

ARTICLE

Received 28 Jan 2014 | Accepted 16 Jul 2014 | Published 27 Aug 2014

DOI: 10.1038/ncomms5726

OPEN

# A genome-wide map of hyper-edited RNA reveals numerous new sites

Hagit T. Porath<sup>1</sup>, Shai Carmi<sup>2</sup> & Erez Y. Levanon<sup>1</sup>

Adenosine-to-inosine editing is one of the most frequent post-transcriptional modifications, manifested as A-to-G mismatches when comparing RNA sequences with their source DNA. Recently, a number of RNA-seq data sets have been screened for the presence of A-to-G editing, and hundreds of thousands of editing sites identified. Here we show that existing screens missed the majority of sites by ignoring reads with excessive ('hyper') editing that do not easily align to the genome. We show that careful alignment and examination of the unmapped reads in RNA-seq studies reveal numerous new sites, usually many more than originally discovered, and in precisely those regions that are most heavily edited. Specifically, we discover 327,096 new editing sites in the heavily studied Illumina Human BodyMap data and more than double the number of detected sites in several published screens. We also identify thousands of new sites in mouse, rat, opossum and fly. Our results establish that hyper-editing events account for the majority of editing sites.

<sup>1</sup>The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 5290002, Israel. <sup>2</sup>Department of Computer Science, Columbia University, New York, New York 10027, USA. Correspondence and requests for materials should be addressed to E.Y.L. (email: Erez.Levanon@biu.ac.il).

**R**NA editing is a frequent post-transcriptional modification that leads to nucleotide changes at specific locations. As such, it can be detected directly by sequencing, allowing large-scale screening of editing events. Adenosine-to-inosine (A-to-I) editing is the most common form of editing in mammals, and is catalysed by the adenosine deaminases ADAR1 and ADAR2 proteins, whose target is double-stranded (ds) RNA<sup>1–4</sup>. *Alu* repeats are short ( $\approx 300$  bp) mobile elements widespread in primates, with over a million copies in the human genome. Due to their density, nearby *Alu* repeats of reverse orientation are common, and if transcribed, can form dsRNA. Therefore, adenosines in inverted pairs of *Alu* repeats are the majority of known RNA-editing sites, where they are edited irregularly<sup>5–9</sup>. Consistent editing of a handful of coding nucleotides is also known, with the editing of some sites being essential. Recently, the effect of RNA editing has been gradually elucidated and its connections to various diseases have been established<sup>10–12</sup>. It is, therefore, of importance to be able to accurately detect RNA editing in a sample.

Naively, detection of editing sites in RNA-seq data should be straightforward, because sequencing of an edited cDNA will read edited adenosines (A) as guanosines (G). Thus, by simply aligning RNA-seq reads to the reference genome and searching for A-to-G mismatches, one should easily detect editing sites. Several computational schemes have exploited this idea, with increasing levels of success<sup>9,13–18</sup>. However, there are several other sources of discrepancies between RNA-seq reads and the reference DNA, making the identification of genuine editing events challenging. Confounders include sequencing errors, polymorphisms and, most importantly, an incorrect alignment of the RNA-seq reads to their genomic location. Specifically, it was found that careful alignment is necessary to reduce systematic errors due to ‘alignment hazards’, such as duplications, splicing, repeats and pseudogenes. To maintain high standards of alignment, current approaches allow only a small number of mismatches between the reads and the genome (these mismatches later become editing site candidates). Thus, although it is well known that ADAR tends to edit sites in clusters (‘hyper-editing’)<sup>5,6,8,19–22</sup>, current methods miss the heavily edited short reads.

Here, we present an approach that overcomes this obstacle (by pre-masking potential A-to-G-editing sites in unaligned reads<sup>19</sup>) and enables the large-scale detection of hyper-edited reads in RNA-seq studies. By applying our approach to a number of recent (human and non-human) RNA-seq data sets, we show that our method discovers, with high specificity, a much larger number of editing sites than previous RNA-editing detection approaches, in some cases by orders of magnitude. Our results thus establish that screening for hyper-editing is crucial in order to obtain a reliable picture of RNA editing.

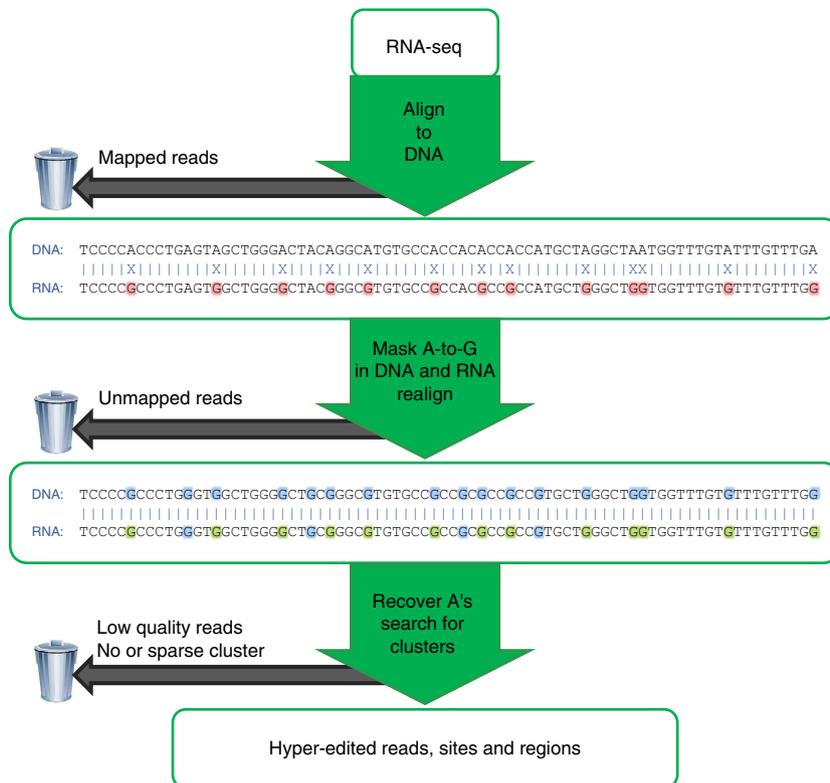
## Results

**Detecting hyper-edited reads in RNA-seq data.** Editing detection methods rely on mapping RNA sequences to the genome using tools that allow only a limited number of mismatches. This prevents the mapping of hyper-edited (short) reads, and therefore, their editing sites go undetected. To identify extensively edited regions, we apply a simple but effective four-step approach: (1) collect all unmapped reads from the initial alignment; (2) transform all As to Gs in both the unmapped reads and the reference genome; (3) realign the transformed RNA reads and the transformed reference genome; and (4) recover the original sequences and search for dense clusters of A-to-G mismatches. An overview of the pipeline is presented in Fig. 1; complete details are provided in Methods.

We previously introduced and applied an early version of this approach to detect hyper-edited human ESTs (expressed sequence tags)<sup>19</sup>. Other groups have recently employed similar approaches to study editing in *Drosophila melanogaster*<sup>23</sup> and *Caenorhabditis elegans*<sup>24</sup>. Here, we describe the results of an extended and improved pipeline that is aimed at detecting hyper-editing in RNA-seq data. In contrast to our previous method, the new pipeline addresses the challenges associated with the numerous and much shorter sequences derived from RNA-seq data, such as ambiguously aligned reads, incorrect mapping due to splicing, paired-end (PE) reads and homopolymeric sequences (see Methods for more details). The main data set we considered was the Illumina Human BodyMap 2.0, consisting of  $\approx 5 \times 10^9$  reads (of lengths 50, 75 and 100 bp), of which  $\approx 788 \times 10^6$  reads remained unmapped after running BWA. After some initial filtering (Methods), we applied our pipeline to detect hyper-edited reads, defined here as reads with  $\geq 5\%$  density of high-quality A-to-G mismatches (Methods). Note that we did not attempt to discover moderately edited reads, as those are usually mapped in the first step and thus not considered by our pipeline.

Our final hyper-editing set consisted of 637,426 reads, containing 649,756 unique A-to-G-editing sites. In the remainder of the paper, we report the analysis of hyper-editing in the 75-bp subset of the Illumina BodyMap data consisting of  $\approx 1.3 \times 10^9$  single-end (SE) reads, of which  $\approx 114 \times 10^6$  reads were unmapped. We discovered 390,881 hyper-edited reads (Supplementary Data 2), containing 455,014 unique A-to-G-editing sites (Supplementary Data 1) (2,080,519 in total) at 62,860 distinct regions (Fig. 2a and Supplementary Table 1; see Methods for definitions). For comparison, careful screening for editing in the pool of the mapped reads in the Illumina BodyMap data was previously carried out by Ramaswami *et al.*<sup>17</sup> These authors detected (in the 50 and 75 bp data sets) 370,623 sites, but only 59,361/455,014 (13.0%) of the sites that we detected when we used a subset of the same data (namely, the 75-bp data set). These results suggest that extensive editing by ADAR enzymes is not only highly abundant but also missed by existing computational screens.

**The hyper-editing detection screen is highly specific.** The specificity of an editing detection screen is usually gauged by rerunning the same pipeline, but when looking for dense clusters of mismatches other than A-to-G<sup>5,6,8,20</sup>. Since non-A-to-G editing is expected to be extremely rare<sup>25–27</sup>, the fraction of A-to-G clusters to all clusters represents a useful measure of the screen’s specificity (note that for data sets where the sequenced strand is random, such as the 75-bp subset of the BodyMap, when we refer to A-to-G-editing sites, they may actually be either A-to-G or T-to-C). Our screen for hyper-editing in the Illumina BodyMap data achieved high specificity, with 97.25% of the unique hyper-editing sites being A-to-G (Supplementary Table 1), higher than in previous screens of other data sets (summarized in Table 1 in ref. 16; when not limited to *Alu*). Beyond A-to-G, the other mismatch type with the most hyper-editing sites was G-to-A, with only 1.64% of the total number of unique editing sites (7,664/467,894) (Fig. 2a and Supplementary Table 1). Even this remarkable low level of noise is likely an overestimate, since a few of the G-to-A events may be genuine DNA or RNA editing by APOBEC proteins<sup>28–30</sup>. The ratio of A-to-G to G-to-A hyper-edited reads was 20.4, compared with 59.4 for the ratio of A-to-G to G-to-A unique-editing sites (Supplementary Table 1), suggesting that many G-to-A events are concentrated at a small number of likely erroneous sites. Additionally, A-to-G sites overlapping with a known genomic SNP (single-nucleotide polymorphism), which are unlikely to be



**Figure 1 | Overview of our hyper-editing detection pipeline.** In the first step, an RNA-seq data set is aligned to the reference genome and mapped reads are discarded. Hyper-edited reads do not map in this step since they contain too many mismatches (see a hypothetical example in the top alignment; edited nucleotides highlighted in red). To detect potential RNA editing in the unmapped reads, we realign them after masking the editing sites by transforming all As to Gs in both RNA and DNA (bottom alignment; transformed nucleotides in blue, edited nucleotides in green). For those reads that realign, the original sequences are recovered and examined for high-quality clusters of A-to-G mismatches. Reads with large, dense clusters of A-to-G differences are designated as hyper-edited, and their editing sites are recorded. See complete details in Methods.

true editing sites<sup>31</sup>, were just 1.64% of all sites, much less than for the other mismatch types (Supplementary Table 1). Finally, even more support for the specificity of our method comes from the strand information. The RNA-seq data we used (the 75-bp reads subset) is strand-indifferent, and therefore, even for true A-to-G sites, the observed mismatches are either A-to-G or T-to-C with roughly equal amounts (Supplementary Fig. 1). However, the Illumina BodyMap project also contains a data set of strand-informative 100-bp reads, in which the sequenced strand is identical to the expressed strand. In that data set, A-to-G is the only type of mismatch that is (almost) exclusively found on the expressed strand (Supplementary Fig. 1), suggesting that other ‘editing’ types are due to technical errors. Similarly, in 191,774/225,879 (84.9%) of the A-to-G 75-bp hyper-edited reads that intersected with RefSeq sequences, the sense strand mismatch was A-to-G (compared with 3,832/9,495 (40.4%) for G-to-A hyper-edited reads). The remaining A-to-G edited reads (15.1%) may be due to editing of transcripts expressed from the antisense strand<sup>32,33</sup>.

**Most of the detected editing sites are novel.** The majority of the editing sites detected by our method were novel. Of 455,014 unique sites, only 42,775 (9.4%) and 126,983 (27.9%) appeared in DARNED<sup>34</sup> (total 322,415 sites) and RADAR<sup>35</sup> (total 1,379,403 sites), respectively, the latter supposedly being the most comprehensive database of human RNA editing to date. The number of sites that are novel with respect to both databases is 327,096/455,014 (71.9%) and the number of novel regions (not

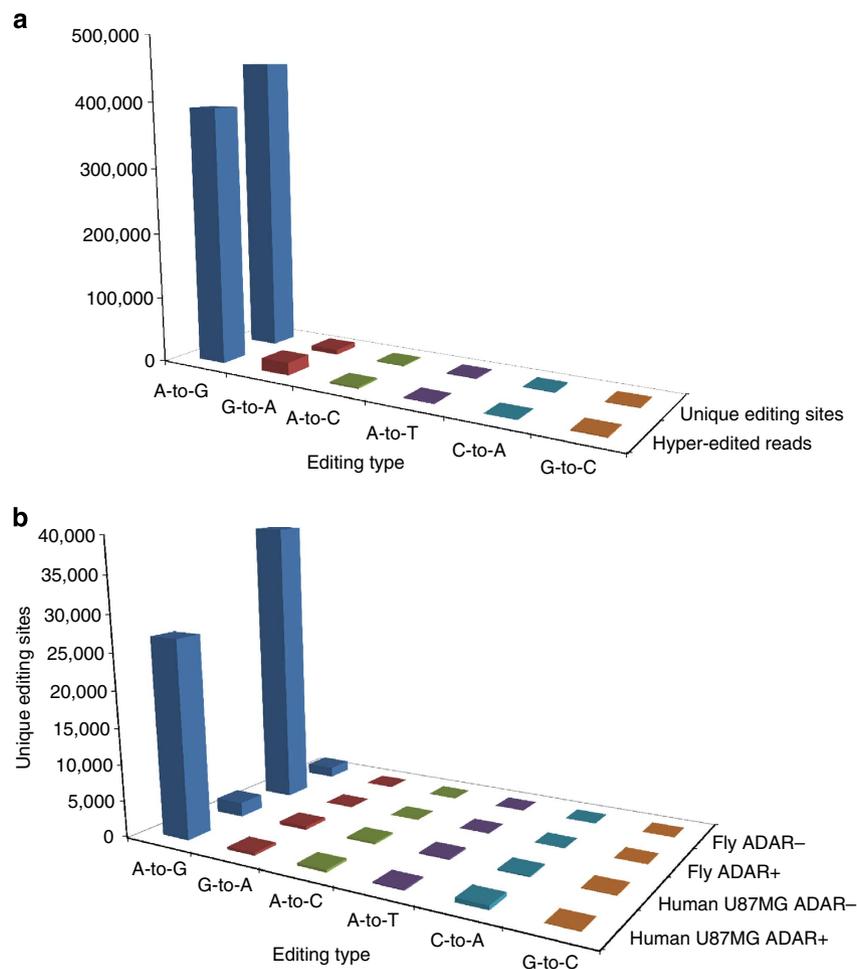
containing any known editing sites) is 25,170/62,860 (40.0%). Our results, therefore, suggest that the majority of our detected regions—precisely those that are most excessively edited—are missed by traditional approaches.

#### Sequence context and tissue of origin of hyper-editing sites.

Examining the sequence context of our detected sites showed that Gs are depleted one base upstream and enriched one base downstream of the editing sites (Fig. 3), in agreement with the known ADAR sequence preference<sup>36</sup> and the known ADAR targets<sup>17,37</sup>. In order to study differential hyper-editing among tissues, we assigned each tissue an enrichment factor based on its relative editing level (see details in Supplementary Table 2). The brain had the highest enrichment factor, as well as the largest number of unique hyper-edited regions and sites, consistently with previous studies<sup>5,6,20,38</sup>. Interestingly, the brain did not have a particularly large number of hyper-edited reads (Supplementary Table 2), suggesting that while editing in the brain encompasses more hyper-editing sites than in other tissues, many of them are expressed and/or edited to low levels.

#### Genomic localization of hyper-edited regions.

The majority of the hyper-edited regions overlapped with genes (41,531/62,860; 66%), usually with introns (38,534/41,531; 93%). As previously found for most editing targets<sup>5,6,8,20</sup>, hyper-edited regions localized mostly in *Alu* repeat elements (55,106/62,860; 88%). These *Alu* repeats likely form a dsRNA structure, the ADAR target, by hybridizing with nearby, oppositely oriented *Alus*. Only



**Figure 2 | Counts of hyper-editing events.** (a) Hyper-editing identified in the Illumina BodyMap 75-bp SE read set. Most of the detected hyper-edited reads (390,881/414,898; 94.2%) were of A-to-G type. A total of 455,014 unique A-to-G-editing sites were discovered, 97.2% of all the detected unique sites. (b) Most editing sites were eliminated in ADAR<sup>-</sup> samples. We detected hyper-editing sites in *Drosophila* nascent-RNA-Seq data from either wild-type (ADAR<sup>+</sup>) flies or from ADAR-null (ADAR<sup>-</sup>) flies. The number of A-to-G sites detected in the ADAR<sup>+</sup> sample was  $\approx 20$ -fold larger than in the ADAR<sup>-</sup> sample (39,472 versus 1,436; there were  $2 \times 10^{-4}$  editing sites per (mapped) read in ADAR<sup>+</sup> compared with  $9 \times 10^{-6}$  in ADAR<sup>-</sup>). Similarly, we detected hyper-editing sites in the human U87MG cell line, either with (ADAR<sup>-</sup>) or without (ADAR<sup>+</sup>) siRNA-induced silencing of ADAR1. The number of sites in the ADAR<sup>+</sup> sample was much higher than in the ADAR<sup>-</sup> sample (27,124 versus 1,992, or  $3 \times 10^{-4}$  versus  $2 \times 10^{-5}$  sites per mapped read). The number of non-A-to-G-editing sites is also presented. In the ADAR<sup>+</sup> samples, more than 94% of the detected sites were A-to-G. In the ADAR<sup>-</sup> samples, while the number of A-to-G sites significantly decreased, the counts of the other mismatches were almost indifferent to the absence of ADARs.

3,539/62,860 regions (5.6%) did not overlap with any documented repetitive element (see more below).

**Coding and microRNA-related hyper-editing sites.** To detect coding hyper-editing sites, we used the RefSeq annotation as well as additional filtering of genomic SNPs and A-to-G mismatches that were not on the (RefSeq-annotated) expressed strand. Overall, we detected 728 coding sites (Supplementary Data 3) in 140 genes (out of a total of 32,693 exonic sites), of which 507 were non-synonymous and three were non-stop (Annovar<sup>39</sup>). These coding hyper-editing sites exhibited the known ADAR motif (depletion of G one base upstream of the site), but with a somewhat weaker magnitude (not shown). Only 75 of the coding sites (48 of the non-synonymous) appeared in RADAR<sup>35</sup>.

Editing of microRNAs (miRNAs) or their precursors may impact miRNA maturation and function<sup>40</sup>. Our analysis of the BodyMap data predicted 20 novel editing sites (total 29; Supplementary Data 4) in six primary transcripts of known

miRNA (miRBase<sup>41</sup>), with 11 novel sites (total 14) localized in five mature miRNA regions. Moreover, we detected 61 new editing sites in conserved mammalian miRNA regulatory targets in 3'-untranslated regions (predicted by TargetScanHuman<sup>42</sup>).

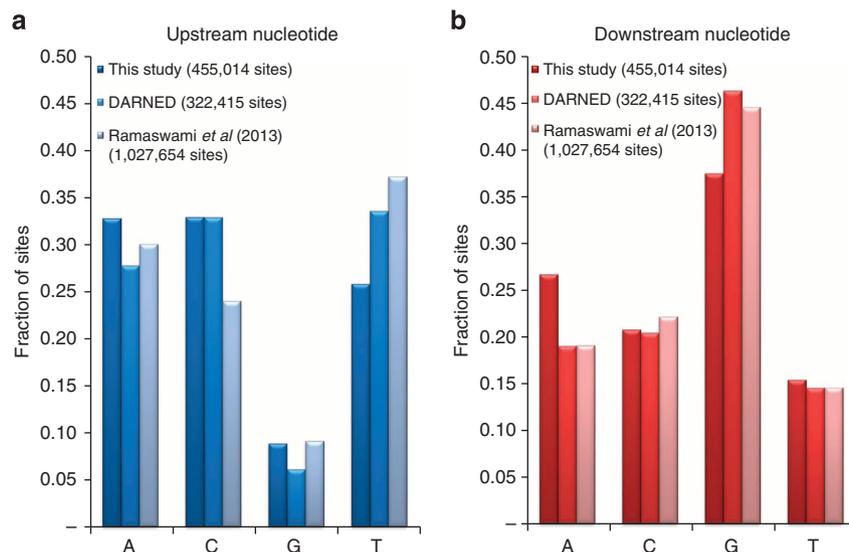
Pinto *et al.*<sup>43</sup> reported 59 evolutionary conserved mammalian-editing sites in 33 genes. While we detected only two of these sites, we detected additional 505 hyper-editing sites in 20 out of these 33 genes (only 150 of which were in RADAR), 1.6-fold more than in all other genes (8,678/23,000). This supports the hypothesis that hyper-editing can induce site-selective editing<sup>44</sup>.

**Hyper-edited *Alus* are part of long dsRNA structures.** Our previous, smaller-scale study of hyper-editing in ESTs found that hyper-edited *Alus* are predicted to form particularly stable dsRNA, more than in selectively edited RNAs. To explore whether a similar trend exists here, we compared the 49,264 hyper-edited *Alus* to 234,503 *Alus* that were identified as edited<sup>17</sup> but were not hyper-edited. Indeed, the average distance between a

**Table 1 | Published methods discard the vast majority of editing sites.**

Data set description*	Source reads <sup>†</sup>	Editing sites, as reported by authors	Hyper-editing sites we detected (% A-to-G of total) (novel)	Ratio of number of sites: our screen /published article
<i>HUMAN</i>				
Illumina Human BodyMap 2.0 75 bp SE and 50 bp PE <sup>17</sup>	3,821,002,610	370,623	546,732 (95.71%) (478,739)	1.48
ENCODE project cell line GM12878 75 bp PE <sup>15</sup>	424,451,564	1,716	157,077 (96.04%) (156,104)	91.54
ENCODE project cell line GM12891 75 bp PE <sup>15</sup>	199,366,592	1,885	97,681 (98.28%) (96,712)	51.82
ENCODE project cell line GM12892 75 bp PE <sup>15</sup>	270,307,390	843	125,146 (98.35%) (124,626)	148.45
Lymphoblastoid cell line of YH PolyA + 75 bp PE and 100 bp PE <sup>18</sup>	319,075,474	10,343	73,463 (98.73%) (73,462)	7.10
Lymphoblastoid cell line of YH PolyA- 90 bp PE <sup>18</sup>	843,673,098	10,770	270,853 (98.81%) (270,850)	25.15
Human U87MG cell line ADAR + (ctrl) 60 bp PE <sup>14</sup>	111,160,762	4,141	27,124 (94.64%) (25,945)	6.55
<i>MOUSE</i>				
Mouse (C57BL/6NJ) brains 76 bp PE <sup>46</sup>	114,374,684	4,869	11,849 (96.38%) (11,811)	2.43
<i>FLY</i>				
<i>Drosophila</i> head nascent RNA 101 bp SE <sup>49</sup>	257,255,489	1,350	39,472 (99.87%) (39,334)	29.24

PE, paired end; SE, single end.  
\*Additional details are given in Supplementary Table 3.  
<sup>†</sup>Paired-end reads were counted as two single-end reads.



**Figure 3 | Sequence context of hyper-editing.** The bars correspond to the fraction of editing sites with each type of (genomic) nucleotide one base (a) upstream and (b) downstream of the site. Results are shown for sites from our study, the DARNED database, and Ramaswami *et al.*<sup>17</sup> The sequence near the hyper-editing sites is depleted of Gs upstream and enriched with Gs downstream.



even a single read. In other words, these regions were not expressed as unedited or even lightly edited. We speculate that this is the result of these regions being edited so excessively, that the chances of sampling an unedited (or little edited) molecule are extremely small. Alternatively, these regions may be particularly hard to sequence and map. Whatever the reason is, this result highlights the fact that a large fraction of the hyper-edited reads would be completely missed by traditional approaches, as we discuss next in more detail.

#### Published methods discard the vast majority of editing sites.

Computational editing detection screens usually begin by aligning the RNA sequences to the reference genome. However, as explained above, most published methods, in order to eliminate noisy reads, allow only a handful of mismatches (for example, three mismatches in ref. 15; three or four mismatches in ref. 18) and thus do not allow the detection of more heavily edited molecules. We showed above, as a proof-of-concept, that hyper-editing can be detected and is highly abundant in the Illumina BodyMap data. Here, we apply our detection pipeline in comparison with nine published RNA-seq data sets (from six publications; including the BodyMap data), where editing sites were previously identified by other groups. The results (Table 1) are remarkable: in all cases, our screen dramatically increased the number of detected sites. Compared with the most comprehensive other studies<sup>17</sup>, we more than doubled the number of editing sites. Compared with the ENCODE study<sup>15</sup>, we strikingly detected 148 times more editing sites. The specificity of our screen (the number of A-to-G-editing sites divided by the number of sites of all types of mismatches) was high, at 94.6% or higher, and the overlap between the sets of previously detected sites and the hyper-edited sites was usually small (at most 62%). As expected<sup>19</sup>, the number of detected hyper-editing sites generally increased with the total number of reads. Taken together, these results suggest that screening for hyper-editing is absolutely crucial for editing detection, or otherwise the majority of sites are overlooked. These two approaches are complementary, that is, both the traditional approach (searching for moderately edited reads) and our approach (screening unmapped reads for hyper-editing) are required to fully characterize the editing spectrum. Alternatively, a new pipeline that combines these two approaches should be developed, based on more sensitive aligners.

**Identification of hyper-edited RNAs in other species.** We next applied our editing detection screen to published RNA-seq data sets from mouse, rat, opossum, fly and platypus<sup>46–49</sup> and detected, with high specificity, thousands of editing sites (11,849, 21,761, 1,563 and 39,472, respectively; nearly no editing sites were detected in the platypus brain) (Table 2)<sup>48</sup>. As in the human data sets, most of the sites we discovered were

previously unknown, highlighting the importance of screening for hyper-editing. The mouse and fly data sets have been previously screened for editing, but not for hyper-editing<sup>17,23,46,49</sup>, and we have dramatically increased the number of editing sites in these species. In rat and opossum, we are the first to screen an RNA-seq data set and report thousands of editing sites. To compare the frequency of editing across species, we computed, for each data set, the number of editing sites per mapped read as an approximation of the true editing rate (Table 2). The human editing level was  $\approx 10$ -fold higher than in non-humans<sup>50</sup>; the other three mammals we examined had a similar editing rate, while the rate in fly was  $\approx 50\%$  higher.

**Lack of hyper-editing in ADAR1 knockdown RNA-seq.** Finally, we applied our hyper-editing detection pipeline to RNA-seq from human and fly cells in which the editing enzymes (ADAR1 and ADAR, respectively) were eliminated (Fig. 2b). Such an experiment can serve two purposes: (i) validate that the hyper-editing sites discovered by our method are indeed due to editing by ADAR; and (ii) (in humans) determine the contribution of specific ADAR proteins to hyper-editing. We first analysed nascent RNA from fly heads<sup>49</sup>, from either wild-type or ADAR-null mutant flies (there is a single ADAR protein in *Drosophila*). The near-complete elimination of hyper-editing from the ADAR-null sample supports the identification of our discovered sites as true products of editing by ADAR (Fig. 2b). We then considered human RNA-seq data from the U87MG cell line, with or without siRNA-induced silencing of ADAR1 (ref. 14). The reduction in ADAR1 fragments per kilobase per million mapped reads (FPKM) decreased from 44.01 and 42.72 in the two control samples to 6.17 and 5.33 in the siRNA samples; ADAR2 FPKM decreased from 3.71 and 3.88 to 1.98 and 2.04) eliminated most of the hyper-editing sites, suggesting that hyper-editing is catalysed mostly by ADAR1.

#### Discussion

Editing events that result in isolated codon changes have been the focus of intensive research in the past two decades. However, recent studies pointed out that most ADAR activity takes place in clusters<sup>6,9,22,51</sup>. In this paper, we implemented an algorithm (improving on our previously published approach<sup>19</sup>) to detect hyper-edited reads from large-scale RNA-seq data. We showed that our method is highly specific and that the detected sites have the known features of ADAR targets, such as a sequence motif, dsRNA structure, and localization mostly in *Alu* elements. In humans, we discovered about half a million new sites, a few hundreds of them recoding and hence of particular interest. In other species, we discovered thousands of sites, in some species for the first time on such a large scale. The enormous number of editing sites that we discovered were missed by the traditional

**Table 2 | Identification of hyper-edited RNAs in different species.**

Species* (reference)	Source reads <sup>†</sup> (Mapped reads)	Hyper-editing sites (% A-to-G of total)	Hyper-editing sites per mapped read
Human <sup>‡</sup> (Illumina Human BodyMap project 2.0)	64,313,204 (59,130,196)	75,984 (98.93%)	$1.3 \times 10^{-3}$
Mouse <sup>46</sup>	114,374,684 (103,507,244)	11,849 (96.38%)	$1.1 \times 10^{-4}$
Rat <sup>47</sup>	238,077,800 (199,514,577)	21,761 (93.61%)	$1.1 \times 10^{-4}$
Opossum <sup>48</sup>	22,273,667 (14,411,371)	1,563 (89.88%)	$1.1 \times 10^{-4}$
<i>Drosophila</i> <sup>49</sup>	257,255,489 (248,984,465)	39,472 (99.87%)	$1.6 \times 10^{-4}$

\*Additional details are given in Supplementary Table 4.

<sup>†</sup>Paired-end reads were counted as two single-end reads.

<sup>‡</sup>Values are reported only for the brain-derived subset of reads, in order to facilitate comparison with the other species.

approaches, even though, by definition, those sites are located in the most heavily edited regions of the genome. These results suggest that our procedure should be made an integral component of any editing screen.

While it is easy to understand the functional consequences of an amino-acid change on the protein product of a gene, the biological impact of hyper-edited regions is less intuitive, specifically as most hyper-editing sites reside in introns and are therefore limited to pre-mRNA. Several studies implicated hyper-edited transcripts in cellular functions that involve, for example, their nuclear retention or degradation<sup>52–57</sup>. One of the most studied mechanisms showed that hyper-edited RNAs bind a protein complex that is largely consist of components of cytoplasmic stress granules, including ADAR1 and Tudor-SN (tudor staphylococcal nuclease)<sup>56,58,59</sup>. Hyper-edited RNAs then interact with Tudor-SN, which promotes their cleavage<sup>55</sup>. Surprisingly, hyper-edited dsRNAs have also been shown to downregulate both endogenous and reporter gene expression in *trans* by having global activity to reduce gene expression and inhibit translation initiation<sup>56</sup>. Hyper-edited RNAs are also sufficient to suppress interferon induction and apoptosis in mammalian cells<sup>60</sup>, suggesting that they play an important role in stress response. Finally, a large number of editing events may alter the local dsRNA structure, which, in turn, will have an impact on the ability of the edited RNA to bind other RNAs or RNA-binding proteins. A pre-requisite to a more thorough investigation of any of these mechanisms is the precise identification of the hyper-edited regions. The data set of hyper-edited regions we created addresses this need. For species, tissues, or conditions we did not cover, our method can be applied using our publicly available code.

We note that our method is complementary to traditional approaches and is designed to identify only sites in hyper-edited regions. Additionally, our final detected set is sensitive to the precise alignment procedure used. Other parameters that influence the final results are the minimum number of A-to-G differences and the quality score required to call hyper-edited reads, as well as the parameters of the various filters used to remove false positives. There are, therefore, many editing sites that likely escaped our detection. Also, we note that our approach does not provide an estimate of the editing level at each site, which is probably rather low in most cases, and for that purpose, one may use the detected sites as an input to other tools, such as REDITools<sup>61</sup>.

Even with those limitations, we discovered a considerable number of sites. However, with ever increasing sequencing depths, the picture emerging from recent studies is that virtually every adenosine in a dsRNA-forming *Alu* can be edited (although usually at very low levels)<sup>9</sup>. Hence, future studies of RNA editing are expected to shift from mere discovery of editing sites to understanding of editing regulation and function. To this end, hyper-edited RNAs will be important as the result of the most extreme editing process. Therefore, the importance of our method for future research goes beyond the discovery of additional sites, and lies in making possible the characterization of this important class of edited sequences.

## Methods

The pipeline scripts are available at <http://levanonlab.lsbu.ac.il/resources/zip>.

**Data sets.** The Illumina Human BodyMap 2.0 data set (Gene Expression Omnibus accession code GSE30611) consists of  $\approx 5 \times 10^9$  reads generated from 16 tissues. Each tissue was sequenced at both 50-bp PE reads and 75-bp SE reads. These sets were polyA selected, generated with random priming and not stranded. RNA from a mixture of the same 16 tissues was also sequenced, at 100 bp SE reads, and stranded. The majority of our analysis was carried out on the 75-bp reads data set ( $\approx 1.3 \times 10^9$ ).

**Identifying hyper-edited reads.** To find hyper-edited reads, we first aligned all reads ( $\approx 5 \times 10^9$ ) to the human reference genome (hg19) using BWA aln 0.6.2 (default parameters) and mem 0.7.4 (minimum seed length 50). We then considered the unmapped reads,  $\approx 788 \times 10^6$  in the entire data set and  $\approx 114 \times 10^6$  in the 75-bp SE data set. While most of the unmapped reads probably represented sequencing artifacts and polymorphisms, we hypothesized that many reads were unmapped due to excessive A-to-G editing<sup>19</sup>. We removed reads with potential sequencing errors<sup>19</sup>: particularly large ( $> 60\%$ ) or small ( $< 10\%$ ) percentage of a single nucleotide, more than 10% of ambivalent nucleotides (non-(ACGT)), long stretches of simple repeats ( $> 10$  repeats, or  $> 20$  for a single nucleotide) or an average Phred quality score below 25 (after removing the lowest 10% of the scores). In order to be able to realign reads with a large number of mismatches caused by A-to-G-editing sites, we transformed every A to G in both the remaining RNA reads and in the genome. We aligned the transformed reads to the transformed genome, again using BWA aln 0.6.2 (maximum edit distance 2, no gaps allowed and iterative mode). The original (four-letter) sequences of the reads that aligned (after the transformation) were recovered and the mismatches between the reads and the reference genome were examined. Note that using the original full sequences again eliminates the problem of lower sequence complexity (and hence, lower alignment accuracy) in the transformed reads. Reads that mapped to multiple locations were separately treated by selecting the location with the largest fraction of A-to-G to all mismatches, provided that this fraction was  $\geq 10\%$  higher than in all other locations (otherwise the read was discarded; the read was also discarded if it mapped to  $> 50$  locations). We then searched for dense clusters of high-quality (Phred  $\geq 30$ ) A-to-G mismatches, in which the number of A-to-G mismatches was  $\geq 5\%$  of the read length and  $> 60\%$  (80% for read lengths  $\leq 60$  bp) of the total number of mismatches. As an additional filtering step, we considered the portion of the read bound by the first and the last A-to-G mismatches. We discarded potentially low-quality/mismatched reads having (i) a too dense A-to-G mismatch cluster (length  $< 10\%$  of the read length); (ii) a cluster too close to the ends of the read (completely contained within either the first or last 20% of the read; those clusters may result from improper alignment to the genome due to splicing); or (iii) a cluster with a particularly large percentage ( $> 60\%$ ) of a single nucleotide. For edited PE reads (which were treated throughout the pipeline as two separate single-end reads), we also required that the mate read was mapped to a nearby region (500 kbp) and in an opposite orientation. Reads that passed all filters were finally designated as hyper-edited. An overview of the pipeline is presented in Fig. 1.

In order to determine the pipeline's specificity, we repeated it when searching for other types of editing events (for example, A-to-C, G-to-A and so on)<sup>19</sup>, which involved modifying the transformation and the definition of the editing mismatches accordingly, but was otherwise identical to the A-to-G screen. There are 12 possible single-nucleotide mismatches. However, since the RNA-seq reads could be either sense or antisense, we could not distinguish between a given mismatch and its complementary one. For example, our detected A-to-G edited reads represent both clusters of A-to-G mismatches and clusters of T-to-C mismatches. The same is true for the other mismatch types, and we therefore report results for only six categories of editing events.

For the subset of the Illumina BodyMap containing the 100-bp reads, we again ran the same pipeline. However, since the 100-bp data set was stranded, we could treat A-to-G and T-to-C mismatches separately (Supplementary Fig. 1).

**Comparison with a previous approach.** The method proposed here is a natural follow-up of an approach we previously developed to detect hyper-editing in EST libraries<sup>19</sup>, in that it detects editing by realigning hard-to-map reads through masking of potential editing sites. However, the present method was reimplemented from the ground up to address the specific difficulties associated with next-generation sequencing data, where reads are numerous and much shorter. Our remodelled pipeline includes a new and much faster aligner (BWA, reducing running time of the whole pipeline by over 1,000-fold), fine tuning of the parameters of existing filters and introduction of several new pre- and post-processing steps designed for short read data. The new pipeline improves the processing of reads that were aligned to more than one region, takes advantage of information in PE reads and per-base-pair sequencing quality, and filters out reads with potential incorrect mapping at splice sites or with decreasing quality at the ends. We successfully applied our new method to a large number of RNA-seq data sets from several species, and detected several orders of magnitude more sites than was possible with the previous approach. The implementation of our new pipeline has been made public, and is expected to be used for analysing the constantly growing number of RNA-seq data sets.

**Annotations.** We annotated genomic regions using RefSeq for known genes and RepeatMasker for known repeats, both from the UCSC Genome Browser<sup>62</sup> downloaded in April 2013. To detect sites overlapping with SNPs, we used genomic SNPs from dbSNP135 (The UCSC Genome Browser). Non-coding RNAs (Supplementary Data 1) were annotated using GENCODE v16. Functional annotation of the predicted editing sites in coding regions was done using ANNOVAR<sup>39</sup>. Human miRNA coordinates were taken from miRBase v20, downloaded from <http://www.mirbase.org/ftp.shtml><sup>44</sup>. The coordinates of the conserved mammalian miRNA regulatory targets of conserved miRNA families in

the 3'-untranslated regions were predicted by TargetScanHuman v.5.1 (ref. 42) (also obtained from the UCSC Genome Browser).

**Defining hyper-edited regions and dsRNA structure.** As explained above, we define the cluster of editing mismatches as the portion of the edited read from the first to the last (high quality (Phred  $\geq$  30)) A-to-G mismatch. To identify the hyper-edited regions, we first merged (the corresponding genomic coordinates of) all overlapping edited clusters, and then set the boundaries of the region from the first base of the most upstream cluster to the last base of the most downstream cluster. Construction and analysis of the edited regions were carried out using BEDTools<sup>63</sup>.

To detect potential dsRNA structure formed by hyper-edited RNAs, the DNA sequences of the hyper-edited regions were aligned to the sequences 10 kbp upstream and 10 kbp downstream of the regions. We used *bl2seq*<sup>45</sup> with parameters -F -W 7 -r 2, and considered a match only for alignment with 90% identity along 80% of the hyper-edited region length.

**Calculating expression levels.** Expression level of ADAR1 and ADAR2 at the human U87MG cell line (with or without siRNA against ADAR1)<sup>14</sup> was calculated using RSEM<sup>64</sup> as either FPKM or transcripts per million. Both measures gave the same fold change between the control and the siRNA samples for both ADAR1 and ADAR2.

**Direct sequencing validation of hyper-editing.** To experimentally confirm our computational editing prediction, we chose seven hyper-edited regions and PCR-amplified genomic DNA and cDNA from a human brain from same donor (obtained from the NICHHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, USA). Primers were designed using Primer3 (ref. 65). The PCR reaction was performed in C1000 Thermal Cycler (Bio-Rad) using Phusion Green Hot Start II High-Fidelity DNA Polymerase (Thermo Scientific) and in a 20- $\mu$ l reaction volume, according to the manufacturer's instructions. We ran the following steps: 98°C  $\times$  30 s, 35 cycles of (98°C  $\times$  10 s, annealing Tm°C  $\times$  30 s (listed in Supplementary Table 5), and 72°C  $\times$  30 s), and 7 min in 72°C. PCR products were run on a 1% agarose gel, extracted using MinElute gel extraction kit (Qiagen), and sequenced in Hy Laboratories. Chromatograms of matched DNA and cDNA were aligned and editing sites were identified. Counts of observed editing sites in each region appear in Supplementary Table 5. Annotated chromatograms for two of the regions appear in Supplementary Fig. 3.

## References

- Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
- Li, J. B. & Church, G. M. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. *Nat. Neurosci.* **16**, 1518–1522 (2013).
- Savva, Y. A., Rieder, L. E. & Reenan, R. A. The ADAR protein family. *Genome Biol.* **13**, 252 (2012).
- Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* **79**, 321–349 (2010).
- Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2**, e391 (2004).
- Levanon, E. Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22**, 1001–1005 (2004).
- Kim, D. D. Y. *et al.* Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.* **14**, 1719–1725 (2004).
- Blow, M., Futreal, P. A., Wooster, R. & Stratton, M. R. A survey of RNA editing in human brain. *Genome Res.* **14**, 2379–2387 (2004).
- Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* **24**, 365–376 (2013).
- Slotkin, W. & Nishikura, K. Adenosine-to-inosine RNA editing and human disease. *Genome Med.* **5**, 105 (2013).
- Maas, S., Kawahara, Y., Tamburro, K. M. & Nishikura, K. A-to-I RNA editing and human disease. *RNA Biol.* **3**, 1–9 (2006).
- Tariq, A. & Jantsch, M. F. Transcript diversification in the nervous system: a to I RNA editing in CNS function and disease development. *Front. Neurosci.* **6**, 99 (2012).
- Li, J. B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
- Bahn, J. H. *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* **22**, 142–150 (2012).
- Park, E., Williams, B., Wold, B. J. & Mortazavi, A. RNA editing in the human ENCODE RNA-seq data. *Genome Res.* **22**, 1626–1633 (2012).
- Ramaswami, G. *et al.* Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods* **9**, 579–581 (2012).
- Ramaswami, G. *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods* **10**, 128–132 (2013).
- Peng, Z. *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* **30**, 253–260 (2012).
- Carmi, S., Borukhov, I. & Levanon, E. Y. Identification of widespread ultra-edited human RNAs. *PLoS Genet.* **7**, e1002317 (2011).
- Kim, D. D. Y. *et al.* Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.* **14**, 1719–1725 (2004).
- Barak, M. *et al.* Evidence for large diversity in the human transcriptome created by Alu RNA editing. *Nucleic Acids Res.* **37**, 6905–6915 (2009).
- Neeman, Y., Levanon, E. Y., Jantsch, M. F. & Eisenberg, E. RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA* **12**, 1802–1809 (2006).
- St Laurent, G. *et al.* Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in *Drosophila*. *Nat. Struct. Mol. Biol.* **20**, 1333–1339 (2013).
- Wu, D., Lamm, A. T. & Fire, A. Z. Competition between ADAR and RNAi pathways for an extensive class of RNA targets. *Nat. Struct. Mol. Biol.* **18**, 1094–1101 (2011).
- Lin, W., Piskol, R., Tan, M. H. & Li, J. B. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. *Science* **335**, 1302 (2012).
- Pickrell, J. K., Gilad, Y. & Pritchard, J. K. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. *Science* **335**, 1302 (2012).
- Kleinman, C. L. & Majewski, J. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. *Science* **335**, 1302 (2012).
- Rosenberg, B. R., Hamilton, C. E., Mwangi, M. M., Dewell, S. & Papavasiliou, F. N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat. Struct. Mol. Biol.* **18**, 230–236 (2011).
- Carmi, S., Church, G. M. & Levanon, E. Y. Large-scale DNA editing of retrotransposons accelerates mammalian genome evolution. *Nat. Commun.* **2**, 519 (2011).
- Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
- Eisenberg, E. *et al.* Identification of RNA editing sites in the SNP database. *Nucleic Acids Res.* **33**, 4612–4617 (2005).
- Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
- Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21**, 379–386 (2003).
- Kiran, A. M., O'Mahony, J. J., Sanjeev, K. & Baranov, P. V. DARNED in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.* **41**, D258–D261 (2013).
- Ramaswami, G. & Li, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* **42**(Database issue): D109–D113 (2013).
- Egginton, J. M., Greene, T. & Bass, B. L. Predicting sites of ADAR editing in double-stranded RNA. *Nat. Commun.* **2**, 319 (2011).
- Kiran, A. & Baranov, P. V. DARNED: a Database of RNA Editing in humans. *Bioinformatics* **26**, 1772–1776 (2010).
- Paul, M. S. & Bass, B. L. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J.* **17**, 1120–1127 (1998).
- Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
- Yang, W. *et al.* Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat. Struct. Mol. Biol.* **13**, 13–21 (2006).
- Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
- Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
- Pinto, Y., Cohen, H. Y. & Levanon, E. Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome Biol.* **15**, R5 (2014).
- Daniel, C., Silberberg, G., Behm, M. & Ohman, M. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biol.* **15**, R28 (2014).
- Tatusova, T. A. & Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250 (1999).
- Danecek, P. *et al.* High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol.* **13**, 26 (2012).
- Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593–1599 (2012).
- Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
- Rodriguez, J., Menet, J. S. & Rosbash, M. Nascent-seq indicates widespread cotranscriptional RNA editing in *Drosophila*. *Mol. Cell* **47**, 27–37 (2012).

50. Eisenberg, E. *et al.* Is abundant A-to-I RNA editing primate-specific? *Trends Genet.* **21**, 77–81 (2005).
51. Morse, D. P., Aruscavage, P. J. & Bass, B. L. RNA hairpins in noncoding regions of human brain and *Caenorhabditis elegans* mRNA are edited by adenosine deaminases that act on RNA. *Proc. Natl Acad. Sci. USA* **99**, 7906–7911 (2002).
52. Scadden, A. D. & Smith, C. W. Specific cleavage of hyper-edited dsRNAs. *EMBO J.* **20**, 4243–4252 (2001).
53. Chen, L. L. & Carmichael, G. G. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol. Cell* **35**, 467–478 (2009).
54. Hundley, H. A., Krauchuk, A. A. & Bass, B. L. *C. elegans* and *H. sapiens* mRNAs with edited 3' UTRs are present on polysomes. *RNA* **14**, 2050–2060 (2008).
55. Scadden, A. D. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nat. Struct. Mol. Biol.* **12**, 489–496 (2005).
56. Scadden, A. D. Inosine-containing dsRNA binds a stress-granule-like complex and downregulates gene expression in trans. *Mol. Cell* **28**, 491–500 (2007).
57. Prasanth, K. V. *et al.* Regulating gene expression through RNA nuclear retention. *Cell* **123**, 249–263 (2005).
58. Weissbach, R. & Scadden, A. D. J. Tudor-SN and ADAR1 are components of cytoplasmic stress granules. *RNA* **18**, 462–471 (2012).
59. Anderson, P. & Kedersha, N. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nat. Rev. Mol. Cell Biol.* **10**, 430–436 (2009).
60. Vitali, P. & Scadden, A. D. J. Double-stranded RNAs containing multiple IU pairs are sufficient to suppress interferon induction and apoptosis. *Nat. Struct. Mol. Biol.* **17**, 1043–1050 (2010).
61. Picardi, E. & Pesole, G. REDIttools: high-throughput RNA editing detection made easy. *Bioinformatics* **29**, 1813–1814 (2013).
62. Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
64. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
65. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
66. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
67. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).

### Acknowledgements

We thank Nurit Paz-Yaakov for help with experimental procedures, Lily Bazak for assistance with the *Alu* analysis, Shahar Alon for assistance with the miRNA data set and Oliver Keller for early work on the data. We thank Eli Eisenberg and Jin Billy Li for a critical reading of an early version of the manuscript. We thank the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, USA for providing a human brain tissue sample. S.C. thanks the Human Frontier Science Program for financial support. This work was supported by the Legacy Heritage Biomedical Science Partnership, Israel Science Foundation (grant no. 1466/10), the European Research Council (grant no. 311257) and the I-CORE Program of the Planning and Budgeting Committee in Israel, and the Israel Science Foundation (grants no. 41/11 and 1796/12).

### Author contributions

E.Y.L., S.C. and H.T.P. designed the algorithms and experiments. H.T.P. implemented the algorithms and analysed the results. S.C. wrote the manuscript with input from H.T.P. and E.Y.L.. E.Y.L. supervised the work.

### Additional information

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Porath, H. T. *et al.* A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* 5:4726 doi: 10.1038/ncomms5726 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>