

SOFTWARE

Open Access



JACUSA: site-specific identification of RNA editing events from replicate sequencing data

Michael Piechotta^{1†}, Emanuel Wyler^{2†}, Uwe Ohler², Markus Landthaler² and Christoph Dieterich^{3,4*} 

Abstract

Background: RNA editing is a co-transcriptional modification that increases the molecular diversity, alters secondary structure and protein coding sequences by changing the sequence of transcripts. The most common RNA editing modification is the single base substitution ($A \rightarrow I$) that is catalyzed by the members of the Adenosine deaminases that act on RNA (ADAR) family. Typically, editing sites are identified as RNA-DNA-differences (RDDs) in a comparison of genome and transcriptome data from next-generation sequencing experiments. However, a method for robust detection of site-specific editing events from replicate RNA-seq data has not been published so far. Even more surprising, condition-specific editing events, which would show up as differences in RNA-RNA comparisons (RRDs) and depend on particular cellular states, are rarely discussed in the literature.

Results: We present JACUSA, a versatile one-stop solution to detect single nucleotide variant positions from comparing RNA-DNA and/or RNA-RNA sequencing samples. The performance of JACUSA has been carefully evaluated and compared to other variant callers in an in silico benchmark. JACUSA outperforms other algorithms in terms of the F measure, which combines precision and recall, in all benchmark scenarios. This performance margin is highest for the RNA-RNA comparison scenario.

We further validated JACUSA's performance by testing its ability to detect $A \rightarrow I$ events using sequencing data from a human cell culture experiment and publicly available RNA-seq data from *Drosophila melanogaster* heads. To this end, we performed whole genome and RNA sequencing of HEK-293 cells on samples with lowered activity of candidate RNA editing enzymes. JACUSA has a higher recall and comparable precision for detecting true editing sites in RDD comparisons of HEK-293 data. Intriguingly, JACUSA captures most $A \rightarrow I$ events from RRD comparisons of RNA sequencing data derived from *Drosophila* and HEK-293 data sets.

Conclusion: Our software JACUSA detects single nucleotide variants by comparing data from next-generation sequencing experiments (RNA-DNA or RNA-RNA). In practice, JACUSA shows higher recall and comparable precision in detecting $A \rightarrow I$ sites from RNA-DNA comparisons, while showing higher precision and recall in RNA-RNA comparisons.

Keywords: SNV, RNA editing, ADAR, APOBEC3, Variant calling

*Correspondence: christoph.dieterich@uni-heidelberg.de

[†]Equal contributors

³Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology at the Department of Internal Medicine III, University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany

⁴German Center for Cardiovascular Research (DZHK) - Partner site Heidelberg/Mannheim, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany
Full list of author information is available at the end of the article

Background

RNA editing refers to co-transcriptional RNA base modifications that increase transcript sequence diversity without changing the underlying genome. Two types of single base modifications, namely adenosine to inosine conversions ($A \rightarrow I$) and cytidin to uridine ($C \rightarrow U$) conversions, have been characterised in detail over decades of research [1]. Both conversions are executed by two specific classes of RNA binding proteins (RBPs) that interact with their respective RNA targets: Adenosine deaminase acting on RNA (ADAR) catalyses $A \rightarrow I$ conversions, whereas APOBEC1 family members catalyse $C \rightarrow U$ conversions.

ADAR mediates the more frequent $A \rightarrow I$ editing by binding to double-stranded RNA and subsequent hydrolytic deamination of adenosine residues [1]. Most functional editing sites described so far are found in transcripts for neuronal transporters and channel proteins in the brain [2]. Herein, editing is critical for normal brain development and function. Specifically, ADAR-mediated editing of the GluA2 subunit of the mammalian AMPA receptor is an essential event [3]. Generally, inosine is interpreted as guanosine by the translation machinery, which may lead to codon substitutions in protein-coding sequences. Almost 100% of the human GluA2 transcripts are edited at codon position 607 which leads to a substitution of glutamine (CAG codon) with arginine (CIG codon) in the polypeptide chain. The introduction of a positive charge reduces calcium permeability in the mammalian AMPA receptor. In human, aberrant editing of the Q/R sites has been associated with death of motor neurons [4].

However, the vast majority of editing events takes place outside of coding regions [2]. Repetitive elements as well as 5' and 3' untranslated regions (UTRs) are the most frequent targets of RNA editing [5]. Especially Alu elements are targets of positionally unspecific abundant editing events [5, 6]. Alu repeats are short (≈ 300 bp) mobile elements that are widespread in primates. Alu elements often co-occur in inverted pairs and form double-stranded RNA molecules after transcription, which constitute a favourable substrate for ADAR family members.

Taken together, site-specific RNA editing events may lead to amino acid substitutions by changing codons in coding sequences. Apart from its role in coding regions, RNA editing may also influence transcript splicing and structure and could have an effect on mRNA stability and nuclear export [2].

Identification of RNA editing sites

The previously introduced RNA editing events are single nucleotide variants that can be detected from comparing genomic and transcriptomic sequencing data. RNA-DNA differences (RDDs) of the nucleotide frequency spectrum at a given location are the most direct way of

identifying editing sites, whereas RNA-RNA comparison may pinpoint differential editing events across samples and conditions (RNA-RNA differences, in short RRDs). The availability of deep next-generation sequencing data enabled the transcriptome-wide discovery of RNA editing events. A direct comparison of gDNA and cDNA sequencing data has been proposed early on [7]. However, these early attempts suffered from the inherent artefacts of short read sequencing data and ambiguities in read mapping. For example, a re-analysis of the primary data of [7] revealed that close to 90% of the reported sites were false positives due to mapping and sequencing artifacts [8]. It was noted specifically that false editing calls were predominantly originating from base calls close to the start or end of reads, whereas true positives did not show this positional bias. Sequencing errors, read mapping errors and library preparation biases, which were introduced by ligation or amplification steps, all contribute to the high false positive rate. It is therefore essential to take these confounding factors into account or to remove them in a pre-processing step.

Several software solutions have been suggested for calling SNV sites: SAMtools/BCFtools [9], REDIttools [10] and others (e.g. [11] and [12]). One particular common procedure for the identification of RNA editing is based on arbitrary thresholds for the number of minimal variant reads and minimal variant frequency (=10%) while at least a coverage of 10 reads is required [13].

Based on our previous experience from developing ACCUSA2 [14], we implemented a new software package, the JAVA framework for accurate SNV assessment (JACUSA). JACUSA is a fast and precise solution for quantitative single nucleotide variant detection in RNA-DNA or RNA-RNA comparisons. JACUSA is primarily designed for the detection of position-specific editing events and readily integrates information from replicate experiments.

In the next sections, we will present our statistical framework and data processing steps in detail. We benchmark JACUSA on simulated data sets and compare its performance to other available and popular variant callers: REDIttools [10], MuTect [15], and SAMtools/BCFtools [9]. We will then discuss the performance of JACUSA and the other tested variant callers in a controlled biological setting using sequencing data from ADAR knockdown experiments with human embryonic kidney (HEK-293) cells. Herein, several gDNA and cDNA libraries were sequenced to facilitate RNA-DNA and RNA-RNA comparisons based on Illumina sequencing data. Moreover, we made use of published RNA-seq data from *Drosophila melanogaster* fly brains that either originate from a wild type strain or a strain with a genetically ablated *dADAR* gene. With the *Drosophila* samples, we specifically look at the identifiability of editing events in

protein coding exons in neuronal tissue where they have been reported previously [16].

Implementation

In the following, we present the JACUSA software in detail, discuss the test statistics that supports replicate experiments and a set of positional filters that enable the pruning of false positive variants for a more accurate detection of SNVs. Equally important, we have implemented parallel and memory-efficient read processing routines for better performance and usability.

Objective

The JACUSA software predicts single-nucleotide variant positions from head-to-head comparisons of read stacks/pileups from Illumina sequencing. In this manuscript, we focus on identifying nucleotide-level differences in RNA-DNA and RNA-RNA comparisons (see Fig. 1a). Our method is robust to differences in read coverage, takes replicate information into account and avoids false calls by removing typical artifacts from short read data. We discuss the power of our approach specifically in the context of RNA editing.

Statistical model

Previously, it has been shown that allele frequencies/counts are not accurately modeled by over-simplistic statistical models (e.g. a multinomial distribution) [17]. Typically, the observed variance will be higher than the theoretically expected variance in a multinomial model. This phenomenon is called overdispersion and will lead to false positive calls in variant detection. Therefore, we model DNA and RNA sequencing data with the Dirichlet-Multinomial distribution that accounts for overdispersion [18]. In the following, we use formulas and nomenclature defined in [19] adjusted to the alphabet of nucleotides $\Sigma = \{A, C, G, T\}$. We define $\mathbf{p} = (p_A, p_C, p_G, p_T)$ to be a random probability vector, such that $p_k : k \in \Sigma$ represents the base or allele probability for base k and the elements sum to 1. We can model \mathbf{p} with a Dirichlet distribution \mathcal{D} that has the parameter vector $\boldsymbol{\alpha} = (\alpha_A, \alpha_C, \alpha_G, \alpha_T)$:

$$p(\mathbf{p}) \sim \mathcal{D}(\alpha_A, \alpha_C, \alpha_G, \alpha_T) \quad (1)$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1} \quad (2)$$

$$\text{where } p_k > 0 \quad (3)$$

In [14], we estimated $\boldsymbol{\alpha}$ from base calls and their respective base call quality score using an empirical Bayesian method. The Dirichlet distribution \mathcal{D} is a conjugate of the multinomial distribution \mathcal{M} . Let $\mathbf{x} = (x_A, x_C, x_G, x_T)$ represent the sum of base calls at some location and let \mathbf{x} follow a multinomial distribution $\mathcal{M}(n, \mathbf{p}) = p(\mathbf{x}|\mathbf{p})$ where

$n = \sum_k x_k$ is the total number of observed bases. By integrating over \mathbf{p} we can combine \mathbf{x} and \mathbf{p} into the compound distribution that is called the Dirichlet-Multinomial:

$$\begin{aligned} \text{DirMult}(\mathbf{x}, \boldsymbol{\alpha}) &:= p(\mathbf{x}|\boldsymbol{\alpha}) = \int p(\mathbf{x}|\mathbf{p})p(\mathbf{p}|\boldsymbol{\alpha})d\mathbf{p} \\ &= \frac{(n!) \Gamma(\alpha_0)}{\Gamma(n + \alpha_0)} \prod_{k \in \Sigma} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(x_k + 1) \Gamma(\alpha_k)}, \end{aligned} \quad (4)$$

where $\alpha_0 = \sum_{k \in \Sigma} \alpha_k$.

An alternative interpretation of the Dirichlet-Multinomial is that of a hierarchical model:

$$\mathbf{p} \sim \mathcal{D}(\boldsymbol{\alpha})$$

$$\mathbf{x} \sim \mathcal{M}(\mathbf{p})$$

Let $D = \{\mathbf{x}_1, \mathbf{x}_i, \dots, \mathbf{x}_N\} : i \in \{1, \dots, N\}$ represent the base count vectors in N replicates and let \mathbf{x}_i be identically and independently distributed. Then $\boldsymbol{\alpha}$ can be estimated from D by maximum likelihood estimation of \mathcal{L} :

$$\mathcal{L}(\boldsymbol{\alpha}; D) = p(D|\boldsymbol{\alpha}) = \prod_i p(\mathbf{x}_i|\boldsymbol{\alpha})$$

In order to model uncertainty of $\boldsymbol{\alpha}$ we add a pseudocount term \mathbf{x}_p to the base call count vector: $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{x}_p$. The pseudocount term \mathbf{x}_p is calculated as a sum from observed quality score q_{BC} (i.e. variable terms) and a fixed noise term $\epsilon (=0.01)$ which models sequencing independent errors, which were derived empirically. q_{BC} is reported per base call as Phred quality score q_{BC} , which is logarithmically related to the base-calling error probability e_{BC} [20]:

$$e_{BC} = Pr\{\text{wrong BC}\} \quad (5)$$

$$q_{BC} = -10 \log_{10} e_{BC} \quad (6)$$

$$1 - e_{BC} = p_{BC} = Pr\{\text{called base}\} \quad (7)$$

In JACUSA, we assume that the error probability e_{BC} is independent of the called base. That is why, the error probability of an uncalled base is given by:

$$\frac{e_{BC}}{3} = Pr\{\text{uncalled base}\} \quad (8)$$

Using these considerations and referring to a specific base call by l , we define \mathbf{x}_p as:

$$\mathbf{x}_p = \sum_{1 \leq l \leq n} \begin{cases} \epsilon + \frac{e_{BC}^l}{3} & \text{for each uncalled base} \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Statistical test

We define our test statistic as a likelihood ratio of two samples $j \in \{I, II\}$ where the data of each sample is defined as the pseudocount adjusted base call vectors $\tilde{D}^j = \{\tilde{\mathbf{x}}_1^j, \tilde{\mathbf{x}}_i^j, \dots, \tilde{\mathbf{x}}_{N_j}^j\}$. We use the Dirichlet-Multinomial distribution to model \tilde{D}^j and estimate $\boldsymbol{\alpha}^j$ as explained in

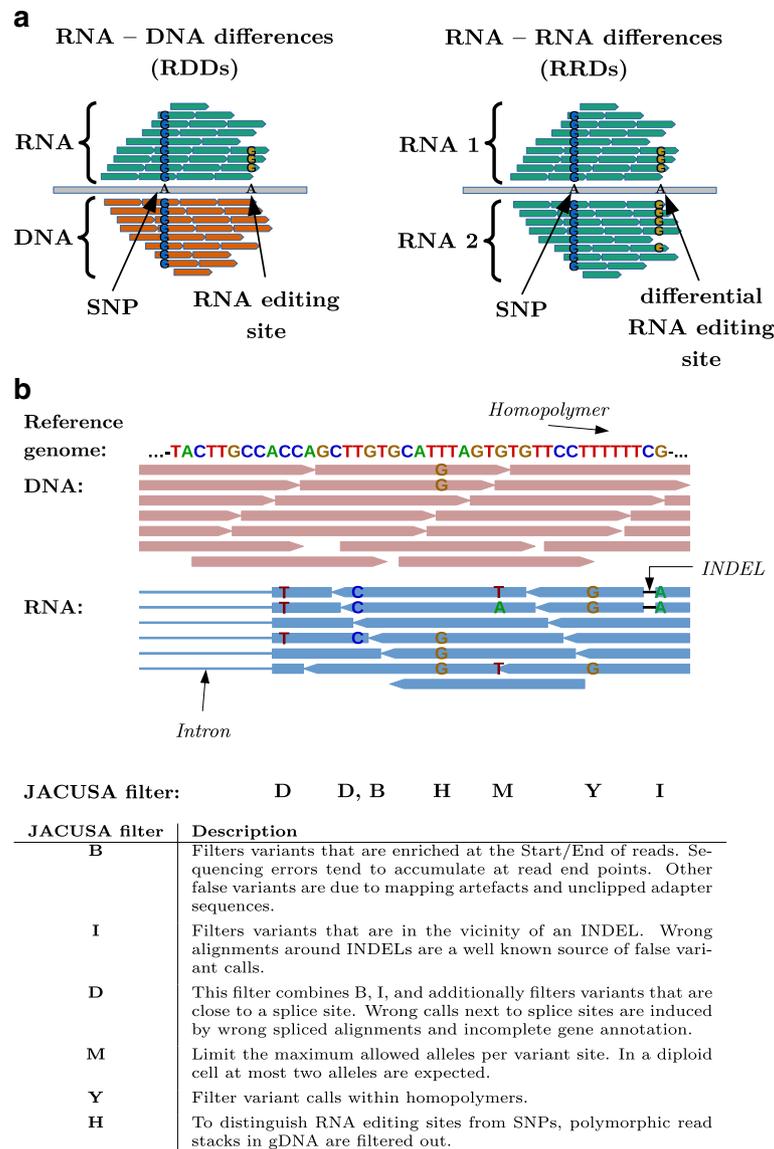


Fig. 1 Possible nucleotide comparisons and implemented JACUSA filter. **a** Graphical representation of RNA-DNA differences (RDDs) and RNA-RNA-differences (RRDs) in head-to-head comparisons of sequencing data. **b** Possible sequencing artifacts and their respective JACUSA filters

the previous section. We test against the null hypothesis H_0 that both samples originate from the same underlying distribution. The log-likelihood score function z (Eq. 10) will have higher values, the better each of the parameter vectors α^I and α^{II} represent the underlying data implying that each sample I and II has a different underlying distribution.

$$z = \log \frac{DirMult(\alpha^I; \tilde{D}^I) \cdot DirMult(\alpha^{II}; \tilde{D}^{II})}{DirMult(\alpha^{I,II}; \tilde{D}^I) \cdot DirMult(\alpha^{I,II}; \tilde{D}^{II})} \quad (10)$$

The coverage between two pileups may differ extremely between RNA-seq samples. This will sometimes lead to an overestimation of confidence in the base call vector \tilde{x} for the sample with higher coverage. We mitigate

this phenomenon by adjusting the underlying read stacks. In essence, large coverage differences between a single nucleotide count vector \tilde{D}_{homo} and count vectors with two or more nucleotides \tilde{D}_{hetero} are evened out by replacing the original \tilde{D}_{homo} with a copy of \tilde{D}_{hetero} where all variant positions have been replaced by the reference nucleotide. Depending on the encountered read stacks, JACUSA automatically switches to the optimal comparison mode.

Implemented filters

Many false positive RDD calls in RNA editing studies are related to mapping artefacts [8]. Short read mappers tend to produce incorrect alignments around INDEL positions that may be falsely identified as variant sites. Tools such

as GATK [21] allow to adjust for this effect by sensitive local realignment of reads that contain INDELS. Other false variant calls originate from uneven base call error distributions along short reads. This may be related to sequencing technology where base calls at read ends are less reliable. In JACUSA, we have implemented a panel of simple threshold based filters to remove the aforementioned and other artefacts (see Fig. 1). Our filters (D, B, I, Y) monitor the distance d of a given candidate site to relevant read features such as start/end, INDEL positions, homopolymeric regions, and splice sites and remove the candidate site from further consideration if a proportion r of all reads falls below the given distance cutoff $\leq d$.

Generally, it is common practice to define RDDs for homozygous genomic positions (filter H) and with less than three distinct base types (filter M). Moreover, we strongly recommend to remove PCR-duplicate reads from the input read sets to minimize biases, which are introduced by PCR amplification biases, before the actual JACUSA run (see Additional file 1: Section 4.4).

In silico benchmark

We define two benchmark scenarios (Fig. 2): 1) gDNA vs. cDNA simulates data for the identification of RNA-DNA differences (RDDs) and 2) cDNA vs. cDNA generates data for the identification of RNA-RNA differences (RRDs). The gDNA vs. cDNA represents the typical setup

for the detection of RNA editing sites. In this scenario, editing sites have been only implanted into the cDNA BAM file(s). In the cDNA vs. cDNA data setup, both data sources may contain base substitutions at different frequencies. This scenario can be interpreted as allele-specific expression or dynamic RNA editing changes. Herein, variants with pairwise different base frequencies ($\Delta > 0.1$) have been implanted into each corresponding cDNA BAM file. Additionally, to make the identification of variants more challenging, SNPs with pairwise similar base frequencies have been included into each cDNA BAM file (see Fig. 2). We use the human reference genome (hg19, chromosome 1) as a template to simulate genomic DNA (gDNA) and RNA-Seq reads. In total, 60,000 non-overlapping sites have been randomly chosen based on sufficient read coverage $5 \geq c \geq 1000$ and read mapping quality ≥ 20 in all simulated BAM files. The initial candidate set of non-overlapping sites has been divided into 30,000 variant and SNP sites, respectively. Each site is modeled with a variant target frequency as shown in Tables 1 and 2.

Moreover, we introduce additional variability by sampling the target variant frequencies from a Beta-Distribution with concentration parameter $\beta \in \{10, 50, 100\}$ representing sites with high, medium, and low variability around the expected target frequencies. Another parameter of our benchmark is the number of RNA-seq replicates. We benchmarked all scenarios

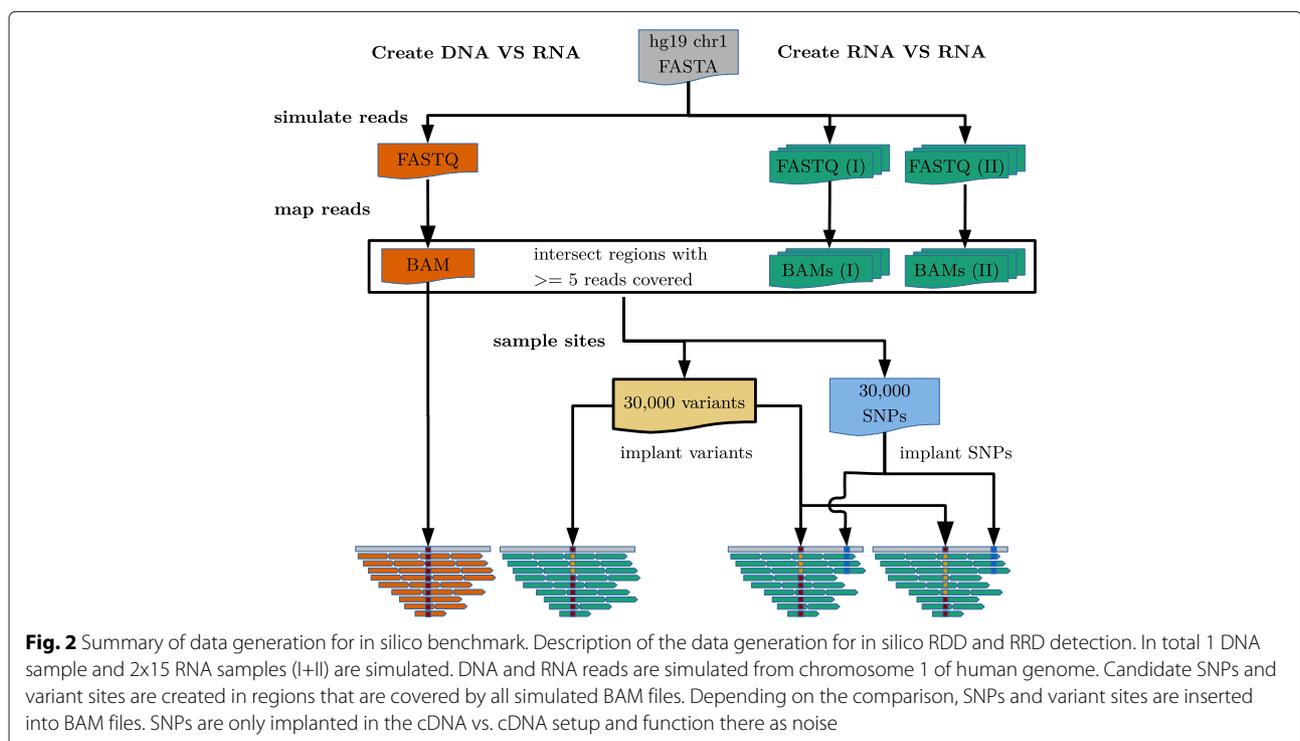


Table 1 Detailed statistics of the implanted sites for RNA-DNA-difference (RDD) benchmark setup

	gDNA	cDNA
Variants	0	30,000 positions
Δ variant freq.		≥ 0.01

with cDNA samples from 1 to 5 replicates. Each replicate setup is simulated 3 times, which amounts to 15 RNA-seq FASTQ files per benchmark (see Table 3 for details).

Additional details on the benchmark setup are given in Additional file 1: Section 3 and Table 4.

Sequencing the HEK-293 genome and transcriptome

HEK-293 genome sequencing

Genomic DNA was isolated from Flp-In T-REx HEK-293 cells (Invitrogen) using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich). DNA was fragmented in the BioRuptor Plus (Diagenode, setting “high” in a total volume of 150 μl (concentration 25 ng/μl), with 24 cycles (30 seconds on, 30 seconds off) in a 4 °C water bath, including a brief centrifugation after 12 cycles. The resulting fragmented DNA was converted to a sequencing library using the TruSeq DNA kit (Illumina) with PCR enrichment and sequenced on a Illumina HiSeq 2500 machine. In total > 10⁹ reads have been sequenced (see Additional file 1: Table S3). The gDNA-seq data have been deposited in the NCBI SRA under accession SRP050149.

HEK-293 transcriptome sequencing

HEK-293 strand-specific RNA-Seq data from [22] has been downloaded and processed as explained in Fig. 5. We used the hg19 human genome and ENSEMBL 75 annotation for mapping. The TopHat2 [23] mismatch parameter was set to 10 and reads with more than 5 mismatches were filtered subsequently (see Additional file 1: Section 4.3). Alu regions have been download and extracted from RepeatMasker annotation Ver. 4.0.2. [24].

RNA-seq data from Drosophila fly heads

We obtained published replicate paired-end RNA-seq data from Drosophila fly heads [25] (2 x 100nt, unstranded, accessions codes: NCBI SRA SRR485862-5).

Table 2 Detailed statistics of the implanted sites for RNA-RNA-difference (RRD) benchmark setup

	cDNA (I)	cDNA (II)
SNPs		30,000 positions
Δ variant freq.		≈ 0
Variants		30,000 positions
Δ variant freq.		≥ 0.1

Table 3 Description of in silico samples

	gDNA	cDNA (I)	cDNA (II)
Library type	gDNA, paired-end	RNA-Seq, paired-end	
Read length	2x100nt		2x100nt
Read count/coverage	30x	15,000,000 raw reads	
# of FASTQ files	1		3x5

The *Drosophila melanogaster* genome carries only a single copy *dADAR* gene. Two replicate RNA samples were generated from flies with wildtype and null alleles of *dADAR* (2 replicates each, FM7a strain background). We processed the data in the same way as the HEK-293 data sets using the Ensembl 75 Drosophila genome and annotations.

Results

In silico benchmark

We use two benchmark scenarios (Fig. 2) to compare JACUSA with other popular variant callers: REDIttools, SAMtools/BCFtools, and MuTect. The gDNA vs. cDNA scenario works with all variant callers while the cDNA vs. cDNA comparison scenario could be only tested with SAMtools/BCFtools and JACUSA. Equally important, SAMtools/BCFtools and JACUSA are the only two variant callers that support replicates in our benchmark. More details on the benchmark setup and how others and our software were used are given in section 3.1 of the Additional file 1.

Detection of SNVs in RNA-DNA comparisons

When no replicates are used, JACUSA shows a 6 – 10% higher true positive rate (TPR) as compared to the other tested methods while being competitive at the level of precision (see Fig. 3a, b). The single replicate scenario is highly relevant in practice, as RNA-seq replicate counts are typically low in RDD studies in the clinics. We specifically used the accuracy measure and the F-score to evaluate the balance between precision and true positive rate (see Additional file 1: Section 3.2). The main difference

Table 4 Summary of variant callers used for available benchmarks and their support for replicates

Variant caller	Support for replicates	gDNA vs. cDNA	cDNA vs. cDNA
SAMtools/BCFtools	x	x	x
REDIttools		x ^a	
MuTect		x	
JACUSA	x	x	x

^aThe REDIttools package supplies multiple methods to identify RNAediting sites. We employed the REDIttoolDenovo.py script because it provides a test-statistic. This method only utilizes RNA-Seq reads

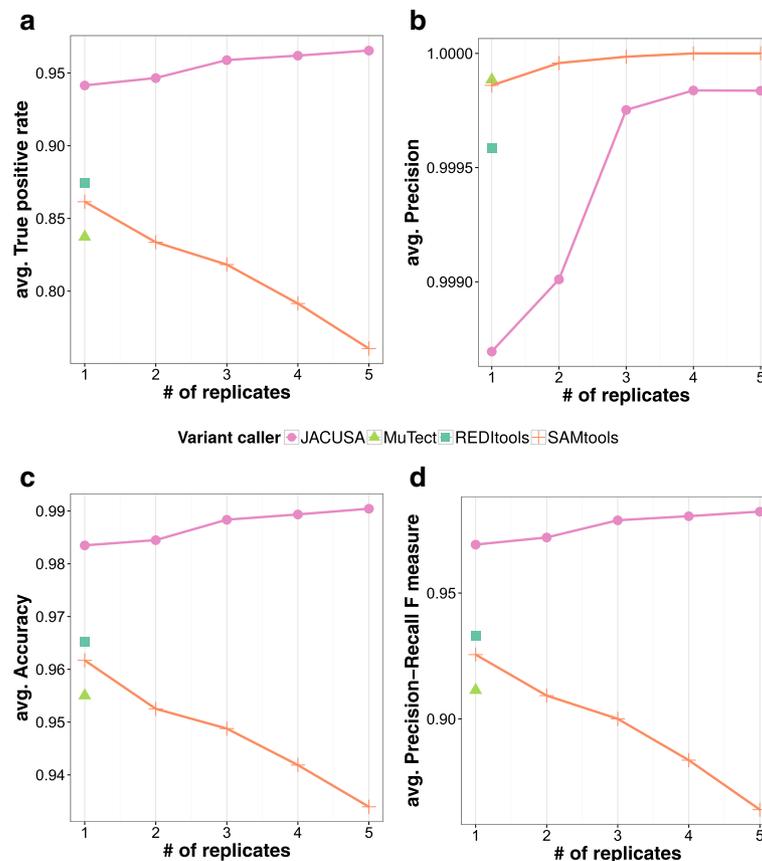


Fig. 3 Benchmark results for in silico RDD detection. **a** True positive rate, **b** Precision, **c** Accuracy plot, and **d** F-measure

between these performance measures is that the accuracy measure includes the number of true negatives.

Of all tested methods, JACUSA scores the highest in terms of accuracy and F-score (see Fig. 3). The trade-off between TPs and FPs can be nicely observed for the comparison of MuTect and REDIttools. While REDIttools shows a higher TPR (87,45% compared to 83,73% of MuTect, Fig. 3a), the precision is slightly higher for MuTect (99,99% compared to 99,96% for REDIttools, Fig. 3b). SAMtools/BCFtools scores third in terms of TPR and achieves together with MuTect the highest precision of 99,99%.

JACUSA takes advantage of replicate information and shows a steady increase in performance with the number of employed replicates. SAMtools/BCFtools on the other hand displays only growing precision with increasing number of replicates and the remaining performance measures are decreasing. The drop in performance is highest for 5 replicates and amounts to more than 15% of TRP. JACUSA consistently performs better than SAMtools/BCFtools in terms of TPR, F-score, and accuracy (see Fig. 3a-d).

Additional results and details are given in Additional file 1: Section 3.3.

Detection of SNVs in RNA-RNA comparisons

In the cDNA vs. cDNA scenario we replace the single gDNA sample by one or more cDNA samples with variant sites where the target frequency differs by more than 10% between both cDNA pools. We introduce polymorphic positions of equal target frequency into both samples. The goal of this benchmark is to test the ability of the respective variant caller to distinguish between variant sites with a target frequency difference of $\Delta > 0.1$ and polymorphic positions with equal target frequency.

As before, we evaluate the variant callers by comparing overall performance measures such as F-score and accuracy. A general observation is the lower accuracy in Fig. 4a for calling variant sites in cDNA vs cDNA comparisons. In essence, it is a much harder task than contrasting gDNA vs. cDNA samples. In terms of true positive rate, JACUSA outperforms SAMtools/BCFtools in this scenario by at least 40% when replicates are available and by over 35% when no replicates are available (see Fig. 4a).

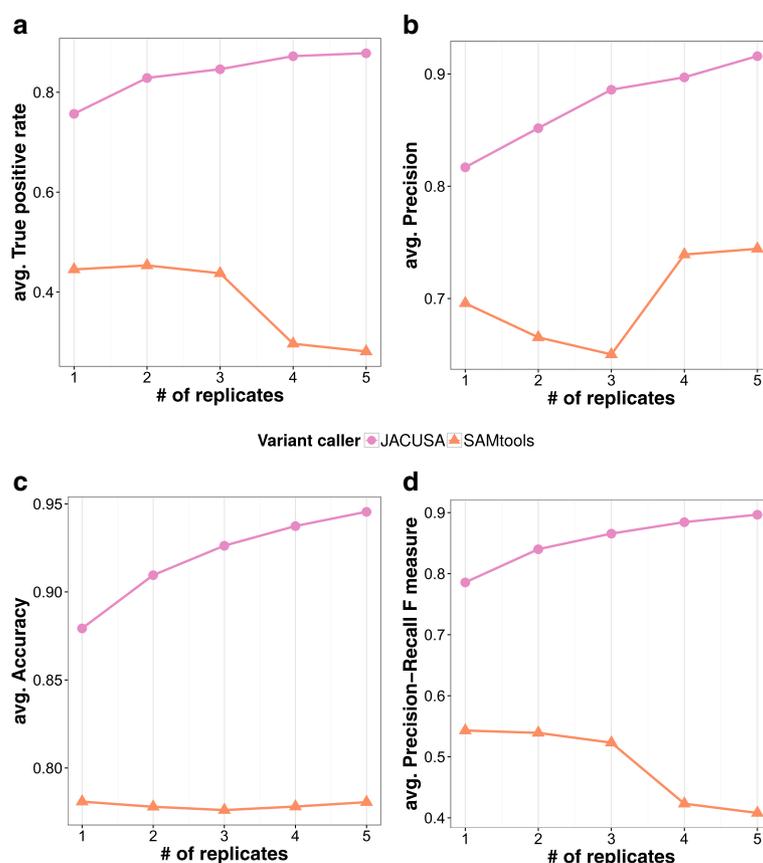


Fig. 4 Benchmark results for in silico RRD detection. **a** True positive rate, **b** Precision, **c** Accuracy plot, and **d** F-measure

Next, we combined true positives (TPs) and false positives (FPs) into composite measures and observed 10–16% better average accuracy for JACUSA (see Fig. 4c). This is even more pronounced for the F-score measure, where JACUSA is performing at least 20% better in all tested replicate scenarios (see Fig. 4d).

Additional results and details are given in Additional file 1: Section 3.4. A general overview on the single thread runtime of each tested software is shown in Additional file 1: Section 3.6.

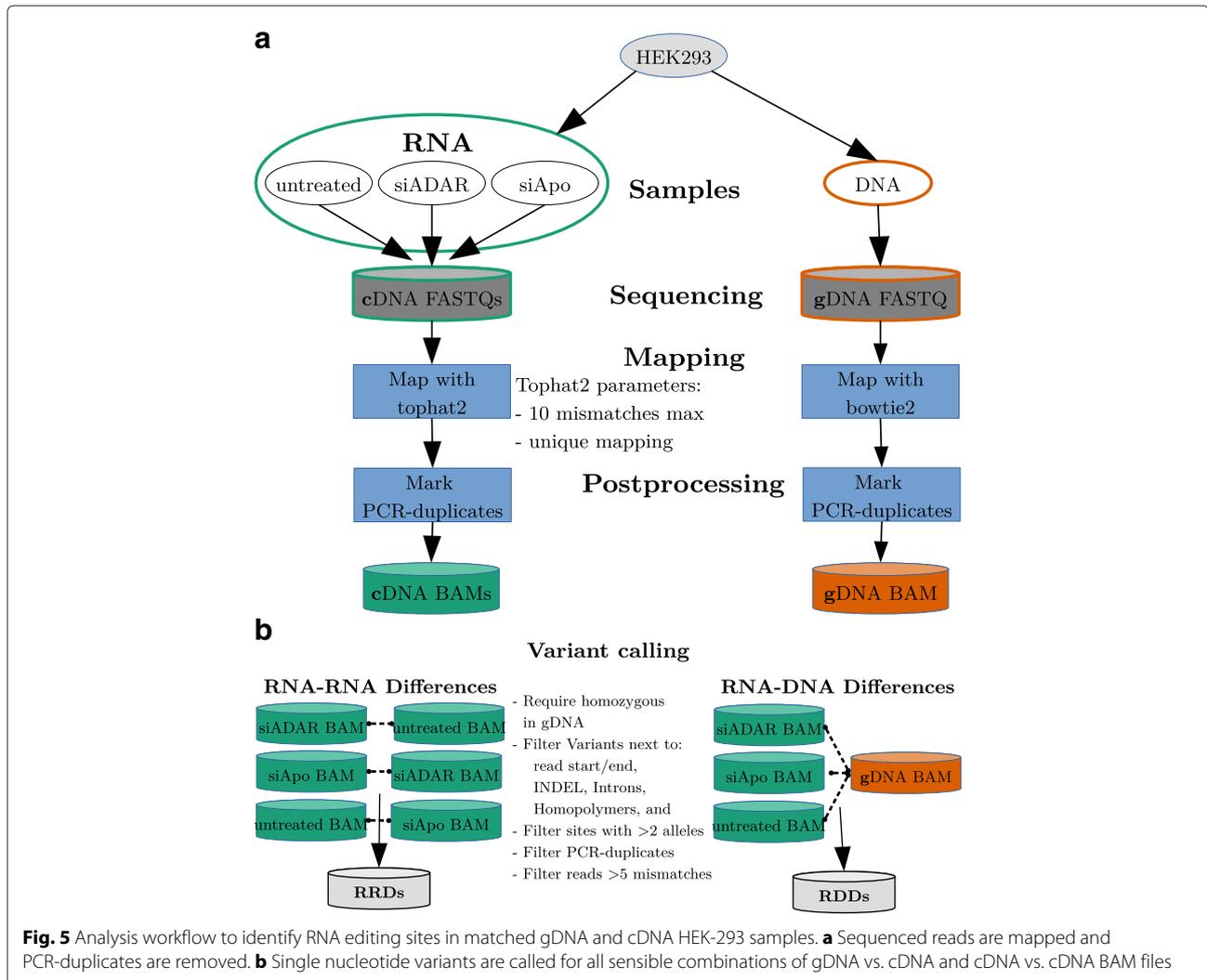
Editing in HEK-293 cells

To assess the performance of JACUSA in practice, we designed a controlled experiment to generate sequencing input data from cell culture experiments (see Fig. 5a). Briefly, we resequenced the genome of HEK-293 cells to an average coverage of 30x (gDNA data). We obtained matching cDNA data from our previously published study [22]. Cells were either untreated or have been subjected to siRNA knockdown experiments targeting either ADAR 1+2 (siADAR) or APOBEC3 B,C, and F (siAPOBEC3). The ADAR and APOBEC3 family members have been previously observed as mRNA-binding proteins in a

transcriptome-wide proteomics screen of the same cell type [26]. However, the APOBEC3 family members did not show significant C-to-U RNA editing activity in our assays.

Subsequently, we conducted gDNA vs. cDNA comparisons on the aforementioned data sets and predicted RNA editing sites with SAMtools/BCFtools, MuTect, and JACUSA. For each variant caller, we selected optimal thresholds for the HEK-293 data set based on our results from the in silico data set: gDNA vs. cDNA score threshold is 1.15 and cDNA vs. cDNA threshold is 1.56. Additional details on selecting score thresholds are given in Additional file 1: Section 3.5.

For MuTect and REDIttools we adopted a strategy presented in [12] to utilize replicate information by first calling variants on pooled biological replicates and finally filtering and requiring that the primarily identified variants are present in all replicates. We used JACUSA as explained in Fig. 5b to detect RNA editing sites utilizing replicates. Additional details on the workflow and results are given in Additional file 1: Section 4 and following. All editing site predictions are listed in Additional file 2.



Calling RDDs from HEK-293 data

In total, 2 biological replicates have been created per condition and were sequenced twice to assess the biological and technical variability. By computing RDDs on each replicate with JACUSA, we could show an excellent agreement among replicates from the same condition (see Fig. 6a). Subsequently, we merged all technical replicates and identified our definite list of RDDs from comparing one gDNA vs. two biological replicate samples for each condition.

Our comparison of gDNA vs. RNA from untreated cells yielded 15,461 variant sites for JACUSA (with a proportion of 92.2% $A \rightarrow G$ sites, see Fig. 6b and Table 5). This number drops to 8722 for the siAPOBEC3 RNA samples (91.2% $A \rightarrow G$) and, as expected, to 3371 sites for the siADAR RNA samples.

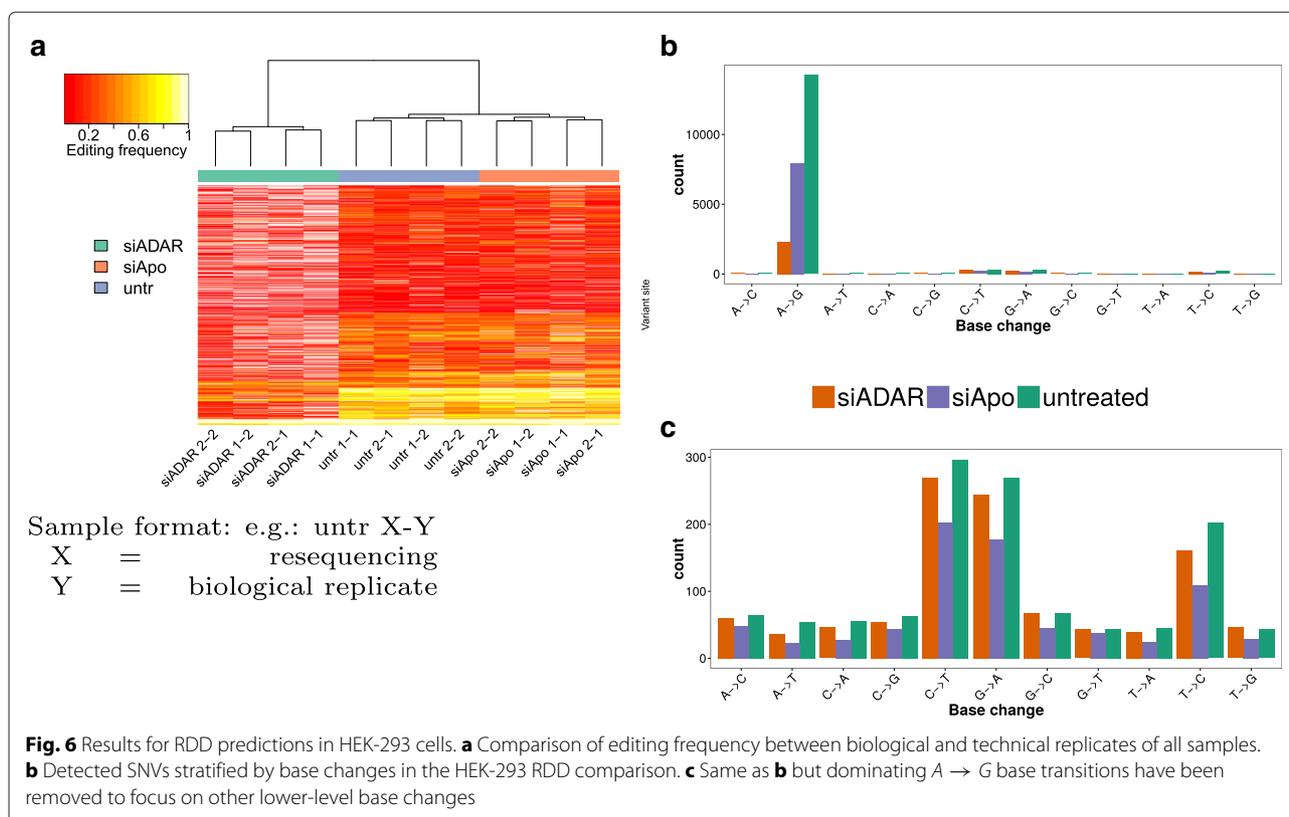
The siAPOBEC3 transfection experiment (mock) already leads to a reduction of editing sites. Editing levels

are further reduced by targeting the correct enzyme class (siADAR experiment).

Interestingly, the non $A \rightarrow G$ sites identified by JACUSA (1203 in total, Fig. 6c) consist mainly of three base substitutions: $C \rightarrow T$ (24.7%) editing is a known but rare modification that is mediated by APOBEC1 [27] and $T \rightarrow C$ and $G \rightarrow A$ variants (39.2%), which are the reverse complement versions of the canonical editing events.

JACUSA identified the highest number of RDDs (15,461 vs 11,191 for SAMtools/BCFtools) and showed a comparable fraction of $A \rightarrow G$ sites (92.2% vs 93.3% for SAMtools/BCFtools) among all tested variant callers (see Table 5).

MuTect identified far fewer RDDs ($\approx 25\%$ less) in comparison to the other variant callers while achieving second highest fraction of $A \rightarrow G$ sites (92.5%). This is in line with the in silico benchmark results on MuTect indicating a high precision but a lower recall. In summary, all variant



callers identify RDDs with a fraction of A → G sites in the range of 90.1 and 93.3%, while the total number of variants varies greatly from 7605 (MuTect) up to 15,461 (JACUSA).

Agreement between RDD calls

All four software solutions report a set of 6064 shared RDD sites for the untreated RNA sample, which show a high proportion of A → G sites (94.6%) (see Fig. 7a). The second largest overlap of 3314 RDD predictions is seen for SAMtools/BCFtools, REDtools, and JACUSA (94.2% A → G sites). Strikingly, JACUSA identifies 2,634 additional RDDs, which are not reported by any other software tool and yet attain a proportion of 87.9% A → G sites. This is far more than for sites that were exclusively reported

by SAMtools/BCFtools (59.5% A → G sites), REDtools (35.8%) or MuTect (42.3%).

Moreover, a detailed assessment of RDDs sites, which have been exclusively reported by JACUSA, shows a low mean editing level and mean coverage (see Fig. 7b and c).

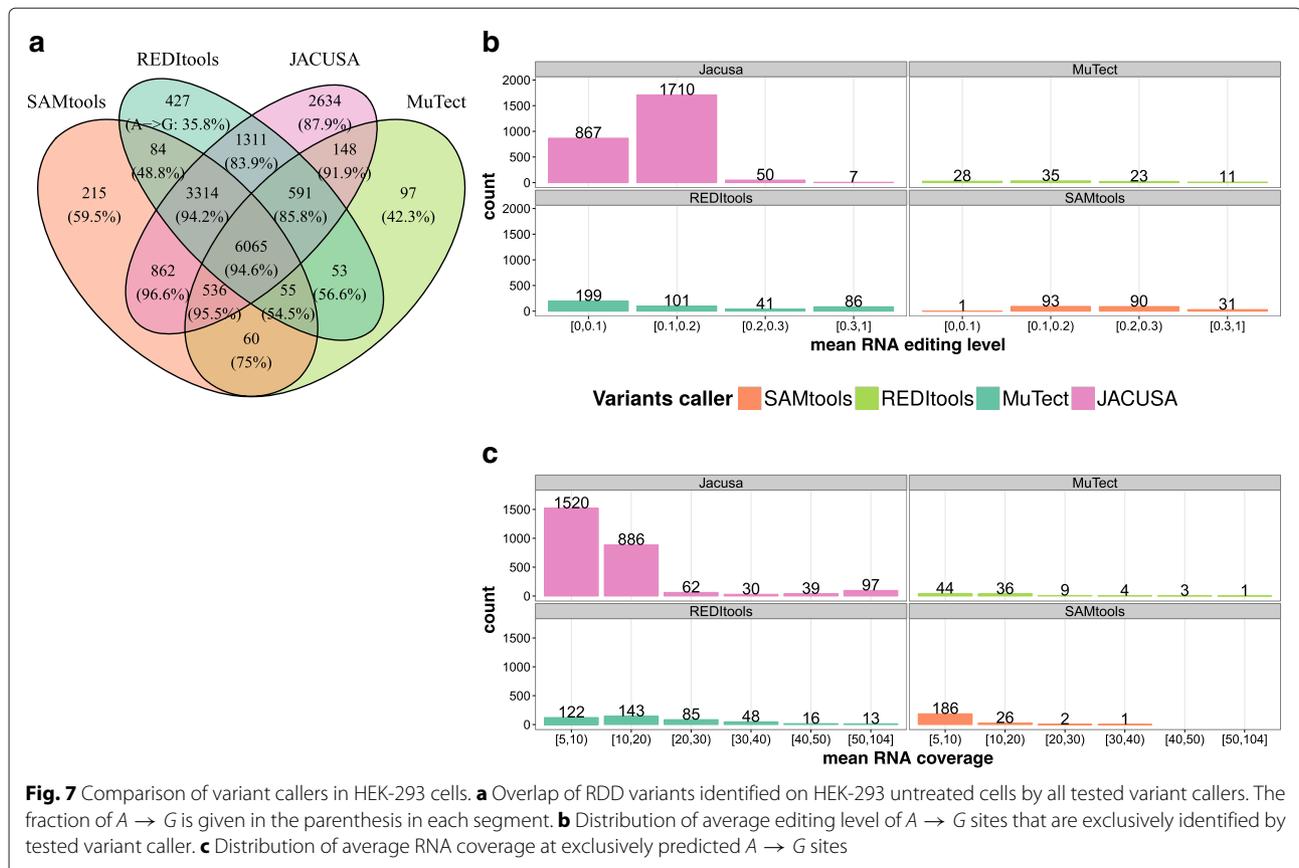
Response of RDD calls in ADAR knockdown

To control the effect of any siRNA knockdown treatment on RNA editing levels (see Fig. 6b), we contrast editing levels of RDDs between siAPOBEC3 and siADAR samples. As mentioned earlier, JACUSA had identified 8722 RDDs in cells treated with siAPOBEC3 (siApo) of which 7953 were A → G substitutions. We classify these A → G sites as true positives if they show a drop in their editing frequency in an siADAR vs. siAPOBEC3 knockdown.

As shown in Table 6, we could assess editing level changes on 7084 RDD sites that had sufficient read coverage in both siRNA knockdown data sets (5 reads per position per replicate in siAPOBEC3 and siADAR samples). JACUSA identifies the highest number of RNA editing sites (6,466) out of which ≈ 98% show lower editing levels in siADAR samples than in samples from siAPOBEC3 treated cells. This means that JACUSA reports 6375 true positive A → G sites out of a set of 6,466 predicted sites, the highest among all tested variant callers. Figure 8a and c depict this important result for each individual site. The clear shift of editing frequency was specific to A → G

Table 5 Predicted RDDs for each treatment of HEK-293 cells. Fraction of A → G RNA editing sites is provided in parenthesis

Variant caller	gDNA vs. treatment		
	untreated	siADAR	siAPOBEC3
SAMtools/BCFtools	11,191 (93.3%)	2117 (68.9%)	6423 (92%)
MuTect	7605 (92.5%)	1793 (69.4%)	4181 (90.2%)
REDtools	11,900 (90.1%)	2729 (59.7%)	6985 (88.6%)
JACUSA	15,461 (92.2%)	3371 (68.3%)	8722 (91.2%)



and could not be observed for any other base substitution (see Table c in Fig. 8). In summary, JACUSA identifies at least > 20% more editing sites than any tested variant caller while its editing sites show an equal responsiveness to ADAR knockdown treatment.

Detection of differential RNA editing from RNA-RNA comparisons

Another JACUSA application is to detect sites of differential RNA editing from RNA-seq data only. This could be effected through a direct assessment of RNA-RNA differences (RRD) in the absence of genomic sequencing data. We reasoned that one way to validate RRD site detection and ultimately differential A → G editing is to use our available RNA/DNA-seq data in the following way:

We screen our samples from siADAR and siAPOBEC3 knockdowns for RRDs. Our assumption is that APOBEC3 family members do not influence A → G editing and siRNA transfection effects cancel out in this comparison. “True” A → G editing sites should show a lower editing frequency in the siADAR knockdown. For the siADAR vs. siAPOBEC3 comparison, SAMtools/BCFtools predicts 6368 RRD sites and JACUSA predicts 5366 RRD sites (see Table 7). Out of these, 3352 RRDs are predicted by both SAMtools/BCFtools (52.6% of all SAMtools/BCFtools predictions) and JACUSA (64.5% of all JACUSA predictions) (see Fig. 9a).

Subsequently, we retained RRDs that had at least 10x read coverage in the gDNA sample and checked if predicted sites are homozygous in the genome.

Table 6 Comparison of average editing levels of detected RRDs on siADAR and siAPOBEC3 (siApo) treated HEK-293 cells

Variant caller	RDDs in gDNA vs. siApo	Covered in siADAR vs. siApo	A → G Editing Sites	Avg. Editing level siADAR < siApo
SAMtools/BCFtools	6423	5066 (78.87%)	4691 (92.60%)	4630 (98.700%)
MuTect	4181	3415 (81.68%)	3087 (90.40%)	3043 (98.575%)
REDIttools	6985	5823 (83.36%)	5180 (88.96%)	5099 (98.436%)
JACUSA	8722	7084 (81.22%)	6466 (91.28%)	6375 (98.593%)

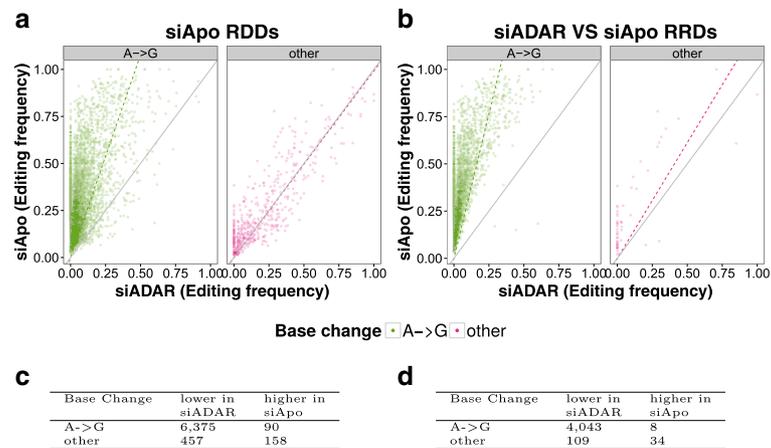


Fig. 8 Properties of RDDs in HEK-293 cells. **a** Comparison of editing frequency of siADAR samples and RDDs detected in siAPOBEC3 (siApo) treated cells. (Dashed line(s) correspond(s) to regression line(s)) **b** Editing frequency of sites that are identified as divergent in RRD comparison of treatments. Tables **c** + **d** show details of editing frequencies statistics for scatterplots **a** and **d**, respectively

Sites that are not homozygous in DNA represent putative SNPs and are typically removed from the candidate set when identifying RNA editing sites in RDD comparisons. As this information is not visible to SAMtools/BCFtools and JACUSA, we reasoned that a lower fraction of SNP sites among identified RRDs would indicate a better performance on calling differential RNA editing events. In essence, JACUSA precision is at 83.0% (4284 true sites vs 5161 candidate sites) Table 8 while SAMtools/BCFtools attains only 67.8% (4088 true sites vs. 6026 candidate sites).

We compared the fraction of RRDs that after coverage filtering were potential SNPs and found that SAMtools/BCFtools predictions contained 15% more putative polymorphic sites than JACUSA (see Fig. 9b).

In summary, RRDs predicted by JACUSA showed a lower overlap with potential polymorphic sites and the fraction of $A \rightarrow G$ editing sites was higher than the candidates called by SAMtools/BCFtools. The editing frequency of 4,043 $A \rightarrow G$ sites was smaller in siADAR treated cells whereas only 8 would show a higher editing frequency in siADAR treated cells (see Fig. 8b). The clear shift of editing frequency was specific to $A \rightarrow G$ and could not be observed for any other base substitution (see Tables in Fig. 8b+d).

Table 7 Summary of all detected RRDs for all possible treatment combinations on HEK-293 cells

Variant caller	siADAR vs. siApo	siADAR vs. untreated	siApo vs. untreated
SAMtools/BCFtools	6368	8195	7462
JACUSA	5366	6977	2701

Editing events across genomic features

Another important aspect is the genomic distribution of our editing predictions. We stratified our RDD and RRD predictions by gene-centric (exon, introns, etc.) and repeat-centric categories (Alu, non-Alu and no repeat regions). As expected, most RDD predictions are made in regions that are annotated as Alu repeats. Prediction accuracy drops dramatically for non-Alu repeat regions and even more so for non-repeat regions. For details see Additional file 1: Tables S5-S8. This holds true for all four tested SNV callers. This effect seems to be independent of gene-centric features and strongly correlates with repeat type. We observed that most RDD sites in non-repeat regions cannot be explained by $A \rightarrow G$ editing. We also cannot exclude the possibility that HEK-293 cells generally show very little RNA editing in non-Alu regions. Nevertheless, JACUSA identifies most $A \rightarrow G$ sites in absolute numbers.

The same phenomenon becomes more evident for the RRD comparisons (see Additional file 1: Tables S9 and S10). Herein, hardly any $A \rightarrow G$ sites are predicted in non repeat regions, by both SAMtools and JACUSA.

Differential editing in Drosophila fly heads

We reasoned that our HEK-293 cell data sets could be complemented by an independent data set with a controlled experimental design for testing RRD site discovery. To this end, we analysed published RNA-seq data from Drosophila fly heads [25]. Rodriguez et al. use a genetic approach to ablate the activity of the single copy *dADAR* gene in the fruit fly (human *ADARBI* homolog). This is a favorable system for fine-mapping editing sites as editing activity depends only on a single enzyme in the fruit fly. Moreover, editing in coding exons, which are

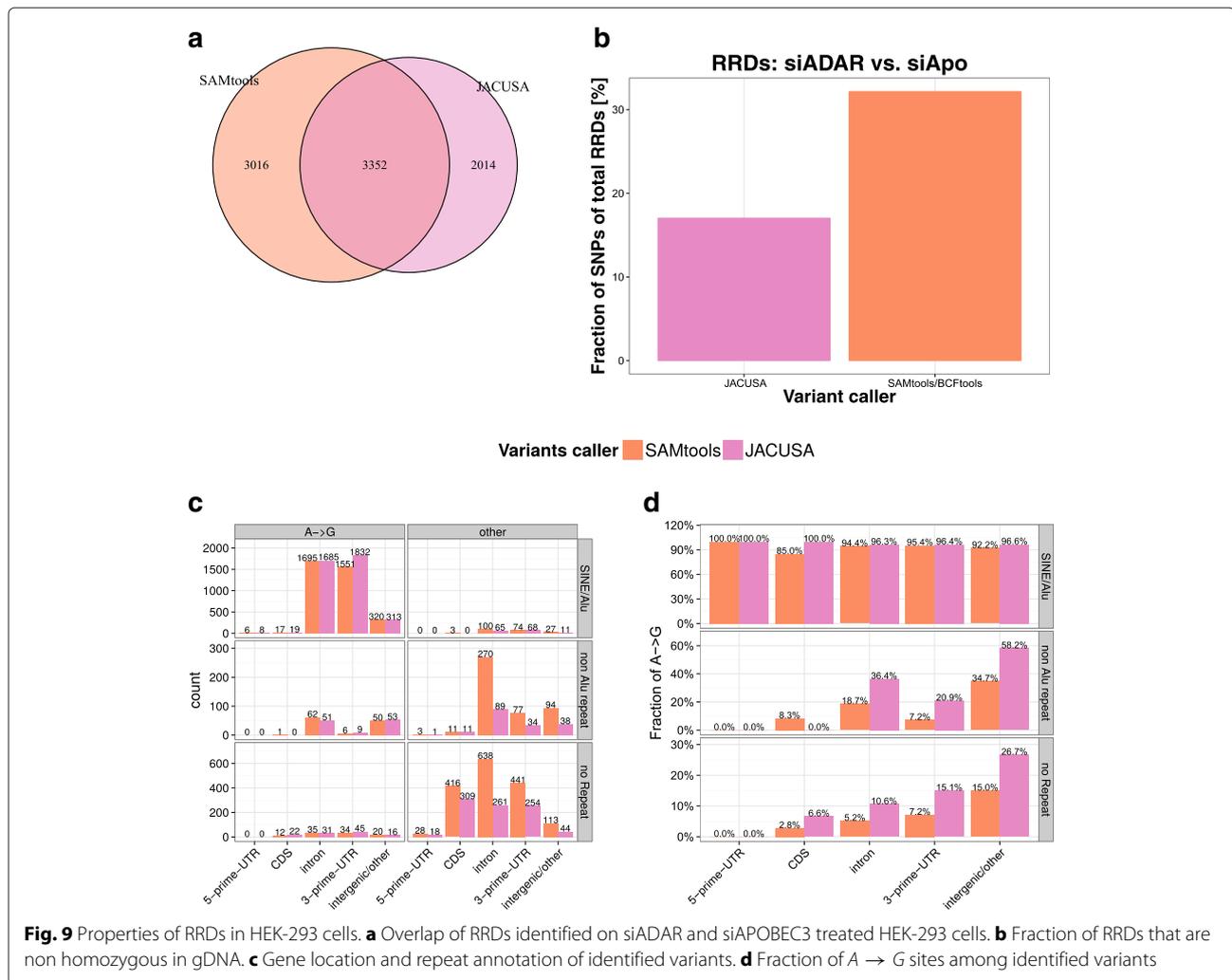


Fig. 9 Properties of RRDs in HEK-293 cells. **a** Overlap of RRDs identified on siADAR and siAPOBEC3 treated HEK-293 cells. **b** Fraction of RRDs that are non homozygous in gDNA. **c** Gene location and repeat annotation of identified variants. **d** Fraction of A → G sites among identified variants

expressed in the fly brain, has been described previously [16]. In summary, JACUSA detected 931 RRD candidate sites (see Additional file 1: Tables S11 and S12) while SAMtools/BCFtools predicted 781 RRD candidates. However, while the vast majority (92.1%) of JACUSA RRD sites are A → G sites, just 86.3% of all SAMtools/BCFtools predictions are (see Fig. 10). Overall, JACUSA predicted 383 RRD sites in coding exons. A closer inspection showed that 336 (87.7%) of these are bona fide A → G sites (see Additional file 1: Tables S11 and S12). This analysis demonstrated that JACUSA is able to accurately predict editing events in RNA-RNA comparisons on an

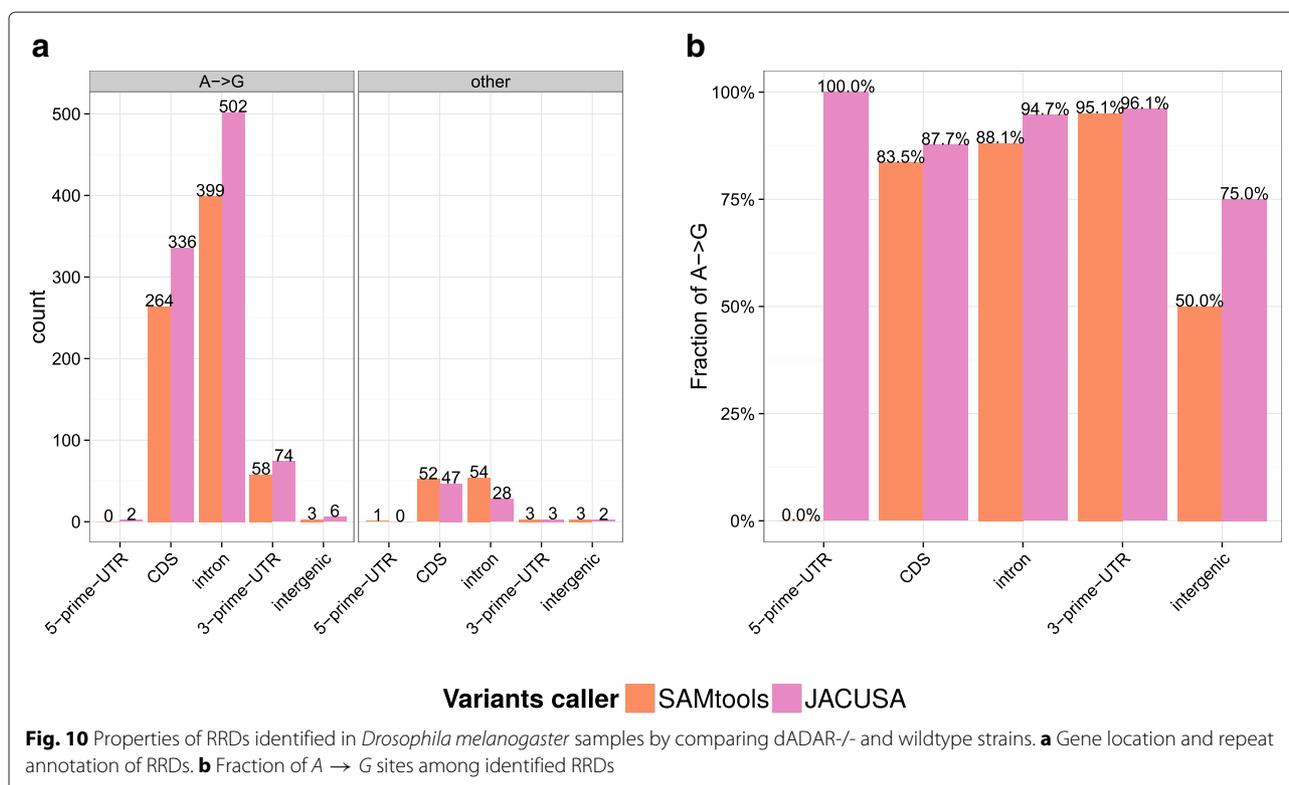
independent data set as well. All editing site predictions are listed in Additional file 3.

Conclusion

In this manuscript, we have presented JACUSA as an accurate and fast one-stop solution to identify site-specific SNV events in matched sequencing samples. JACUSA outperformed other SNV callers in an in silico benchmark that assessed SNV calling performance in terms of identifying site-specific RNA-DNA differences (RDDs) and RNA-RNA differences (RRDs). While the first benchmark is the typical scenario for identifying RNA editing sites

Table 8 Comparison of average editing levels of RNA editing sites that have been identified as RRDs in siADAR and siAPOBEC3 (siApo) treated HEK-293 cells

Variant caller	RRDs in siADAR vs. siApo	Covered in gDNA	Homozygous in gDNA	A → G Editing Sites	Avg. Editing level siADAR < siApo
SAMtools	6368	6026 (94.6%)	4088 (67.8%)	3838 (93.9%)	3731 (97.2%)
JACUSA	5366	5161 (96.2%)	4284 (83.0%)	4051 (94.6%)	4043 (99.8%)



from homozygous genomic positions, the second benchmark represents another interesting case of identifying condition specific changes in editing frequencies.

JACUSA shows the best recall and competitive precision in comparison to all tested software solutions. The performance gain over its competitors is especially visible for the detection of RNA-RNA differences. In terms of recall, JACUSA outperforms SAMtools/BCFtools in the RRD scenario by at least 40% when replicates are available and by over 35% when no replicates are available. Intriguingly, this is not at the expense of precision which is at least 10% better over all tested number of replicates.

In practice, we tested JACUSA in a controlled experimental setup where we generated DNA and RNA-seq data from HEK-293 cells. Similar to the in silico benchmark, we first identified candidate sites of RNA editing via RDD comparisons and checked if their editing frequency would respond to changes in ADAR protein levels by siRNA knockdown experiments. With this setup, we could nicely demonstrate that JACUSA has a better recall and comparable precision to other tested variant callers in identifying A → G editing sites in RNA-DNA comparisons.

Subsequently, we assessed the RRD or differential editing scenario by predicting SNVs between replicate siAPOBEC3 and siADAR RNA samples. Again, JACUSA overall predicts more sites in homozygous DNA positions

and a greater proportion of A → G editing sites than SAMtools (83.0% vs 67.8%) in this RNA-RNA comparison scenario on HEK-293 RNA-seq data.

These results were further corroborated by looking at an independent RNA-seq data set from *Drosophila melanogaster* heads. Herein, JACUSA reports the highest number of RNA editing sites (857 vs 674) with much higher precision (92.1% vs 86.3% of all RRD sites).

In summary, JACUSA is a versatile software for the precise and sensitive detection of single nucleotide level differences in DNA-RNA as well as RNA-RNA comparisons from Illumina sequencing data. In this manuscript, we have specifically explored its excellent ability to detect site-specific RNA editing events.

Availability and requirements

gDNA-seq data have been deposited in the NCBI SRA under accession SRP050149.

Project name: JACUSA

Project home page: <https://github.com/dieterich-lab/JACUSA>

Operating system(s): N/A

Programming language: JAVA 1.6

Other requirements: none

License: GPL-3.0

Additional files

Additional file 1: Supplementary Text. (PDF 2816 kb)

Additional file 2: Excel Spreadsheet with predictions on HEK-293 DNA and RNA sequencing data. (XLS 8867 kb)

Additional file 3: Excel Spreadsheet with predictions on Fly head RNA sequencing data. (XLS 259 kb)

Abbreviations

ADAR: Adenosine deaminases that act on RNA; RDD: RNA-DNA-differences; RRD: RNA-RNA-differences; RBP: RNA binding protein; SNP: Single nucleotide polymorphism; SNV: Single nucleotide variant

Acknowledgments

All authors would like to thank all members of the Dieterich and Landthaler Labs for numerous valuable discussions.

Funding

CD and MP acknowledge funding by the Max Planck Society.

Authors' contributions

CD conceived and designed the study. MP implemented JACUSA. MP and CD performed data analyses and wrote the manuscript with input from all other authors. EW carried out all experiments in HEK-293 cells. ML and UO provided materials and valuable feedback in the course of this project. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Max Planck Institute for Biology of Ageing, Joseph-Stelzmann Str. 9b, 50931 Cologne, Germany. ²Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany. ³Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology at the Department of Internal Medicine III, University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany. ⁴German Center for Cardiovascular Research (DZHK) - Partner site Heidelberg/Mannheim, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany.

Received: 26 November 2016 Accepted: 16 December 2016

Published online: 03 January 2017

References

- Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Ann Rev Biochem.* 2010;79:321–49. doi:10.1146/annurev-biochem-060208-105251.
- Alseth I, Dalhus BR, Bjørås M. Inosine in DNA and RNA. *Curr Opin Genet Dev.* 2014;26:116–23. doi:10.1016/j.gde.2014.07.008.
- Slotkin W, Nishikura K. Adenosine-to-inosine RNA editing and human disease. *Genome Med.* 2013;5(11):105. doi:10.1186/gm508.
- Kawahara Y, Ito K, Sun H, Aizawa H, Kanazawa I, Kwak S. Glutamate receptors: RNA editing and death of motor neurons. *Nature.* 2004;427(6977):801. doi:10.1038/427801a.
- Daniel C, Lagergren J, Öhman M. RNA editing of non-coding RNA and its role in gene regulation. *Biochimie.* 2015. doi:10.1016/j.biochi.2015.05.020.
- Deininger P. Alu elements: know the sines. *Genome Biol.* 2011;12(12):236. doi:10.1186/gb-2011-12-12-236.
- Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. *Science (New York, N.Y.)* 2011;333(6038):53–8. doi:10.1126/science.1210624.
- Kleinman CL, Adoue V, Majewski J. RNA editing of protein sequences: a rare event in human transcriptomes. *RNA.* 2012;18(9):1586–96.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England).* 2011;27(21):2987–93. doi:10.1093/bioinformatics/btr509.
- Picardi E, Pesole G. REDtools: High-throughput RNA editing detection made easy. *Bioinformatics.* 2013;29(14):1813–1814. doi:10.1093/bioinformatics/btt287.
- Porath HT, Carmi S, Levanon EY. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat Commun.* 2014;5:4726. doi:10.1038/ncomms5726.
- Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods.* 2013;10(2):128–32.
- Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression. *Cell Rep.* 2013;5(3):849–60. doi:10.1016/j.celrep.2013.10.002.
- Piechotta M, Dieterich C. ACCUSA2: Multi-purpose SNV calling enhanced by probabilistic integration of quality scores. *Bioinformatics.* 2013;29(14):1809–10. doi:10.1093/bioinformatics/btt268.
- Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013;31(3):213–9. doi:10.1038/nbt.2514. NIHMS150003.
- Li JB, Church GM. Deciphering the functions and regulation of brain-enriched a-to-i RNA editing. *Nat Neurosci.* 2013;16(11):1518–22. doi:10.1038/nn.3539.
- Heinrich V, Stange J, Dickhaus T, Imkeller P, Krüger U, Bauer S, Mundlos S, Robinson PN, Hecht J, Krawitz PM. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.* 2012;40(6):2426–31.
- Poortema K. On modelling overdispersion of counts. *Statistica Neerlandica.* 1999;53(1):5–20. doi:10.1111/1467-9574.00094.
- Minka T. Estimating a Dirichlet distribution. Technical report, MIT. 2000.
- Ewing B, Green P. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res.* 1998;8(3):186–94.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303. doi:10.1101/gr.107524.110.
- Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, et al. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.* 2015;10(2):170–7.
- Kim D, Perte G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14(4):36. doi:10.1186/gb-2013-14-4-r36.
- Smit A, Hubley R, Green P. RepeatMasker open-4.0. 2013–2015. <http://www.repeatmasker.org>.
- Rodriguez J, Menet JS, Rosbash M. Nascent-seq indicates widespread cotranscriptional RNA editing in *Drosophila*. *Mol Cell.* 2012;47(1):27–37.
- Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, Wyler E, Bonneau R, Selbach M, Dieterich C, Landthaler M. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell.* 2012;46(5):674–90. doi:10.1016/j.molcel.2012.05.021.
- Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN. Transcriptome-wide sequencing reveals numerous apobec1 mRNA-editing targets in transcript 3'UTRs. *Nat Struct Mol Biol.* 2011;18(2):230–6.