Plonski *et al. BMC Bioinformatics* 2020, **21**(Suppl 18):578
https://doi.org/10.1186/s12859-020-03888-6

BMC Bioinformatics

SOFTWARE

Open Access

Automated Isoform Diversity Detector (AIDD): a pipeline for investigating transcriptome diversity of RNA-seq data

Noel-Marie Plonski[1,2], Emily Johnson[1], Madeline Frederick[1], Heather Mercer[1,3], Gail Fraizer[1,2], Richard Meindl[2,4], Gemma Casadesus[1,2,5,6] and Helen Piontkivska[1,2,5*]

From 8th Workshop on Computational Advances in Molecular Epidemiology (CAME 2019) Niagara Falls, NY, USA. 07 September 2019

*Correspondence:
opiontki@kent.edu
[1] Department of Biological Sciences, Kent State University, 256 Cunningham Hall, Kent, OH 44242, USA
Full list of author information is available at the end of the article

Abstract

Background: As the number of RNA-seq datasets that become available to explore transcriptome diversity increases, so does the need for easy-to-use comprehensive computational workflows. Many available tools facilitate analyses of one of the two major mechanisms of transcriptome diversity, namely, differential expression of isoforms due to alternative splicing, while the second major mechanism—RNA editing due to post-transcriptional changes of individual nucleotides—remains under-appreciated. Both these mechanisms play an essential role in physiological and diseases processes, including cancer and neurological disorders. However, elucidation of RNA editing events at transcriptome-wide level requires increasingly complex computational tools, in turn resulting in a steep entrance barrier for labs who are interested in high-throughput variant calling applications on a large scale but lack the manpower and/or computational expertise.

Results: Here we present an easy-to-use, fully automated, computational pipeline (Automated Isoform Diversity Detector, AIDD) that contains open source tools for various tasks needed to map transcriptome diversity, including RNA editing events. To facilitate reproducibility and avoid system dependencies, the pipeline is contained within a pre-configured VirtualBox environment. The analytical tasks and format conversions are accomplished via a set of automated scripts that enable the user to go from a set of raw data, such as fastq files, to publication-ready results and figures in one step. A publicly available dataset of Zika virus-infected neural progenitor cells is used to illustrate AIDD's capabilities.

Conclusions: AIDD pipeline offers a user-friendly interface for comprehensive and reproducible RNA-seq analyses. Among unique features of AIDD are its ability to infer RNA editing patterns, including ADAR editing, and inclusion of Guttman scale patterns for time series analysis of such editing landscapes. AIDD-based results show importance of diversity of ADAR isoforms, key RNA editing enzymes linked with the innate

immune system and viral infections. These findings offer insights into the potential role of ADAR editing dysregulation in the disease mechanisms, including those of congenital Zika syndrome. Because of its automated all-inclusive features, AIDD pipeline enables even a novice user to easily explore common mechanisms of transcriptome diversity, including RNA editing landscapes.

Keywords: High-throughput sequencing, Analysis of RNA-seq, Transcriptome, Editome, RNA editing, Isoform, Differential expression, Sequencing variants, Adenosine deaminases acting on RNA (ADAR)

Background

Transcriptome complexity and diversity, including patterns of differential isoform expression, non-canonical transcripts, diversity of non-coding RNAs, and regulation of RNA editing, including editing by adenosine deaminases acting on RNA (ADAR) enzymes resulting in A to I substitutions, play fundamental roles in both normal physiological function and disease mechanisms [1–4]. Due to advances in deep sequencing technologies, RNA-seq experiments have become a more affordable and therefore popular tool for studying intricacies of molecular processes [5–8]. In fact, currently RNA-seq can be considered almost routine if not for the still substantial costs of experiments and subsequent *in-silico* analyses [9], including those associated with data storage and handling [10]. This, along with explosive increases in available volumes of data generated in large-scale RNA-seq experiments, contributes to an ongoing demand for universal, easy-to-use computational tools capable of user-specific customization.

One of the widely used workflows available for high-throughput RNA-seq analyses is Galaxy, which is a reproducible and collaborative analytic platform that offers developers a framework for integrating and sharing their tools and workflows [11, 12]. Yet, although Galaxy is designed to be relatively easy to use, even for a beginner, performing more in depth analysis with multi-step workflows often requires that a user possesses and/or has access to a specialized bioinformatics expertise. Other challenges are related to sharing potentially large-scale analyses on a public webserver, which can become time-consuming, e.g., with time to completion increasing during high peak usage hours. Further, while there are hundreds of workflows currently accessible on Galaxy, many of these are quite complex and have a substantial learning curve to perform analyses and/or often require user knowledge of reference genomes and file formats. This limits the types of datasets that can be analysed without deploying a custom Galaxy instance, which in turn requires specialized skills. Likewise, for tasks beyond the basic transcriptome discovery analysis the user would need to know how to install and utilize additional tools in the Galaxy instance, somewhat hampering its usability to the potential user with only the basic computing skills. We would like to note that Galaxy Training Network (https://training.galaxyproject.org/, accessed 12 August 2020) already provides a variety of excellent tutorials to help inexperienced Galaxy users to performed complex analyses [13]. These tutorials nonetheless require substantial time and effort investments from users, which may exclude small labs lacking necessary manpower or somewhat limit Galaxy's usability in the classrooms. In the past few years several toolboxes have been released in an effort to address such

challenges with using Galaxy [14–19]. Yet, these toolkits are often designed to analyse only one specific dimension of transcriptome diversity, and/or not fully automated and require some prior knowledge of R command line script [20].

Implementation:

AIDD features overview

To help overcome some of these limitations, our pipeline—Automated Isoform Diversity Detector (AIDD)—has been designed implicitly with a novice user in mind, and thus, can be used, for example, as an educational tool for RNA-seq-based laboratory exercises in the classroom setting with a minimal prior user training. Because the pipeline is packaged in a VirtualBox environment, it is easy to install on essentially any operating system and/or a broad range of hardware (Windows, Linux, MacOS) that is capable of handling a VirtualBox installation without concerns for compatibility. Yet despite the seeming simplicity of installing it, our AIDD pipeline is powerful enough to handle a broad range of RNA-seq analyses, spanning from differential gene and isoform expression, to variant calling, and RNA editing analysis using dimension reduction and machine learning approaches, including Guttman scale patterns [21] for time series analysis of ADAR editing landscapes. Unlike comparable tools, AIDD offers a fully automated data analysis pipeline with a simple setup and one-click execution, while still allowing for easily customizable options to account for a wide range of experimental conditions that users may wish to include. AIDD incorporates GATK haplotype caller [22], which is currently not available from Galaxy, as a variant caller for RNA editing prediction, customizable R and bash scripts for detailed statistical analyses of the transcriptome, including RNA editing patterns as well as transcriptome-level differential expression combined with gene enrichment and pathway analysis. SnpEff [23]is used to add depth to the complete transcriptome analysis by predicting the impact of RNA editing on protein structure and function. AIDD also performs data visualization as part of the automated pipeline and produces publication-ready heatmaps, volcano and violin plots, bar charts and Venn diagrams.

AIDD availability and hardware requirements

The AIDD pipeline is built in an Oracle VirtualBox (https://www.oracle.com/virtualization/virtualbox/index.html, accessed 12 August 2020) virtual machine based on Ubuntu 18.04.2 LTS (Bionic Beaver) 64-bit PC (AMD64) desktop image (http://releases.ubuntu.com/18.04/, accessed 12 August 2020) and contains all tools necessary for transcriptome-level analysis (Fig. 1). The distributed VirtualBox image is ~20 Gb in size and is publicly available for download via GoogleDrive link (https://drive.google.com/open?id=1XOWh9H-v1nA6_Vl53PI6G2gKaVoZX6ls, accessed 12 August 2020). The up-to-date detailed description of included software tools, AIDD manual and step-by-step tutorial for AIDD are distributed via our GitHub site (https://github.com/RNAdetective/AIDD, accessed 12 August 2020).

Implicitly tailored toward a novice user with no or minimal experience in computational analyses, AIDD is designed to run automatically with limited user input through a customizable bash script that controls multiple computational tools, including HISAT2 and GATK, among others, to comprehensively analyse RNA-seq datasets. AIDD can

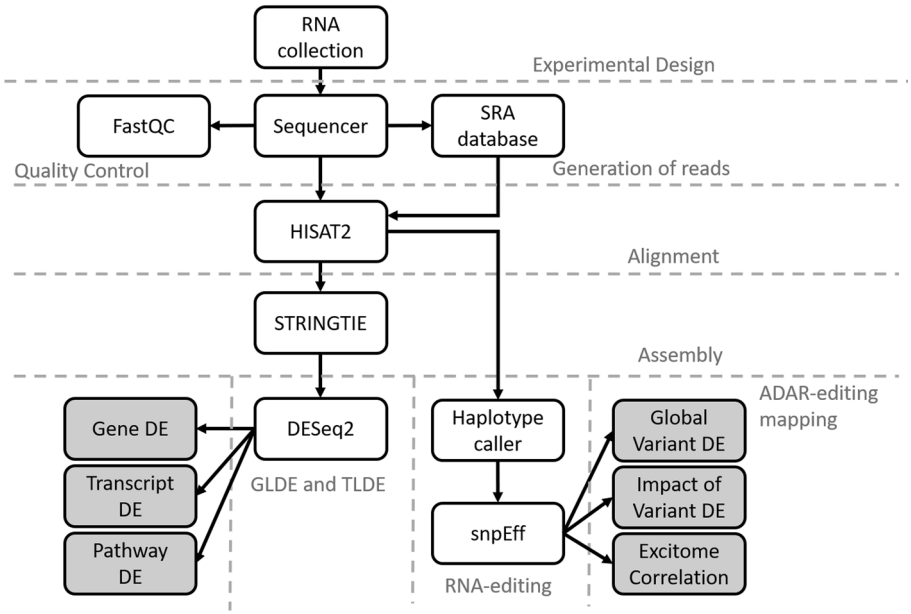

Fig. 1 Flow chart of the tools and steps used in the automated workflow carried out by AIDD pipeline. The analysis begins from gathering relevant RNA-seq data files from the NCBI SRA database, followed by reads alignment using HISAT2 with Ensembl annotations. Transcriptome assembly is then performed by Stringtie. Downstream expression analysis can be performed using multiple tools, including DESeq2, edgeR and topGO. Variant calling to detect RNA-editing events, including A-to-I editing, is performed using tools implemented in GATK; and statistical analysis of the effect of RNA editing is performed using custom R scripts

be deployed on almost any modern laboratory, classroom or office computer capable of running Ubuntu 18.04 in a VirtualBox environment. To shortcut the early learning curve, the pipeline is set up to run with default parameters directly "out of the box", and includes commented out examples in the form of R markdown file that the user can choose to deploy as a step-by-step tutorial.

The minimum recommended hardware specifications include 4 GHz dual-core processor (or better), 8 to 12 GB system memory available to the virtual environment, and 50 GB of free hard drive space (https://www.ubuntu.com/download/desktop, accessed 12 August 2020), although at least 16 GB system memory is recommended, and some applications may require more. For example, STAR alignment tool needs at least 10 times more memory bytes than the target genome, which for human genome translates into at least 32 GB and upwards if annotations are needed [24].

Included example datasets: transcriptomes of ZIKV-infected neural progenitor cell lines and importance of ADAR gene family

To illustrate the AIDD capabilities, we use a publicly available dataset from a study by McGrath et al. [25] that contains RNA-seq data from three genetically distinct neural progenitor cell (NPC) lines infected with Zika virus (ZIKV). The authors found varying degrees of severity of symptoms associated with congenital Zika syndrome (CZS), including decreased differentiation and proliferation, and increased signs of apoptosis [25]. McGrath et al. also reported increased expression of genes involved in innate

immune response, including interferon alpha (IFNA) and adenosine deaminase acting on RNA (ADAR) during ZIKV infection (Additional file 2: Table 1 in McGrath et al. 2017) [25]. The ADAR gene family consists of three genes, namely, ADAR (also referred to as ADAR1), ADARB1 (ADAR2), and ADARB2 (ADAR3). Only ADAR and ADARB1 have proven deaminase activity [26–28] catalyzing the deamination of adenosine (A) to inosine (I) transition seen in RNA editing [29, 30]. ADARB2 is thought to play a regulatory role through competition with other ADARs for substrate binding [29, 31]. ADARs play a prominent role in the nervous system [30, 32, 33], specifically in the brain [34, 35], where the majority of ADAR editing target genes are expressed [20, 26, 34, 36], including during development [37].

Running AIDD: uploading RNA-seq data into AIDD

AIDD is designed to automatically download and convert RNA-seq datasets from the SRA accession numbers that user defines in the experimental conditions table. For the example analysis discussed here, a subset of Bioproject PRJNA360845 [25] was downloaded and converted to fastq format. Once converted to fastq format, fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed 12 August 2020) is used for quality control. Upon user assessment of quality of files, fastx-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/, accessed 12 August 2020) is used to trim fastq files to assure best quality for alignment. In addition to downloading and preparing sequences, AIDD also automatically downloads and formats all necessary default references and indexes for human genome to run the tools. There are also options for user-defined reference sets, e.g., if RNA-seq data comes from mouse rather than human. AIDD can also run from locally stored fastq or standard alignment SAM/BAM files.

In addition to PRJNA360845 RNA-seq data [25], the included tutorial uses a second dataset from Bioproject PRJNA313294 [38]. While PRJNA313294-based results are not discussed here, they are available through the AIDD manual and in the distributed AIDD image (https://github.com/RNAdetective/AIDD, accessed 12 August 2020).

Running AIDD: reads alignment and assembly

Once the RNA-seq data and the reference files have been downloaded, the reads are aligned to the chosen reference (GRCh37_snp_tran is used as a default, and in this example). The pipeline uses HISAT2 [39] as a default alignment tool. SALMON [40] and STAR [24] aligners are also available as options. The HISAT2 (https://ccb.jhu.edu/software/hisat2/index.shtml, accessed 12 August 2020) aligner is a low-memory yet sensitive alignment program that allows for comparable results to other slow and more memory intensive aligners such as STAR [24, 39]. Once the reads have been aligned, the output files (SAM format) are converted into BAM format using Picard tools (http://broadinstitute.github.io/picard/, accessed 12 August 2020) in preparation for variant calling and transcriptome analysis. The pipeline saves these intermediate files should the user ever need to use them for additional analyses.

Next, the transcriptome is reconstructed using Stringtie [41], with cufflinks available as an option (https://software.broadinstitute.org/cancer/software/genepattern/modules/docs/Cuffdiff/7, accessed 12 August 2020), with output generated as raw counts (Fragments Per Kilobase Million (FKPM), Transcripts Per Kilobase Million (TPM), and

coverage) in the "counts" folder, and gene transfer format (GTF) files. The latter are then automatically modified into the count matrix for subsequent input into DESeq2 [42, 43], using the coverage correction for raw counts unique to Stringtie. The conversion step is performed by a Python script available from the Stringtie website (https://ccb.jhu.edu/software/stringtie/, accessed 12 August 2020).

Running AIDD: differential expression analysis

Once reads have been mapped, DESeq2 [42] and other dependent packages are used to generate gene-level and transcript-level differential expression outputs, including results of the principal component analysis. The latter can be used as a quality control or as an exploratory analysis step, to verify the similarity among samples or treatments, and to identify outliers. DESeq2 uses empirical Bayes shrinkage approach to take into account within-group variation as well as fold change estimation to control for variance observed in the low read count genes [44]. This approach allows for increased sensitivity and decreased false positive rate [44].A user supplied gene list, for example, a Gene Ontology (GO)-based list, can be used to create pathway expression heatmaps and volcano plots to visualize significantly differentially expressed genes involved in those user-defined pathways, along with the default pathways for GO terms involved in neural development, proliferation, differentiation and signalling as well as the gene list of the innate interferon pathway that we used to explore the role of ADAR editing in CZS (Additional file 2: Table 1, Additional file 3: Table 2, Additional file 4: Table 3, Additional file 5: Table 4, Additional file 6: Table 5). Additional pathway enrichment analysis is automatically performed using included R package topGO [45]. Alternatively, generated gene and transcript lists can be used with outside gene enrichment analysis tools such as PANTHER [46] or DAVID [47].

Running AIDD: variant calling

While the state of the art identification of genomic variants that can be linked to phenotypic variation is based upon whole-genome (WGS) or whole-exome sequencing (WES) [48], much broader availability (and affordability) of transcriptome sequencing data makes it another appealing source of variants discovery [49]. Furthermore, some mechanisms of variants generation—such as RNA editing and splice-site variation—can only be studied at the transcriptome level. Thus, our pipeline includes tools enabling variant discovery from transcriptome data, with the focus on ADAR-mediated RNA editing.

GATK haplotype caller [50] is the tool used in AIDD to infer potential RNA editing events, based upon the best practice settings as defined by the GATK developers as of March 2019 (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891, accessed 12 August 2020). Picard tools are used for quality control and proper formatting of input files. Haplotype caller is used twice in the pipeline, along with filtering steps to control for both false positives and false negatives. Specifically, to filter for false positives Qualality by Depth is set to 2.0, Fisher Strand is set to 60.0, and RMS Mapping Quality is set to 40.0, and to control for strand bias Strand Odds Ratio is set at 4.0, respectively, as default filtering parameters in GATK. SnpEff is then used to predict consequences on protein structure and function for the inferred variants [23]. Once a final list of potential variants is generated, these are then processed using R scripts

to demonstrate both global and local view of RNA editing. Additional set of R scripts will then compare differential ADAR editing landscapes between conditions. It should be noted that here we focus on potential editing events within coding regions, and thus, we are not considering hyperediting events [51]. Likewise, genomic polymorphisms can appear as potential editing events in RNA-seq, and thus we include an annotation of detected edited site candidates with available polymorphism data (where applicable). Figure 2 and Additional file 7: Table 6 outline various tools, used, as well as folders and files generated by the pipeline.

Running AIDD: reproducibility and customization

Although designed with the simplicity of user experience in mind, through inclusion of developer defined or best-practice default parameters, AIDD pipeline can be customized to address user's specific needs, including changes in various mapping and variant calling parameter, and reference genomes. The user manual offers detailed instructions on how to change parameters for every tool used in the pipeline, including GATK haplotype caller, with specific directions on which line(s) of code to change. Furthermore, upon completion of a job, the AIDD Results directory contains not only the desired outputs, such as mapped reads and gene counts, but also a folder with copies of all the scripts used for that particular job as well as log files generated during execution of that job in an effort to enable reproducibility and to facilitate troubleshooting.

Results and discussion

To illustrate AIDD's capabilities, we describe results from the included tutorial that uses Bioproject PRJNA313294 data from [25]. Using PRJNA313294 data, AIDD mapped reads and then computed normalized and transformed gene and transcript count matrices for differential expression (DE) analysis using DESeq2 with a multivariate model for infection status taking into account cell-line identity. Principle

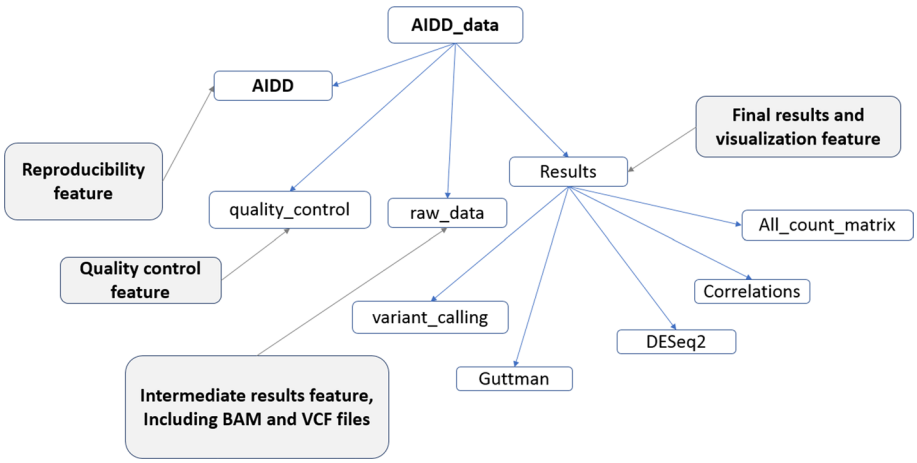

Fig. 2 Flow chart showing directory structure created by AIDD. The main folder is AIDD_data and contains 4 folders including (i) AIDD, containing all scripts used in analysis for reproducibility, (ii) quality control files, (iii) intermediate files, including BAM, GTF and VCF files, (iv) results of statistical analysis and data visualization including differential isoform expression and ADAR editing landscapes

component analysis (PCA) of the top 500 expressed genes showed that ~47% of the variance is explained by the first principle component, which separated cell lines by fetal age, with K048 cell line derived from the 9 week old fetal tissue being separated from the 13 weeks old fetal tissue of G010 and K054 cell lines. The second principle component explained ~27% of the variation, and clustered ZIKV-infected cells from the mock infected cells, except in the case of the G010 cell line (Fig. 3a). The pipeline also generated a heatmap of the top 60 differentially expressed genes with hierarchal clustering that showed clustering of samples by infection status, except for the G010 cell line (Fig. 3b). This latter phenomenon is consistent with reported findings of McGrath et al. [25] that showed that G010 cells exhibited the least amount of cytopathic effects, if any, due to ZIKV infection, potentially reflecting genetic

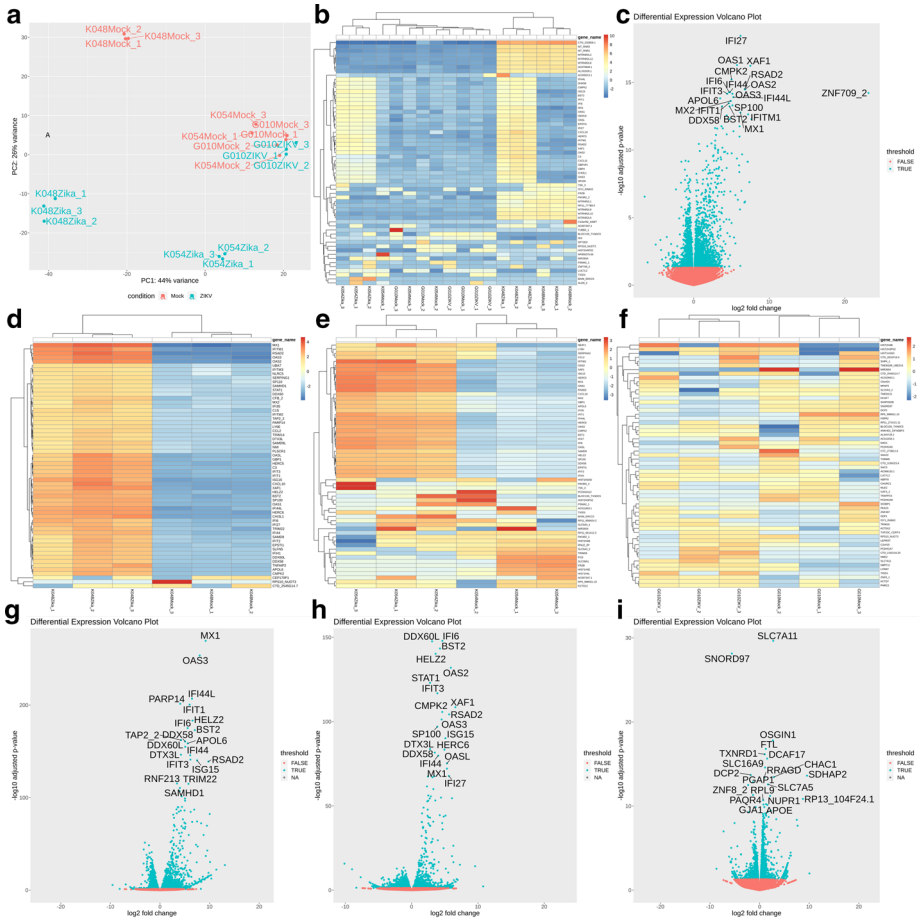

Fig. 3 Visualization of differential expression analysis using AIDD. **a** Principle component analysis of the top 500 expressed genes counts show 47% of the variance in the system is attributed to differences in cell lines and 27% of the variance is attributed to ZIKV infection status. **b** The top 500 hierarchal clustering also shows clustering of CSZ phenotype cell line (K048 & K054) ZIKV infected cells and normal phenotype cells (G010) regardless of ZIKV infection status clustered with the CSZ phenotype cell line mock infections. **c** The top 20 differentially expressed genes during ZIKV infection taking into account genetic cell line differences highlight the innate immune activation. When looking at each cell line independently, K048 (**d**) and K054 cells (**e**) have clear pattern of differentially expressed genes during ZIKV infection, whereas G010 cells **f** shows less of a pattern of differentially expressed genes. Panels **g**–**i** show that when the top 20 differentially expressed genes are considered, each genetically distinct cell line shows a differentially gene expression response to ZIKV infection

heterogeneity across studied cells. Figure 3c shows generated volcano plots that visualize the top 20 differentially expressed genes between ZIKV and mock infections taking into account differences in cell-lines. AIDD generates clustering heatmaps for each cell line, which showed that while both K048 and K054 exhibit clear differences between mock and ZIKV infections consistent with the phenotypic differences between the two conditions (Fig. 3d,e), G010 cells showed no significant difference between ZIKV and mock infected cells, consistent with McGrath et al. (2017) results (Fig. 3f). By looking at each cell line individually, AIDD is able to highlight differential effects of ZIKV infection in combination with host genetics that are consistent with results originally reported by McGrath et al. [25] (Fig. 3g–i).

Pathways analysis

The gene pathways exploration tool included in AIDD was used to examine differential expression in neurodevelopmental pathways during ZIKV infection. Using gene list supplied by the user, AIDD will generate customized heatmap, volcano plot, and data table with differential expression results for genes of interest. Gene ontology (GO) terms "innate immunity"," brain development", "central nervous system development", "neurological development", and "peripheral nervous system" are already included as default pathways. We also included a custom gene list for genes in the interferon alpha pathway (Additional file 2: Table 1). AIDD results showed that ZIKV-infected cells showed increased expression of innate immune genes (Fig. 4a), as well as those in the interferon alpha pathway, including ADAR (Fig. 4b), except for the G010 cells. Consistent with McGrath et al. findings [25], cell lines that have CZS-like phenotypic appearance if ZIKV infected (namely, K048 and K054) have significant differential expression in the majority of the genes involved in the interferon alpha pathway (Fig. 4c,d), whereas G010 cells that appear to be essentially normal phenotypically showed only a few significantly differentially expressed genes in the interferon alpha pathway (Fig. 4e), pointing to potential involvement of interferon alpha response in ZIKV infection and CZS-like symptoms. On the other hand, only cell line-associated differences but not the ZIKV infection-mediated differences were observed for genes associated with GO terms of brain development (Fig. 4f), central nervous system development (Fig. 4g), neurological development (Fig. 4h), and peripheral nervous system development (Fig. 4i).

Mapping ADAR expression and editing landscapes

To explore the potential role of ADAR enzymes and ADAR editing, AIDD allows us to focus on expression of ADAR genes and editing patterns (Additional file 8: Table 7, Additional file 9: Table 8), including applying Guttman scale patterns to identify temporal changes in ADAR editing landscapes (Additional file 1: Fig. 1). The results showed that ADAR1p150 isoform-specific expression was significantly higher in ZIKV infected cells with the CZS phenotype (K048 and K054), while not being significantly different in G010 cells (Fig. 5a). Interestingly, ADARB1 showed the opposite pattern, being significantly upregulated in G010 cells, but not in cells with CZS-like phenotype (Fig. 5b). Because ADARB1 and ADAR both share some overlapping editing targets as well as have gene-specific ones [52, 53], this expression pattern suggests that both ADAR genes may play complementary roles in the differential response to

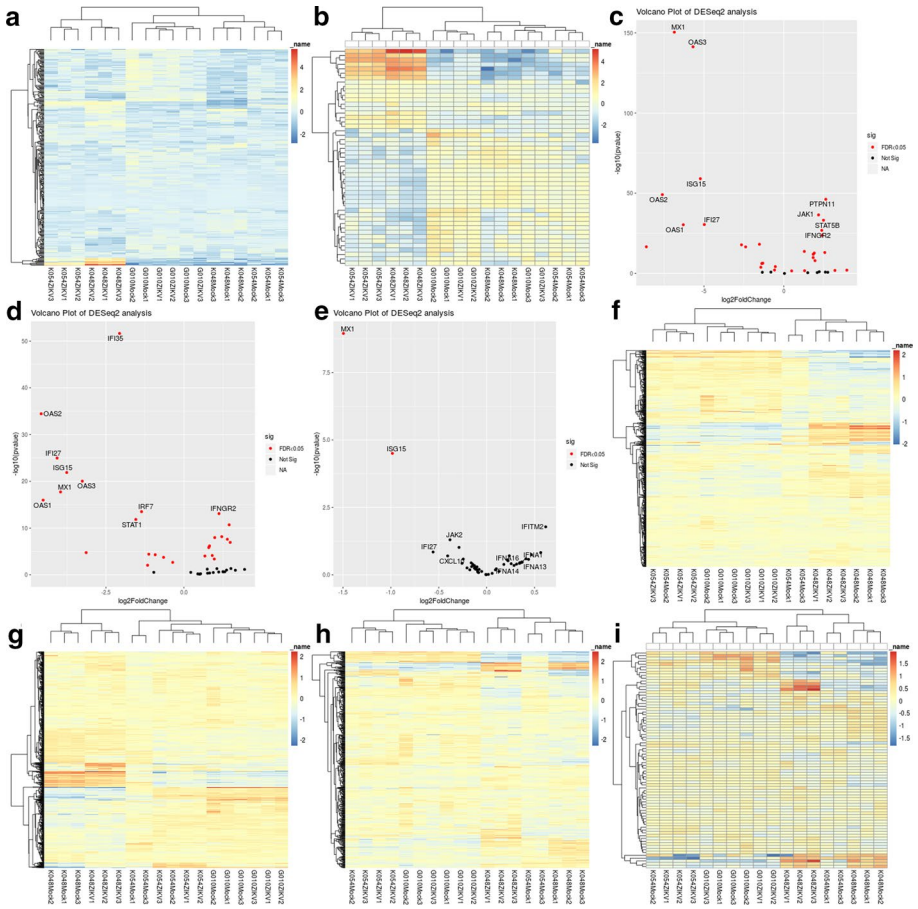

Fig. 4 Results of AIDD pathway expression analysis. **a** Gene Ontology term "innate immune system" shows clustering of ZIKV infected cells with the CSZ phenotype (K048 & K054) and clustering of normal phenotype (G010) with the mock infected cells of all 3 cell lines. **b** Customized "interferon alpha pathway" list shows similar clustering pattern as (**a**). The CZS phenotype cell lines (K048 & K054) show the top 10 differentially expressed genes with gene products induced by interferon alpha pathway, including OAS1 and 2, and intermediary genes in the interferon alpha pathway, including STAT1 (**c**, **d**, respectively). On the other hand, phenotypically normal cell line (G010) has only 2 differentially expressed genes, which are not part of the interferon alpha pathway (**e**). Gene ontology terms "brain development" (**f**), "CNS development" (**g**), "neurological development" (**h**), and "PNS development" (**i**) exhibit differential expression patterns that can be attributed to genetic differences among cell lines, but not associated with ZIKV infection

ZIKV infection [54]. This would be consistent with prior suggestions that ADARB1 contributes to dysregulation of RNA editing in many diseases [55–58].

AIDD also allows the user to map ADAR editing landscapes by performing variant calling to identify potential A to G substitutions. Globally, we found that the total numbers of A to G substitutions are higher in ZIKV-infected in both the G010 and K048 cell lines but not in the K054 line (Fig. 5c). However, when the potential impact of these substitutions on protein structure and function is examined, cell lines with the CZS-like phenotype (K048 and K054) had more of high and moderate impact variants detected in ZIKV infection, while seemingly normal G010 cells had smaller number of potentially impactful changes in ZIKV infection (Fig. 5d–f).

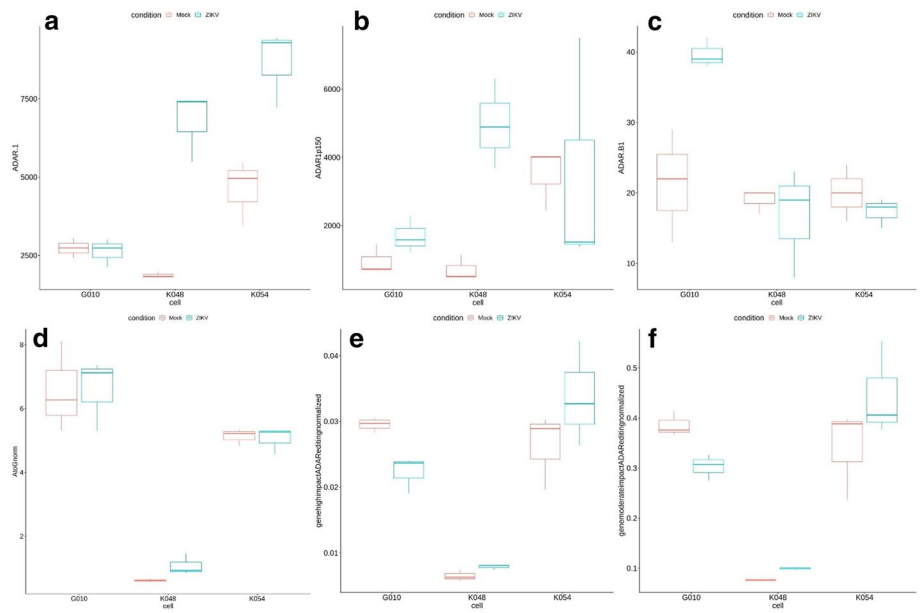

Fig. 5 Visualization of ADAR expression and ADAR editing landscapes. **a** ADAR expression is significantly increased in CZS phenotype cell lines K048 (ANOVA F-statistic (F) = 58.396, significance *p*-value (*p*) = 0.001575) and K054 (F = 18.516, *p* = 0.01261), but not in phenotypically normal G010 (F = 0.1219, *p* = 0.7446) cells. **b** ADAR1p150 expression is significantly higher in K048 (F = 29.497, *p* = 0.005576), but not in K054 (F = 2e−04, *p* = 0.9902) or G010 (F = 3.4772, *p* = 0.1357) cells. **c** ADARB1 expression is not significantly different in K048 (F = 0.2579, *p* = 0.6383) or K054 (F = 1.0492, *p* = 0.3636) cells, but is significantly higher in G010 (F = 14.684, *p* = 0.01859). **d** The numbers of A to G substitutions were somewhat elevated in K048 (F = 6.0422, *p* = 0.06984), but not in G010 (F = 6e−04, *p* = 0.9813) or K054 cells (F = 0.0648, *p* = 0.8116). **e** The numbers of A to G substitutions with predicted high impact on protein structure and function were significantly lower in G010 (F = 17.498, *p* = 0.01388), but somewhat higher in K048 cells (F = 6.3489, *p* = 0.06538); there were no changes in K054 cells (F = 1.7384, *p* = 0.2578). Likewise, moderate impact substitutions were also significantly lower in G010 (F = 15.737, *p* = 0.01658) and significantly higher in K048 (F = 157.23, *p* = 0.0002328) cells, while were not changed in K054 cells (F = 1.9198, *p* = 0.2381) (**f**)

It should be noted that one major challenge of using variant calling methods for detecting RNA editing events is the need to have a sufficient coverage depth (of at least 50 million reads or higher per sample) to accurately detect editing events when editing frequencies are low. AIDD attempts to correct for this by normalizing substitution counts by the read depth as determined from alignment algorithms. Therefore, these observed editing differences among cell lines could be attributed to interactions between ADAR family members as well as ADAR preferences at the editing sites, and spatio-temporal regulation of editing.

We were also interested in editing events at known editing sites in ion channels and transporters that are known to be associated with fine-tuning of neural signalling, including excitotoxicity, brain development and neural plasticity [33, 59, 60]. To define the excitome, computationally-predicted ADAR editing sites found in psychiatric disorders confirmed with PCR [61] were combined with editing sites from RADAR database that were previously examined in Alzheimer's disease [62] to create a list of 151 editing sites located in 91 genes (Additional file 7: Table 8). In part because of relatively low coverage in all three cell lines as well as rather drastic differences in fetal age, the editing patterns at specific sites varied both between different cell lines and between infected

and uninfected cells. ZIKV infected K048 cells showed likely editing events at multiple sites, including at two ion channel receptors (namely, GRIA3 and GRIN3B). Other ZIKV-induced editing events were detected at IGFBP7, KIF20B and SRP9 genes, responsible for controlling cellular metabolism, vesicular transport, and proper protein storage and transport respectively [63–66]. There was also an increased editing detected at the ATXN7 gene that is implicated in degenerative ataxia [67]. ZIKV infected K054 cells showed likely editing events in PTPRN2, GRIA2 Q/R site, GRIA3 and IGFBP7, whereas uninfected cells showed editing events in ATXN7, BEST1, BLCAP, and KIF20B. ZIKV infected G010 cells exhibited increased editing in ATXN7, KIF20B, and PTPRN2, and decreased editing at the NEIL1 genes. Changes in editing landscapes can also be visualized with Guttman scale patterns, where differences between distinct cell lines as well as mock and infected cells are shown for individual editing events/residues (Additional file 1: Fig. 1). However, further transcriptomics studies – including at much higher read depth—are needed to fully elucidate the changes in editing patterns that can be induced by viral infections.

Conclusions

A fully automated pipeline, Automated Isoform Diversity Detector (AIDD), has been developed to facilitate RNA-seq analyses focused on changes in transcriptome diversity, including isoform expression ratios and ADAR-editing events. A publicly available dataset of human neural progenitor cells [25] is used to demonstrate how AIDD pipeline can be used to robustly and reproducibly analyse transcriptome diversity and to infer RNA editing patterns from RNA-seq data. Presented results illustrate the importance of examining both the gene-level and the isoform-level expression differences, as well as exploring RNA editing aspects of transcriptome diversity and their potential association with pathogenicity mechanisms.

AIDD pipeline has additional benefits of being novice-user friendly and completely automated for highly reproducible results. Briefly, AIDD incorporates multiple steps needed for using RNA-seq data to study transcriptome diversity, and offers an easy-to-use pipeline for mapping and contrasting genome-wide RNA editing patterns, with focus on protein-coding transcripts (Additional file 9: Table 8). Once reads have been mapped to the reference genome, AIDD uses DESeq2 to infer patterns of differential expression at both gene and transcript levels. For users—such as ourselves—interested in patterns of editing of excitome-related genes, AIDD will summarize the expression of the excitome gene members, including ADARs and other genes with known editing sites. AIDD will further summarize global RNA editing patterns and infer correlations between edited sites and ADAR expression patterns. Lastly, lists of genes involved in ADAR editing landscape changes are produced and can be used as potential biomarkers for diagnostic and prognostic purposes.

The distributed pipeline image includes a user-friendly tutorial written in R markdown that can be used to illustrate AIDD features in a classroom setting as teaching tool and/or to generate hypotheses for future experimental validation, or both. The ZIKV infection-associated example described in this paper further highlights the ability of AIDD to conduct complicated analyses within the constraints of a small

research laboratory. Future work includes testing AIDD's accuracy against simulated reads with known editing sites and across various read depths per sample, as well as expanding AIDD's ability for variant calling by incorporating other methods (such as Freebayes) [68]. AIDD can also be used in meta-analysis of publically available RNA-seq datasets to comprehensively map ADAR editing landscapes across different cells and organisms, and to facilitate discovery of novel diagnostic and prognostic biomarkers and potential targets for drug therapies.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03888-6.

Additional file 1. **Figure 1**: Guttman scale patterns (Proctor 1970) were used to order and group ADAR editing sites based on the frequency of samples that had editing at those sites. ADAR editing landscapes are differentially edited in both order and groupings based on cell line and ZIKV infection. (A) The expression and editing events are ordered by normal phenotype cell line G010 shown in blue, with cell lines K048 and K054 shown in green and red, respectively. The mock-infected cells are shown with solid lines and ZIKV-infected cells are shown with dashed lines. (B) The mean editing frequencies differ between mock- and ZIKV-infected cells at several sites including; (i) AZIN1 at amino acid position 367 (F=7.1095, p=0.00263), (ii) CRB2 at amino acid position 969 (F=3.2, p=0.04584), (iii) IGFBP7 at amino acid position 95 (F=40.651, p=4.09e-07), (iv) SRP9 at amino acid position 75 (F=3.5131, p=0.03459), and (v) UQCRHL at amino acid position 53 (F=8.796, p=0.00105). Changes in editing patterns were also detected at ADAR1 at amino acid position 427 (F=2.9571, p=0.05749), CCN1 at amino acid position 75 (F=2.5546, p=0.08504), and GRIA3 at amino acid position 775 (F=2.5515, p=0.08531), respectively.

Additional file 2. **Table 1**: IFNA pathway. Custom list of genes in the interferon alpha pathway genes containing intermediate pathway genes along with interferon stimulated genes (ISG)s including ADAR1p150.

Additional file 3. **Table 2**: GO terms "brain development". Gene ontology term containing 706 genes involved in brain morphogenesis, brain segmentation, central complex development, forebrain, hindbrain, midbrain development and mushroom body development, subarachnoid space development, trigeminal sensory nucleus development, and ventricular system development.

Additional file 4. **Table 3**: GOterms "central nervous system development". Gene ontology term containing 900 genes involved in astrocyte differentiation, brain development, central nervous system maturation, morphogenesis, differentiation, and segmentation, larval central nervous system remodeling, microglia differentiation, preganglionic parasympathetic fiber, spinal cord, ventral cord ventral cord and ventral midline development, vertebrate eye-specific patterning..

Additional file 5. **Table 4**: GO terms "neurological development". Gene ontology term containing 526 genes involved in neurological development.

Additional file 6. **Table 5**: GOterms "peripheral nervous system development". Gene ontology term containing 82 genes involved in establishment of blood-nerve barrier, lateral line ganglion development, peripheral nervous system neuron differentiation, postganglionic parasympathetic fiber development and Schwann cell differentiation.

Additional file 7. **Table 6**: Quality control and other features implemented in the AIDD pipeline (https://github.com/RNAdetective/AIDD).

Additional file 8. **Table 7**: All_count_matrix.csv. Coverage counts from Stringtie for important IFNA genes and ADAR and APOBEC gene family. This is combined with variant calling substitution matrix for nucleotide and amino acid substitutions and the number of genes with editing sites that have a high or moderate impact on protein structure and function..

Additional file 9. **Table 8**: Excitome_freq_matrix .csv. Computationally predicted ADAR editing sites found in Psychiatric Disorders study and confirmed with PCR (Zhu et al. 2012) were combined with editing sites from RADAR database that were previously compared in Alzheimer's disease (Khermesh et al., 2016) to create a list containing 151 editing sites located in 91 genes. If a specific editing site is found in the dbSNP database, a reference number (rs) is included, along with the genomic location, strand orientation, annotation containing details about type of amino acid substitution, amino acid position, codon position on the mRNA, and two columns showing minimum cutoff value for expression and editing read depth for determining accuracy of variant calling.

Abbreviations

ADAR: Adenosine deaminases acting on RNA; AIDD: Automated Isoform Diversity Detector; CZS: Congenital Zika syndrome; GO: Gene Ontology; IFNA: Interferon alpha; NPC: Neural progenitor cell; WES: Whole-exome sequencing; WGS: Whole-genome sequencing; ZIKV: Zika virus.

Acknowledgements

Not applicable.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 21 Supplement 18, 2020: Proceedings from the 8th Workshop on Computational Advances in Molecular Epidemiology (CAME 2019). The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-21-supplement-18.

Authors' contributions
NMP designed and implemented the pipeline, and wrote the manuscript. EJ, MF, HM and GF contributed to conceptualization of pipeline features, testing of code components and validation, and provided manuscript feedback. RM and GC contributed to conceptualization of pipeline features and analysis steps. HP conceived the pipeline, supervised the project, helped with code and testing, and wrote the manuscript. All authors read and approved the final manuscript.

Funding
Publication of this supplement is funded by a Kent State University Research Council Seed Award, Brain Health Research Institute Pilot Award, and the National Institutes of Health (NIA award R21AG064479-01). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
Supplementary Tables are available at GitHub, https://github.com/RNAdetective/AIDD/tree/master/AIDD_supplFiles.ST1-8_and_SF1. The datasets used in this current study are publicly available in the NCBI SRA/BioProject repository, at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA360845/ and https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA313294.

Availability and requirements
Project name: Automated Isoform Diversity Detector (AIDD) pipeline. Project home page: https://github.com/RNAdetective/AIDD. Operating system(s): Platform independent. Programming language: R. Other requirements: R environment. Tested on R version 3.6.1. License: GNU GPL. Any restrictions to use by non-academics: none.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Biological Sciences, Kent State University, 256 Cunningham Hall, Kent, OH 44242, USA. [2] School of Biomedical Sciences, Kent State University, PO Box 5190, Kent, OH 44242, USA. [3] University of Mount Union, 1972 Clark Ave, Alliance, OH 44601, USA. [4] Department of Anthropology, Kent State University, Kent, OH 44242, USA. [5] Brain Health Research Institute, Kent State University, Kent, OH 44242, USA. [6] Present Address: Department of Pharmacology & Therapeutics, College of Medicine, University of Florida, Gainesville, FL 32610, USA.

Received: 16 November 2020 Accepted: 18 November 2020
Published: 30 December 2020

References

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306:636–40.
2. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16:197–212.
3. Ardlie KG, Guigó R. Data resources for human functional genomics. Curr Opin Syst Biol. 2017;1:75–9.
4. Gallo A, Vukic D, Michalík D, O'Connell MA, Keegan LP. ADAR RNA editing in human disease; more to it than meets the I. Hum Genet. 2017;136:1265–78.
5. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12:87–98.
6. Conesa A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
7. Wang, Z. & Ma'ayan, A. An open RNA-Seq data analysis pipeline tutorial with an example of reprocessing data from a recent Zika virus study. F1000 Research 2016;5:1574.
8. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83.
9. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13:599–604.
10. Kwon T, Yoo WG, Lee W-J, Kim W, Kim D-W. Next-generation sequencing data analysis on cloud computing. Genes Genomics. 2015;37:489–501.
11. Goecks J, Nekrutenko A, Taylor J, The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11:R86.
12. Afgan E, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44:W3–10.
13. Batut B, et al. Community-driven data analysis training for biology. Cell Syst. 2018;6:752-758.e1.
14. Grüning, B. A. et al. Enhancing pre-defined workflows with ad hoc analytics using Galaxy, Docker and Jupyter. 2016. doi:https://doi.org/10.1101/075457.
15. Hung, L.-H. et al. Building containerized workflows using the BioDepot-workflow-builder (Bwb); 2017. doi:https://doi.org/10.1101/099010.

16. Meiss T, Hung L-H, Xiong Y, Sobie E, Yeung KY. Software solutions for reproducible RNA-seq workflows. 2017. doi:https://doi.org/10.1101/099028.
17. Tithi SS, Lee J, Zhang L, Li S, Meng N. Biopipe: a lightweight system enabling comparison of bioinformatics tools and workflows. 2017; doi:https://doi.org/10.1101/201186.
18. Beccuti M, et al. SeqBox: RNAseq/ChIPseq reproducible analysis on a consumer game computer. Bioinformatics. 2018;34:871–2.
19. Hung L-H, Kristiyanto D, Lee SB, Yeung KY. GUIdock: Using Docker containers with a common graphics user interface to address the reproducibility of research. PLoS ONE. 2016;11:e0152686.
20. Li JB, Church GM. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. Nat Neurosci. 2013;16:1518–22.
21. Proctor CH. A probabilistic formulation and statistical analysis of guttman scaling. Psychometrika. 1970;35:73–8.
22. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
23. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
24. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
25. McGrath EL, et al. Differential responses of human fetal brain neural stem cells to Zika virus infection. Stem Cell Rep. 2017;8:715–27.
26. Chen CX, et al. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. RNA. 2000;6:755–67.
27. Jin Y, Zhang W, Li Q. Origins and evolution of ADAR-mediated RNA editing. IUBMB Life. 2009;61:572–8.
28. Walkley CR, Liddicoat B, Hartner JC. Role of ADARs in mouse development. Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing. vol. 2011. Springer, Berlin.
29. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.
30. Savva YA, Rieder LE, Reenan RA. The ADAR protein family. Genome Biol. 2012;13:252.
31. Hardt O, et al. Gene expression analysis defines differences between region-specific GABAergic neurons. Mol Cell Neurosci. 2008;39:418–28.
32. Maas S, Rich A, Nishikura K. A-to-I RNA editing: recent news and residual mysteries. J Biol Chem. 2003;278:1391–4.
33. Tan BZ, Huang H, Lam R, Soong TW. Dynamic regulation of RNA editing of ion channels and receptors in the mammalian nervous system. Mol Brain. 2009;2:13.
34. Mehler MF, Mattick JS. Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiol Rev. 2007;87:799–823.
35. Liscovitch N, Bazak L, Levanon EY, Chechik G. Positive correlation between ADAR expression and its targets suggests a complex regulation mediated by RNA editing in the human brain. RNA Biol. 2015;11:1447–56.
36. Gonzalez C, Lopez-Rodriguez A, Srikumar D, Rosenthal JJC, Holmgren M. Editing of human K V 1.1 channel mRNAs disrupts binding of the N-terminus tip at the intracellular cavity. Nat. Commun. 2011; 2:436.
37. Wahlstedt H, Daniel C, Enstërö M, Ohman M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. Genome Res. 2009;19:978–86.
38. Tang H, et al. Zika virus infects human cortical neural precursors and attenuates their growth. Cell Stem Cell. 2016;18:587–90.
39. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
40. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
41. Pertea M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
42. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
43. Varet H, Brillet-Guéguen L, Coppée J-Y, Dillies M-A. SARTools: a DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-seq data. PLoS ONE. 2016;11:e0157022.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15.
45. Alexa A, Rahnenfuhrer J. topGO: Enrichment analysis for gene ontology. (Bioconductor version: Release (3.11), 2020). doi:https://doi.org/10.18129/B9.bioc.topGO.
46. Mi H, et al. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res. 2010;38:D204–10.
47. Huang DW, et al. DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35:W169–75.
48. Piskol R, Ramaswami G, Li JB. Reliable Identification of Genomic variants from RNA-seq data. Am J Hum Genet. 2013;93:641–51.
49. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinforma Biol Insights. 2015;9:29–46.
50. McKenna A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
51. Porath HT, Carmi S, Levanon EY. A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat Commun. 2014;5:4726.
52. Lehmann KA, Bass BL. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities†. Biochemistry. 2000;39:12875–84.
53. Riedmann EM, Schopoff S, Hartner JC, Jantsch MF. Specificity of ADAR-mediated RNA editing in newly identified targets. RNA. 2008;14:1110–8.

54. Piontkivska H, Plonski N-M, Miyamoto MM, Wayne ML. Explaining pathogenicity of congenital Zika and Guillain-Barre syndromes: Does dysregulation of RNA editing play a role? BioEssays News Rev Mol Cell Dev Biol. 2019;41:e1800239.
55. Amore M, et al. Sequence analysis of ADARB1 gene in patients with familial bipolar disorder. J Affect Disord. 2004;81:79–85.
56. Cenci C, et al. Down-regulation of RNA editing in pediatric astrocytomas ADAR2 editing activity inhibits cell migration and proliferation. J Biol Chem. 2008;283:7251–60.
57. Hideyama T, et al. Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. Neurobiol Dis. 2012;45:1121–8.
58. Karanović J, et al. Joint effect of ADARB1 gene, HTR2C gene and stressful life events on suicide attempt risk in patients with major psychiatric disorders. World J Biol Psychiatry. 2015;16:261–71.
59. Hood JL, Emeson RB. Editing of neurotransmitter receptor and ion channel RNAs in the nervous system. Curr Top Microbiol Immunol. 2012;353:61–90.
60. Eran A, et al. Comparative RNA editing in autistic and neurotypical cerebella. Mol Psychiatry. 2013;18:1041–8.
61. Zhu H, et al. Quantitative analysis of focused A-To-I RNA editing sites by ultra-high-throughput sequencing in psychiatric disorders. PLoS ONE. 2012;7:e43227.
62. Khermesh K, et al. Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. RNA. 2016;22:290–302.
63. Godfried Sie C, Hesler S, Maas S, Kuchka M. IGFBP7's susceptibility to proteolysis is altered by A-to-I RNA editing of its transcript. FEBS Lett. 2012;586:2313–7.
64. Ivanova E, Berger A, Scherrer A, Alkalaeva E, Strub K. Alu RNA regulates the cellular pool of active ribosomes by targeted delivery of SRP9/14 to 40S subunits. Nucleic Acids Res. 2015;43:2874–87.
65. Lee S-H, et al. Identification of diverse adenosine-to-inosine RNA editing subtypes in colorectal cancer. Cancer Res Treat. 2017;49:1077–87.
66. McNeely KC, Little JN, Dwyer ND. Cytokinetic abscission dynamics in neuroepithelial stem cells during brain development. bioRxiv 529164; 2019. doi:https://doi.org/10.1101/529164.
67. Clark L, Ye X, Liu X, Mirzozoda K, Louis E. Genetic analysis of ten common degenerative hereditary ataxia loci in patients with essential tremor. Parkinsonism Relat Disord. 2015;21:943–7.
68. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio (2012).

Publisher's Note