates 2 and 3. From these data, we conclude that both RPS12 and ND7-5′ mRNAs contain IPSs, implying that there are inherent limitations to the 3′ to 5′ progression of full editing that lead to accumulation of these partially edited transcripts in the steady-state RNA population.

Characteristics of IPSs

Having revealed the presence of IPSs in both RPS12 and ND7-5′ mRNAs, we next analyzed the characteristics of these sites to provide insight into potential limitations in full editing progression. We first asked whether the distribution of IPSs provides evidence that gRNA exchange constitutes the barrier to progression of full, canonical editing. If this were the case, we would expect that the distribution of IPSs would be dispersed at a distance roughly the size of the coverage region of a single gRNA, and would often correspond to the 3′ ends of known gRNAs (Koslowsky et al. 2014). Alternatively, there may exist barriers to utilization of a given gRNA. It is thought that progressive realignment of the gRNA/mRNA duplex facilitates the progression of editing through each gRNA (Koslowsky et al. 1991; Maslov and Simpson 1992). How the different RECC, MRB1, or other components exchange or interact to facilitate this process is not known, and barriers to these processes may also cause intrinsic pausing. In this case, we would expect to observe IPSs spaced more closely than if gRNA exchange was limiting. To distinguish between these alternatives, we examined the locations of IPSs relative to edited RNA sequences and the complete families of gRNAs recently reported to direct their editing (Figs. 3, 4; Koslowsky et al. 2014). For RPS12, we observed many IPSs in close proximity to one another (e.g., ES 15, 17, 21, and 26; ES 32, 35, 38, and 39; ES 66, 67, and 68) (Fig. 3). These data suggest that barriers to utilization of a single gRNA contribute to intrinsic pausing. Likewise, for ND7-5′, IPSs were clustered at ES 45, 48, and 53, and again at ES 76 and 79 (Fig. 4). In RPS12 (but not ND7-5′), we also observed intrinsic pause sites corresponding or adjacent to 3′ ends of reported cognate gRNAs, although given the high degree of gRNA heterogeneity and redundancy, it is difficult to determine the significance of this observation. Collectively, the spacing of IPSs suggests that barriers to editing progression often arise during utilization of a single gRNA.

Since editing appears to pause at specific ESs within a given gRNA-defined block, we asked whether IPSs are enriched at ESs with specific characteristics. We used a Fisher's exact test

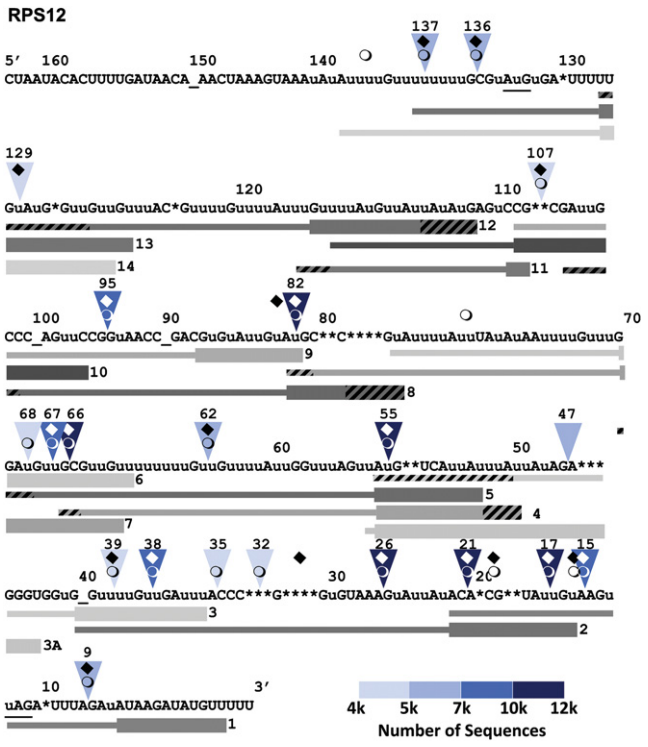

FIGURE 3. Locations of IPSs in RPS12 relative to edited mRNA sequence and known gRNAs. The sequence of canonical, fully edited RPS12 is displayed. Small u's denote uridines added to the sequence, (*) denotes encoded uridines that have been deleted, and large U's are encoded uridines that were untouched by the editing process. Gray bars *below* the sequence represent gRNAs as previously published (Koslowsky et al. 2014). The wider segments of the bars denote the predicted anchor sequence based on gRNA overlap and regions requiring no editing. The hashed segments represent regions where some gRNAs in a given family were reported to be shorter than the most abundant sequence. The blue arrowheads represent the IPSs in replicate 1 and the shade of blue corresponds to the normalized number of sequences that have their editing stop site at these pause sites. The circle denotes IPSs in replicate 2, and the diamond represents IPSs in replicate 3. These are shown in contrasting colors for better visibility. The numbers *above* the triangle indicate the editing site to which they are pointing.

on each of the three replicates independently and considered an observation true only when significant *P*-values were obtained from two or more of the three replicates. First, we examined the IPSs themselves to ascertain whether there was a significant over- or underrepresentation of sites with U insertion, U deletion, or no action required (i.e., pre-edited and edited sequence are identical) and observed no correlations with any specific action in either RPS12 or ND7-5′ mRNA (*P*-values are shown in Supplemental Tables S1, S2). We next analyzed the characteristics of the ESs directly 5′ to IPSs to ask whether the requirement for specific editing actions at the proximal site may constitute barriers to the progression of full editing. In RPS12, we found no enrichment for any specific required actions at ESs 5′ of IPSs vs. non-IPSs (Supplemental Table S1). From these data we conclude that, with respect to RPS12, difficulty in executing the proximal editing action is not the cause for pauses in full editing, implicating other factors in the generation of IPSs. In ND7-5′, we observed an enrichment for 5′ proximal sites that would require U insertion to achieve full editing in two of the three replicates (Supplemental Table S2). This suggests that there may be a slight, transcript-specific obstacle in executing downstream U insertions; however, it is also possible that undetermined factors at play in RPS12 may also be affecting the progression of editing in ND7-5′ and that these coincide with U insertion sites in this transcript. We next asked if specific nucleotides are enriched abutting IPSs and found no consistent enrichment of A, C, or G immediately 5′ or 3′ of an IPS in RPS12 and only a significant enrichment of 3′ Gs in ND7-5′ mRNA (Supplemental Tables S1, S2). We also saw a significant de-enrichment of 3′ As in RPS12 mRNA. While these observations could suggest that downstream G:C base-pairing plays a role in IPS formation, the preponderance of G:U base pairs in gRNA/mRNA interactions lessens this possibility. Collectively, these data suggest that features immediately surrounding IPSs contribute minimally to pausing, thereby implicating longer-range sequence characteristics and/or RNA–RNA interactions as barriers to the 3′ to 5′ progression of editing.

Identification of junctions and junction lengths

Though IPSs represent the 5′ boundaries of full canonical editing, the process of U insertion/deletion typically progresses past this site, generating junction regions containing, mis-edited sequence (Fig. 1A; Sturm and Simpson 1990; Koslowsky et al. 1991; Ammerman et al. 2010). The design of TREAT allows us to define junction sequences and examine junction

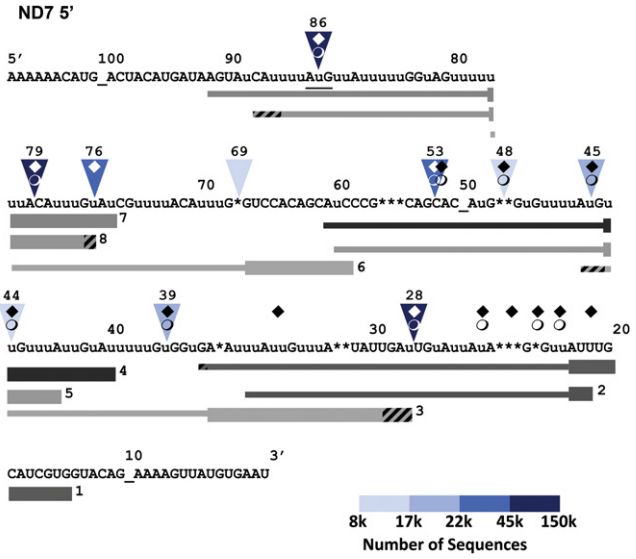

FIGURE 4. Location of IPSs in ND7-5′ relative to edited mRNA sequence and known gRNAs. The sequence of canonical, fully edited ND7-5′ is displayed. All symbols as in Figure 3.

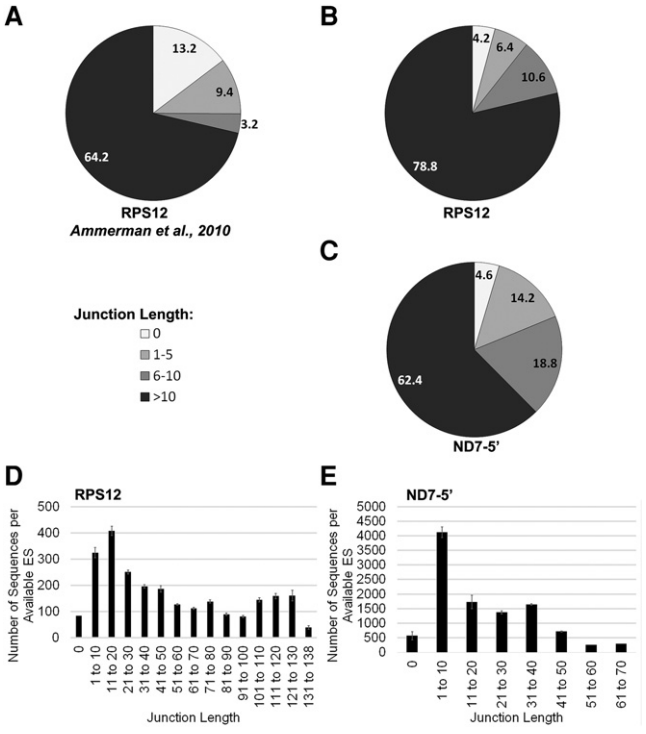

FIGURE 5. Analysis of junction lengths across all partially edited transcripts. (*A*) The percentage of RPS12 sequences with each junction length as reported (Ammerman et al. 2010). The percentage of RPS12 (*B*) and ND7-5′ (*C*) sequences with each junction length in replicate 1. A detailed breakdown of junction populations in RPS12 (*D*) and ND7-5′ (*E*) for replicate 1. The counts in *D* and *E* were normalized to the number of possible editing sites that could generate each junction length.

characteristics at a population level, potentially illuminating junction origins and functions. First, we characterized junction regions by length and compared this parameter to previously published data derived by conventional sequencing (Koslowsky et al. 1991; Ammerman et al. 2010). Previous data report the percentages of sequences lacking junctions as 13.2% in RPS12 (Ammerman et al. 2010), 13.3% in ND7 (combined 5′ and 3′ domains) (Koslowsky et al. 1991) and 8.7% for ATPase subunit 6 (Koslowsky et al. 1991). Our data confirm that the percentage of sequences with junction length zero is small, constituting 4.2% in RPS12 and 4.6% in ND7-5′ for replicates 1 (Fig. 5B,C) and between 8.5 and 11% in replicates 2 and 3 for both RPS12 and ND7-5′. Because junctions are present in the majority of partially edited mRNAs, these data support the hypothesis that mis-editing is a natural part of the editing process and likely has a specific function (Koslowsky et al. 1991; Ammerman et al. 2010). We next compared sequences with specific junction length ranges across the population of partially edited RPS12 sequences in our data set with that published in Ammerman et al. (2010) (Fig. 5A,B). While the proportion of RPS12 sequences with junction lengths of 10 ES or less in our data set (25.1%, 29.7% and 27.3% in three replicates) was quite similar to the 25.8% reported previously (Ammerman et al. 2010), our data show a somewhat different distribution of junction lengths within this range. To compare across transcripts, we also analyzed the same junction lengths in the ND7-5′ population and found a marked consistency in the percentage of sequences with no junction and a slightly larger proportion of sequences with junction lengths less than 10 ES (49.0%, 37.5% and 41.8% in three replicates). To determine whether junctions can arise from mis-utilization of a single gRNA, we next analyzed junction lengths by nucleotide (Supplemental Fig. S3). The reported gRNA populations range from ~50–70 bp in length, and we observe a substantial number of junctions of lengths >70 bp, especially in RPS12 mRNA, suggesting that a sub-population of junctions are generated through utilization of more than one gRNA. Overall, for both RPS12 and ND7-5′ RNAs, we found that >89% of partially edited sequences in all three replicates contain junctions. Moreover, our data show that the proportion of long-to-short junctions remains relatively consistent across transcripts and confirm previously published junction lengths.

Given that the majority of sequences recovered for both RPS12 and ND7-5′ had junctions greater than 10 ESs long, we examined the longer junctions in detail to determine whether some junction lengths were more prominent than others. As the maximum possible junction length is dependent on the location of the editing stop site (i.e., 10 times as many ESs in RPS12 can generate junctions of length 11 than can generate those of length 110), we first scaled the normalized sequence counts by dividing them by the number of possible ESs that could generate each junction length. From this, we pooled the junctions into ranges of ten ESs to see whether a particular population of longer junction predominated (Fig. 5D,E). In all three replicates, for both RPS12 and ND7-5′, we found that the shorter of the junction populations are highly abundant relative to how often they could theoretically be generated. In RPS12 replicate 1, the peak populations fell between junction lengths 1 and 50 with a slight peak in junctions of 101 to 130 nt that was reproducible in replicates 2 and 3 (Fig. 5D). In ND7-5′ replicate 1, the peak populations spanned junctions length 1–40 in all three replicates, and no second peak of longer junctions was observed (Fig. 5E). From this we conclude that shorter junctions form more readily relative to how often they can form. This is consistent with the hypothesis that junctions are regions of active editing that are being modified to match the canonically edited sequence.

Identification and characteristics of MJESs

We next asked whether junction end sites are concentrated at certain ESs or whether the 5′ boundaries of junctions are equally likely to arise at any ES. MJESs were identified using the same outlier strategy we used to identify IPSs (see Materials and Methods), and the orange lines in Figure 6 delineate the thresholds defining MJESs for both the RPS12

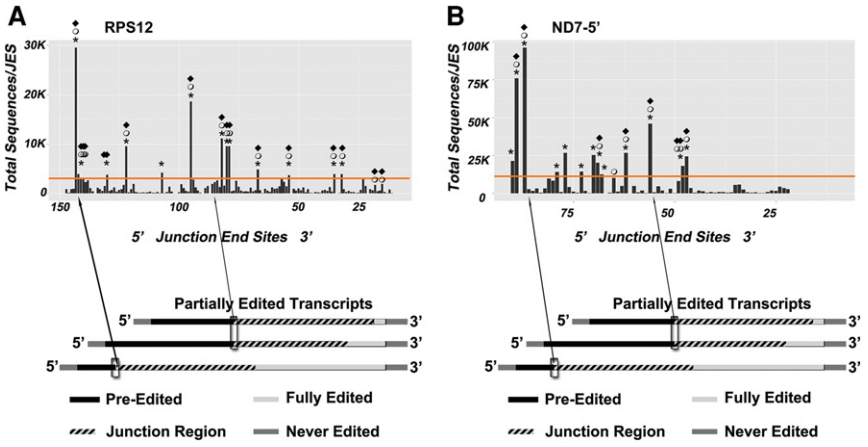

FIGURE 6. Amplitude of junction end sites across RPS12 and ND7-5′ populations of partially edited transcripts. Graphs show the total number of sequences (normalized) per junction end site across the RPS12 (*A*) and ND7-5′ (*B*) transcripts in replicate 1. The orange lines denote the outlier thresholds for each transcript, and asterisks *above* peaks denote MJESs in replicate 1. Open circles denote MJESs in replicate 2, and black diamonds are MJESs in replicate 3.

and ND7-5′ populations. MJESs recovered from all three replicates were markedly consistent as, with one exception, all MJESs were found in at least two of the three replicates of each transcript. Taking replicate 1 as an example, this analysis revealed 13 MJESs in RPS12 (9.2% of total ESs) and 13 MJESs in ND7-5′ (17.3% of total ESs) (Fig. 6). Thus, junction end sites are not equally likely to form at any ES but rather cluster at specific ESs on a transcript.

As junction end sites represent the 5′ most ES at which any editing action has occurred, sequence characteristics enriched at MJESs may illuminate factors that cause overall editing action to stall (Fig. 1C). Alternatively, if junctions arise due to alternative gRNA usage, MJESs may reflect the ends of alternative gRNAs. To begin to understand the factors that contribute to the generation of MJESs, and thus define barriers to overall 3′ to 5′ progression of editing, we examined the positions of these sites with respect to known gRNAs and the edited RNA sequence (Figs. 7A, 8A). With respect to known gRNAs, we observe that MJESs cluster together in both RPS12 and ND7-5′ transcripts, much like the IPS (Figs. 3, 4), suggesting they arise independently of gRNA exchange. Therefore, we next examined sequences characteristics surrounding MJESs (Supplemental Tables S1, S2). Because visual inspection suggested that ESs that correspond to deletion sites in canonical fully edited RNA sequences often constitute MJESs (Figs. 7, 8), we statistically analyzed whether MJESs are enriched at ES that would require U insertions, U deletions, or no action to shift from pre-edited to fully edited sequence. Remarkably, RPS12 RNA populations showed highly significant enrichment for deletion sites in all three replicates ($P = 1.6 \times 10^{-5}$; 1.0×10^{-3}; 2.4×10^{-4}), and ND7-5′ mRNAs showed significant enrichment for U deletion sites at MJESs in two of the three replicates ($P = 0.10$; 0.03; 0.02) (Figs. 7B, 8B). We further examined the MJESs that correspond to canonical deletion sites to determine the extent of deletion that had actually taken place. That is, in a junction region, these sites might have the fully edited number of U's deleted, have partial U deletion compared to fully edited sequence, or have U addition. For RPS12, using replicate 1 as our representative data set, we examined all sequences ending at the seven RPS12 MJESs that require deletions for full editing and found that 87-99.5% of these sequences had undergone some U deletion at that site, though not always the canonical deletion (Supplemental Fig. S4A,B). Similarly, in ND7-5′ replicate 1, 76%–96% of MJESs at canonical U deletion sites had undergone some U deletion (data not shown). We next asked whether the extent of U deletion differed between those canonical U deletion sites that constitute MJESs and those that do not, and found that the extent of actual U deletion at these sites was similar (Supplemental Fig. S4C). For example, of the six RPS12 ES requiring U deletions which were not MJESs, 75.2% –98.7% of all sequences whose junctions end at these sites exhibited some level of U deletion (Supplemental Fig. S4C). These data indicate that the 3′ to 5′ progression of editing has a strong propensity to stall following an executed U deletion.

Because not all U deletion sites constitute MJESs, additional factors must also contribute to stalls in editing progression. To further define these factors, we asked whether the nucleotides abutting MJESs were enriched for A, C, or G. In RPS12 mRNA we found five MJESs with 5′ Gs, four with 5′ As, and four with 5′ Cs, suggesting no bias in the 5′ nucleotide overall. However, all four of the Cs occurring 5′ of the MJESs occur following deletion sites. Thus, the combination of a canonical U deletion site and a 5′ C appears especially prone to stalling in RPS12. In ND7-5′, we observed eight MJESs with 5′ Gs, five with 5′ As, and none with 5′ Cs. Of these, the two deletion sites that are MJESs have 5′ Gs; however, ND7-5′ mRNA contains no deletion sites bounded by a 5′ C (see Fig. 8A). The full reports of *P*-values for tested conditions with MJESs in RPS12 and ND7-5′ are shown in Supplemental Tables S1 and S2. Overall, we conclude that the trend of U deletions leading to pausing in the 3′ to 5′ editing progression is conserved in RPS12 and ND7-5′ mRNAs, and that MJESs are more dependent on mRNA sequence than on gRNA exchange in both RPS12 and ND7-5′ mRNAs.

Evidence that IPSs and MJESs arise independently

Next, we asked whether the MJESs represent a few junction sequences present in large numbers or a more diverse population of junction sequences. We reasoned that if there are alternative editing pathways that lead to translatable alternative sequences, these will likely appear as junction sequences with

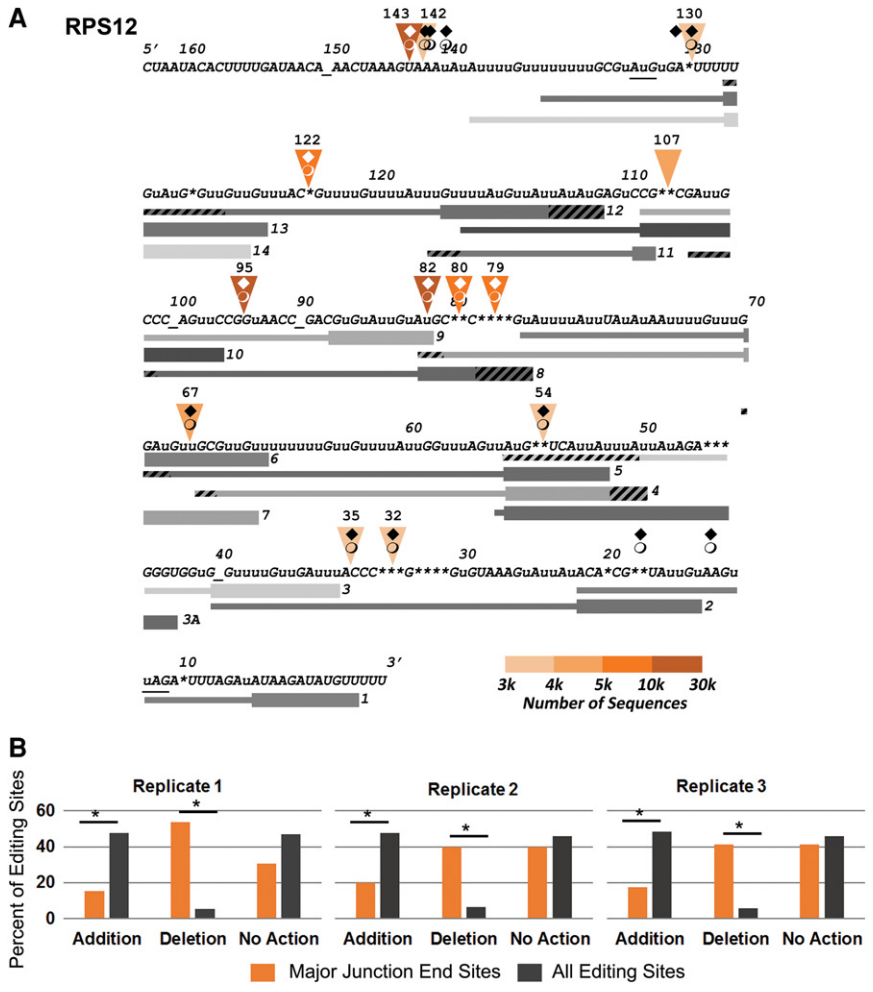

FIGURE 7. Location of MJESs in RPS12 relative to edited mRNA sequence and known gRNAs. (*A*) The fully edited sequence of RPS12 is shown. Orange triangles denote MJESs in replicate 1, and the shade of orange corresponds to number of sequences. Open circles denote MJESs in replicate 2 and diamonds the MJESs in replicate 3. The numbers *above* the triangles indicate the editing site to which they are pointing. (*B*) Percentage of MJESs that require addition, deletion, or no action in fully edited RNA (orange), and the percentage of all editing sites that require these actions (gray). All three replicates are shown and asterisks denote significance as determined by a Fisher's exact test ($P < 0.01$) between the MJESs and non-MJES editing sites.

consistent start and end points. If the MJESs are reflecting these kinds of sequences, we would expect to see a lower diversity of sequences represented at MJESs compared with non-MJESs. If there were a higher diversity of sequences present at MJESs, this would imply that they arise through diverse editing actions including RECC error, misalignment of the gRNA/mRNA duplex, or use of aberrant gRNAs. To test this, we asked whether sequences whose junctions end at a specific MJES share common editing stop sites that are distinct from the editing stop sites giving rise to junctions that end at other MJESs. If we see that each MJES correlates with a specific editing stop site, this implies that the cause for the full editing to stop is related to the specific junction end, and thus an alternative pathway that could be due to gRNA utilization. For the sequences whose junctions end at MJESs, we used the outlier threshold method to determine the most abundant corresponding editing stop sites for each MJES. These analyses showed that each MJES can have multiple abundant editing stop sites, as shown by the examples in Figure 9 (compare bars A, B, C and bars D, E, and F). The most common editing stop sites for a given MJES are often IPSs. We then analyzed these IPSs to define their corresponding MJESs, and found that each IPS corresponds to multiple MJESs (Fig. 9, compare bars A and E, B and F, and C and D) and that these typically span the majority of the MJESs 5′ of that IPS. Thus, MJESs and their corresponding major editing stop sites are not linked by the presence of a consistent alternative gRNA utilization path, indicating that MJESs and IPSs arise through distinct mechanisms.

Evidence for alternative editing in the mRNA

Upon examination of the junction sequences, we found that longer junctions often comprise two stretches of junction sequences with an intervening extended sequence that matches the fully edited transcript (Fig. 10A). The fully edited portion arising mid-junction can include nearly the full remaining length of the transcript and sometimes includes the canonical start codon (Fig. 10B). This observation, which was noted across several different editing stop sites, suggests that there are regions of RNAs that are consistently mis-edited such that full, canonical editing can continue unhindered. This editing pattern may represent regions where nonfidelity is tolerated at the sequence level and may give rise to alternative protein sequences. For example, Figure 10B illustrates a family of RPS12 sequences defined by a short alternatively edited region following ES21 which returns to canonical fully edited sequence and generates the canonical start codon with some variations in the 5′ UTR. In replicate 1, we observed 338 reads within this family. An additional 1137 reads contain the 3′ alternatively edited region shown in Figure 10B but do not form the canonical start codon, instead forming other junction sequences of varying lengths. This suggests that the 3′ alternatively edited region shown here is consistently generated in our cell population. If translated, the sequence in Figure 10B would generate a 91 amino acid protein that is slightly longer than the 82 amino acid

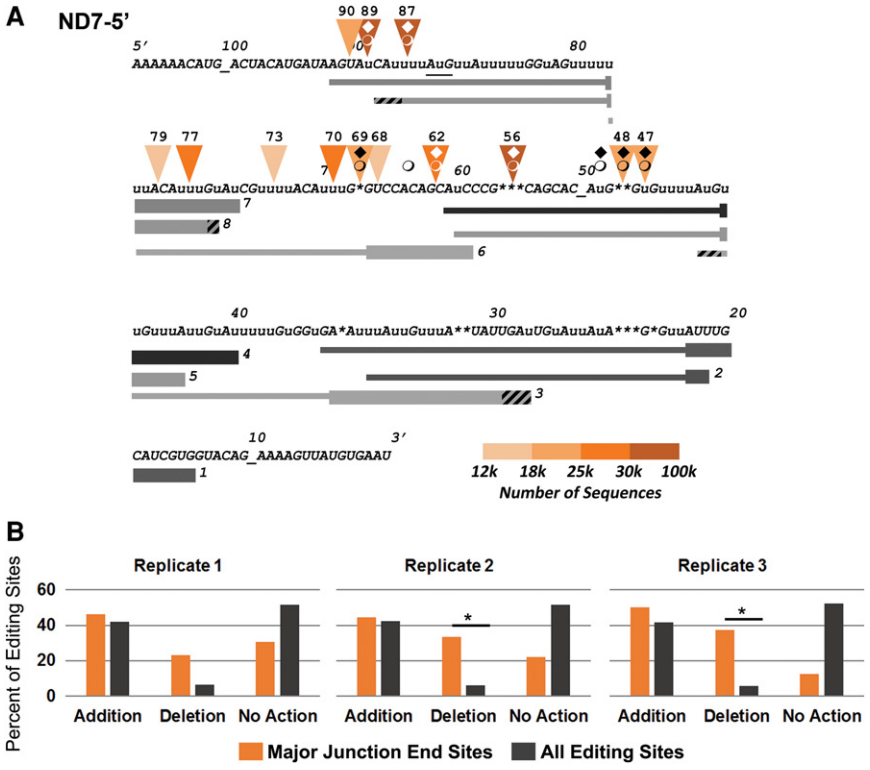

FIGURE 8. Location of MJESs in ND7-5′ relative to edited mRNA sequence and known gRNAs. (*A*) The fully edited sequence of ND7-5′ is shown. Symbols as in Figure 7A. (*B*) Percentage of MJESs that require addition, deletion, or no action (orange) and the percentage of all editing sites that require these actions (gray). All three replicates are shown and asterisks denote significance as determined by a Fisher's exact test ($P < 0.05$) between the MJESs and non-MJES editing sites.

canonical RPS12 with an alternate, extended C terminus (Fig. 10B, lower). Because we observe 338 such alternative reads that generate the canonical start code, compared to a total of 14,385 reads in which the canonical ORF is generated and only the 5′ UTR contains variation, the alternative sequence presented in Figure 10B generates a substantial pool of potentially translatable RNA (2.3% as abundant as canonical sequence), with the potential to generate an alternative protein.

A second example of alternative editing identified by TREAT is shown in Figure 10C. Here, an alternatively edited sequence beginning at ES95 is, like the example above, followed by canonical fully edited sequence and the canonical start codon is generated. However, in silico translation shows that, although an alternate mRNA sequence is generated, the resulting mRNA still predicts the canonical protein sequence (Fig. 10C, lower). This alternatively edited region is present in 3025 sequences, 62% of which are edited to generate the full canonical ORF; thus, the fully edited alternative sequence is present at ~13% of the level of canonical edited RPS12 sequence. Together, these data provide evidence that alternative gRNAs direct editing leading to silent variations in mRNA sequence, and demonstrate the power of TREAT to uncover natural variants in the edited RNA population.

Evidence for utilization of alternative gRNAs

Finally, we asked if there is evidence that published alternative gRNAs are utilized in these cells and whether editing progresses beyond these regions in a way that would allow for translation of the alternative sequence (Madej et al. 2008; Koslowsky et al. 2014; Madina et al. 2014). The published gRNAs we include here are likely only a small fraction of total alternative gRNAs in any given cell. We use them as a base case to gather evidence for the utilization of alternative gRNAs and to examine how that affects subsequent editing action. Further investigation of the gRNA pool is likely to reveal other candidate alternative gRNAs that can be tested using the same methodology. To this end, we utilized the multiple template functionality in TREAT and included templates constructed to represent alternatively edited sequences wherein the alternative gRNA is utilized and the remainder of the sequence is edited canonically. Sequences from the data sets were batched as alternatively edited if and only if they contained canonical edited sequence up to the region covered by the alternative gRNA, and this was then followed by edited sequence matching that which would be directed by the alternative gRNA. The sequence matching the alternative gRNA must be edited with full fidelity up to the 5′ end of the region covered by that alternative gRNA. Three published alternative gRNAs for RPS12 and one for ND7-5′ were examined in this manner (Madej et al. 2008; Koslowsky et al. 2014; Madina et al. 2014). This analysis revealed no evidence for utilization of the ND7-5′ gRNA. However, we did find evidence for utilization of all three alternative RPS12 gRNAs, encompassing a total of 760 unique sequences (2234 reads) meeting the above criteria. Thus, alternative RPS12 gRNAs are utilized, albeit infrequently as these reads represent <1% of the total reads. We next asked how many of these sequences return to fully edited sequence such that the canonical start codon is generated to measure the potential translatability of the alternatively edited RNAs. Of the RPS12 sequences, only 3%–5% of RNAs containing sequence generated from the three alternative gRNAs had no editing action beyond the 5′ boundary of the alternative gRNAs, suggesting that these are not dead-end products. The majority of these alternatively edited regions were followed by mis-edited junctions, implying their continued editing. However, a small number (17, 7, and 45 for the three alternative gRNAs) returned to canonical editing such that

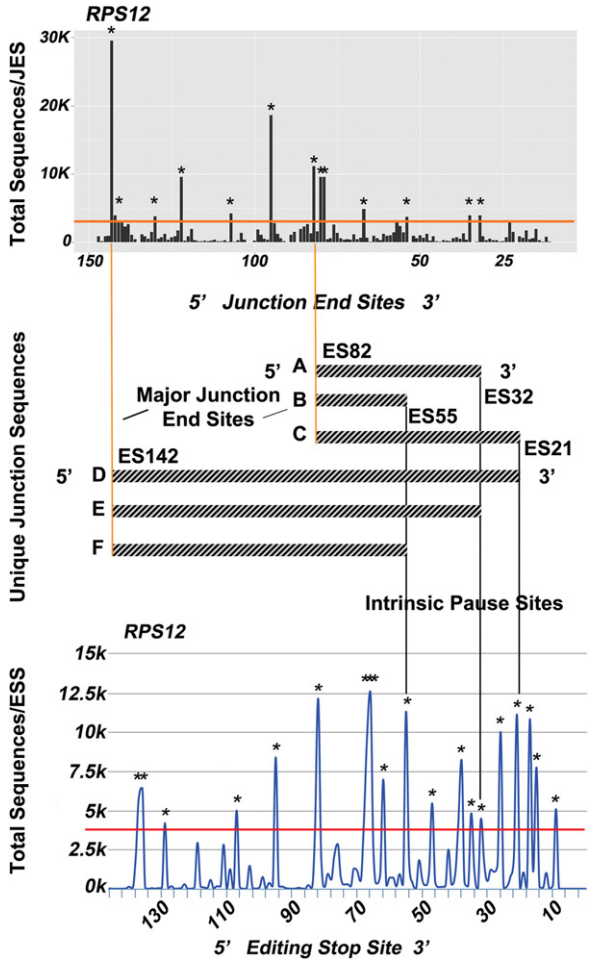

FIGURE 9. MJESs and IPSs arise through independent mechanisms. The *top* panel shows the abundance of sequences per junction end site in RPS12 (replicate 1) with MJESs denoted by an asterisk (from Fig. 6A). The *bottom* panel shows the abundance of pausing at each editing site in RPS12 (replicate 1), and intrinsic pause sites are denoted by an asterisk (from Fig. 2A). The *middle* segment shows schematics of several unique junction sequences, A through F, demonstrating that sequences with different junction end sites (orange vertical lines) share the same editing stop sites (compare A and E; B and F; C and D) and those correspond with IPSs (black vertical lines). Junctions with different editing stop sites can also share the same junction end sites (compare A, B, and C; compare D, E, and F).

the canonical start codon was created, although there were variations in the 5′UTRs. This is compared to the 14,389 total of canonical RPS12 sequences that generate the start codon but contain variable 5′ UTRs. Thus, these alternative gRNAs appear to be capable of generating translatable alternative mRNAs, although these represent a small proportion of the overall translatable mRNA pool.

DISCUSSION

In *Trypanosoma brucei*, U insertion/deletion RNA editing is required to create functional open reading frames in many mitochondrial mRNAs. This process generates a diverse array of partially edited mRNAs, which comprise the majority of the steady state mitochondrial mRNA pool. Because the numbers of U's in any given ES are variable in partially edited mRNAs, these sequences are difficult to align using conventional tools. Here, we describe the Trypanosome RNA Editing Alignment Tool (TREAT), which permits alignment of large populations of sequences ignoring a single base, and generates a user-friendly database that allows users to analyze high-throughput sequencing data sets for common editing characteristics. Using this tool, we analyzed two pan-edited RNA transcripts, RPS12 and ND7-5′, in strain 29–13 procyclic form *T. brucei*. Our data reveal that the general 3′ to 5′ progression of editing pauses via at least two distinct mechanisms. The first generates MJESs and marks the end of all editing action on a transcript. Generation of MJESs appears to be related to the local sequence and is enriched after a U deletion action. The second generates the IPSs and marks the 5′ boundary of correctly edited sequence. In contrast to the MJESs, the positions of IPSs suggest that this mechanism of pausing is independent of the local mRNA sequence. Analysis of pause sites in partially edited sequences in the context of known gRNAs (Koslowsky et al. 1991, 2014) indicates that both IPSs and MJESs arise through mechanisms related to utilization of a single gRNA, although we cannot rule out an additional contribution of impaired gRNA exchange. Finally, our study provides strong evidence that alternative editing pathways and alternative gRNAs are being utilized during editing of RPS12 RNA, albeit to a relatively small extent compared to pathways resulting in canonical sequence. Overall, this work demonstrates the power of TREAT, in conjunction with high-throughput sequencing, to analyze large populations of partially edited RNAs and thereby enhance our understanding of numerous aspects of kinetoplastid RNA editing.

The finding that common pauses in the generation of correctly edited sequence, termed here IPSs, arise independently of the local RNA sequence suggests that RNA structure impacts RECC utilization and editing progression at specific sites. The myriad, dynamic RNA structural changes necessary for proper editing create the opportunity for multiple mRNA/gRNA duplexes as well as intra-mRNA and intra-gRNA secondary structures to arise. Especially strong intramolecular interactions could render gRNAs or mRNAs relatively refractory to unwinding, whereas weaker interactions could cause premature dissociation of mRNA/gRNA duplexes. Both situations could also lead to a disruption in the proper alignment of mRNA/gRNA duplexes that would either inhibit editing or lead to improper editing and the formation of a junction region. Resolution of both inter- and intramolecular RNA structure has long been recognized as essential for proper 3′ to 5′ progression of editing, and several protein factors that may contribute to this process have been described. These include the helicases REH1 (Li et al. 2011) and REH2 (Hernandez et al. 2010; Madina et al. 2014, 2015), the

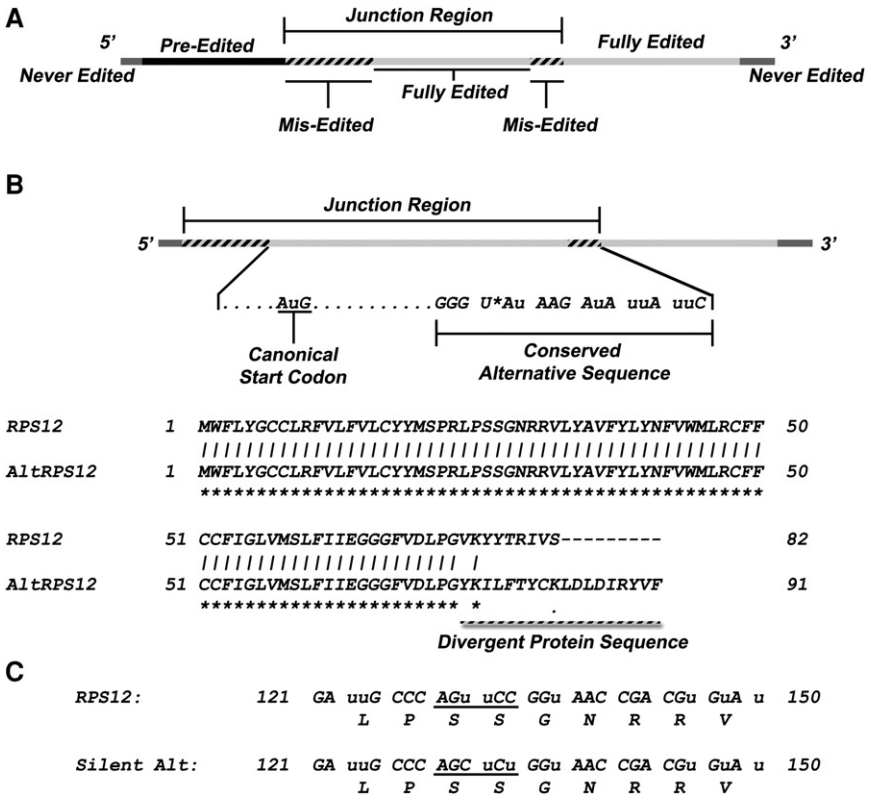

FIGURE 10. Evidence for alternative editing. (*A*) Schematic demonstrating the makeup of many longer junctions wherein two regions of mis-edited sequence flank an extended region matching fully edited sequence. (*B*) A sequence of the type shown in *A* encodes a predicted alternative protein. The conserved alternative 3′ sequence is followed by a return to canonical fully edited sequence (dotted line) and this often extends to include the canonical start codon. Shown *below* the mRNA sequence is the predicted alternative protein sequence (AltRPS12) aligned with the canonical RPS12 sequence (RPS12). (*C*) An alternatively edited RPS12 mRNA encodes the canonical RPS12 protein. Canonical RPS12 mRNA and protein sequences encoded between nucleotides 121 and 150 (*top*). The alternatively edited region (Silent Alt; underlined codons) still encodes the canonical protein sequence (aligned *below*).

essential MRB1 protein TbRGG2 (Fisk et al. 2008; Ammerman et al. 2010; Foda et al. 2012), and nonenzymatic RECC components (Kala and Salavati 2010). REH1 has been implicated in removal of gRNAs from fully edited mRNA (Li et al. 2011) and REH2 with the assembly of RNA with the MRB1 core (Madina et al. 2015). TbRGG2 is an MRB1 component with RNA melting and annealing activity hypothesized to facilitate the progressive realignment of mRNA/gRNA duplexes during the editing process (Fisk et al. 2008; Ammerman et al. 2010; Foda et al. 2012). KREPA4 is a RECC component with affinity for structured gRNA in vitro that exhibits in vitro RNA annealing activity (Kala and Salavati 2010). The IPSs identified here reflect natural limiting steps in RNA editing, and as such could also function as points of regulation. In future studies, we will define predicted RNA structures surrounding IPSs. We will further utilize TREAT to analyze changes in partially edited sequences in cells depleted for, overexpressing, or expressing mutated versions of specific RNA editing factors to determine the contributions of these proteins to the generation of correct edited mRNA sequences.

MJESs represent common positions constituting the 5′-most editing action in any given sequence, and thus these ESs appear to be those beyond which it is most difficult for editing to progress. The close distribution of MJESs in RPS12 and ND7-5′ RNAs points away from gRNA exchange as the primary mechanism for their accumulation and indicates that there are intra-gRNA barriers to continued editing. The striking correlation of MJESs with deletion activity in both RPS12 and ND7-5′, as well as with 5′ adjacent C's in RPS12, demonstrates that mRNA sequence character affects the successful 3′ to 5′ continuation of all editing action. Distinct RECC variants catalyze U insertion and U deletion (Panigrahi et al. 2006; Carnes et al. 2008, 2011). These variants comprise 12 common proteins associated with different endonucleases and partner proteins, although it is not clear whether these are stable complexes or whether insertion and deletion modules shuttle on and off of the 12-protein base. Regardless of the mechanism, our data support a model in which the switch from a deletion RECC to an insertion RECC poses more of a barrier to editing progression than does use of the same RECC type at adjacent sites or switching from insertion RECC to deletion RECC. The deletion to insertion switch may also be confounded by additional factors such as G:C base-pairing 5′ of the deletion site creating stronger mRNA/gRNA duplexes or mRNA secondary structure that inhibits editing. Interestingly, RECC has never been shown to stably associate with mRNA or gRNA (Rusche et al. 1997; Cifuentes-Rojas et al. 2005; Alatortsev et al. 2008; Carnes et al. 2011; Aphasizheva et al. 2014), and the current model proposes that the MRB1 complex comprises the portion of the editing holoenzyme mediating correct mRNA/gRNA duplex formation and RECC association with RNA (Hashimi et al. 2013; Aphasizheva et al. 2014; Madina et al. 2014, 2015; Read et al. 2016). Consistent with this model, depletion of the MRB1 protein, TbRGG2, increases the number of sequences with no junction in the steady state population of RPS12, indicating an important role in overall editing progression (Ammerman et al. 2010). Thus, we expect that examination of editing defects caused by depletion of other MRB1 components, as well as non-MRB1 accessory editing proteins, using TREAT will be informative with regard to the regulation of 3′ to 5′ editing progression, and the specific aspects of progression regulated by distinct factors.

One of the most striking findings in this study was the identification of consistently generated alternatively edited RNA sequences arising from previously reported alternative gRNAs as well as novel, consistent alternatively edited regions. The identification of junction sequences has long sparked a debate over whether these mis-edited sequences lead to the generation of distinct proteins, much like splicing does in other organisms, or whether they represent an editing intermediate that either becomes corrected or degraded. Either scenario suggests the possibility of dynamic regulation within the cell. Ochsenreiter and colleagues reported the presence of several mRNAs with alternative editing that generate plausible open reading frames in silico, and they provided evidence for the generation of an alternative protein, AEP-1, from an alternatively edited COXIII mRNA (Ochsenreiter and Hajduk 2006; Ochsenreiter et al. 2008a,b). Our data suggest that mis-edited sequences can generate both potentially translatable sequences and sequences that are intermediates in the editing process. The repeated generation of the alternative RNA sequence shown in Figure 10B, as well as the evidence for utilization of alternative gRNAs, suggests that there are regions where alternative editing is consistently tolerated and can generate alternative proteins. However, even the majority of sequences bearing these alternative regions contain additional junctions between the alternative fully edited and pre-edited sequence that vary even when the alternative segment is conserved. Thus, it appears that junction sequences that will either be corrected during editing or lead to degradation, rather than to a productive alternatively edited RNA, are features of most partially edited RNAs and are likely integral to the editing process. Current hypotheses speculate that these alternative or mis-edited sequences arise due to aberrant gRNA utilization, misalignment of canonical gRNAs with the mRNA, or RECC error. Though our current data do not distinguish between these possibilities directly, we hypothesize that multiple mechanisms contribute to junction creation and that the utilization of gRNAs generating translatable alternative sequences are likely regulated in a way distinct from other aberrant gRNAs. Analysis of additional mitochondrial RNAs using high-throughput sequencing combined with TREAT analysis will increase our understanding of the frequency of alternative RNA editing. Further identification of plausible alternative ORFs will allow us to distinguish between mis-editing that represents potential protein diversification and that which arises as an intermediate in the natural editing process.

Though we have identified potential alternative protein ORFs in this data set, it remains unknown whether these sequences are translated. The RNA editing process is apparently coupled to the translation process through the addition of a long poly(A/U) tail, although the mechanism(s) by which fully edited RNAs are identified and then trafficked to the mitochondrial ribosomes is not well understood (Militello and Read 1999; Etheridge et al. 2008; Aphasizhev and Aphasizheva 2011a,b, 2013; Aphasizheva et al. 2011). As editing begins at the 3′ end of the transcript and the poly(A/U) tail appears to be added primarily to fully edited transcripts, it is plausible that changes at the 5′ end of the sequence are necessary for post-editing RNA processing (Etheridge et al. 2008; Aphasizheva et al. 2011). This makes the 5′ UTR a candidate for regulatory function. In both RPS12 and ND7-5′, we found only a small number of sequences with both the canonical ORF and the canonical 5′ UTR. For example, these sequences represented only 0.007% of total reads in RPS12 replicate 1. In contrast, when we considered sequences that had the full canonical ORF but variations in the 5′ UTR, the number of sequences that could potentially be translated into canonical RPS12 protein increased over 700-fold, and these reads comprised 5.7% of the total reads. Similarly, the majority of alternative sequences that maintained the canonical start codon had variable 5′ UTRs. Though we do not know what impact 5′ UTR variations will have on the processing and translatability of these transcripts, TREAT provides us with a valuable tool for examining 5′ UTR variation in vivo. Characterization of 5′ UTRs using TREAT could easily be used to complement in vivo studies of polyadenylation or ribosome bound transcripts to provide greater insight into the regulation of RNA translation in mitochondria.

Of the genes requiring RNA editing in trypanosomes, several are differentially edited between the procyclic and bloodstream forms of the parasite (Feagin et al. 1987; Koslowsky et al. 1990; Read et al. 1992, 1994; Souza et al. 1992, 1993; Corell et al. 1994; Stuart et al. 1997; Schnaufer et al. 2002). Recent studies of two nonenzymatic RECC components revealed differential functionality in the bloodstream and procyclic forms of the parasite, although the mechanisms of action behind these differences are unknown (McDermott et al. 2015a,b). The potential roles of other factors, such as components of MRB1 complex, in developmental regulation of RNA editing are as yet unknown. Further analysis of partially edited sequences in parental cells in the different life cycle stages, as well as the differential effects of editing proteins in these two stages, is warranted to explore the mechanisms of RNA editing regulation. Overall, our data demonstrate the power of using high-throughput sequencing and our Trypanosome RNA Editing Alignment Tool (TREAT) to study the editing process in kinetoplastids. This approach has great potential to be applied in cells where critical editing factors have been depleted or mutated so that their roles in the progression of editing can be more clearly understood. The observations derived from TREAT-based analysis have the potential to be both conclusive and hypothesis generating with further confirmation possible through in vitro or in vivo techniques. We anticipate that TREAT will have increasing value as a tool for studying RNA editing and aim to expand its capabilities through planned additions as well as refinement following usage by other groups and complementary biochemical studies.

MATERIALS AND METHODS

Samples for high-throughput sequencing

RNA was harvested from mid-log *Trypanosoma brucei brucei* strain 29-13 procyclic cells using TRIzol per manufacturer's instructions, followed by phenol: chloroform extraction. DNA was removed from the samples using the DNA Free Kit (Ambion). First-strand cDNA was synthesized using 1.2 µg of RNA in a 20 µL reaction and Superscript III Reverse Transcriptase using the 3′ primers specific for each gene and containing an ID tag to permit binning following sequencing (Supplemental Table S3). The linear range of each PCR reaction was determined by qRT-PCR using final primer concentrations of 1.2 µM and 1.5 µL of cDNA reaction in a 25 µL PCR reaction. To generate samples for MiSeq sequencing, cDNA was amplified by RT-PCR, using the same primer concentrations and cycle number that was determined to be in the linear range of the PCR reaction for each sample. The PCR amplicons were then purified using the Illustra GFX PCR cleaning kit and eluted in 10 µL of 10 mM Tris–HCl, pH 8.0 elution buffer provided by the company.

Library construction and sequencing

cDNA resulting from PCR reactions described above was quantified using the Picogreen Assay (Invitrogen), and the sizes of products were confirmed using the Agilent Bioanalyzer DNA high Sensitivity chip (Agilent). A 50 µL index PCR reaction was carried out to attach dual indices and Illumina sequencing adapters. Twenty-five microliters of 2× KAPA HiFi HotStart Ready mix was combined with 5 µL Nextera XT Index primer 1 (N7xx) and Index primer 2 (S5xx) and added to 2 ng of cDNA for the PCR reaction. AMPure XP beads (Beckman Coulter Genomics) were used to purify the final libraries. Libraries were then quantified using the Picogreen assay and Library Quantification kit (Kapa Biosystems). Agilent Bioanalyzer DNA high sensitivity chip (Agilent) was used to confirm the sizes of the cDNA libraries. The libraries were normalized and pooled. The pooled libraries were then sequenced using Illumina MiSeq 300 cycle paired-end sequencing.

Preprocessing of RNA-seq paired-end reads

Paired-end sequencing reads from the Illumina MiSeq were obtained in FASTQ format. The FASTQ files were merged using PEAR (Paired-End reAd mergeR) (Zhang et al. 2014). The resulting reads were then merged using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) utility program fastq-collapser, which collapses identical sequences into a single sequence while maintaining read counts. Unique sequences and their merged read counts were loaded using TREAT. The merge count for each unique sequence was normalized using the ratio of the total reads returned for each sample to the average of the total reads across all samples in a technical replicate.

Trypanosome RNA Editing Alignment Tool (TREAT)

TREAT is a multiple sequence alignment and visualization tool written in Go and freely available under the GPLv3 license at https://github.com/ubccr/treat. The core functionality of TREAT consists of a special purpose multiple sequence aligner. TREAT aligns sequences using three bases and assembles editing sites to detect the extent of editing of the fourth base, called the edit base. The edit base is configurable in TREAT and by default uses "T." TREAT requires as input two template sequences in FASTA format representing the fully edited and pre-edited mRNA transcripts. Optionally, one or more alternatively edited templates can be provided which define special regions of the transcript where alternate editing mechanisms can generate a specific alternative sequence. All template sequences must be identical with respect to the non-edited bases. An editing site can occur before or after any non-edited base in the template sequence. Editing sites are numbered in the 3′ to 5′ direction (0 based) and non-edited bases are numbered in the 5′ to 3′ direction (0 based). TREAT aligns an input sequence to the templates by first performing a global alignment on the non-edited bases using the Needleman–Wunsch algorithm (Needleman and Wunsch 1970) and then adding back in the edited bases found at each editing site, as shown in Figure 1. The global alignment on the non-edited bases accounts for any possible sequencing errors or mutations in the input sequence compared to the templates. If a mutation in the non-edited bases is found, TREAT will flag the alignment and record the type of mutation (indel or SNP) for downstream analysis. Because this analysis is done on the sequence from which the T's have been removed, changes from a non-edited base (C/G/A) to T will appear as a deletion of the C/G/A and thus are eliminated from the primary data pool, regardless of whether they are internal or terminal. Any variation found in the user-designated primer binding region is ignored. In the data sets reported here, for RPS12 57.9% (R1), 90.26% (R2) and 89.08% (R3); for ND7-5′ 45.66% (R1), 62.82% (R2), and 49.72% (R3) were removed as they had non-T sequence errors. It is assumed that the input sequence represents the entire mRNA transcript; i.e., TREAT performs a global alignment and does not currently support sequencing reads shorter than the template sequences. Input sequences shorter than the templates will be reported as deletions and thus only visible when the "mutant" box is checked. It is assumed paired-end sequencing reads have been preprocessed with tools like PEAR for merging paired-end reads and fastx-collapser for collapsing unique sequences.

TREAT alignment definition

Let n equal the number of editing sites and m be the number of template sequences where $m \geq 2$, representing the pre-edited, fully edited, and any alternatively edited templates the user provides. Let $t = \{t_{0\ldots}t_{n-1}\}$ be an n vector where t_j is the count of edited bases for editing site j in the input sequence. Let T be an $m \times n$ matrix where $T_{i,j}$ equals the count of edited bases for template i at editing site j. Let us further define row T_0 be the fully edited template, row T_1 be the pre-edited template, and rows $T_2 \ldots T_{m-1}$ be any alternatively edited templates. Let A be an $m \times n$ binary matrix where

$$A_{i,j} = \begin{cases} 1 & if\ T_{i,j} = t_j \\ 0 & \text{otherwise} \end{cases}.$$

Let I be an $m \times n$ unit matrix where every element is equal to one. We compute $A' = A \oplus \mathrm{I}$, the exclusive disjunction (XOR) between A and I. Using A' we can find all editing sites in the input sequence that do not match a given template. Let find (v,x) be a function that returns a vector containing the linear indices of each element in

vector v with value x. The junction start site (JSS) is defined as

$$\text{JSS} = \begin{cases} |A_0| & \text{if } |A_0| = |\,\text{find}(A_0, 1)| \\ -1 & \text{if } 0 = |\,\text{find}(A_0, 1)|\,. \\ \min(\,\text{find}(A_0', 1)) & \text{otherwise} \end{cases}$$

The junction end site (JES) is defined as

$$\text{JES} = \begin{cases} -1 & \text{if } |A_1| = |\,\text{find}(A_1, 1)| \\ |A_1| & \textit{if } 0 = |\,\text{find}(A_1, 1)|\,. \\ \min(\,\text{find}(A_1', 1)) & \text{otherwise} \end{cases}$$

The editing stop site is simply ESS = JSS−1. We further require JES ≥ JSS otherwise the junction region is not defined. The junction length is equal to JES − JSS and when JES = JSS the junction length is zero (i.e., there is no junction). A negative value for the JSS or JES indicates the site potentially falls within a primer region that was trimmed during preprocessing of the input sequence. The additional rows $A_2 \ldots A_{m-1}$ contain alternatively edited templates which are used to potentially shift the JSS depending on the presence of alternative editing and handled as a special case described in the next section.

Detection of alternative editing

If any alternatively edited templates were provided, TREAT will attempt to detect regions of alternative editing during the alignment. Each alternatively edited template requires the start and end editing site numbers to be specified in the FASTA header file. These define the region within the alternative edited templates that TREAT will compare against the junction region of the input sequences. If the junction start site matches the first editing site of the alternative region specified in the alternatively edited template and the edited bases of the junction match the alternatively edited sequence through the full region specified in the FASTA header file with fidelity, the determination of the junction start site will be shifted to the region beyond the alternatively edited region and the sequence is flagged as alternatively edited.

Storing results in a database

TREAT can be configured to store the alignment results in a database for fine-grained analysis. TREAT uses Bolt (https://github.com/boltdb/bolt) a low-level key/value datastore written in Go. Bolt stores data in buckets, which are collections of key/value pairs within the database. A database is represented as a single file on a disk, thus no server is required. TREAT databases can be easily copied, shared, and backed up as they are just files. The schema for TREAT is described in Figure 11. TREAT organizes data by gene and sample. Genes correspond to the RNA transcripts given by the template sequences. Each gene can have one or more samples which contain RNA-seq reads. As shown in Figure 11, the database consists of three buckets: templates, alignments, and fragments. Templates are stored in the templates bucket keyed by gene name. Alignment and fragment data are stored in the alignments and fragments buckets, respectively, keyed by the composite key gene + sample. Searching is performed using Bolt's prefix scans, which allow efficient lookup of keys using a prefix string.

Visualizing results

TREAT includes a built-in web server and provides a robust web-based interface for viewing and analyzing the alignment results. After loading the templates for the gene or genes of interest and their

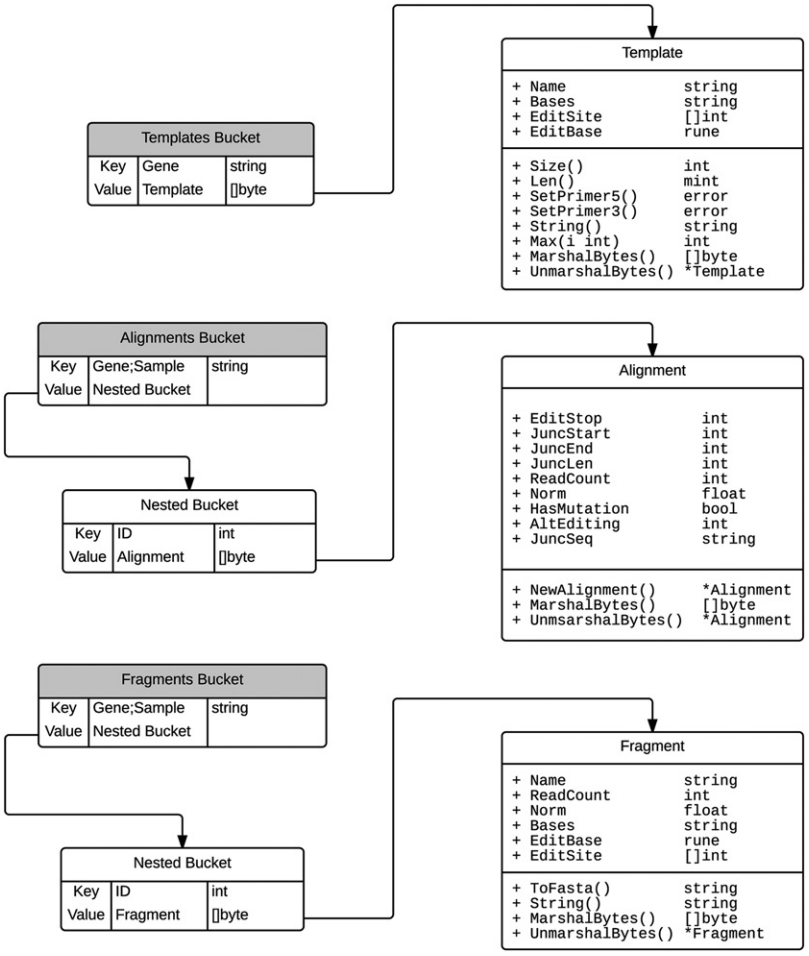

FIGURE 11. TREAT database schema. TREAT uses a key-value database (bolt) for storing alignment results (https://github.com/boltdb/bolt). Data in TREAT are organized into buckets (collections of key/value pairs) by gene and sample. Each gene has a single Template object that captures data on the fully edited, pre-edited, and any alternative edited templates. A serialized form of the Template object is stored in the templates bucket keyed by gene. The alignments and fragments buckets store the alignment and associated raw sequence data keyed by the composite key gene + sample. The value for these buckets is a nested bucket that stores a serialized form of the Alignment and Fragment objects keyed by a unique identifier. Nested buckets allow for the grouping of alignment data by gene and sample for more efficient searching. A TREAT database is represented as a single file stored on disk.

corresponding sample data (RNA-seq reads), TREAT can be run in server mode pointing to a database file or optionally a directory of database files. The web interface will then be accessible using a web browser. The overview page shows the distribution of editing stop sites, junction length, and junction end sites by gene and sample, as shown in Supplemental Figure S1A. The histograms can be dynamically filtered using the search form and data points in the chart are clickable allowing the user to drill down to specific sites. The search results view shown in Supplemental Figure S1B displays detailed information about the editing characteristics for alignments matching the given search criteria. The raw data can be exported in comma separated format (CSV) for viewing in third party applications. The alignment view shown in Supplemental Figure S1C displays the multiple sequence alignment for an individual input sequence.

Determination of intrinsic pause sites

Here, we define an "editing stop site" as the 5′ boundary of a run of fully edited sequence that begins at the 3′ end of the editing domain. The editing stop site itself is defined as the 5′ most correctly edited ES moving from the 3′ to 5′ direction and it is abutted at its 5′ end by the presence of an ES that does not match fully edited sequence. We define an "intrinsic pause site (IPS)" as an editing stop site present in the population in a very high abundance relative to the majority of editing stop sites. To determine whether IPSs were present in our data sets, we examined the normalized number of reads at which full editing stopped for each ES. Using a histogram and the distance between the mean and median of the amplitude of editing stop sites, we determined that the data was sufficiently skewed, not conforming to a bell curve, that it was not normally distributed. Due to this, we could not use a normal distribution to determine which levels of pausing were significant relative to an average. Instead, we posited that if there is a "background" level of pausing, it would remain centralized around a normal distribution and our major peaks may be far enough outside of this distribution to qualify as outliers. We determined outliers as those above the Outlier threshold (oThresh) using the following formula (Crawley 2011):

$$\text{Outlier threshold} = (1.5 \times \text{IQR}) + 3\text{Q},$$

where IQR is the interquartile range and 3Q is the third quartile value. Once calculated, it was determined that multiple sites existed that qualified as outliers in our data set. Thus, we defined that any ES with a level of pausing greater than this threshold is an IPS. The oThresh was used as a marker for determining major peaks of other measures in our data including MJESs and major editing stop sites corresponding to MJESs.

Correlation of editing stop sites with sequence characteristics

To determine whether IPSs correlate with specific sequence characteristics, we used a Fisher's exact test. For a given RNA, we categorized each ES as having either (i) U addition, (ii) U deletion, or (iii) no action based on the action needed to go from pre-edited to canonical fully edited transcripts. We tested for a correlation between the Editing Stop Site and the action taken at the subsequent editing site. Additionally we categorized each ES as being flanked by A, C, or G in both the 5′ and 3′ direction and tested for a correlation between editing stop sites and flanking nucleotides. The Fisher's exact test was done using the default 2×2 test in *R* where the (central or noncentral) hypergeometric distribution is used to determine *P*-values. In this manner, we determined whether the proportion of ES with or without these characteristics differed significantly between the set of intrinsic pause sites and the non-pause sites.

Analysis of junction length

The junction is the region of partially edited RNA found between the fully edited portion at the 3′ end and the pre-edited portion at the 5′ end. RNA sequence in the junction region has undergone editing, but its sequence matches neither fully edited nor pre-edited sequence. The junction length is the number of ES contained within a junction. The average junction length was determined across all sequences where editing stopped at a given editing site. Taking the set of sequences with a common editing stop site, the distribution of junction length was determined. The overall distribution of junction lengths across the population was determined by examining the total number of sequences, regardless of their editing stop site, with a junction of a given length. Although they lack junctions, completely pre-edited RNAs were not included in the totals of RNAs having junction length zero, since the absence of a junction reflects a biologically distinct mechanism in these RNAs compared to RNAs already having begun editing.

Analysis of alternative gRNA utilization

Templates for alternatively edited sequences were generated using the published gRNAs that could code for alternative editing (Madej et al. 2008; Koslowsky et al. 2014; Madina et al. 2014). These sequences can be examined by checking the alternative editing box in the web interface. Additionally, TREAT includes a drop down menu to look at sequences that match each alternatively edited template. Sequences are accessible via this menu only if they match the alternatively edited sequence with full fidelity through the full region covered by the alternative gRNA, thus most likely to represent true usage of these alternative gRNAs.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant RO1 AI061580 to L.K.R. We thank the University at Buffalo Genomics and Bioinformatics Core, especially Sujith Valiyaparambil, for deep sequencing. We are also grateful to Natalie McAdams for critical reading of the manuscript and to Steve Gill for assistance in the early stages of this project.

Received November 4, 2015; accepted January 28, 2016.

REFERENCES

Acestor N, Panigrahi AK, Carnes J, Zikova A, Stuart KD. 2009. The MRB1 complex functions in kinetoplastid RNA processing. *RNA* **15:** 277–286.

Alatortsev VS, Cruz-Reyes J, Zhelonkina AG, Sollner-Webb B. 2008. *Trypanosoma brucei* RNA editing: coupled cycles of U deletion reveal processive activity of the editing complex. *Mol Cell Biol* **28:** 2437–2445.

Ammerman ML, Presnyak V, Fisk JC, Foda BM, Read LK. 2010. TbRGG2 facilitates kinetoplastid RNA editing initiation and progression past intrinsic pause sites. *RNA* **16:** 2239–2251.

Ammerman ML, Hashimi H, Novotna L, Cicova Z, McEvoy SM, Lukes J, Read LK. 2011. MRB3010 is a core component of the MRB1 complex that facilitates an early step of the kinetoplastid RNA editing process. *RNA* **17:** 865–877.

Ammerman ML, Downey KM, Hashimi H, Fisk JC, Tomasello DL, Faktorova D, Kafkova L, King T, Lukes J, Read LK. 2012. Architecture of the trypanosome RNA editing accessory complex, MRB1. *Nucleic Acids Res* **40:** 5637–5650.

Ammerman ML, Tomasello DL, Faktorova D, Kafkova L, Hashimi H, Lukes J, Read LK. 2013. A core MRB1 complex component is indispensable for RNA editing in insect and human infective stages of *Trypanosoma brucei*. *PLoS One* **8:** e78015.

Aphasizhev R, Aphasizheva I. 2011a. Mitochondrial RNA processing in trypanosomes. *Res Microbiol* **162:** 655–663.

Aphasizhev R, Aphasizheva I. 2011b. Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer. *Wiley Interdiscip Rev RNA* **2:** 669–685.

Aphasizhev R, Aphasizheva I. 2013. Emerging roles of PPR proteins in trypanosomes: switches, blocks, and triggers. *RNA Biol* **10:** 1495–1500.

Aphasizhev R, Aphasizheva I, Nelson RE, Gao G, Simpson AM, Kang X, Falick AM, Sbicego S, Simpson L. 2003. Isolation of a U-insertion/deletion editing complex from *Leishmania tarentolae* mitochondria. *EMBO J* **22:** 913–924.

Aphasizheva I, Maslov D, Wang X, Huang L, Aphasizhev R. 2011. Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes. *Mol Cell* **42:** 106–117.

Aphasizheva I, Zhang L, Wang X, Kaake RM, Huang L, Monti S, Aphasizhev R. 2014. RNA binding and core complexes constitute the U-insertion/deletion editosome. *Mol Cell Biol* **34:** 4329–4342.

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. 1986. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46:** 819–826.

Bhat GJ, Koslowsky DJ, Feagin JE, Smiley BL, Stuart K. 1990. An extensively edited mitochondrial transcript in kinetoplastids encodes a protein homologous to ATPase subunit 6. *Cell* **61:** 885–894.

Bilbe G. 2015. Infectious diseases. Overcoming neglect of kinetoplastid diseases. *Science* **348:** 974–976.

Carnes J, Trotter JR, Peltan A, Fleck M, Stuart K. 2008. RNA editing in *Trypanosoma brucei* requires three different editosomes. *Mol Cell Biol* **28:** 122–130.

Carnes J, Soares CZ, Wickham C, Stuart K. 2011. Endonuclease associations with three distinct editosomes in *Trypanosoma brucei*. *J Biol Chem* **286:** 19320–19330.

Cifuentes-Rojas C, Halbig K, Sacharidou A, De Nova-Ocampo M, Cruz-Reyes J. 2005. Minimal pre-mRNA substrates with natural and converted sites for full-round U insertion and U deletion RNA editing in trypanosomes. *Nucleic Acids Res* **33:** 6610–6620.

Corell RA, Myler P, Stuart K. 1994. *Trypanosoma brucei* mitochondrial CR4 gene encodes an extensively edited mRNA with completely edited sequence only in bloodstream forms. *Mol Biochem Parasitol* **64:** 65–74.

Crawley MJ. 2011. *Statistics: an introduction using R*. Wiley, Hoboken.

Cruz-Reyes J, Zhelonkina AG, Huang CE, Sollner-Webb B. 2002. Distinct functions of two RNA ligases in active *Trypanosoma brucei* RNA editing complexes. *Mol Cell Biol* **22:** 4652–4660.

Ernst NL, Panicucci B, Igo RP Jr, Panigrahi AK, Salavati R, Stuart K. 2003. TbMP57 is a 3′ terminal uridylyl transferase (TUTase) of the *Trypanosoma brucei* editosome. *Mol Cell* **11:** 1525–1536.

Etheridge RD, Aphasizheva I, Gershon PD, Aphasizhev R. 2008. 3′ adenylation determines mRNA abundance and monitors completion of RNA editing in *T. brucei* mitochondria. *EMBO J* **27:** 1596–1608.

Feagin JE, Jasmer DP, Stuart K. 1987. Developmentally regulated addition of nucleotides within apocytochrome b transcripts in *Trypanosoma brucei*. *Cell* **49:** 337–345.

Feagin JE, Abraham JM, Stuart K. 1988. Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. *Cell* **53:** 413–422.

Fisk JC, Ammerman ML, Presnyak V, Read LK. 2008. TbRGG2, an essential RNA editing accessory factor in two *Trypanosoma brucei* life cycle stages. *J Biol Chem* **283:** 23016–23025.

Foda BM, Downey KM, Fisk JC, Read LK. 2012. Multifunctional G-rich and RRM-containing domains of TbRGG2 perform separate yet essential functions in trypanosome RNA editing. *Eukaryot Cell* **11:** 1119–1131.

Guo X, Ernst NL, Carnes J, Stuart KD. 2010. The zinc-fingers of KREPA3 are essential for the complete editing of mitochondrial mRNAs in *Trypanosoma brucei*. *PLoS One* **5:** e8913.

Hashimi H, Zikova A, Panigrahi AK, Stuart KD, Lukes J. 2008. TbRGG1, an essential protein involved in kinetoplastid RNA metabolism that is associated with a novel multiprotein complex. *RNA* **14:** 970–980.

Hashimi H, Cicova Z, Novotna L, Wen YZ, Lukes J. 2009. Kinetoplastid guide RNA biogenesis is dependent on subunits of the mitochondrial RNA binding complex 1 and mitochondrial RNA polymerase. *RNA* **15:** 588–599.

Hashimi H, Zimmer SL, Ammerman ML, Read LK, Lukes J. 2013. Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex. *Trends Parasitol* **29:** 91–99.

Hernandez A, Madina BR, Ro K, Wohlschlegel JA, Willard B, Kinter MT, Cruz-Reyes J. 2010. REH2 RNA helicase in kinetoplastid mitochondria: ribonucleoprotein complexes and essential motifs for unwinding and guide RNA (gRNA) binding. *J Biol Chem* **285:** 1220–1228.

Huang Z, Faktorova D, Krizova A, Kafkova L, Read LK, Lukes J, Hashimi H. 2015. Integrity of the core mitochondrial RNA-binding complex 1 is vital for trypanosome RNA editing. *RNA* **21:** 2088–2102.

Jensen RE, Englund PT. 2012. Network news: the replication of kinetoplast DNA. *Annu Rev Microbiol* **66:** 473–491.

Kafkova L, Ammerman ML, Faktorova D, Fisk JC, Zimmer SL, Sobotka R, Read LK, Lukes J, Hashimi H. 2012. Functional characterization of two paralogs that are novel RNA binding proteins influencing mitochondrial transcripts of *Trypanosoma brucei*. *RNA* **18:** 1846–1861.

Kala S, Salavati R. 2010. OB-fold domain of KREPA4 mediates high-affinity interaction with guide RNA and possesses annealing activity. *RNA* **16:** 1951–1967.

Koslowsky DJ, Bhat GJ, Perrollaz AL, Feagin JE, Stuart K. 1990. The MURF3 gene of *T. brucei* contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. *Cell* **62:** 901–911.

Koslowsky DJ, Bhat GJ, Read LK, Stuart K. 1991. Cycles of progressive realignment of gRNA with mRNA in RNA editing. *Cell* **67:** 537–546.

Koslowsky DJ, Reifur L, Yu LE, Chen W. 2004. Evidence for U-tail stabilization of gRNA/mRNA interactions in kinetoplastid RNA editing. *RNA Biol* **1:** 28–34.

Koslowsky D, Sun Y, Hindenach J, Theisen T, Lucas J. 2014. The insect-phase gRNA transcriptome in *Trypanosoma brucei*. *Nucleic Acids Res* **42:** 1873–1886.

Leung SS, Koslowsky DJ. 1999. Mapping contacts between gRNA and mRNA in trypanosome RNA editing. *Nucleic Acids Res* **27:** 778–787.

Leung SS, Koslowsky DJ. 2001a. Interactions of mRNAs and gRNAs involved in trypanosome mitochondrial RNA editing: structure probing of an mRNA bound to its cognate gRNA. *RNA* **7:** 1803–1816.

Leung SS, Koslowsky DJ. 2001b. RNA editing in *Trypanosoma brucei*: characterization of gRNA U-tail interactions with partially edited mRNA substrates. *Nucleic Acids Res* **29:** 703–709.

Li F, Herrera J, Zhou S, Maslov DA, Simpson L. 2011. Trypanosome REH1 is an RNA helicase involved with the 3′-5′ polarity of multiple gRNA-guided uridine insertion/deletion RNA editing. *Proc Natl Acad Sci* **108:** 3542–3547.

Madej MJ, Niemann M, Huttenhofer A, Goringer HU. 2008. Identification of novel guide RNAs from the mitochondria of *Trypanosoma brucei*. *RNA Biol* **5:** 84–91.

Madina BR, Kumar V, Metz R, Mooers BH, Bundschuh R, Cruz-Reyes J. 2014. Native mitochondrial RNA-binding complexes in kinetoplastid RNA editing differ in guide RNA composition. *RNA* **20:** 1142–1152.

Madina BR, Kumar V, Mooers BH, Cruz-Reyes J. 2015. Native variants of the MRB1 complex exhibit specialized functions in kinetoplastid RNA editing. *PLoS One* **10:** e0123441.

Maslov DA, Simpson L. 1992. The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing. *Cell* **70:** 459–467.

McDermott SM, Carnes J, Stuart K. 2015a. Identification by random mutagenesis of functional domains in KREPB5 that differentially affect RNA editing between life cycle stages of *Trypanosoma brucei*. *Mol Cell Biol* **35:** 3945–3961.

McDermott SM, Guo X, Carnes J, Stuart K. 2015b. Differential editosome protein function between life cycle stages of *Trypanosoma brucei*. *J Biol Chem* **290:** 24914–24931.

Militello KT, Read LK. 1999. Coordination of kRNA editing and polyadenylation in *Trypanosoma brucei* mitochondria: complete editing is not required for long poly(A) tract addition. *Nucleic Acids Res* **27:** 1377–1385.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443–453.

Ochsenreiter T, Hajduk SL. 2006. Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. *EMBO Rep* **7:** 1128–1133.

Ochsenreiter T, Anderson S, Wood ZA, Hajduk SL. 2008a. Alternative RNA editing produces a novel protein involved in mitochondrial DNA maintenance in trypanosomes. *Mol Cell Biol* **28:** 5595–5604.

Ochsenreiter T, Cipriano M, Hajduk SL. 2008b. Alternative mRNA editing in trypanosomes is extensive and may contribute to mitochondrial protein diversity. *PLoS One* **3:** e1566.

Panigrahi AK, Schnaufer A, Ernst NL, Wang B, Carmean N, Salavati R, Stuart K. 2003. Identification of novel components of *Trypanosoma brucei* editosomes. *RNA* **9:** 484–492.

Panigrahi AK, Ernst NL, Domingo GJ, Fleck M, Salavati R, Stuart KD. 2006. Compositionally and functionally distinct editosomes in *Trypanosoma brucei*. *RNA* **12:** 1038–1049.

Read LK, Myler PJ, Stuart K. 1992. Extensive editing of both processed and preprocessed maxicircle CR6 transcripts in *Trypanosoma brucei*. *J Biol Chem* **267:** 1123–1128.

Read LK, Wilson KD, Myler PJ, Stuart K. 1994. Editing of *Trypanosoma brucei* maxicircle CR5 mRNA generates variable carboxy terminal predicted protein sequences. *Nucleic Acids Res* **22:** 1489–1495.

Read LK, Lukes J, Hashimi H. 2016. Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip Rev RNA* **7:** 33–51.

Reifur L, Koslowsky DJ. 2008. *Trypanosoma brucei* ATPase subunit 6 mRNA bound to gA6–14 forms a conserved three-helical structure. *RNA* **14:** 2195–2211.

Reifur L, Yu LE, Cruz-Reyes J, Vanhartesvelt M, Koslowsky DJ. 2010. The impact of mRNA structure on guide RNA targeting in kinetoplastid RNA editing. *PLoS One* **5:** e12235.

Rusche LN, Cruz-Reyes J, Piller KJ, Sollner-Webb B. 1997. Purification of a functional enzymatic editing complex from *Trypanosoma brucei* mitochondria. *EMBO J* **16:** 4069–4081.

Rusche LN, Huang CE, Piller KJ, Hemann M, Wirtz E, Sollner-Webb B. 2001. The two RNA ligases of the *Trypanosoma brucei* RNA editing complex: cloning the essential band IV gene and identifying the band V gene. *Mol Cell Biol* **21:** 979–989.

Schmid B, Riley GR, Stuart K, Goringer HU. 1995. The secondary structure of guide RNA molecules from *Trypanosoma brucei*. *Nucleic Acids Res* **23:** 3093–3102.

Schnaufer A, Domingo GJ, Stuart K. 2002. Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA. *Int J Parasitol* **32:** 1071–1084.

Seiwert SD, Stuart K. 1994. RNA editing: transfer of genetic information from gRNA to precursor mRNA in vitro. *Science* **266:** 114–117.

Souza AE, Myler PJ, Stuart K. 1992. Maxicircle CR1 transcripts of *Trypanosoma brucei* are edited and developmentally regulated and encode a putative iron-sulfur protein homologous to an NADH dehydrogenase subunit. *Mol Cell Biol* **12:** 2100–2107.

Souza AE, Shu HH, Read LK, Myler PJ, Stuart KD. 1993. Extensive editing of CR2 maxicircle transcripts of *Trypanosoma brucei* predicts a protein with homology to a subunit of NADH dehydrogenase. *Mol Cell Biol* **13:** 6832–6840.

Stuart K, Allen TE, Heidmann S, Seiwert SD. 1997. RNA editing in kinetoplastid protozoa. *Microbiol Mol Biol Rev* **61:** 105–120.

Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK. 2005. Complex management: RNA editing in trypanosomes. *Trends Biochem Sci* **30:** 97–105.

Stuart K, Brun R, Croft S, Fairlamb A, Gurtler RE, McKerrow J, Reed S, Tarleton R. 2008. Kinetoplastids: related protozoan pathogens, different diseases. *J Clin Invest* **118:** 1301–1310.

Sturm NR, Simpson L. 1990. Partially edited mRNAs for cytochrome b and subunit III of cytochrome oxidase from *Leishmania tarentolae* mitochondria: RNA editing intermediates. *Cell* **61:** 871–878.

Trotter JR, Ernst NL, Carnes J, Panicucci B, Stuart K. 2005. A deletion site editing endonuclease in *Trypanosoma brucei*. *Mol Cell* **20:** 403–412.

Weng J, Aphasizheva I, Etheridge RD, Huang L, Wang X, Falick AM, Aphasizhev R. 2008. Guide RNA-binding complex from mitochondria of trypanosomatids. *Mol Cell* **32:** 198–209.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614–620.

High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing

Rachel M. Simpson, Andrew E. Bruno, Jonathan E. Bard, et al.

RNA 2016 22: 677-695 originally published online February 23, 2016
Access the most recent version at doi:10.1261/rna.055160.115