Published in final edited form as:

Nat Protoc. 2018 August ; 13(8): 1829–1849. doi:10.1038/s41596-018-0020-y.

Identification of RNA-Binding Protein Targets with HyperTRIBE

Reazur Rahman, Weijin Xu, Hua Jin, and Michael Rosbash

Department of Biology, Howard Hughes Medical Institute and National Center for Behavioral Genomics, Brandeis University, Waltham, Massachusetts, 02453, USA.

Abstract

RNA binding proteins (RBPs) accompany RNA from birth to death, affecting RNA biogenesis and functions. Identifying RBP-RNA interactions is essential to understand their complex roles in different cellular processes. However, detecting in vivo RNA targets of RBPs, especially in a small number of discrete cells, has been a technically challenging task. We have previously developed a novel technique called TRIBE (Targets of RNA-binding proteins Identified By Editing) to overcome this problem. TRIBE expresses a fusion protein consisting of a queried RBP and the catalytic domain from RNA editing enzyme ADAR (ADARcd), which marks target RNA transcripts by converting adenosine to inosine near the RBP binding sites. These marks can be subsequently identified via high-throughput sequencing. In spite of its usefulness, TRIBE is constrained by a low editing efficiency and editing-sequence bias from the ADARcd. So, we developed HyperTRIBE by incorporating a previously characterized hyperactive mutation, E488Q, into the ADARcd. This strategy increases the editing efficiency and reduce sequence bias, which dramatically increased sensitivity of this technique without sacrificing specificity. HyperTRIBE provides a more powerful strategy to identify RNA targets of RBPs with an easy experimental and computational protocol at low cost not only in flies but also in mammals. The HyperTRIBE experimental protocol described below can be carried out in cultured *Drosophila* S2 cells in one week using tools available in a common molecular biology laboratory, and the computational analysis requires 3 more days.

EDITORIAL SUMMARY

HyperTRIBE uses a hyperactive RNA editing enzyme fused to an RNA-binding protein (RBP), to mark the target RNA transcripts of the RBP by converting adenosine to inosine near the binding sites with increased efficiency and reduced sequence bias.

Corresponding author: Michael Rosbash, rosbash@brandeis.edu.

AUTHOR CONTRIBUTIONS

R.R., W.X., H.J. and M.R. wrote the manuscript. W.X. and H.J. developed the experimental protocol and wrote the relevant sections of the protocol. R.R. developed software and wrote the bioinformatics sections of the protocol.

COMPETING FINANCIAL INTERESTS

The authors declare that a PCT patent application (PCT patent application no.: PCT/US2016/05425) has been filed based on the HyperTRIBE method described in this paper.

CODE AVAILABILITY

The HyperTRIBE software is available on GitHub at <https://github.com/rosbashlab/HyperTRIBE>. This code repository will be maintained and has been assigned the following DOI: 10.5281/zenodo.1203743.

Keywords

RNA-binding protein; RBP; Cell-specific method; ADAR; CLIP; TRIBE; HyperTRIBE

INTRODUCTION

RNA-binding proteins (RBPs) are a class of proteins that physically bind to RNAs to carry out their intrinsic functions. RBPs are important for regulating many post-transcriptional RNA events, including pre-mRNA splicing, polyadenylation, mRNA export from the nucleus to the cytoplasm, localization, stability and translation^{1–5}. Disruption of RBP function has been associated with many human diseases, e.g., loss of FMRP causes Fragile X syndrome and TDP43 mutations can contribute to Amyotrophic Lateral Sclerosis (ALS)^{6, 7}.

To thoroughly understand the role of RBPs, it is necessary to identify their target transcripts^{8, 9}. However, a limited number of tools are available for this purpose¹⁰. In addition, the cell specificity of RBPs produces additional challenges as different cell types can express different RBPs. Moreover, the same RBP can bind different target transcripts within different cell types. In fact, many single cell sequencing studies have indicated that profound heterogeneity in transcriptional and translational profiles occurs at the single cell level^{11, 12}. Therefore, it would be ideal for targets of RBPs to be identified in a cell-specific manner.

Traditional methods to identify in vivo RBP targets typically rely on immunoprecipitation of the RNA-RBP complex. These experiments are either performed with a cross-linking step beforehand, as in CLIP¹³ (Cross-Linking and ImmunoPrecipitation) and its variants¹⁴ (referred to as CLIP henceforth for simplicity), or without cross-linking, as in RIP¹⁵ (RNA ImmunoPrecipitation). The low efficiency of these antibody-based biochemical procedures necessitate large amounts of starting material (thousands if not millions of cells)¹⁰. Since it is nearly impossible to purify millions of cells of a single type, previous studies have compromised by using tissues with mixed cell types. Consequently, results from these studies face potential problems of low signal:noise and a high false-positive rate^{10, 16}.

These problems become exacerbated when the cells of interest make up a very small part of the tissue e.g., the 16 circadian pacemaker neurons (PDF neurons) buried in one 100,000-cell *Drosophila* brain. To understand mRNA regulation within these circadian neurons, we were interested in developing a tool capable of identifying RBP targets in specific cells.

Development of TRIBE and HyperTRIBE

The successful implementation of DamID¹⁷ and TaDa (Targeted DamID)¹⁸ methods, in which the Dam methyltransferase was fused to a chromatin protein or polymerase to leave methylation marks on target DNA regions, provided some conceptual inspiration to our design. There are several enzymes in living organisms that can leave comparable marks on RNAs, the best characterized of which is ADAR¹⁹ (Adenosine Deaminase Acting on RNA).

We wondered whether fusing an RBP of interest with the ADAR enzyme would work in similar fashion.

The ADAR enzyme is comprised of two functionally independent modules, the double-strand RNA binding motifs (dsRBMs) and the catalytic domain (ADARcd)^{20–22}. The dsRBMs are responsible for target RNA selection and binding, whereas the ADARcd catalyzes the adenosine-to-inosine deamination²⁰. Because of its structural similarity to guanosine, inosines resulting from deamination are recognized as guanosines by cellular enzymes and can be identified as A-to-G mutations by high-through sequencing. In other words, replacing the dsRBMs with the RBP of interest should direct the fusion ADAR enzyme towards the specific target mRNAs of the RBP and leave A-to-G mutation marks on these mRNA transcripts. We call this method Targets of RNA-binding proteins Identified By Eding (TRIBE)²³.

Indeed, TRIBE worked as expected in *Drosophila* with the three different RBPs tested: Hrp48, dFMRP and NonA²³. TRIBE was able to induce a large number of editing sites around the binding sites of the RBP, and no above-background spontaneous editing is observed with ADARcd alone²³. CLIP data and TRIBE data from the same cell context are similar in both the identity of the target mRNAs and the preferred binding region of the RBPs. Most importantly, we successfully performed TRIBE experiments in a small number of specific *Drosophila* neurons (even as scarce as PDF neurons), which was technically impossible to accomplish with CLIP²³.

Nevertheless, TRIBE was not perfect. Even with great sequence depth, Hrp48 TRIBE only identifies approximately 25% of the target mRNAs identified via Hrp48-CLIP, suggesting TRIBE may have a false-negative problem²³. This phenomenon is likely caused by the low editing rate and sequence bias of the ADARcd. It has been found to prefer certain sequence (UAG for human ADAR2 and *Drosophila* ADAR) and a double-stranded structure surrounding the deaminating adenosine^{24–26}. Increasing the editing rate and decreasing the sequence bias of the ADARcd should ameliorate these issues.

Fortunately, Brenda Bass and her colleagues had already come up with a solution – a hyperactive E488Q mutation in the ADARcd conferring faster editing speed and less of a preference for UAG neighboring sequence surrounding the editing site²⁷. We introduced this mutation into the TRIBE construct, producing HyperTRIBE²⁸. In fact, HyperTRIBE dramatically increased the detection sensitivity of this method without sacrificing target specificity, both in cultured cells and in fly neurons²⁸. Indeed, the UAG sequence bias at the editing sites is significantly reduced²⁸. Through these improvements, HyperTRIBE can identify the targets of RBPs with lower sequencing depths and false-negative rates. We therefore contend that HyperTRIBE is a superior version of TRIBE and the go-to approach for RBP target identification in specific cells.

Comparison with other methods

A major competitor of HyperTRIBE is CLIP, the current gold standard method for identifying RBP targets. Generally, there are advantages to each of the two approaches in different situations. To identify cell-specific targets of RBPs, HyperTRIBE is obviously the

better solution because it requires very low amounts of starting material^{23, 28}. However, CLIP is a better tool for investigating the specific binding sequence of the RBP on its target RNAs. This is because CLIP reveals the exact target RNA binding site; this region is protected by the protein-RNA interaction and later sequenced. In HyperTRIBE experiments, the ADARcd can barely access the exact binding sites (due to occupation by the RBP itself) and edits on adenosines surrounding these regions; 72% of the editing sites are within 500nt of CLIP sites²⁸. This is a major limitation of HyperTRIBE, which cannot be circumvented. Advantages of HyperTRIBE include 1) an efficient and simple experimental procedure 2) a simple way to normalize binding to the level of gene expression and 3) cost savings from sequencing libraries with less depth.

Another interesting and recent method for RBP target identification, published at almost the same time as TRIBE, is RNA Tagging developed by the Wickens lab²⁹. This method adopts a strategy similar to TRIBE – fusing the RBP with an enzyme that leaves marks on the target RNAs, with poly (U) polymerase (PUP) in this case³⁰. PUP adds U-tags of varying lengths to the 3' end of the target RNA molecules, which can be identified via RNA-seq²⁹. One major caveat with the RNA Tagging method is that the 3'-polyuridylation might trigger the U-tail-dependent degradation pathway in some circumstances or cell types, confounding the results of a tagging experiment³¹. Another issue is that the library preparation procedure and computational analysis are quite complicated, making it hard to set up de novo. It is also unclear whether RNA tagging can be applied to RBPs that have binding sites far from 3' end of RNA. To our knowledge, this method has only been successfully applied to RBPs in *S. cerevisiae*.

To summarize, HyperTRIBE is simple and accurate in identifying RNA targets of RBPs unless identification of the specific RBP binding sites is desired. Hyper TRIBE has the added advantage of working in a cell type specific manner and requiring a small amount of starting material. In situations where both HyperTRIBE and CLIP are applicable, doing both experiments will be ideal as they will provide strong orthogonal validation of the targets.

Applications

In addition to *Drosophila*, we have been successful in using HyperTRIBE to study several RBPs in cultured mammalian cells (unpublished data from H.J. and W.X.). Theoretically, HyperTRIBE could be applied to any organism and system where induction of fusion protein expression is possible. As for the RBP of interest, HyperTRIBE should be compatible with most RBPs or even RNA-associated proteins without direct RNA contact if they are in close proximity to RNA.

Limitations

Exact binding sites.—As mentioned above, HyperTRIBE cannot identify the exact binding sites on the target RNAs because of the fusion protein spatial organization. That being said, we should be able to locate the approximate binding region with HyperTRIBE, as it is often flanked by clusters of editing sites.

Overexpression.—HyperTRIBE experiments were carried out by overexpression of the fusion protein, either by copper induction of the metallothionein (MT) promoter in cultured cells or using the Gal4-UAS system in *Drosophila*. Since RBPs have important cellular functions, overexpression of queried RBP could lead to non-physiological changes in cell behavior and gene expression profiles. An excess of the RBP might also result in promiscuous binding to non-specific targets after saturation of real RNA targets. The same problem can exist in RNA Tagging, and we are making efforts to address this problem with the CRISPR/Cas9 knock-in strategy.

ADAR editing preference.—Although the neighbor sequence preference is substantially reduced in HyperTRIBE, it is still somewhat biased. For example, a 5' guanosine, the least favored neighbor nucleotide (only found in 2% of all editing sites), would interfere with the ADAR catalytic domain reaching the adenosine²⁵. However, HyperTRIBE does not mark the exact binding sites so this is not a problem. Further reducing the neighbor sequence preference would probably require systematic combination screening of several mutations in the catalytic domain to search for an even less biased version.

RNA stability.—There is possibility that editing by ADAR on the mRNA could lead to change in its stability, which could potentially introduce bias to TRIBE/HyperTRIBE results.

Experimental design

HyperTRIBE construct cloning.—All TRIBE/HyperTRIBE fusion protein constructs that we have generated so far place RBPs at the N-terminus and the dADARcd at the C-terminus (Fig. 1). However, each individual RBP differs in folding and spatial conformation, so there is no reason to believe that the other orientations would not work. Readers are advised to experimentally ensure that fusion to ADARcd does not disrupt the intrinsic binding of the RBP, such as CLIP experiments to compare the targets of RBP-ADARcd and RBP²³. As a background control, a construct of Hyper-dADARcd alone should be cloned in parallel. The constructs for cultured S2 cell expression were cloned into pMT vectors with a MT promoter (pMT-Hrp48-ADARcd-V5), whereas the *Drosophila* transgenic constructs were driven by a 20X UAS promoter (Figure 1). A stringent peptide linker requirement between the RBP and the ADARcd is not necessary, as the fusion protein editing efficiency is only marginally affected by length and properties of the linker²⁸. However, an oversized linker, like the 200 AA linkers tested²⁸, is not recommended as it should give enhanced flexibility to the ADARcd and complicate data interpretation. To allow easy detection by Western blotting, we also attached a V5 tag to the C-terminus of the fusion protein (Fig. 1). In our TRIBE/HyperTRIBE constructs, two small original introns were included in the ADARcd to possibly aid gene expression²³. These introns will be probably problematic in organisms other than *Drosophila* and should therefore be removed for other applications. The step-by-step procedures for plasmid cloning are not provided in this protocol, the readers are recommended to follow the instructions found in the Gibson Assembly® Master Mix manual (<https://www.neb.com/-/media/catalog/datacards-or-manuals/manuale2611.pdf>).

Fusion protein expression.—(i) Stable expression in cultured S2 cells: The pMT plasmids include a blasticidin resistance gene, which can be used as a selection marker for stable cell lines. Blasticidin selection can be initiated 3–4 days after transfection of HyperTRIBE constructs. Although this process can take several weeks, stable cell lines can be stored in liquid nitrogen and conveniently revived later²³.

(ii) Transient expression in cultured S2 cells: We failed to generate effective stable lines for Hrp48 HyperTRIBE despite multiple attempts, perhaps due to cell toxicity of HyperTRIBE excessive editing. We therefore turned to transient expression for HyperTRIBE experiments. Because of the low transfection efficiency of S2 cells (~1% transfection efficiency), we FACS-sorted the successfully transfected cells. Thus, HyperTRIBE constructs were co-transfected with a constitutively expressing EGFP plasmid and were induced with copper 24 hours prior to FACS-sorting. Cells transfected would uptake both plasmids if they are thoroughly mixed. For cells with high transfection efficiency, FACS-sorting is not necessary. The transient expression strategy is described in detail in the Procedure.

(iii) Tissue-specific expression in *Drosophila*: In vivo HyperTRIBE constructs were placed under UAS promoter control, transgenic flies were generated, which can be crossed to flies with a tissue-specific *Gal4* driver to induce expression in specific tissues. For our proof of principle HyperTRIBE experiment²³, we used a pan-neuronal and drug inducible *Elav-gsg-Gal4* driver to prevent potential developmental defects caused by early-stage expression of HyperTRIBE. These flies also contain a UAS-EGFP transgene driven by the same *Elav-gsg-Gal4* driver, which labels all the HyperTRIBE expressing cells with green fluorescence. We chose to collect the GFP-positive neurons with an established manual-sorting protocol³². Because of the limited number of neurons that we could collect manually, we performed extra amplification of their RNA content before standard RNA-seq library preparation. However, a *Gal4* driver that expresses in more cells, or the use of FACS sorting, should allow the assay of larger number of cells with simple TRIzol RNA extraction.

RNA-seq library preparation.—For RBPs expected to bind RNA exons, we used standard Illumina® TruSeq RNA library preparation v2 kit and protocol to generate our RNA-seq libraries. For RBPs that possibly bind to introns, Nascent-seq should be used for target identification^{23, 33}. If the RBP of interest is believed to bind to long non-coding RNAs or non-polyadenylated RNAs, preparing an RNAseq library with ribosomal RNA depletion may be an option. Readers can follow protocols provided by rRNA depletion kits available on the market from various suppliers. For quality control and quantification, we chose the Agilent bioanalyzer 2100 using DNA 1000 or DNA high-sensitivity chips because of its ability to examine the size of the library fragments. DNA 1000 chips are designed for quantifying higher concentration libraries, such as the undiluted sequencing libraries. However, the DNA high-sensitivity chips are designed for quantifying low concentration ones, such as the sequencing libraries diluted for sequencing, typically at 2nM concentration. The quantified libraries were then sequenced on Illumina® NextSeq 500 platform with High Output v2 kits (75 cycles). For HyperTRIBE experiments, we pooled together 20–24 libraries with different barcodes and sequenced them on one flow-cell, which generated ~20 million reads for each library. We picked this lower sequence depth based on two considerations: 1) Our Hrp48 HyperTRIBE data indicate that 20 million reads are

adequate for identifying a similar number of targets compared to CLIP; 2) The 200 million reads coverage adopted in the original TRIBE paper may prevent many labs from trying this technique from a financial standpoint.

Genomic DNA library preparation.—Genomic DNA from *Drosophila* or cultured cells with the same genetic background should be prepared using the following protocol³⁴ and constructed into gDNA sequencing libraries³⁵.

Overview of bioinformatics analysis.—The bioinformatics software for HyperTRIBE compares RNA-seq libraries from a HyperTRIBE construct with a control genomic DNA (gDNA) library or a wild type mRNA library (wtRNA) from the same genetic background to identify a precise set of nucleotides that have been edited by HyperTRIBE. After filtering the sites for potential single nucleotide polymorphisms and background editing, all RBP target transcripts are identified.

The bioinformatics analysis of HyperTRIBE begins with quality control of the sequenced DNA-seq and RNA-seq libraries. The reads from these libraries are trimmed to remove 6 nucleotides from their 5' end because of potential sequencing errors due to random hexamer mispriming³⁶. The reads are filtered to eliminate low quality reads and trimmed to remove low quality bases from either 5'- or 3'- end of the reads. RNA-seq libraries are then mapped to the transcriptome using STAR, allowing a maximum of 5 mismatches in 75 nucleotide reads to accommodate multi-editing events by HyperTRIBE. DNA-seq libraries are mapped to the genome using bowtie2 with parameters “--sensitive” (Figure 1). The resulting alignment files are filtered to remove low quality alignments, as they are more likely to have errors and therefore introduce noise into the dataset. Next, we computationally remove PCR duplicates using aligned reads, because PCR duplicates originate from sequencing two or more copies of the same DNA/RNA molecule and may introduce bias into the editing percentage of HyperTRIBE data. Computational methods to remove PCR duplicates start by identifying multiple reads that map to the same exact location based on the 5' coordinate with the same strand orientation. The read with the highest sequence quality score is then retained and the others discarded. This approach is not optimal as it risks discarding many reads originating from different molecules that just happen to terminate at the same positions. We therefore recommend using library preparation kits with unique molecular indices (UMIs), which can differentiate thousands of molecules with the identical sequence.

Once PCR duplicates are removed, we convert the alignments in SAM format into “matrix” files, which record the frequency for each type of nucleotide from aligned reads at each position in the genome. These files are then uploaded to a MySQL database so that nucleotide distribution at each position in the genome for each sequenced library can be easily queried and subsequently compared to identify the editing sites. A perl script is used to call RNA editing sites where the frequency of A is greater than 80% and the frequency of G is zero in the gDNA (or wtRNA) and the frequency of G is greater than 0 in the RNA (using the reverse complement if annotated gene is in the reverse strand). In order to remove potential noise from sequencing errors, we require each editing site to have at least 10% editing and a minimum of 20 reads at that site. Background editing and single nucleotide polymorphism are removed by subtracting the editing sites found in Hyper-ADARcd alone.

HyperTRIBE experiments should be performed with at least two replicates so that transient editing sites can be filtered out by only considering sites that are present in all replicates (Figure 1). All editing sites are then pooled to create a list of transcripts that are potential targets of the RBP. Transcripts that are bound more stably by queried RBP may have multiple editing sites²⁸.

More recently, we have discovered that a RNA-seq library prepared from wild type mRNA of the same genetic background can be used as reference to identify editing instead of gDNA library. This provides significant cost savings since genomic DNA must be sequenced at high depth. We tested this idea by assuming that editing sites identified in a HyperTRIBE RNA to gDNA (gDNA-RNA) comparison are 100% accurate. We then compared these sites with those identified in the HyperTRIBE RNA to wild type RNA (wtRNA-RNA) comparison. This second approach was able to identify editing sites with a sensitivity (TP/(TP+FN)) of 90% and specificity (TP/(TP+FP)), where TP is true positive, FP is false positive, and FN is false negative. of 95% (Figure 2), which demonstrates its broad applicability. The non-overlapping editing sites between the two approaches are likely due to lack of sequence coverage or nucleotide polymorphisms. The limitation of the wtRNA-RNA comparison is that it cannot be used to find endogenous editing sites. This alternate strategy uses the same computational software as above with only minor modifications; we provide options for using either gDNA-RNA or wtRNA-RNA as a reference to identify editing in the Procedure (Steps 78–84).

Downstream analysis.—Downstream analysis of HyperTRIBE results will depend on the specific goals of the experiments. HyperTRIBE provides a list of editing sites along with a list of genes that are targets of the RBP of interest. Bedtools is a powerful toolkit that is capable of analyzing multiple aspects of the editing data. For example, it can analyze the distribution of editing sites among different mRNA regions, i.e., 3'UTR, 5'UTR or coding sequence (CDS), by intersecting the editing sites with the annotation files of the corresponding regions. The enrichment of editing sites in these regions reflect the binding specificity of the RBP. Further, the distribution of editing sites obtained from HyperTRIBE can be compared with distribution of RNA-seq reads among different mRNA regions to rule out potential bias from sequencing reads. We used the online tool Venny 2.1 to examine the overlap among the lists of edited genes and to generate the corresponding venn diagrams. The editing sites can also be visualized as tracks in the Integrative Genomics Viewer (IGV) or University of California, Santa Cruz (UCSC) Genome Browser along with RNA expression data, predicted regulatory motifs and conservation tracks. Furthermore, HyperTRIBE results can be compared with CLIP results by using bedtools to examine the intersection of the gene lists produced from the two methods.

Software requirements.—The computational pipeline for HyperTRIBE is available on GitHub (<https://github.com/rosbashlab/HyperTRIBE>) along with instructions on its installation and use. This software is an updated version of our previously released software called TRIBE and is designed to work seamlessly in Linux and Unix-like operating systems, including Mac OS X. The software uses shell scripting, Perl and Python, all which are preinstalled in a Linux-based operating system. In case you need to install them, please

download perl at <https://www.perl.org/get.html> and Python at <https://www.python.org/downloads/>. The software uses some commonly used linux based genomics tools, which include bedtools (<http://bedtools.readthedocs.io/>), and samtools (<http://samtools.sourceforge.net/>). We use trimmomatics (<http://www.usadellab.org/cms/?page=trimmomatic>) for trimming low quality bases from both ends of the reads before genome alignment. The program STAR (<https://github.com/alexdobin/STAR>) is used to align reads to the transcriptome and bowtie2 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>) to align reads to the genome. MySQL database is used to store nucleotide frequency from mapped reads at each position in the genome and transcriptome. MySQL database installation files can be downloaded from <https://downloads.mariadb.org/>. Basic familiarity with command line tools and making simple edits to the shell scripts using a text editor, like the built-in text editor nano, are assumed.

MATERIALS

REAGENTS

! CAUTION All cell culture experiments must be performed under certified hood with BSL2 or higher. Reagents with environmental hazard should be disposed in accordance with regulations.

- Gibson Assembly® Master Mix (New England Biolabs, E2611L)
- pMT-Hrp48-ADARcd-V5 (Addgene, Plasmid #81172)
- pAc5.1B-EGFP (Addgene, Plasmid #21181)
- Restriction endonucleases (NotI-HF, New England Biolabs, R3189S; KpnI-HF, New England Biolabs, R3142S)
- Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, F530S or F530L)
- SFX-Insect cell culture media (HyClone, SH30278.02)
- Fetal Bovine Serum (Gibco, 10082147)
- Antibiotic-Antimycotic (Gibco, 15240062)1
- Cellfectin II Reagent (Invitrogen, 10362100)
- Copper(II) sulfate (Sigma-Aldrich, C1297–100G)
- 1M Tris pH 7.0 (Ambion, AM9850G)
- 1M Tris pH 8.0 (Ambion, AM9855G)
- 0.5M EDTA pH 8.0 (Ambion, AM9260G)
- cOmplete™, Mini Protease Inhibitor Cocktail (Roche, 04693159001)
- NuPAGE™ 4–12% Bis-Tris Protein Gels, 1.0 mm, 10-well (Invitrogen, NP0321BOX)
- NuPAGE™ MOPS SDS Running Buffer (20X) (Invitrogen, NP0001)

- Monoclonal Anti-V5 antibody produced in mouse (Sigma, V8012–50UG)
- Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare, RPN2232)
- TRIzol LS Reagent (Ambion, 10296010)

! CAUTION TRIzol is a serious health hazard. Use in a fume hood with suitable eye protection and gloves.

- Chloroform, Certified ACS grade (Fisher Scientific, C298–500)

! CAUTION Chloroform fumes are hazardous to the skin, eyes and airways. It should be handled in a fume hood with appropriate protective equipment, including eye protection and gloves.

- Isopropanol, Molecular Biology Grade (Fisher Scientific, BP2618–1)
- Ethanol, 200 proof (100%) (v/v), (Decon Laboratories, Inc., 2716)

! CAUTION Ethanol is highly flammable.

- GlycoBlue™ Coprecipitant (Thermo Fisher Scientific, AM9516)
- TruSeq RNA Library Preparation Kit v2 (Illumina, RS-122–2001)
- Agencourt AMPure XP beads (Beckman Coulter, A63880)
- SuperScript™ II Reverse Transcriptase (Thermo Fisher Scientific, 18080093)
- Agilent DNA 1000 Kit (Agilent, 5067–1504)
- Agilent High Sensitivity DNA Kit (Agilent, 5067–4626)
- NextSeq 500/550 High Output v2 kit, 75 cycles (Illumina, FC-404–2005)
- S2 cell line (Invitrogen, R69007)

! CAUTION The cell line should be regularly checked to ensure they are authentic and are not infected with mycoplasma.

EQUIPMENT

Nalgene™ Rapid-Flow™ Sterile Disposable Filter Units with CN Membrane (Thermo Fisher Scientific, 450–0045)

Nalgene™ 25mm Syringe Filters, 0.2µm (Thermo Fisher Scientific, 723–2520)

Falcon Test Tube with Cell Strainer Snap Cap (Corning, 352235)

HyBlot CL® Autoradiography Film, 5 × 7 inch (Denville Scientific Inc., E3012 (1001364))

Nonstick, RNase-free Microfuge Tubes, 1.5 ml (Ambion, AM12450)

BD FACSAria II (BD Biosciences)

Thermomixer R (Eppendorf)

NanoDrop ND-1000 (Thermo Fisher Scientific)

Agilent 2100 Bioanalyzer (Agilent)

NextSeq 500 (Illumina)

Software

Any Linux/Unix environment

HyperTRIBE software: <https://github.com/rosbashlab/HyperTRIBE>

Perl (5.8.8, 5.12.5, 5.22.1): <https://www.perl.org/get.html>

Python (2.7.2 or later): <https://www.python.org/downloads/>

Perl Module: DBI.pm (1.631, 1.636) and DBD:mysql (4.042)

bedtools suite (v. 2.16.2, 2.26.0 or later): <http://bedtools.readthedocs.io/>

samtools (1.3.1): <http://samtools.sourceforge.net/>

Trimmomatic (0.36): <http://www.usadellab.org/cms/?page=trimmomatic>

STAR(2.5.2b): <https://github.com/alexdobin/STAR>

bowtie2 (2.1.0, 2.2.9): <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

MySQL (MariaDB v10.1 or later): <https://downloads.mariadb.org/>

REAGENT SETUP

S2 cell culture media: Mix 445 ml of SFX-Insect cell culture media with 50 ml of heat-inactivated FBS and 5 ml of Antibiotic-Antimycotic, then filter through a Nalgene™ Rapid-Flow™ Sterile Disposable Filter. Store at 4 °C for up to 6 months.

0.5 M copper sulfate solution: Filter the solution through a Nalgene™ 25mm syringe filter. Store at 4 °C for up to 1 year.

TE buffer, pH 7.5—(10 mM Tris, 1mM EDTA): To make a 50 ml solution of TE buffer, mix 150 µl of 1M Tris pH 8.0, 350 µl of 1M Tris pH 7.0, 100 µl of 0.5M EDTA pH 8.0, and 49.4 ml of nuclease-free water. Store at 25 °C for up to 1 year.

RIPA buffer: 10 mM Tris pH 7.5, 150 mM NaCl, 1% (v/v) NP-40, 0.5% (w/v) sodium deoxycholate, 0.1% (w/v) SDS and 2 mM EDTA, dissolve in ddH₂O; add 1 tablet of protease inhibitor to 10 ml of RIPA buffer right before use. Store at 25 °C for up to 1 year.

5× SDS sample buffer: 0.05% (w/v) bromophenol blue, 5% (w/v) beta-mercaptoethanol, 30% (v/v) glycerol, 10% (w/v) SDS (sodium dodecyl sulfate), 0.25 M Tris-HCl (pH 6.8). Store at −20 °C for up to 6 months.

PROCEDURE

Transiently transfecting S2 Cells with HyperTRIBE plasmids ● TIMING 2 days

1. Plate 1.5×10^6 S2 cells/well in 6-well format in 2 ml of S2 cell culture media. Incubate cells at 25 °C overnight.
2. For each sample, prepare transfection mix by following Steps 2–4. Firstly, add 0.8 µg of pActin-eGFP (pAc5.1B-EGFP) and 0.8 µg of pMTA-HyperTRIBE into 100 µl of plain SFX-insect media without FBS and antibiotics. Vortex briefly to mix.

CRITICAL STEP: The transient transfection efficiency of S2 cells is low, so it is necessary to co-transfect the HyperTRIBE construct with a pActin-eGFP plasmid into S2 cells and sort GFP-positive cells by FACS.

3. Vortex Cellfectin® II before use and add 8 µl into 100 µl of plain SFX-insect media without FBS and antibiotics. Vortex briefly to mix.
4. Combine the solutions from Steps 2 and 3 to a total volume of 200 µl. Vortex briefly to mix and incubate for 30 minutes at room temperature, which should be between 20 °C to 25 °C. The solution may appear cloudy.
5. Remove the growth medium from the cells from Step 1 and wash with plain SFX-insect media once.
6. Add 0.8 ml of plain SFX-insect media to the mixture from Step 4, mix gently by pipetting and add to the cells from Step 5. Incubate cells at 25 °C for 5 hours.
7. Remove the transfection mixture with vacuum and replace with 2 ml of S2 cell culture media. Incubate cells at 25 °C for 24 hours.
8. Induce expression of HyperTRIBE protein by adding 1:1000 volume of 0.5 M copper sulfate to the medium to reach a final concentration of 500 µM. Incubate cells at 25 °C overnight. To verify expression of the TRIBE construct, protein samples can be collected at this point and assayed with western blotting (Box 1). This step can be skipped after verification at the first experimental attempt.

?TROUBLESHOOTING

Sorting eGFP-positive S2 cells by FACS ● TIMING 5 hours

9. Wash the transfected S2 cells with 1×PBS and resuspend the cells with 1 ml of 1×PBS.
10. Add 0.5 ml of 1×PBS to the cells and harvest them into 1.5 ml tubes by scraping and pipetting. Keep cells at room temperature before sorting.
 - ▲ CRITICAL STEP Pipette cells several times to resuspend them as single cells.
11. Filter cells using a 35 µm cell strainer cap to remove cell aggregates.
12. FACS sort the cells for eGFP and proper size: follow instructions in BD FACS Aria II user guide. Gate FITC parameter (GFP intensity) at larger than 10^3 , appropriate FSC-A to SSC-A range (Figure 3) and collect cells directly into 0.75

ml of TRIzol LS. More than 10,000 cells are recommended for RNA-seq library preparation. Homogenize the sample by pipetting, then keep on dry ice.

- ▲ **CRITICAL STEP** Use wild type S2 cells as a negative control to check the background level of FITC.
- **PAUSE POINT** The collected samples can be stored at -80°C for up to a year.

RNA isolation using TRIzol LS ● TIMING 4 hours—! CAUTION Use TRIzol in a chemical hood. Always use gloves and eye protection and avoid contact with skin or clothing and breathing vapor. All pipette tips and tubes with TRIzol contact must be collected in a designated container and disposed of according to regulations.

▲ **CRITICAL** RNA isolation (Steps 13–27) follows supplier's (Invitrogen) instructions with minor modifications, as detailed in the following steps. Use DNase and RNase-free materials in Steps 13–60.

13. If the samples were stored at -80°C , thaw them at room temperature. Add nuclease-free water to the samples to adjust final volume to 1 ml.
14. Vortex the samples thoroughly and incubate them at room temperature for 5 minutes to allow complete dissociation of the nucleoprotein complex.
15. Add 0.2 ml of chloroform to each sample. Cap sample tubes securely.
16. Vortex samples vigorously for 20 seconds and incubate them at room temperature for 5 minutes to allow phase separation.
17. Centrifuge the samples at $12,000 \times g$ for 15 minutes at 4°C . Following centrifugation, the mixture separates into a lower red phenol-chloroform phase, an interphase, and a colorless upper aqueous phase. RNA remains exclusively in the aqueous phase.
18. Carefully transfer $\sim 500\ \mu\text{l}$ of upper aqueous phase into 1.5 ml nonstick microfuge tubes without disturbing the interphase.
 - ▲ **CRITICAL STEP** Contamination from other phases causes degradation of RNA.
19. Add 0.5 ml isopropyl alcohol and 1 μl Glycoblue to the collected aqueous phase. Mix by inverting the tubes several times. Incubate samples at -80°C for >2 hours to increase final RNA yields.
 - ▲ **CRITICAL STEP** Incubating samples at -80°C increases the precipitation efficiency of RNAs but also results in more salts in precipitates.
20. Centrifuge at $12,000 \times g$ for 30 minutes at 4°C . RNA precipitate often forms a blue pellet on the side and bottom of the tube.
21. Discard supernatant completely by pipetting carefully without disturbing the pellet.

22. Wash the RNA pellet by adding 1 ml of 75% (v/v) ethanol to the tubes and mix the samples by inverting. Centrifuge at $12,000 \times g$ for 5 minutes at 4 °C and remove supernatant.
23. Repeat Step 22 once more and remove ethanol.
24. Centrifuge at $12,000 \times g$ for 5 minutes at 4 °C. Remove all leftover ethanol by pipetting. Be careful not to remove the RNA pellets.
25. Air-dry the RNA pellets for 3–5 minutes.
▲ **CRITICAL STEP** It is important not to let the RNA pellet dry completely as this will greatly decrease its solubility. Partially dissolved RNA samples have an A260/A280 ratio < 1.6.
26. Dissolve RNA in 20 µl RNase-free TE buffer (pH 7.5) or RNase-free ddH₂O. Incubate the tubes at room temperature for 10 minutes at 1,400 r.p.m. in a Thermomixer.
27. Measure RNA concentration with NanoDrop. The A260/A280 ratio should be ~2.0. **PAUSE POINT:** RNA samples can be stored in –80°C freezer **for six months**.

?TROUBLESHOOTING

Preparing RNA-seq libraries ● TIMING ~2 days—CRITICAL: Steps 28–58 are adapted from TruSeq RNA Library Preparation Kit v2 manual (Illumina, RS-122–2001) using one-third reaction volumes. Visit the following website for detailed step-by-step protocol: (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqrna/truseq-rna-sample-prep-v2-guide-15026495-f.pdf).

28. *Poly A select mRNA (Steps 28–40).* Use 50–150 ng of total RNA as the starting material. Adjust the volume to 16.67 µl with ddH₂O.
29. Vortex RNA Purification Beads and add 16.67 µl to each RNA sample. Mix until beads are in a homogenous suspension.
30. Incubate in a thermocycler: 65 °C for 5 minutes, then hold at 4 °C.
31. Place samples at room temperature for 5 minutes, then place them on a magnetic rack for 5 minutes, remove and discard all supernatant.
32. Add 66.7 µl of Bead Washing Buffer, remove from magnetic rack, and mix by pipetting. Place the samples on the magnetic rack for 5 minutes. Remove and discard the supernatant.
33. Add 16.67 µl of Elution Buffer, remove from magnetic rack, and mix well by pipetting.
34. Incubate in thermocycler at 80 °C for 2 minutes, then hold at 25 °C.
35. Remove samples from the thermocycler when they reach 25 °C and keep at room temperature.

36. Add 16.67 μ l of Bead Binding Buffer and mix well by pipetting. Incubate at room temperature for 5 minutes, then place on magnetic rack for 5 minutes. Remove and discard all supernatant.
37. Repeat the wash in step 32.
38. Add 6.5 μ l of Elute, Prime, Fragment Mix and homogenize.
39. Incubate in the thermocycler at 94 °C for 8 minutes, then hold at 4 °C.
40. Place samples in the magnetic rack for 5 minutes. Transfer 5.67 μ l of the supernatant to a new 0.2 ml PCR tube.
41. *First strand synthesis (Steps 41–42)*. Add 2.67 μ l of First Strand Master Mix/ Super Script II mix to each sample.
42. Incubate in the thermocycler: 25 °C for 10 minutes, 42 °C for 50 minutes, 70 °C for 15 minutes and hold at 4 °C.
43. *Second strand synthesis (Steps 43–46)*. Add 8.33 μ l of Second Strand Master Mix to each sample.
44. Incubate in the thermocycler at 16 °C for 1 hour.
45. Remove samples from the thermocycler and allow them to warm to room temperature.
46. Purify with 30 μ l of Ampure XP Beads and elute with 20 μ l of Resuspension Buffer. Each round of Ampure XP Beads purification is performed as following: Mix Ampure Beads with sample by pipetting and incubate for 15 minutes at 25°C; Pellet beads by placing the mixture on magnetic racks and pipette out the supernatant; Keep samples on magnetic racks and wash the beads pellet with 200 μ l of 80% (v/v) ethanol for two times; To elute the samples, remove samples from racks and add indicated volume of resuspension buffer, thoroughly mix by pipetting, incubate for 2 minutes at 25°C; Place samples back on magnetic racks for 5 minutes and collect indicated amount of supernatant for following steps.
47. *End repair (Steps 47–49)*. Add 13.3 μ l of End Repair Mix to each sample.
48. Incubate at 30 °C for 30 minutes.
49. Purify with 30 μ l of Ampure XP Beads as described in Step 46 and elute with 5.83 μ l of Resuspension Buffer.
50. *Add 'A' bases to 3' ends (Steps 50–51)*. Add 4.17 μ l of A-Tailing Mix to each sample.
51. Incubate at 37 °C for 30 minutes, 70 °C for 5 minutes and hold at 4 °C.
52. *Ligate adapters (Steps 52–55)*. To each sample, add 0.83 μ l of DNA Ligase Mix, 0.83 μ l of Resuspension Buffer and 0.83 μ l of RNA Adapter Index.

▲ CRITICAL STEP Most library preparation kits provide multiple sample indices to add barcodes into different samples that can be used to demultiplex

sequencing reads when multiple samples are included in the sequencing run. For example, TruSeq kit has 24 different barcodes which means at most 24 samples can be given a unique barcode and ran in the same sequencing run. Successful demultiplexing requires different indices for each different sample.

53. Incubate at 30 °C for 10 minutes.
54. Add 1.67 µl of Stop Ligase Mix immediately after ligation.
55. Purify twice with Ampure XP Beads following instructions in Step 46: Purify with 14 µl of Ampure XP Beads and elute with 16.67 µl of Resuspension Buffer. Purify with 16.67 µl of Ampure XP Beads and elute with 6.67 µl of Resuspension Buffer.
56. *Amplify library by PCR (Steps 56–58).* Add PCR mix and primers to the 6.67 µl of Adapter-ligated DNA from step 55: 1.67 µl of PCR Primer Cocktail and 8.33 µl of PCR Master Mix provided by TruSeq library preparation kit.
57. Amplify with the following PCR protocol:

Initial denaturation	98 °C for 30 seconds	
Denature	98 °C for 10 seconds	Repeat for 5–18 cycles
Anneal	60 °C for 30 seconds	
Extend	72 °C for 30 seconds	
Final extension	72 °C for 5 minutes	

▲ **CRITICAL STEP** Using fewer PCR cycles will decrease PCR duplication in sequencing results. If the recommended amount of total RNA is used for library preparation, 12 cycles should provide sufficient amplification.

58. Purify with 16.67 µl of well-mixed AMPure XP beads following Step 46 and elute with 10 µl of Resuspension Buffer. **PAUSE POINT:** Libraries can be stored safely in –20 freezer for more than a year.

?TROUBLESHOOTING

Sequencing RNA-seq libraries ● TIMING 12 hours

59. Quantify RNA-seq libraries with Agilent 2100 Bioanalyzer system using Agilent DNA 1000 Kit or Agilent High Sensitivity DNA Kit (Figure 4). For detailed protocol, visit the following websites: https://www.agilent.com/cs/library/usermanuals/Public/G2938-90014_DNA1000Assay_KG.pdf
60. https://www.agilent.com/cs/library/usermanuals/Public/G2938-90321_SensitivityDNA_KG_EN.pdf

Perform High-throughput RNA-seq using the NextSeq 500 (Illumina) system with the NextSeq 500/550 High Output v2 kit. For detailed protocol, refer to user guides provided by Illumina.

Software installation and data retrieval ● TIMING 4 h

- 61.** *Install software for data analysis.* Download the software from github to a desired location in your linux machine:

```
cd /path_from_root/desired_location

git clone https://github.com/rosbashlab/HyperTRIBE
```

HyperTRIBE software is the updated version of the previously released TRIBE software package. It uses the latest version of publicly available tools, and includes some bug fixes. Should you encounter any potential bugs, please use the “Issues” tab on our github page for the software to report them.

Documentation for the HyperTRIBE software is available here <http://hypertribe.readthedocs.io/en/latest/index.html>.

▲ **CRITICAL STEP** HyperTRIBE has the following software dependencies (the tested software version is in brackets): 1. Trimmomatics (0.36); 2. Bowtie2 (2.1.0, 2.2.9); 3. STAR (2.5.2b); 4. Perl (5.8.8, 5.12.5, 5.22.1); 5. Perl Module: DBI.pm (1.631, 1.636) and DBD:mysql (4.042); 6. MySQL database (MySQL, MariDB); 7. Python (2.7.2 or later); 8. bedtools suite (v. 2.16.2, 2.26.0 or later); 9. samtools (1.3.1); 10. Operating systems tested: RHEL 5.11 and RHEL 7.2. HyperTRIBE is likely to work seamlessly with the latest version of these software as well.

- 62.** *Install MySQL database and perl modules.* We recommend using MariaDB, which is an open source version of MySQL database, but other flavors of MySQL will work just as well. Here are some useful instructions for setting up MariaDB on Ubuntu: (<http://hypertribe.readthedocs.io/en/latest/mariadb.html>). Once root password is set up using `mysql_secure_installation`, log in to MySQL using your root password, using the following command:

```
mysql -h localhost -u root -p
```

▲ **CRITICAL STEP** If MySQL is installed on a different machine, then replace “localhost” with the IP address of that machine.

- 63.** Create a MySQL database which will be used by HyperTRIBE software by using the following command:

```
CREATE DATABASE dmseq;
```

▲ **CRITICAL STEP** The scripts “load_matrix_data.pl” and “find_rnaeditsites.pl” assume that a MySQL database called “dmseq” has already been created. If you want to create a different database for this purpose, update the scripts to change the default value of the “\$database” variable on line 23 and line 24 in respective perl scripts.

- 64.** After exiting MySQL, install MariaDB-devel using the following command:

```
yum install mariadb-devel
```

- 65.** Install the perl modules cpan or cpanm using the following command unless they are not pre-installed in your operating system:

```
cpan DBI
cpan DBD:mysql
```

- 66.** Edit the perl scripts “load_matrix_data.pl” and “find_rnaeditsites.pl” to update the mysql- related variables that are needed to communicate with the MySQL database tables. To do so, provide MySQL username (in this case “root”), password and name of database, which was created earlier (Steps 62–63), in the perl scripts load_matrix_data.pl (line 22–25) and find_rnaeditsites.pl (line 23–26). If the MySQL database is hosted on a different machine, change the “\$host” variable from “localhost” to the IP address of that machine.

- 67.** (Optional) Download a sample dataset of five sequencing libraries needed for analysis of HyperTRIBE experiment from NCBI GEO: GSE102814. This dataset is used to demonstrate the workflow of HyperTRIBE computational analysis for both gDNA-RNA and wtRNA-RNA approaches. SRA Toolkit is used to download the libraries based on their SRA accession number, which is subsequently converted to fastq format. The SRA accessions for the sequence libraries along with identifiers are: 1) S2 Genomic DNA: SRR3177714; 2) S2 WT mRNA: SRR6426146; 3) Hrp48 HyperTRIBE Replicate 1: SRR5944748; 4) Hrp48 HyperTRIBE Replicate 2: SRR5944749; 5) HyperADARcd alone: SRR5944750.

Download the SRA file using the following command:

```
prefetch SRR3177714
```

- 68.** Convert SRA to fastq format using the following command:

```
fastq-dump SRR3177714.sra
```

This produces a fastq file called SRR3177714.fastq.

- 69.** Rename the file using the following command:

```
mv SRR3177714.fastq S2_gDNA.fastq
```

- 70.** Repeat Steps 67–69 for the other files to produce these fastq files: S2_wtRNA.fastq, HyperTRIBE_rep1.fastq, HyperTRIBE_rep2.fastq, and HyperADARcd_rep1.fastq

Download annotations and genome sequence ● TIMING 1 h

- 71.** *Download annotations for the transcriptome and create bowtie/STAR indices for the reference genome.* To download refseq annotation from UCSC Genome Browser (<https://genome.ucsc.edu/index.html>), use the following parameters for

Drosophila (dm6) as an example: Tools:TableBrowser; clade=Insect; genome=D.melanogaster; assembly=dm6; group:Genes and Gene Predictions; track=RefSeq Genes; table=refFlat; output format=all field from selected tables; output file: choose a filename (dm6_refFlat.txt).

▲ CRITICAL STEP: This file is provided as part of HyperTRIBE code distribution as an example of required annotation file for *Drosophila* (dm6).

▲ CRITICAL STEP: HyperTRIBE requires transcriptome annotation in two formats (RefSeq annotation in refFlat format from UCSC genome browser and gene transfer format (GTF)) and genome sequence in fasta format. These files should be downloaded by the user based on the organism and genome build of interest.

72. Optional: To obtain the refseq annotation file if the “Refseq Genes” track is not available as an option in UCSC Genome Table Browser, download the refseq annotation for human genome, for example, from UCSC Genome Browser (<https://genome.ucsc.edu/index.html>) using the following parameters: Tools:TableBrowser; clade=Mammal; genome=Human; assembly=GRCh38/hg38; group:Genes and Gene Predictions; track=NCBI RefSeq; table=RefSeq Curated; output format=all field from selected tables; output file: choose a filename (hg38_refseq.txt). Since the organization of the columns in this file is different from refFlat format, rearrange the columns using the following command:

```
awk '{print $13"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9"\t"$10"\t"$11}' hg38_refseq.txt > hg38_ncbi_refseq_curated.txt
```

73. Download refseq annotation in GTF format from Illumina’s Igenome page (https://support.illumina.com/sequencing/sequencing_software/igenome.html). This file can alternatively be downloaded from UCSC browser using the instructions from Step 71 with one small change, choose output file format as “GTF – gene transfer format”.

▲ CRITICAL STEP: This file is also provided for *Drosophila* (dm6) in HyperTRIBE code distribution with the name “genes.gtf”.

74. Download the genome sequence in fasta format from UCSC genome browser or from any other appropriate place.
75. Create bowtie2 indices for the genome using the following command:

```
cd /location_of_genome/
bowtie2-build genome.fa genome
```

76. Create STAR indices for the transcriptome using the following command:

```
STAR --runThreadN 4 --runMode genomeGenerate --genomeDir
star_indices --genomeFastaFiles genome.fa --sjdbGTFfile genes.gtf
```

The genes.gtf file downloaded in Step 73 is used during the creation of STAR indices. The star indices are created at “/location_of_genome/star_indices”.

Prepare shell scripts to run HyperTRIBE ● TIMING 30 mins

77. Update Shell Scripts to reflect the location of software, annotation and genome indices. Open the shell scripts with a text editor like nano or emacs and update the following lines of code with the location of HyperTRIBE code, annotation files, Bowtie2 and STAR indices. Edit the file “trim_and_align.sh” to update these variables:

```
star_indices="/path_from_root/star_indices"
TRIMMOMATIC_JAR="/path_from_root/trimmomatic.jar"
PICARD_JAR="/path_from_root/picard.jar"
```

Edit the file “trim_and_align_gDNA.sh” to update these variables:

```
bowtie_indexes="/path_from_root/genome"
TRIMMOMATIC_JAR="/path_from_root/trimmomatic.jar"
```

Edit the file “load_table.sh” to update this variable:

```
HyperTRIBE_DIR="/path_from_root/HyperTRIBE/CODE"
```

Trim and align sequence libraries ● TIMING 4–24 h

78. Trim low quality bases from either ends of reads in genomic DNA libraries (fastq files), and then align to reference genome. The alignments are sorted by using samtools. Run the “trim_and_align_gDNA.sh” shell script for gDNA library:

```
nohup /path_from_root/HyperTRIBE/CODE/trim_and_align_gDNA.sh
S2_gDNA.fastq &
```

The output of this shell script is “S2_gDNA.sort.sam” which records the alignment to the genome in SAM format. **CRITICAL STEP:** This step is not required for users who are only interested in doing wild type mRNA to HyperTRIBE RNA comparison in order to identify editing sites.

79. Similarly, trim low quality bases from either ends of reads in RNA libraries (fastq files), and then align to reference transcriptome. The alignments are sorted by using samtools. Run the “trim_and_align.sh” shell script for each RNA library using the following command:

```
nohup /path_from_root/HyperTRIBE/CODE/trim_and_align.sh
S2_wtRNA.fastq &
```


The output of this shell script is “S2_wtRNA.sort.sam” which records the alignment to the transcriptome in SAM format. This script also uses Picard to remove PCR duplicates. Repeat this step for the other RNA libraries.

?TROUBLESHOOTING

Load alignments to MySQL ● TIMING 3–8 h

80. Convert alignment in SAM format to a matrix format, where nucleotide frequency at each position in the transcriptome/genome is recorded from aligned reads. This matrix file is then uploaded to a MySQL table based on the provided parameters. Use the shell script “load_table.sh” with the the following command:

```
nohup /path_from_root/]HyperTRIBE/CODE/load_table.sh sam_filename
mysql_tablename expt_name replicate/timepoint &
```

The “load_table.sh” script requires four parameters: 1. SAM file name 2. MySQL table name 3. Experiment name (unique identifier for the sequence library; include alphabets and digits and keep the size to a few characters) 4. An integer reflecting replicate/time-point. The combination of MySQL table name, experiment name, and replicate/time-point integer should be unique for each library. It allows us to deposit the matrix file and subsequently retrieve relevant information from MySQL table very quickly and easily.

Example usage for S2 gDNA-seq library:

```
nohup /path_from_root /HyperTRIBE/CODE/load_table.sh S2_gDNA.sort.sam
s2_gDNA s2_gDNA 25 &
```

Example usage for a RNA-seq library:

```
nohup /path_from_root/HyperTRIBE/CODE/load_table.sh
S2_wtRNA.sort.sam testRNA rnalibs 2 &
```

Repeat this for the other RNA libraries with relevant arguments, for example:

Hrp48_HyperTRIBE_rep1.sort.sam testRNA rnalibs 3; Hrp48_HyperTRIBE_rep2.sort.sam testRNA rnalibs 4; HyperADARcd_rep1.sort.sam testRNA rnalibs 3

▲ **CRITICAL STEP** Carefully note the last three arguments for all the load_table.sh runs for each library because these values are needed in the next step of the analysis. For simplicity, we use the same tablename and experiment name for each RNA library; users are free to choose the desired names.

?TROUBLESHOOTING

81. (Optional) Confirm that all the matrix files have been uploaded to MySQL properly. Edit the perl script “diagnose_mysql_table.pl” on (line 21–25) to provide mysql username (in this case “root”), password and name of database which were created in Steps 62–63.

```
perl diagnose_mysql_table.pl -t testRNA wtRNA.matrix
HyperTRIBE_repl.matrix HyperTRIBE_repl.matrix H
yperADARcd_repl.matrix
```

This script calculates the total number of entries in the given MySQL table and ensures that it is equal to total number of lines of all the matrix files used for that table. If these two numbers do not match, delete the MySQL table and repeat Step 80.

Find RNA editing sites ● TIMING 1–2 h

82. Identify RNA editing sites using either gDNA-RNA (option A) or wtRNA-RNA (option B) approaches.

- A. Identify RNA editing sites using gDNA-RNA approach.

- i. Genomic DNA nucleotide frequency is used as reference against HyperTRIBE RNA nucleotide frequency to identify editing sites. For each position in the transcriptome, the RNA nucleotide distribution is retrieved using MySQL tablename, experiment name and replicate/timepoint integer and is compared with the gDNA nucleotide frequency at the same position to determine the editing sites. To do this, firstly copy the script “rnaedit_gDNA_RNA.sh” to your working directory:

```
cd /directory_of_choice/
cp /path_from_root/HyperTRIBE/CODE/
rnaedit_gDNA_RNA.sh.
```

- ii. Edit the following variables in “rnaedit_gDNA_RNA.sh” by using a text editor like “nano”. Based on how the data is processed in Step 80, it should look like this (line 3, 8–13):

```
HyperTRIBE_DIR="/path_from_root/HyperTRIBE/CODE"
annotationfile="/path_from_root/HyperTRIBE/
annotation/dm6_refFlat.txt"
gDNAtablename="s2_gDNA"
gDNAexp="s2_gDNA"
gDNAtp="25"
RNAtablename="testRNA"
RNAexp="rnalibs"
timepoint=(2 3 4 5)
```

The timepoint array allows the program to loop over the unique integer for replicate/time-point, allowing multiple editing scripts to run on multiple libraries in a convenient way.

- iii. Run the updated shell script from current directory:

```
nohup ./rnaedit_gDNA_RNA.sh &
```

This shell script runs a perl script called “find_rnaeditsites.pl”, which does a pairwise comparison of RNA against gDNA for each nucleotide in the transcriptome to detect a set of editing sites. It then runs the python script “Threshold_editsites_20reads.py” to ensure that the editing sites are required to have at least 10% editing and at least a coverage of 20 reads. The output for this shell script is a list of editing sites in bedgraph format for each pairwise comparison. In this case there will be four bedgraph files with editing sites for: 1) S2_wtRNA: rnalibs_25_2_A2G.bedgraph; 2) HyperTRIBE_rep1: rnalibs_25_3_A2G.bedgraph; 3) HyperTRIBE_rep2: rnalibs_25_4_A2G.bedgraph; and 4) HyperADARcd_rep1: rnalibs_25_5_A2G.bedgraph

?TROUBLESHOOTING

a. Identify RNA editing sites using wtRNA-RNA approach.

- i. For each position in the transcriptome, the HyperTRIBE RNA nucleotide frequency is compared with the wtRNA nucleotide frequency to determine the editing sites. To do this, firstly copy rnaedit_wtRNA_RNA.sh to your working directory:

```
cd /directory_of_choice/
cp /path_from_root/HyperTRIBE/CODE/rnaedit_wtRNA_RNA.sh.
```

- ii. Edit the following variables in “rnaedit_wtRNA_RNA.sh” using a text editor. Based on how we processed the data in Step 80, it should look like this (line 3, 8–13):

```
HyperTRIBE_DIR="/path_from_root/HyperTRIBE/CODE"
annotationfile="/path_from_root/HyperTRIBE/annotation/
dm6_refFlat.txt"
wtRNAtablename="testRNA"
wtRNAexp="rnalibs"
wtRNAtp="2"
RNAtablename="testRNA"
RNAexp="rnalibs"
timepoint=(3 4 5)
```

- iii. Run the updated shell script from current directory:

```
nohup ./rnaedit_wtRNA_RNA.sh &
```

This shell script runs a perl script called “find_rnaeditsites.pl”, which does a pairwise comparison of wtRNA against RNA for each nucleotide

in the transcriptome to call a set of editing sites. It then runs a python script “Threshold_editsites_20reads.py” to ensure that the editing sites are required to have at least 10% editing and at least a coverage of 20 reads. The output for this shell script is a list of editing sites in bedgraph format for each pairwise comparison, in this case there will be three bedgraph files with editing sites for: 1) HyperTRIBE_rep1: rnalibs_2_3_A2G.bedgraph; 2) HyperTRIBE_rep2: rnalibs_2_4_A2G.bedgraph; and 3) HyperADARcd_rep1: rnalibs_2_5_A2G.bedgraph.

?TROUBLESHOOTING

83. Review the bedgraph output files that record the edited sites and the transcripts that are the targets of the RBP. The editing sites listed in these files can be visualized on IGV. Description of column headers in the bedgraph files are provided below: 1) Chr name 2) Start coordinate 3) End coordinate 4) Editing percentage 5) Concatenation of editing percentage and reads 6) Chr name 7) Edit Coordinate 8) Name 9) Type 10) A count 11) T count 12) C count 13) G count 14) Total nucleotide count 15) A count from gDNA/wtRNA 16) T count from gDNA/wtRNA 17) C count from gDNA/wtRNA 18) G count from gDNA/wtRNA 19) Total count from gDNA/wtRNA 20) Editbase Count 21) Total nucleotide count 22) Editbase count from gDNA/wtRNA 23) Total nucleotide count from gDNA/wtRNA 24) Identifier for edit site.

?TROUBLESHOOTING

Post-processing of editing outputs ● TIMING 1 h

84. Create a high confidence set of HyperTRIBE editing sites where the sites are present in both replicates, and background editing sites are removed using Option A for the gDNA-RNA approach or Option B for the wtRNA-RNA approach.

- A. Create a high confidence set of HyperTRIBE editing sites for gDNA-RNA approach.

- i. Use bedtools intersect to find the overlap between two HyperTRIBE replicates, using the following command:

```
bedtools intersect -wa -wb -f 0.9 -r -a
rnalibs_25_3_A2G.bedgraph -b
rnalibs_25_4_A2G.bedgraph > present_both.bedgraph
```

- ii. Remove background (S2 wtRNA) editing sites, using the following command:

```
bedtools intersect -wa -v -f 0.9 -r -a
present_both.bedgraph -b rnalibs_25_2_A2G.bedgraph
> temp.bed
```

- iii. Remove HyperADARcd editing sites, using the following command:

```
bedtools intersect -wa -v -f 0.9 -r -a temp.bed -b
rnalibs_25_5_A2G.bedgraph >
HyperTRIBE_1_2_gDNA.bedgraph
```

These editing sites can be visualized on IGV (Figure 5).

?TROUBLESHOOTING

- iv. Create a list of transcripts that are marked by editing and summarize editing results, using the following command:

```
perl /path_from_root/HyperTRIBE/CODE/summarize_results.pl
HyperTRIBE_1_2_gDNA.bedgraph > HyperTRIBE_results_gDNA.xls
```

There are six columns in the final output file: 1. Gene name; 2. Number of editing sites present in gene; 3. Average Editing percentage; 4. Edit percentage and reads concatenated string separate by “,” for each replicate; 5. Gene feature (EXON/INTRON) separated by comma for each replicate; 6. Concatenated identifier for editing sites.

- B. Create a high confidence set of HyperTRIBE editing sites for wtRNA-RNA approach.
 - i. Use bedtools to find the overlap between two HyperTRIBE replicates, using the following command:

```
bedtools intersect -wa -wb -f 0.9 -r -a
rnalibs_2_3_A2G.bedgraph -b rnalibs_2_4_A2G.bedgraph >
present_both_wtRNA.bedgraph
```

- ii. Remove HyperADARcd editing sites, using the following command:

```
bedtools intersect -wa -v -f 0.9 -r -a
present_both_wtRNA.bedgraph -b rnalibs_2_5_A2G.bedgraph >
HyperTRIBE_1_2_wtRNA.bedgraph
```

These editing sites can be visualized on IGV (Figure 5).

?TROUBLESHOOTING

- iii. Create a list of transcripts that are marked by editing and summarize editing results, using the following command:

```
perl /path_from_root/HyperTRIBE/CODE/summarize_results.pl
HyperTRIBE_1_2_wtRNA.bedgraph > HyperTRIBE_results_wtRNA.xls
```

There are six columns in the final output file: 1. Gene name; 2. Number of editing sites present in gene; 3. Average Editing percentage; 4. Edit percentage and reads concatenated string separate by “,” for each replicate; 5. Gene feature (EXON/INTRON) separated by comma for each replicate; 6. Concatenated identifier for editing sites.

85. Optional: To determine the binding preference of the RBP, use RSeQC (<http://rseqc.sourceforge.net/>) to calculate the distribution of reads in different RNA regions, which can be compared to the distribution of editing sites in these regions.

?TROUBLESHOOTING

Troubleshooting advice can be found in Table 1.

● TIMING

Day 1–2—Steps 1–8, Transiently transfecting S2 Cells with HyperTRIBE plasmids: 2 days

Day 3—Steps 9–12, Sorting eGFP-positive S2 cells by FACS: 5 hours

Steps 13–27, RNA isolation using TRIzol LS: 4 hours

Day 4–5—Steps 28–58, Preparing RNA-seq libraries: ~2 days

Day 6—Steps 59–60, Sequencing RNA-seq libraries: 12 hours

Bioinformatics analysis—The time required for the analysis will vary greatly based on the size of the dataset and size of the transcriptome.

Day 1—Steps 61–70, Software installation and data retrieval: 4 h

Step 71–76, Download annotations and genome sequence: 2 h

Step 77, Prepare shell script to run HyperTRIBE: 30 mins

Step 78–79, Trim and align sequence libraries: 4 –24 h

Day 2—Steps 80–81, Load alignments to MySQL: 3–8 h

Steps 82–83, Find RNA edit sites by using either gDNA-RNA or wtRNA-RNA approaches: 1–2 h

Day 3—Steps 84–85, Post-Processing of editing outputs: 2 h

ANTICIPATED RESULTS

After western blot validation of HyperTRIBE fusion protein expression in the desired cells (Box 1), for example S2 tissue culture cells or specific cells in *Drosophila*, RNA-seq libraries can then be prepared from these cells. When using a FACS machine to sort for

cultured S2 cells, we gated on FITC-A to select for GFP positivity and FSC-A with SSC-A to select for objects with a specific cell size (Figure 3). We can normally recover more than 10,000 cells from one well of transfected S2 cells in 6-well format with these criteria and extract ~300 ng of total RNA from them, which is enough for one library preparation reaction.

A successful library generated with Illumina TruSeq kits v2 should center at ~300 bp without any sharp peaks on Bioanalyzer chips (Figure 4). Concentration of the libraries may vary between samples, but does not strongly indicate library quality. However, it is crucial to determine the exact concentration of each sample before multiplexing them for the sequencing run.

In a good library, 85% of the sequencing reads should uniquely map to the genome, which indicates an effective mRNA selection outcome without rRNA contamination. For Hrp48 HyperTRIBE in S2 cells, we were able to identify 10,689 reproducible editing sites in 3085 genes with two replicates (Figure 5)²⁸.

For the *Drosophila* transcriptome of size 30.1 million bases, we found that RNA-seq libraries with read size of 75 bases and approximately 20 million mapped reads before PCR duplicate removal (5 million mapped reads after PCR duplicate removal) gave consistent and reproducible results. Using the Lander/Waterman equation, the estimated required coverage is 50 before PCR duplicates are removed and 12.5 after PCR duplicates are removed.

ACKNOWLEDGEMENTS

We thank Kate Abruzzi and Jennifer Sherk for their helpful comments on improving the manuscript. We thank Joseph Rodriguez for his contribution in developing some of the scripts that are now part of the HyperTRIBE software. This work was supported by the Howard Hughes Medical Institute and by a NIH EUREKA grant (DA037721).

REFERENCES

1. Gerstberger S, Hafner M & Tuschl T A census of human RNA-binding proteins. *Nat Rev Genet* 15, 829–845 (2014). [PubMed: 25365966]
2. Glisovic T, Bachorik JL, Yong J & Dreyfuss G RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 582, 1977–1986 (2008). [PubMed: 18342629]
3. Witten JT & Ule J Understanding splicing regulation through RNA splicing maps. *Trends Genet* 27, 89–97 (2011). [PubMed: 21232811]
4. Zhao J, Hyman LE & Moore C Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol.Mol.Biol.Rev* 63, 405–445 (1999). [PubMed: 10357856]
5. Jensen TH, Patricio K, McCarthy T & Rosbash M A block to mRNA nuclear export in *S. cerevisiae* leads to hyperadenylation of transcripts that accumulate at the site of transcription. *Molecular cell* 7, 887–898 (2001). [PubMed: 11336711]
6. Ito D, Hatano M & Suzuki N RNA binding proteins and the pathological cascade in ALS/FTD neurodegeneration. *Sci Transl Med* 9 (2017).
7. Maziuk B, Ballance HI & Wolozin B Dysregulation of RNA Binding Protein Aggregation in Neurodegenerative Disorders. *Front Mol Neurosci* 10, 89 (2017). [PubMed: 28420962]
8. Darnell JC et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146, 247–261 (2011). [PubMed: 21784246]

9. Garzia A, Morozov P, Sajek M, Meyer C & Tuschl T PAR-CLIP for Discovering Target Sites of RNA-Binding Proteins. *Methods Mol Biol* 1720, 55–75 (2018). [PubMed: 29236251]
10. Darnell RB HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley interdisciplinary reviews. *RNA* 1, 266–286 (2010). [PubMed: 21935890]
11. Hrvatin S et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci* 21, 120–129 (2018). [PubMed: 29230054]
12. Artegiani B et al. A Single-Cell RNA Sequencing Study Reveals Cellular and Molecular Dynamics of the Hippocampal Neurogenic Niche. *Cell Rep* 21, 3271–3284 (2017). [PubMed: 29241552]
13. Ule J, Jensen K, Mele A & Darnell RB CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* 37, 376–386 (2005). [PubMed: 16314267]
14. Hafner M et al. PAR-CLIP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp* 2, (41). 2034 (2010) doi: 10.3791/2034.
15. Gilbert C & Svejstrup JQ RNA immunoprecipitation for determining RNA-protein associations in vivo. *Curr Protoc Mol Biol* Chapter 27, Unit 27 24 (2006).
16. Lambert N et al. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular cell* 54, 887–900 (2014). [PubMed: 24837674]
17. van Steensel B & Henikoff S Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 18, 424–428 (2000). [PubMed: 10748524]
18. Southall TD et al. Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells. *Dev Cell* 26, 101–112 (2013). [PubMed: 23792147]
19. Lehmann KA & Bass BL Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry* 39, 12875–12884 (2000). [PubMed: 11041852]
20. Macbeth MR et al. Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science* 309, 1534–1539 (2005). [PubMed: 16141067]
21. O'Connell MA et al. Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Molecular and cellular biology* 15, 1389–1397 (1995). [PubMed: 7862132]
22. Kim U, Wang Y, Sanford T, Zeng Y & Nishikura K Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proceedings of the National Academy of Sciences of the United States of America* 91, 11457–11461 (1994). [PubMed: 7972084]
23. McMahon AC et al. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* 165, 742–753 (2016). [PubMed: 27040499]
24. Bass BL & Weintraub H An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* 55, 1089–1098 (1988). [PubMed: 3203381]
25. Matthews MM et al. Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. *Nat Struct Mol Biol* 23, 426–433 (2016). [PubMed: 27065196]
26. Eggington JM, Greene T & Bass BL Predicting sites of ADAR editing in double-stranded RNA. *Nature communications* 2, 319 (2011).
27. Kuttan A & Bass BL Mechanistic insights into editing-site specificity of ADARs. *Proceedings of the National Academy of Sciences of the United States of America* 109, E3295–3304 (2012). [PubMed: 23129636]
28. Xu W, Rahman R & Rosbash M Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein. *Rna* 24, 173–182 (2018). [PubMed: 29127211]
29. Lapointe CP, Wilinski D, Saunders HA & Wickens M Protein-RNA networks revealed through covalent RNA marks. *Nat Methods* 12, 1163–1170 (2015). [PubMed: 26524240]
30. Kwak JE & Wickens M A family of poly(U) polymerases. *Rna* 13, 860–867 (2007). [PubMed: 17449726]
31. Lim J et al. Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell* 159, 1365–1376 (2014). [PubMed: 25480299]
32. Abruzzi K, Chen X, Nagoshi E, Zadina A & Rosbash M RNA-seq Profiling of Small Numbers of *Drosophila* Neurons. *Methods Enzymol* 551, 369–386 (2015). [PubMed: 25662465]

33. Khodor YL et al. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes Dev* 25, 2502–2512 (2011). [PubMed: 22156210]
34. Huang AM, Rehm EJ & Rubin GM Quick Preparation of Genomic DNA from *Drosophila*. *Cold Spring Harbor Protocols* 2009, pdb.prot5198 (2009).
35. Rodriguez J, Menet JS & Rosbash M Nascent-seq indicates widespread cotranscriptional RNA editing in *Drosophila*. *Molecular cell* 47, 27–37 (2012). [PubMed: 22658416]
36. van Gurp TP, McIntyre LM & Verhoeven KJF Consistent Errors in First Strand cDNA Due to Random Hexamer Mispriming. *PloS one* 8, e85583 (2014).
37. Mahmood T & Yang P-C Western blot: Technique, theory, and trouble shooting. *North American Journal of Medical Sciences* 4, 429–434 (2012). [PubMed: 23050259]

Box 1. Western blot verification of protein expression ● TIMING 2 days

CRITICAL: The transfection efficiency in S2 cells is low. Alternatively, you can perform Western blotting after FACS. In that case, collect GFP-positive cells in 100 μ l of 2 $\times$ RIPA buffer and keep on dry ice after collecting.

1. To collect protein samples from the copper-induced cells, wash transfected S2 cells with 1 $\times$ PBS, add 200 μ l of RIPA buffer to each well, resuspend the cells in RIPA buffer by pipetting or scraping, and transfer to a new 1.5 ml Eppendorf tube. Keep on dry ice for 10 minutes. **PAUSE POINT:** Protein samples can be frozen in -80 freezers for future experiments.
2. Centrifuge the protein samples at $20,000 \times g$ for 5 minutes at 4°C and transfer 160 μ l of supernatant to a new tube. Add 40 μ l of 5 $\times$ SDS sample buffer to the supernatant, mix by pipetting, and boil samples at 100°C for 5 minutes.
3. Keep the samples at room temperature to cool down, centrifuge at $20,000 \times g$ for 5 minutes at 4°C and load 10–15 μ l of the supernatants on SDS-PAGE gels. Visualize proteins by Western blotting³⁷.

Figure 1. Overview of the HyperTRIBE protocol.

Genetic scheme for building the HyperTRIBE *in vivo* expression construct. The HyperTRIBE construct is expressed in cultured cells (Steps 1–8) with MT promoter induced by Cu^{2+} (magenta ellipse) or in specific tissues in *Drosophila* with UAS promoter induced by Gal4 (orange ellipse). In cultured cells, HyperTRIBE construct is cotransfected with an actin-promoter driven GFP plasmid, while UAS driven GFP is co-activated by Gal4 in specific tissues in *Drosophila*. In this protocol, we only provide procedures for expressing HyperTRIBE construct in cultured cells. The cells expressing the construct together with GFP are sorted by FACS (Steps 9–12) and their transcriptomes are used to prepare RNA-Seq libraries (Steps 28–58). The locations of the nucleotide where the transcripts are edited are marked by I, showing the adenosine to inosine editing by HyperTRIBE construct. Sequencing reads from these libraries are aligned to the transcriptome (Step 78–79), compared to gDNA or wild-type mRNA to identify editing sites and calculate editing percentage (Steps 82–83). We require high confidence editing sites to be present in at least two replicates (Step 84).

Figure 2. Venn diagram shows the comparison of two methods used to identify HyperTRIBE editing sites.

The editing sites are identified by comparing genomic DNA with HyperTRIBE RNA (gDNA-RNA), or by comparing wild type RNA with HyperTRIBE RNA (wtRNA-RNA). 90% of gDNA-RNA sites can be retrieved by the wtRNA-RNA approach with a specificity of 95%.

Figure 3. FACS sorting parameters for selecting eGFP-positive S2 cells.

The cell selection was performed with BD FACSAria II. Laser extension was set at 2 and FSC Area scaling at 0.6. (A) The first sorting gate was set with FSC-A on X-axis and SSC-A on Y-axis, both with linear scale, to select for the particles that have the size of S2 cells and normal internal complexity. Wild-type S2 cells were used to determine the gate and kept for the following experiments. (B) The second gate selects for singlets which uses SSC-H as the X-axis and SSC-W as the Y-axis, both with linear scale. High SSC-W population was discarded. (C) The final gate was set on log scaled GFP signal intensity and SSC-A was simply included for easier data visualization. The cutoff GFP intensity was determined by analyzing wild-type S2 cells to ensure that the threshold is right above the maximum auto-fluorescence from S2 cells. The P1–4 population was the only cell population collected for the following experiments. These criteria select about 1% of the sorted S2 cells, but the actual number varies from experiments due to transfection efficiency fluctuation. Specific parameters may be examined and altered when setting up the experiment for the first time, and the sorting effectiveness should always be verified before the actual sorting by test sorting of samples and viewing by fluorescence microscopy.

Figure 4. Bioanalyzer plot showing recommended RNA-seq library fragment distribution.

The fragment peak should be centered at about 300 bp without any ancillary peaks. For quantification, the entire peak region should be selected. When the peak height dramatically exceeds the reference marker peaks (often on high sensitivity chips), the quantification can be inaccurate. The sample then needs to be diluted and quantified again for best sequencing outcome. Refer to the troubleshooting section for possible solutions if multiple peaks are observed.

Figure 5. HyperTRIBE fusion protein reproducibly edits more sites compared to TRIBE and has a higher overlap with CLIP signals.

The top two panels show wild type mRNA expression and a CLIP profile for a sample gene, CG11357. The panels directly below show (from top to bottom): RNA editing tracks from wild type cells, cells with HyperADARcd alone, high confidence sites from Hrp48 TRIBE, Hrp48 HyperTRIBE replicate 1 and replicate 2, which are followed by high confidence editing sites from Hrp48 HyperTRIBE at the bottom. The HyperTRIBE data were generated by expressing Hrp48 HyperTRIBE with the elav-geneswitch-Gal4 driver and manually sorting for eGFP-positive cells. The criteria of 20 reads and 10% editing percentage were enforced to select for these editing sites. The black bars in the editing track indicate editing events and the height of the bars shows editing percentage at that site. Editing sites reproduced in at least two replicates with no failures are designated high confidence sites.

Table 1.

Troubleshooting table

Step	Problem	Possible reason	Solution
8	Low transfection efficiency	Inappropriate DNA and cellfectin ratio; Poor S2 cell condition	Try different ratio of DNA and cellfectin to find the optimal condition. Guarantee 50%-70% S2 cell confluency at the time of transfection.
27	Nanodrop peak at 270 nm	Phenol contamination	Perform an extra round of ethanol wash. When estimating RNA concentration, take into account of the phenol signal.
58	Extra peaks on Bioanalyzer chips	Primer dimer contamination or beads carry-over	Primer dimers can be removed by an extra round of Ampure beads purification. Bead carry-over can be addressed by placing sample on magnetic rack again and carefully collecting the liquid.
79	Error “./trim_and_align.sh: Permission denied”. Other shell scripts may produce this error	Shell script does not have execute privileges	chmod +x trim_and_align.sh This makes the shell script executable. Repeat this for other scripts if necessary.
80	Error in database creation: DBD::mysql::st execute failed	Unable to create MySQL table	This error message occurs when load_table.sh tries to create the same MySQL table for multiple datasets. This error can be ignored.
82	Editing on HyperTRIBE RNA exists based on visualization of alignment file but is not being identified	Not enough read coverage in gDNA/wtRNA	Increase the coverage of library by sequencing more. The minimum coverage of set is set as 9 bases in the perl script (find_rnaeditites.pl) for gDNA/wtRNA.
83	Many editing sites are 100% edited	The HyperTRIBE mRNA library has large number of SNPs compared to gDNA/wtRNA due to background variation	Remove the editing sites which have approx. 100% editing. Also, consider re-running the analysis with gDNA/wtRNA sequenced from the same background.
84	Editing site or calculated editing percentage is not as expected	Undetermined	Upload the alignment in SAM/BAM format on IGV and visually confirm the result.