RNA editing by adenosine deaminases changes the information encoded in the mRNA from its genomic blueprint. Editing of protein-coding sequences can introduce novel, functionally distinct, protein isoforms and diversify the proteome. The functional importance of a few recoding sites has been appreciated for decades. However, systematic methods to uncover these sites perform poorly, and the full repertoire of recoding in human and other mammals is unknown. Here we present a new detection approach, and analyze 9125 GTEx RNA-seq samples, to produce a highly-accurate atlas of 1517 editing sites within the coding region and their editing levels across human tissues. Single-cell RNA-seq data shows protein recoding contributes to the variability across cell subpopulations. Most highly edited sites are evolutionary conserved in non-primate mammals, attesting for adaptation. This comprehensive set can facilitate understanding of the role of recoding in human physiology and diseases.
adenosine-to-inosine (A-to-I) RNA editing, catalyzed by the ADAR (adenosine deaminases acting on RNA) family of enzymes1-4, is the most common form of RNA editing among animals5. As the translation machinery largely interprets inosines as guanosines6 editing within the coding sequence (CDS) may lead to amino-acid substitution (“recoding”) and diversify the proteome. Recoding of GRIA2 (Glutamate Ionotropic Receptor AMPA Type Subunit 2) Q/R site is essential in mammals7, and recoding of a few additional well-studied targets has been shown to be functionally important. Notable examples are the antizyme inhibitor 1 (AZIN1) ADAR1-dependent recoding that promotes cell proliferation and contributes to cancer progression8, recoding of NEIL1 (Nji Like DNA Glyco- sylase 1) that results in a 30-fold reduction in the thymine glycol cleavage rate when acting on duplex DNA9, and FLNA (Filamin A) recoding which regulates vascular contraction and diastolic blood pressure10. Despite the potential importance demonstrated by these few examples, the full repertoire of human recoding sites is not known, and the scope of its functionality is still unclear.

As part of standard sequencing protocols, inosines in the RNA are reverse-transcribed into guanosines in the cDNA. Thus, aligning the sequencing reads to the reference genome, A-to-I editing is manifested as A-to-G genome-read mismatches. Typically, recoding activity is dwarfed by A-to-I editing events at millions of sites within non-coding regions5, and is therefore much more difficult to detect (Fig. 1a). As a result, systematic analyses of mismatches yielded high-sensitivity identification of sites in human5,11-17, virtually all of them within Alu repeats, but performed poorly in coding regions. The relatively small number of recoding sites14,18 is overshadowed by additional sources of discrepancies between the reads and the reference genome. First, genomic variability between the reference genome and the sampled individuals translates into mismatches between the reference genome and the RNA sequenced from these individuals. Over 500 million human single-nucleotide-polymorphisms (SNPs) have been identified so far, of which almost all are rare. A typical individual genome includes tens of thousands of rare polymorphisms19. Matching DNA-seq data could be used to filter out these sites per sample, but it is usually not available20. Furthermore, even when available, somatic genomic mutations may result in mismatches misidentified as editing events. Second, systematic misalignment of RNA-seq reads to the wrong (but homologous) genomic locus could lead to an apparently consistent mismatch, to be misidentified as an editing event21-23. These are common in the CDS as many proteins has closely-related paralogs, polymorphic duplications, and processed pseudogenes. These mapping problems are enhanced by splicing (especially when genomic polymorphisms reside in proximity to canonical splicing junctions)21,22. Moreover, each individual genome harbors several megabases of non-canonical segments24,25, which commonly include unique duplications and processed pseudogenes. Chimeric sequences and spontaneous deamination contribute to the observed mismatches as well. Finally, technical sequencing errors are customarily cleaned using the quality score. However, the quality score seems to underestimate the error rate in specific locations such as homopolymeric sequences26 or reads’ ends, as well as the vicinity of mispriming events by the random hexamer primers27. Thus, to date, the low specificity and sensitivity of RNA-editing site detection within the coding sequence28 hinder global functional analyses and evolutionary studies of recoding in human.

Here we present a novel de-novo RNA editing detection approach, dedicated and optimized for the coding region. Analyzing 9125 human RNA-seq samples from various tissues (GTEx data)29, we produce a set of 1517 A-to-I RNA editing sites within protein-coding regions, with an estimated false positive rate of only 8%. Surprisingly, recoding is found in most tissues at similar levels and is not enriched specifically in the brain. There is a wide range of editing efficiencies, and a few target sites account for the majority of recoding activity. Yet, the variation of the editing level per-site across individuals is generally low, implying regulation. Most strongly-edited sites are conserved across mammals, and thus recoding as a whole seems to be under positive selection. Unexpectedly, we find a number of large clusters of recoding sites, which could lead to an extensive combinatorial diversity of the affected transcripts. Analyzing mass-spectrometry data, we provide evidence that editing at these sites results in modified proteins. Finally, we show many recoding sites to be differentially edited in cancer and pneumonia, suggesting possible etiological relevance.

Results

A new pipeline for detecting editing in the coding region. Previous searches and existing detection tools (e.g., SPRINT30, REDITools31, or RNAEditor32) for editing in human coding regions have resulted in low sensitivity and precision27 (Fig. 1b and Supplementary Fig. 1), due to the abovementioned difficulties. Precision has been much improved using a more stringent alignment scheme33. However, this approach is highly time-consuming and is not applicable for large datasets such as GTEx. Here we present a detection pipeline designed specifically to overcome these issues (Fig. 1c and “Methods”). Briefly, we focused on non-repetitive protein-coding regions, excluding, in particular, the editing-rich Alu exons, and analyzed 9125 RNA-seq normal human samples from the GTEx project (548 donors, 47 tissue types; Supplementary Data 1)34. We applied strict alignment procedures, discarding mismatches that are likely to be explained by systematic alignment or sequencing errors (Supplementary Fig. 2), and a statistical model that integrates the cumulative profile of mismatches found for all donors of a given tissue type into a single score. Using this score, we filtered out mismatches found in a small number of donors, possibly due to rare SNPs and duplications. We masked genomic loci where multiple closely-located mismatches of different types were observed, as they are suspected to result from misalignments, and used whole-exome-sequencing (WES) data to discard mismatches observed in genomic reads and further suppress false-positive detections. The breadth and depth of GTEx data allow for a reliable and consistent detection of low-level editing, as well as detection in tissues in which editing has not been studied so far. However, sites that are rarely edited (<1% editing level in all tissue types) were discarded, as they are harder to detected reliably and less likely to be functionally relevant.

This pipeline resulted in a reliable set of 1517 human CDS A-to-I RNA sites (Fig. 2a and Supplementary Data 2), showing the familiar 5’ and 3’ neighbor-sequence preferences of mammalian ADARs35 (Fig. 2b). The large number of T-to-C sites (1006) also exhibit the familiar (reverse-complement) motif (Fig. 2c), suggesting that they are mostly due to A-to-I editing of RNA molecules transcribed from the antisense strand, complementary to the annotated coding RefSeq exons (Fig. 2d). To verify that, we analyzed a strand-specific RNA-seq dataset36, for which one can separate the reads based on their genomic strand of origin. As expected, the T-to-C signal observed in GTEx non-stranded data does not show up in reads originating from the coding strand and is fully accounted for by A-to-G editing of reads transcribed from the non-coding strand (Fig. 2e). Based on these results, we estimate that only ~24 of the 1006 T-to-C sites observed are false-positive detections (“Methods”). Note that the editing levels found at the detected T-to-C sites using stranded data are ~3-fold higher than the A-to-G sites, due to a lower detection power in
the antisense strand. Expression of this strand is typically lower compared to the protein-coding strand (see Supplementary Data 3), and thus the editing signal is masked by the many unedited reads coming from the sense strand. Accordingly, the sites that are detected under these conditions are typically more strongly edited, and the editing levels reported for these sites based on non-stranded data are underestimated. Some of the C-to-T sites detected (mostly in blood samples, where APOBEC3A is highly expressed) may reflect DNA or RNA editing by members of the AID/APOBEC family of deaminases. Indeed, one of the detected sites is the only known C-to-U recoding site in APOB (Apolipoprotein B)\textsuperscript{37,38}. Accordingly, we used the next abundant mismatch type, and estimated the false positive rate to be 8% (115 G-to-A sites).

| Mismatch type | Abundance Distribution |
|---------------|------------------------|
| AC            | TG                     |
| AG            | GA                     |
| AT            | TC                     |
| CA            | CT                     |
| CG            | GC                     |
| CT            | GA                     |

### Chart a

- All mismatches, unfiltered
- CDS
- Alu

### Chart b

- Not in RefSeq CDS, 2303
- Genomic, 1
- Other, 28
- Low-level editing sites (<1%), 11
- Misalignments, 193
- CDS RNA-editing, 198
- Alu sites, 600
- RNA editing not detected, 142

### Chart c

1. Reads alignment and preprocessing
2. BAM file
3. Read-level mismatch detection
4. Exclude reads with multi-type mismatches
5. Filtered BAM file
6. Focus on RefSeq CDS regions
7. Exclude error-prone regions
8. Exclude putative misalignment-induced mismatches
9. Removal of repeatMasker Alu sequences
10. 9,125 GTEx RNA-seq samples
11. 609 whole-exome sequencing GTEx samples
12. 1,517A-to-G sites
13. Exclude low-level editing sites (<1%)
14. Exclude sites observed in whole-exome sequencing
15. Discard heterogenous transcript-level clusters
16. Strand-bias-aware filtration
17. Statistical model and score calculation

---

\textsuperscript{37,38}
ADAR enzymes bind to dsRNA secondary structures. Accordingly, the predicted editing sites tend to be part of putative intra-molecular dsRNA structures (Fig. 2f and “Methods”). To test whether editing of these sites translates into novel protein isoforms, we have analyzed 311 human mass-spectrometry test whether editing of these sites translates into novel protein isoforms, we have analyzed 311 human mass-spectrometry data12 is rather low (Supplementary Fig. 3), which we attribute to low specificity of previous approaches. Indeed, for most of the previously published sites we can pinpoint the reason for their low sensitivity of previous approaches. Indeed, for most of the previously published sites, we were able to classify 24 sites for both (Fig. 3e and Supplementary Fig. 4). This includes the editing of the cystic fibrosis transmembrane conductance regulator (CFTR) site. Similarly, for many sites (GRIA2, GRIA3, TMEM63B, CYP1F2, and KCNA1) editing is significantly higher for excitatory neurons compared to oligodendrocyte precursor cells in the brain, but lower at the COPA site. We note that excitatory and inhibitory neurons exhibit markedly different editing levels in the coding sequence (the RNA editing level is 92% for excitatory neurons, GRIA2 RG site (editing level 92% for excitatory neurons, compared to 61% in inhibitory neurons) and KCNA1 (69% vs. 10%). Interestingly, ADAR3 (also known as ADARB2), known to regulate RNA editing levels,14,47, was previously implicated as one of the few genes differentiating between subpopulations of inhibitory neurons.48

Validation of results in additional datasets. In order to further validate our A-to-G results, and further support our conclusion that the large number of T-to-C sites result from A-to-I editing of RNA molecules transcribed from the antisense strand, we have analyzed three additional strand-specific RNA-seq datasets (see “Methods”), together with the previously mentioned dataset.44,45,45 Looking at the coding strand, we find that 1375 of the 1517 A-to-G sites were covered in at least one of the four datasets (>100 reads). Of these, 948 were found to be edited in at least one dataset (FDR corrected binomial p-value < 0.05 for the pooled data, per dataset). Note that sites that did not show evidence of editing could be edited in other tissues (not present in this validation study) or missed due to a low editing level and insufficient coverage. For the 1006 T-to-C sites, 885 were covered but only 40 (4.5%) of them have passed the binomial test with p < 0.05. Looking at the opposite strand, only 76 of the 1517 A-to-G sites were covered in at least one of the four datasets, and only 5 of them (6.6%) have passed the binomial test with p < 0.05, while 582/1006 T-to-C sites were covered, and 545 of them were found to be edited. Taken together, these results demonstrate the accuracy of our set of editing sites, and the proposed explanation for the T-to-C mismatch observed in non-stranded data.

ADAR-specificity of CDS editing sites. ADAR2 (also known as ADARB1) is the main contributor to editing in the coding sequence. Quantifying editing at our sites upon ADARs overexpression in HEK293 cell lines,44,45 we were able to classify 24 sites as ADAR1-dependent, 179 sites as ADAR2-dependent, and 24 sites were shown to be edited by both (Fig. 3e and Supplementary Data 6).
Supplementary Data 8). Of the sites whose editing is conserved in mouse, we found 11 sites in which editing disappears upon ADAR2 knockout (KO), compared to only 3 sites in ADAR1 KO. In 45/58 sites (78%), editing is suppressed more strongly upon ADAR2 KO (Fig. 3f and Supplementary Data 9).

Assuming editing in coding sequences is functionally important, we expect the editing levels to be regulated. Looking at the variability of editing levels across individuals (controlling for the sampling noise; "Methods"), we find that weakly edited sites are generally noisy, with standard deviation as high as the mean.
However, the strong sites mostly show a well-regulated behavior (Fig. 3g and Supplementary Data 10).

Clustering of CDS editing sites. It is well known that editing sites in non-coding regions of the genome are clustered\(^\text{31-53}\). Interestingly, we observe here clustering of many sites in the coding region. About half of the sites belong to dozens of clusters, some of which include over 10 sites each (Fig. 4a). Moreover, for many editing sites, we find a putative editing complementary sequence (ECS) that is located up to 15 kb downstream to the canonical transcript. These structures are mostly overlapping with a paralog gene expressed from the other strand and account for many of the larger clusters (Fig. 4b). We believe these events are due to an extended UTR, not included in the canonical RefSeq transcript, which are known to be expressed in brain tissues\(^\text{54}\). For example, an editing cluster including 19 sites is located in the coding sequence of cldn9 (Claudin 9). A long ECS is found 1741 bp downstream, overlapping the coding sequence of a paralog gene, cldn6 (Claudin 6). One observes a cluster of T-to-C sites within this putative ECS, attesting for expression and editing of this region in the CLDN9 coding strand (opposing CLDN6 coding sequence). Consistently, T-to-C mismatches are over-represented in the brain, especially in cerebellum (Supplementary Fig. 5), where the long UTRs are expressed. This suggests a new model for editing regulation, where editing depends on the expression of the longer UTR isoform, which may be tissue-dependent (Fig. 4c).

The HSPA1L (Heat shock 70 kDa protein 1-a) gene, which has a large UTR, has been validated editing in this gene by sequencing matched DNA and RNA blood samples from three individual persons (Fig. 4d; “Methods”).

Differential editing of specific recoding sites was reported in multiple diseases\(^\text{37,38}\). We have re-analyzed\(^\text{59-61}\) matched normal and cancer samples of the TCGA dataset\(^\text{62}\) and found eight sites from our set which show a significant and appreciable (>10%) change of their editing level in at least one of the nine cancer types studied (Supplementary Data 11). In light of the strong recoding activity in arteries, we have also looked for differential editing in the arteries using GTEx data. Interestingly, of the 27 different diseases and conditions annotated we have found significant and appreciable (>10%) differential editing only in one disease, pneumonia, for which higher editing is observed at five sites (Supplementary Data 12).

Conservation of recoding sites, and signals of adaptation. To gain insight into the evolution of recoding in the mammalian lineage, we have analyzed 5673 RNA-seq samples originating from 21 non-human mammals\(^\text{63-93}\) (Fig. 5a; Supplementary Data 13 and “Methods”) and quantified the editing level at the one-to-one orthologous locations for each of the 1517 human-detected sites (when available). Contrary to previous estimates, we find that a sizable fraction of human sites are conserved across species. As many as 835 editing sites (55% of the set) are conserved in at least one of the groups (Fig. 5a). The fraction of conserved sites increases with the (human) editing level (Fig. 5b). For example, 46% of sites edited to >10% and 75% of sites edited to >30% are conserved out of the primates. On the other hand, only 17 sites can be identified as (probably) human-specific (“Methods” and Supplementary Data 14).

It was previously pointed out that the reported editing sites in human coding sequences are depleted in nonsynonymous sites (ratio of nonsynonymous to synonymous editing incidence rates \(f_N/f_S = 0.6\)) and that editing levels of nonsynonymous sites are generally lower, suggesting an overall deleterious effect\(^\text{94}\). Revisiting this finding with our set of coding-sequences editing sites, we observe neither depletion of nonsynonymous sites nor a lower editing level (abundance \(f_N/f_S = 0.93\); not significantly different from unity, \(p\)-value = 0.20, proportion test; editing level \(p = 0.42\), Mann-Whitney test) (Fig. 5c, d). Moreover, taking into account that neighboring synonymous and non-synonymous sites are not mutually independent, and a strong adaptive nonsynonymous site may be accompanied by synonymous (and nonsynonymous) weak “satellite” sites, we looked again at the abundances of sites, focusing on strong (≥10%) sites and including only the strongest site in each gene. Here, one observes an enrichment and an increased editing level of nonsynonymous editing \((f_N/f_S = 1.5; p\)-value = 0.0028, proportion test; editing level \(p = 0.0014\), Mann-Whitney test) (Fig. 5c, d), indicating positive selection. Following Jiang and Zhang\(^\text{95}\), one may distinguish between restorative and diversifying recoding sites. In the former, recoding introduces an amino acid that was encoded in the genome of an ancestral primate species (“Methods”), whereas diversifying editing introduces a novel protein isoform within the considered phylogenetic tree. Comparing the incidence rates and the editing levels of synonymous, restorative and diversifying editing sites, one finds that although diversifying sites are slightly depleted compared to synonymous editing (incidence rate ratio 0.78, \(p = 0.016\), their mean editing level is 1.5-fold higher \((p = 0.019)\) (Fig. 5e, f). Taken together, these results suggest that the 260 strongly-edited (≥10%) human editing sites are overall adaptive.
Evolution of editing sites. The detailed comparison of human editing sites to other mammals enables us to examine possible evolutionary routes through which novel editing sites may emerge. Interestingly, even a single mutation may be responsible for the introduction of editing sites. For example, an exon of the ASNS gene pairs with an intronic sequence to form a perfectly-paired 50 bp-long dsRNA. This structure is conserved in most primates, but we find no editing (or merely the editing of a single position) out of the Homininae. A single G-to-A mutation in the intron resulted in an A:C mismatch specific to the Homininae. Such a mismatch flanked by long dsRNAs is known to be a preferred ADAR target. Consistently, the mutated intronic A is strongly edited, and four additional editing sites appear in the vicinity of the exonic C (Fig. 6a). On the other hand, a few mutations can abolish strong editing sites, as is demonstrated by the NEIL1 example. Two adenosines are strongly edited (79 and 94%) in NEIL1 coding sequence in human, but not in rodents (no editing in rat; 9 and 0.5% editing in mouse). It was previously suggested that the strong editing is primate-specific and depends on a neighboring Alu sequence. However, these two sites are...
extensively edited in the Alu-lacking sheep (91 and 94%). The secondary structure is similar in all mammals, but a few mutations result in the appearance of two asymmetric internal loops in rat (compared with one in human, and a larger symmetric loop in sheep), which may be responsible for the dramatic suppression of editing (Fig. 6b).

Finally, a human specific editing site appears following a G-to-A mutation within an exon of the sema5B (Semaphorin 5B) gene (Fig. 6c). The location is included within a region that forms a dsRNA structure, and changes a G/G pair into an A/C mismatch, which is favorably edited.

**Discussion**

Reocding by A-to-I editing provides an additional layer of the complex regulation of the proteome. Mapping the landscape of recoding sites enables future studies to reveal the full functional impact of recoding in mammals, across previously unstudied tissues, pathways, and pathological conditions. It also provides a better understanding of the structural and sequence determinant underlying recoding efficiency and regulation which may be then utilized for an improved design of ADAR-based RNA engineering strategies.

Most sites in our list are weakly edited. Furthermore, the number of detected sites depends strongly on the editing level cutoff, and there are thousands of additional sites edited to <1% in all tissues examined. These are less likely to be functionally important. However, it is possible that sites appearing to be weakly edited when averaged over a tissue under normal conditions, exhibit much higher editing levels under specific conditions, in specific subpopulations of cells, or even at a single-cell level. Combining large-scale single-cell sequencing data with the editing detection method presented here could reveal the full scope of recoding and its functional impact.

**Methods**

**Human RNA-sequencing data.** RNA-seq samples from the Genotype-Tissue Expression (GTEX) project\(^{15}\) were used as the core input for our de novo RNA editing detection pipeline. Raw data (700bp-long, paired-end reads) were accessed via the database of Genotypes and Phenotypes (dbGaP, Study Accession: phs000242.v8.p2) and FASTQ files were downloaded via Sequence Read Archive (SRA) using SRA Toolkit (v2.8.0). We used the raw reads and not the alignment files available from GTEX, as they include multi-mapped reads, which hinder de novo discovery of editing sites. A total of 9125 RNA-seq samples from 548 donors and 47 tissues were analyzed (Supplementary Data 1).

**Nonhuman RNA-sequencing data.** Nonhuman mammalian RNA-seq datasets were collected from varied sources\(^{8,30}\) to test the conservation of the de novo detected human editing sites across the mammalian lineage. Altogether, 5673 RNA-seq samples (Supplementary Data 13) from 21 different nonhuman species (Supplementary Data 15) were analyzed.
detections. Alu elements are actually known to be heavily edited\textsuperscript{4,103}, and were excluded to prevent a bias in the accuracy calibration of the pipeline due to the numerous editing events in these regions.

Original dbSNP tables, RefSeq and repeatMasker annotations were downloaded from the UCSC genome browser website\textsuperscript{100} (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/). The RefSeq and the repeatMasker tables are updated as of August 8, 2017 and January 1, 2018, respectively. These filtered regions sum up to 2.1% out of the RefSeq annotation that was used.

**BLAT-based mismatches filtration.** Many mismatches result from erroneous mapping of reads to a highly similar locus in the reference genome. To clean these, we used a sliding window approach, looking each time at a 76 bp-long subsequence (to match reads’ length) of all RefSeq coding sequences with a step of 19 bp (25% of the read length). These 76 bp-long subsequences were aligned to the reference genome using BLAT (BLAST-like Alignment Tool)\textsuperscript{104}, in the search of highly-similar genomic regions (alignment length $\geq$ 61 bp, $\geq$ 94% identity, no indels in query), mainly due to duplicated genomic regions and processed pseudogenes.
**Fig. 4 Many of the CDS editing sites are clustered.** a Bar plot: Distribution of editing clusters by size. Pie chart: Relative abundance of editing sites by cluster size. b Editing requires a dsRNA secondary structure, often depending on a distant editing complementary sequence (ECS). Distribution of putative ECSs location (“Methods”) shows that smaller clusters tend to depend on an intronic ECS, while the ECS for large clusters (≥10) is often found out of the annotated gene boundaries, overlapping an exon of a neighboring downstream gene. c Editing due to a downstream reversely-oriented paralog gene, overlapping an extended 3′UTR. The TOR1B gene contains a 10-site RNA editing cluster in its coding sequence. The corresponding ECS is a closely-related sequence that overlaps a coding exon within the neighboring paralog TORIA gene (genomic coordinates of the last exon of both genes are shown). Presumably, the 3′UTR of edited TORIB transcripts is longer than that of the Refseq canonical transcript, extending to overlap the coding sequence of Torla (on the opposite strand). These extended-3′UTR variants of the Torlb transcripts contain two highly-similar reversely-oriented sequences, and can form a long and stable dsRNA structure, resulting in extensive editing of both the CDS (red lines) and the TORIA-overlapping 3′UTR (yellow lines). A 14-site cluster of T-to-C sites is detected in the coding sequence of TORIA, a hallmark of A-to-I editing in antisense transcripts (see Fig. 2d). See Supplementary Data 3 for further evidence for expression and editing of the extended UTR of TORIB. d Editing at one of the large clusters within the HSPAIL transcript. Up: pile-up of the RNA reads’ coverage, 19 different editing sites are observed in this 391 bp-long segment of the editing cluster (red and blue bars stand for A and G fractions, respectively). Bottom: Matched RNA and DNA sequences (presented is the pooled data for the three individuals, for simplicity). Editing events (A-to-G mismatches) are shown in blue (different colors represent other type of mismatches). The absence of A-to-G mismatches in the matched DNA samples and lack of linkage between neighboring sites both support the sites being RNA-edited.

**Fig. 5 Evolution, conservation, and signs for adaptation in CDS-editing sites.** a Phylogenetic tree of the mammalian species analyzed (branch lengths not to scale). For each of the six evolutionary groups, numbers on left present the number of sites conserved between the human and the last common ancestor (LCA) of the group, and, in parentheses, number of human sites which are determined as not edited in the group (“Methods”). b Fraction of evolutionary-conserved sites whose (human) editing level is above a given cutoff, as a function of this cutoff. c Ratio of incidence rates for synonymous and non-synonymous sites is approximately unity (dashed gray line), for weak (<10%) and strong (≥20%) sites alike. For the strongest site in the gene, weak nonsynonymous editing sites are slightly depleted while strong nonsynonymous sites are enriched, suggesting positive selection and adaptivity. Box-and-whisker plots represent the distribution of ratios over 10^6 random As, controlling for the ±1 bp nucleotide context. d Mean (colored bars) and distribution (box-and whisker) of editing levels. Number of sites: 879 nonsynonymous and 378 synonymous weak sites; 177 and 83 strong sites. Considering the strongest site per gene only: 408 nonsynonymous and 211 synonymous weak sites; 105 and 28 strong sites. e Sites were classified into three categories based on the reconstructed amino acid encoded by ancestral primates (“Methods”): synonymous, restorative (i.e., recoding results in an amino acid encoded by ancestral primate), or diversifying sites (otherwise). The incidence rate (“Methods”) is higher for restorative editing (not significant) and slightly lower for diversifying sites, relative to synonymous ones. Box-and-whisker plots show median (horizontal lines), upper and lower quartiles (box edges), and 1.5 × the interquartile range (whiskers). P-values by randomization test (c, e) or two-sided Mann-Whitney test (d, f). Source data are provided as a Source Data file.
Mismatches found by these BLAT alignments are likely to appear as mismatches between RNA-seq data and the reference genome, due to misalignments of sequencing reads originating from one genomic locus to another, highly-similar, locus. Thus, these mismatches were discarded from the above list of potential editing sites. A-to-G mismatches were detected in ~0.5% of all adenosines in the RefSeq annotation.

**Statistical model and score calculation.** To decide whether a specific genomic position is an editing site, we employed a statistical model that integrates the data from all donors for each tissue, using a maximum likelihood approach. For each donor, three binomial tests and three p-values are evaluated per site, to assess the likelihood of the observed total coverage C and number of reads supporting a mismatch E, assuming editing does not occur at this site:
Fig. 6 Species-specific dsRNA structures contribute to the formation of novel RNA editing sites. a Several hominid-specific exonic RNA editing sites were introduced to the Asn5 gene following a single G-to-A genomic substitution modifying the RNA secondary structure. Two uppermost sequences show a few CDS adenosines being edited in human and chimpanzee. These editing events are not conserved in more distant primates. In non-hominid primates (two lowermost structures), the RNA structure is very similar, except for a single hominid-specific mismatch (internal loop) feature, resulting from a genomic G-to-A substitution in the intronic ECS of hominids. b The extensive editing in NEIL1 CDS seen in primates (top) is conserved in some non-primates as well (bottom). However, editing is much weaker in rodents (middle). Mutations in the rat sequence alter the secondary structure, while the sheep structure is more similar to the human one. c The emergence of human-specific RNA-editing site. A human-specific G-to-A substitution results in an A:C mismatch within a dsRNA structure already present in the ancestral hominid. This newly introduced adenosine is efficiently deaminated by the ADAR enzymes, and an additional weaker site appears in the complementary sequence. In each of the three panels, the top track shows the locus of the edited CDS region and the ECS (gray dashed-line boxes) within the corresponding gene (blue boxes – exons; lines with arrowheads – introns); the bottom track shows the corresponding calculated secondary structures. The topology of the phylogenetic tree for the species presented is shown on the left (branch lengths not to scale). Species names are denoted on the right. Major structure alternations along with indels and single nucleotide substitutions (relative to human) are indicated. Adenoses in all structures are annotated in different colors based on their calculated RNA editing levels (pooling all brainsamples available, per species; Gray = no detectable editing, blue = low editing levels; red = high editing levels). Editing levels were not assessed for sites covered by <50 reads.

1. The probability that the retrieved C and E are merely due to sequencing errors, given a maximal probability of base calling sequencing error of $E_{\text{error}} = 10^{-3}$ (as set by the minimal Phred quality score to call subtraction of $q = 30$):

$$P_{\text{error}}(X \leq E) = \sum_{p=0}^{X} \binom{C}{j} \left(1 - E_{\text{error}}\right)^{C-j} \left(E_{\text{error}}\right)^{j}$$

2. The probability that the retrieved C and E represent a rare heterozygous SNP, given an expected frequency of the alternative allele $p_{\text{hetero}} = 0.5$ and an a-priori probability of such a genomic event of $p_{\text{error}} = 10^{-4}$:

$$P_{\text{hetero}}(X \leq E) = \sum_{p=0}^{X} \binom{C}{j} \left(1 - p_{\text{hetero}}\right)^{C-j} \left(p_{\text{hetero}}\right)^{j}$$

3. The probability that the retrieved C and E represent a rare homozygous SNP, given an expected frequency of the alternative allele $p_{\text{homo}} = 1$ and an estimated probability of such genomic event of $p_{\text{homo}} = (p_{\text{error}})^{C}$:

$$P_{\text{homo}}(X \leq E) = \sum_{p=0}^{X} \binom{C}{j} \left(1 - p_{\text{homo}}\right)^{C-j} \left(p_{\text{homo}}\right)^{j}$$

We choose the most likely of the above three explanations and adopt the lowest of these three $p$-values. Had this been a bona fide $p$-value, one could have then incorporated the multiple $p$-values (one per donor) into one statistical test using Fisher’s method, and define a score $S$ per site based on the full data from N donors to be (natural log):

$$S = -2 \sum_{i=1}^{N} \log \left[ \min \left( P_{\text{error}}, P_{\text{hetero}}, P_{\text{homo}} \right) \right]$$

This statistic should then have a $\chi^2$ distribution with 2N degrees of freedom. However, as we allow for more than one test per site and choose the most likely possibility, the chosen $p$-value per donor is not truly a $p$-value (in the sense that the values are not uniformly distributed along the unity interval given the null hypothesis of no editing). Therefore, $S$ does not follow the $\chi^2$ distribution. Yet, our approach overestimates the true $p$-value, and estimating the likelihood of getting values as high as $S$ by the $\chi^2$ distribution would provide an underestimate of the statistical significance. Note, however, that the statistical significance obtained is only correct if one assumes the mismatches of different reads to be independent. De facto, we have found that many of the false-positive detections are due to systematic misalignments and other factors not satisfying this independence assumption. To avoid confusion and misinterpretation, we therefore chose not to assign a $p$-value to each site, but use the score only as a filtering step. Mismatch sites with a score $S \geq 2N$ were discarded ($2N$ being the mean of the distribution). Another $p$-value-based filtration was aimed to eliminate several sorts of errors that are specific to one sequenced-strand. During the sequencing protocol, the original RNA fragments are reverse-transcribed into double-stranded cDNA which is then amplified. The sequencing machines sequence both strands of these cDNA fragments, regardless of the original expressed strand. Therefore, an RNA editing event could be misidentified as a mismatch showing up in both strands of the cDNA. In contrast, mismatches due to technical sequencing and amplification errors, are often strand specific. Thus, we have pooled the data from all samples and applied a binomial test to the reads aligned to each of the genomic strands, separately, requesting that both tests result in a significant $p$-value rejecting the possibility of sequencing errors (both $p$-values < exp$^{-N/4}$).

Cluster-based filtering. Clusters of mismatch sites, i.e., multiple mismatch sites of different substitution types (e.g., A-to-G and C-to-A, etc.) at the same locus, might point to problematic read alignments (e.g., mapping of reads due to a polymorphic position leading to its original genomic position) or to false-positive detections. To filter these, we calculated the positions of all mismatches relative to the mature mRNA sequences using AnnoV (v2018April16; https://doc.openbio.readthedocs.io/projects/annovar/en/latest/)104 with RefSeq transcripts built database. Using BEDTools cluster (v2.26.0)106, we joined pairs of mismatch-positions at a distance <100 bp apart into clusters, iteratively. Clusters containing more than one type of mismatch were discarded.

Whole-exome-sequencing-based filtering. To further clean up our list from sites with a genomic origin, we analyzed WES data, deposited as a part of the GTEx dataset. Additionally, this analysis may also filter out candidate sites that resulted from alignment errors. A total of 609 WES samples from 602 distinct donors (mostly from whole blood specimens) were used (Supplementary Data 16). Matching WES data was available for 504 of the 548 RNA-seq donors. The other 98 unmatched WES data were used for the analysis as well. We mapped the WES data using the same pipeline applied to the RNA-seq samples, and calculated the WES-mismatch level at each of the putative editing sites. Sites with insufficient (<100) WES reads or a WES-mismatch level exceeding 0.1% (namely, the maximal expected sequencing error rate) were discarded.

Recalculating editing levels and editing-level-based filtration. To further reduce noise in our set of sites we also excluded sites with low editing level, as sites in which a very small fraction of the transcripts were edited are assumed to have only a limited contribution to functionality. To correctly assess the per-site editing levels in each of the tissues, candidate sites from the 47 tissues were combined to a single list. Next, the total coverage and the number of reads supporting the mismatch (i.e., edited reads) were calculated from the original unfiltered BAM files using SAMtools mpileup (v1.2; http://www.htslib.org)101. The reference genome was not supplied and the following options were applied: -B -Q 0 -d 0 -l 1000000 -f SECONDARY,UNMAP,DUP. The per-site-per-sample (i.e., donor in a certain tissue) coverage data were summed. Sites that did not exhibit an editing level >1% in at least one tissue were discarded. The resulting 1517 A-to-G sites, their annotation, and editing level per tissue, appear in Supplementary Data 2.

Maximal editing levels. The maximal editing level was calculated separately for each site. It was defined as the maximum editing level across all tissues with coverage ≥100 reads. Sites’ maximal editing level was used in all non-tissue-specific analyses as a representative editing level value unless otherwise stated.

RNA editing index. The editing index is defined as described in44 for all GTEx samples. Briefly, for a chosen set of A-to-G editing sites, mpileup data was used to determine the number of observed matched As and the number of mismatching Gs at the sites. The index is the percent of Gs out of the number of As and Gs combined. The editing index values were calculated separately for each tissue type unless stated otherwise. The T-to-C index was assessed in the same manner.

Coverage calculation in nonhuman samples. Genomic coordinates of the sites orthologous to the 1517 human RNA editing sites in the 21 nonhuman reference genomes were found using the liftOver program (http://hgdownload.soe.ucsc.edu/).
admin/exe/linux.x86_64/liftOver) with default parameters and the appropriate conversion files (*.over.chain.gtf files). Then, liftOver was used again to nonorthogonally remap the data back to the human reference genome, to verify these are reciprocal best hits, thus a common proxy for real orthologs. For each species, unmapped records, and records for which the orthologous site was not mapped back to the original human coordinates were discarded. Whenver the nucleotide at the orthologous genomic position in a given species was an adenosine, total coverage and a total number of edited reads were calculated for each biological sample, using mpileup as described in the paragraph above.

Identification of conserved RNA editing sites. We used NCBI taxonomy (https://www.ncbi.nlm.nih.gov) to construct a phylogenetic tree for human and the 21 nonhuman mammals. Inner nodes exhibiting multifurcation were manually indexed (http://www.ensembl.org/info/docs/api/), retaining only genes having one-to-one orthologous genes in all the indicated species. Next, mRNA and protein sequences of the Ensembl canonical gene models were aligned using ClustalOmega (https://www.clustal.org/omega/). The tree was visualized using iTOL (ver. 5.6.3) online tool107.

In order to gain sufficient sequencing coverage and minimize multiple-testing (thus increasing statistical power), the 21 individual mammalian species were grouped into 6 main groups according to lineages and species. Primates (Cercopithecinae, non-Catarrhini primates, Glires, Laurasiatheria, and non-Boreoeutheria mammals, Supplementary Data 15). Due to their different editing pattern, samples in each group were further classified into two categories: brain and non-brain samples (Supplementary Data 13). The per-site values of total coverage and number of supporting reads were summed over all samples in each of the six groups of organisms and each species, resulting in 12 different lists. A binomial test was applied for each site, with the null hypothesis that the site is not edited in a given nonhuman group and category, and possible mismatches can be attributed to random sequencing errors (with a rate of $10^{-3}$, corresponding to $q = 30$). The resulting p-values were adjusted using Benjamini and Hochberg's False Discovery Rate (FDR = 0.001) correction, for each of the 12 lists, separately. A human RNA editing site was considered conserved in a certain nonhuman group if it was detected in either brain or non-brain tissues. On the other hand, a site that was not to be conserved and exhibited a non-significant $> 0.05$ FDR and a total coverage $> 500$ in either brain or non-brain categories was considered non-conserved (Supplementary Data 2). To determine the emergence of editing at a given site along mammalian evolution, we look for the last common ancestor (LCA) of human and the most distant mammal exhibiting RNA editing, for which the editing in the most recent outgroup (of the descendants of the LCA) is non-conserved. This LCA is considered to be the one at which the editing of a site has emerged.

To search for human-specific editing sites (Supplementary Data 14), we looked for either (i) edited human adenosines where the reconstructed nucleotide in the ancestral Hominum sequence is not an adenosine (and therefore not edited), or (ii) human sites exhibiting an editing level $> 2\%$ in cerebellum and cortex that were deemed non-conserved (as defined above) in nonhuman-Homininae and in one of Cercopithecinae or non-Catarrhini primates.

Calculating the relative frequency of synonymous and non synonymous editing sites. To estimate the extent to which our detected A-to-G sites are positively selected throughout evolution we used an approach that largely follows94. First, choosing the longest protein-forming transcript per gene from RefSeq annotation (see above), we counted CDS adenosines whose potential editing into inosines would have resulted in synonymous and nonsynonymous changes ($S = 2,512,451$ and $N = 6,185,931$, respectively). Then, we calculated the ratios of the number of actual editing sites in our list to the number of potential sites, resulting in $f_S = 1.83 \times 10^{-4}$ and $f_N = 1.71 \times 10^{-4}$ for synonymous and nonsynonymous editing, respectively. As weak editing sites are expected to have a lesser functional impact, we have repeated the calculation for weak ($< 10\%$) and strong ($> 10\%$) sites separately. In addition, for each of the four sets of weak and strong, synonymous and non-synonymous, sites, we constructed 10,000 sets of i517 randomly chosen CDS adenosines, with the same distribution over chromosomes and sequence contexts (1 bp downstream and upstream), and calculated the mean and SD over the 10,000 random sets of each category. The same methodology was applied to all genes in the RefSeq annotation (see above) regarding all adenosines as putative RNA editing sites in order to assess the frequencies of restorative, diversifying, and synonymous sites relative to their genomic background. Additionally, 10,000 sets of 828 randomly-selected sites were chosen out from the sites in the genomic background with the same distribution over sequence contexts (1 bp downstream and upstream), as a control. Randomization-based p-values were evaluated using these sets.
conversion, 439 sites were excluded due to a non-coherent or a non-A conversion result, leaving 1078 sites. Importantly, none of these sites have shown A-to-G mismatches or an acceptable level in ADAR1 protein loss (Supplementary Data 13). First, biological replicates were pooled, and Fisher’s exact test was applied, comparing the numbers of edited and non-edited reads per site in each targeted ADAR and genetic modification to those in the appropriate WT control. A site was considered as enzyme-specific if the editing level was found to be significantly greater than the respective null (corrected p-value ≤ 0.05) by either the OE or the KO for one of the ADARs but not for the other and if the editing level was changed in agreement with the nature of the modification (i.e., higher editing levels for OE and lower for KO). When significant p-values were obtained for both Adar1 and Adar2 the site was considered as a shared target of both enzymes.

In addition, we estimated the relative influence of Adar1 (I1) on the editing level at site as follows:

\[
I_1 = \frac{\log_2 \left( \frac{W T \text{ Editing level} - 0.001}{\text{Adar1 KO Editing level} + 0.001} \right)}{\log_2 \left( \frac{W T \text{ Editing level} - 0.001}{\text{Adar1 KO Editing level} + 0.001} \right) + \log_2 \left( \frac{W T \text{ Editing level} - 0.001}{\text{Adar1 KO Editing level} + 0.001} \right)}
\]

(5)

Negative values of \(I_1\) were replaced by \(I_1 = 0\).

Identification of ECS. To search for complementary sequences, we retrieved a 41bp-long genomic sequence flanking each of the editing sites (20 bp in each side) and used the BLAST program (release: 2020-03-17; https://www.bioinformatics.org/) to look for a reversely-oriented complementary sequence within 20,001 bp (10k in each side) surrounding the editing site (command: ab-blastn W = 4 Q = 14 R = 4 -matrix=RNA homspimp5). Hits of length ≥30 and identity ≥70%, residing within the same transcript or up to 5 kb downstream to it (accounting for long, unannotated UTRs) were considered as putative ECS. As a control, we randomly chose 1517 exonic adenosines that are not known to be edited. For both editing and control sites, minimum free energy (ΔG) was calculated for the longest BLAST hit (within the transcript or 5000 bp downstream, but regardless of its length and identity) applying the fold program from the RNAs- tructure package113 to the query region and the subject region of the hit, connected with a 100 bp poly(A) linker when needed. Sites for which either the calculated free energy was positive or no hit was found were assigned ΔG = 0.

Multi-species RNA secondary structure comparisons. Multi-species dsRNA structures were calculated for regions within 3 genes: ASNS, NEILL and SEMA3B. Human ECS were detected as mentioned above with the exception of SEMA3B, where the BLAST-dependent analysis was not sensitive enough to detect the ECS. Instead, a 72 bp-long region that harbors the single editing site in the coding sequence of SEMA3B and found to form a dsRNA structure in silico was used. Next, nonhuman homologous sequences of the edited region and ECS were retrieved, using the UCSC multiZwayy (ASNS and SEMA3B) and multiz100way (NEILL) tables. RNA secondary structures were predicted, for each gene separately, using the RNAStructure Multilign program with default parameters. Sequences from species with non-editable CDS variants were folded separately from editable ones. Next, to determine the possibility of mismatches between the homologous sequence we used Clustal Omega (Ver. 1.2.4)108 program to generate their multiple sequence alignment.

Proteomic evidence for editing. In order to detect editing events at the protein level we downloaded mass spectrometry data, available at the PRIDE database40. We analyzed 311 proteomic and phosphoproteomic samples, (Supplementary Data 4) originating from various human tissues, digested using trypsin, and quantified using label-free quantification (LFQ), tandem mass tag (TMT), or isobaric tag for relative and absolute quantification (iTRAQ).

To prepare a database of all possible peptides, including the ones derived from the 1036 nonsynonymous RNA editing sites in our set, we used RefSeq transcripts and considered all possible combinations of edited and non-edited states at all A-to-G editing sites detected by our pipeline. The translated transcripts were in silico trypsin-digested into peptides, taking into account up to two missed cleavages per peptide. We considered potential trypsin cleavages losses and gains due to editing but not stop losses. In addition to this set of peptides, we created two additional peptide databases as controls: (i) A proteome where RefSeq transcripts may be “edited” at the 851 non-synonymous non-A-to-G sites detected by our pipeline (sites we consider as noise). Some of these are likely to be found in the proteome, as they represent genomic variability or misalignment due to genomic duplications. (ii) A proteome where RefSeq transcripts may be “edited” at 1036 randomly-chosen genomic positions to create a non-synonymous change. These random sites were matched to the 1036 detected recoding sites in terms of the edited codon, and (if possible) were chosen within the same exon as the true edited site (829 sites). In cases an identical codon was not found in the same exon, we picked a position in the same codon from another exon of the same gene (187 sites) if possible, or in other genes otherwise (20 sites).

The proteomic samples were compared to this set of potential peptides using MaxQuant (MO)114, with default parameters and several modifications, as follows: for phosphoproteome samples we added phospho-TRY modification to the variable modifications list; LFQ samples were analyzed with the parameter limmode=1 for delayed normalization using MO LFQ algorithm; for TMT and for iTRAQ samples we chose 11plex-TMT and 8plex-TTRAQ in the isobaric labels parameter section, respectively (with reporter ion MS2). Peptides that may have derived from common contaminants (based on MQ Contaminants Fasta file) were excluded. We then considered only peptides that are distinct to either the non-edited or the edited versions. For each editing site and each sample, we recorded peptides that support the genomic version, the edited version, or both. The same search was applied to the two control databases.

Estimating biological variability of editing levels. To assess the variability of editing across donors, we first summed the total coverage and the number of edited reads over a number of tissues. To avoid variability due to a different set of available tissue samples, we wanted to have the exact same set of tissues for all donors considered. Since not all tissues are available for all donors, we had to choose a subset of donors and a subset of tissues, trying to keep both the number of donors and the number of brain regions sufficiently large to gain statistical power. Focusing on the brain, we studied 65 donors for each of which we pooled the coverage and editing data from five brain regions: caudate, nucleus accumbens, cerebellum, cerebellar hemisphere, and putamen.

Next, we used the following model to separate the stochastic sampling noise and estimate the biological variability across samples. For each editing site, denote \(c_i\) and \(g_i\) the total coverage and number of edited reads in sample \(i\), respectively. The observed editing level per sample is therefore \(e_i = g_i / c_i\). We assume \(g_i\) to be distributed binomially \(g_i 
= \binom{c_i}{p_i}\), where \(p_i\) is the underlying editing level per site per sample (i.e., fraction of edited DNA molecules). This level may be different from sample to sample, and we denote its mean and variance over the population by \(p\) and \(\nu\), respectively. The mean observed level is thus

\[
\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n}
\]

where \(\bar{e}\) denotes averaging over both samples (5) and the binomial distribution (B), and we have used the harmonic mean \(\bar{p} = \sum_{i=1}^{n} p_i (c_i) / \sum_{i=1}^{n} c_i\). The variance of the observed editing level (over both samples and binomial distributions) is

\[
\text{Var}(\bar{e}) = \left( \frac{\bar{e}}{c} \right) - \left( \frac{\bar{e}}{c} \right)^2 - \left( \frac{g}{c} \right) + \left( \frac{g}{c} \right)^2 - \left( \frac{g}{c} \right) + \left( \frac{g}{c} \right)^2
\]

(6)

where \(c^{-1} = \frac{1}{c}\) and the penultimate equation assumes that coverage and editing level distributions over samples are uncorrelated. One thus obtains the following for the variance over the population of the underlying editing level:

\[
\text{Var}(\bar{p}) = \left( \frac{\bar{c}}{c} \right) - \left( \frac{\bar{c}}{c} \right)^2 - \left( \frac{\bar{g}}{c} \right) + \left( \frac{\bar{g}}{c} \right)^2 - \left( \frac{\bar{g}}{c} \right) + \left( \frac{\bar{g}}{c} \right)^2
\]

(7)

For a given set of samples, one may estimate \(p\) and \(c^{-1}\) as the mean (over samples) of \(e_i\) and \(c_i\), respectively, and then apply the above formula to obtain an estimate \(\nu\). Note that this estimate may occasionally turn out negative, due to statistical errors in these estimates. In these cases, our best estimate for the variability in the underlying editing level is zero. Finally, we point out that some of the variability across samples of the underlying editing level may be due to technical rather than biological reasons (sample-to-sample differences in the sequencing protocol that may have an effect on the variance). The above estimate is therefore an over-estimate of the biological variability.

Differentially edited sites in cancer and cardiovascular diseases. To look for differential editing in cancer, RNA-seq samples for tumor samples and matching controls were downloaded from The Cancer Genome Atlas42. Following59, we considered the nine cancer types for which healthy controls are available. Reads were re-aligned to the genome and editing levels were calculated per sample per site, as described above (Reads Alignment and BAM processing; and Recalculating Editing Levels and Editing-Level-Based Filtration).

For each donor, we excluded sites for which the number of reads aligned in the normal and tumor samples combined was <40. Then, we calculated the editing level in the pooled normal samples and the pooled tumor samples, separately, and excluded sites for which both editing levels were <0.5% in any of the cancer types. Paired-samples student’s t-test was then applied to the remaining sites, for each tissue type, to compare the editing levels per sample between the matched normal and tumor groups, followed by Benjamini–Hochberg multiple testing correction (FDR = 0.05). We further requested an appreciable >10% difference in the mean editing level between the groups.

Similarly, we looked at the editing levels observed in the three artery tissue types available in GTEx (aorta, coronaries, and tribial). For each tissue and site, we looked at the 27 diseases and conditions annotated in GTEx. We discarded low-coverage
Single-cell differential RNA editing analysis. We used Tabula Muni,[15] a recently published comprehensive mouse single-cell dataset to test for variances in editing levels between different cell populations within the same tissue. Smart-seq2 full-length RNA-seq data from 44,494 different cells (SRA accession: SRP131661) were aligned as performed for other nonhuman samples (see above). Total coverage at the editing sites and editing levels were assessed as described above (Recalculating Editing Levels and Editing-Level-Based Filtration) for the classification by targeting ADAR. Next, per-cell total coverage and editing values were summarized across all cells in each cell population based on the previously-published annotations[14]. As for lung cells, a revised and more accurate annotation was used[14] (Supplementary Data 6). Then, sites that did not exhibit an editing level ≥ 0.1% with minimal coverage of 5 reads in at least one cell population were excluded, and for each site, all possible pairs of cell populations within the same tissue were compared to detect significant differences in editing levels. A total of 14,503 pairs were evaluated in both the cell-population and the single-cell levels. At the level of the assembled editing events controlled by pre-mRNA splicing, the mean editing level between the groups. None of the sites identified is adjacent (<2000 bp distance) to known GWAS loci.  

Reporting summary Furthermore information on research design is available in the Nature Research Reporting Summary linked to this article.  

Data availability The data supporting the findings of this study are available from the corresponding authors upon reasonable request. Human data used in the paper is available upon dbGaP approval (GTEx: phs000424.v8.p2; TCGA: phs000178.v11.p8). Newly generated next-generation sequencing data reported in this study have been deposited in the SRA database, BioProject ID: PRJNA715360. Non-human RNA-seq samples analyzed are available from SRA (see detailed in Supplementary Data 13). Source data are provided with this paper.  

Code availability The scripts used to produce the data are available at GitHub (https://github.com/a2iediting/deNovo-Detect).[17]  

Received: 28 April 2021; Accepted: 27 January 2022; Published online: 04 March 2022  

References 1. Walkley, C. R. & Li, J. B. Rewriting the transcriptome: Adenosine-to-inosine RNA editing by ADARs. Genome Biol. 18, 205 (2017). 2. Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat. Rev. Mol. Cell Biol. 17, 83–96 (2016). 3. Eisenberg, E. & Levanon, E. Y. A-to-I RNA editing—immune protector and transcriptome diversifier. Nat. Rev. Genet. 19, 473–490 (2018). 4. Bass, B. L. RNA editing by adenosine deaminases that act on RNA. Annu. Rev. Biochem. 71, 817–846 (2002). 5. 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 (2014). 6. Basilio, C., Wahba, A. J., Lengyl, P., Speyer, J. F. & Ochoa, S. Synthetic polynucleotides and the amino acid code. Proc. Natl Acad. Sci. USA 48, 613–616 (1962). 7. Higuchi, M. et al. Point mutation in an AMPA receptor gene recues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature 406, 1998–2001 (2000). 8. Chen, L. et al. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. Nat. Med. 20, 209–216 (2013). 9. Yeo, J. Goodman, R. A., Schr, N. T., David, S. S. & Beal, P. A. RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. Proc. Natl Acad. Sci. USA 107, 20715–20719 (2010). 10. Jain, M. et al. RNA editing of Filamin A pre-mRNA regulates vascular contraction and diastolic blood pressure. EMBO J. 37, e94813 (2018). 11. Ramaswami, G. & Li, J. B. RADAR: A rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 42, D109–D113 (2014). 12. Picardi, E., D’Erchia, A. M., Lo Giudice, C. & Pesole, G. REDiPortal: A comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 45, D750–D757 (2017). 13. Mangul, S. et al. ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. Genome Biol. 19, 36 (2018). 14. Tan, M. H. et al. Dynamic landscape and regulation of RNA editing in mammals. Nature 550, 249–254 (2017). 15. Park, E., Williams, B., Wold, B. J. & Mortazavi, A. RNA editing in the human ENCODE RNA-seq data. Genome Res. 22, 1626–1633 (2012). 16. Forath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNAs reveals numerous novel sites. Nat. Commun. 5, 4726 (2014). 17. Zhang, Q. & Xiao, X. Genome sequence-independent identification of RNA editing sites. Nat. Methods 12, 347–350 (2015). 18. Licht, K. A. et al. High resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. Genome Res. 29, 1453–1463 (2019). 19. Carmi, S. et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat. Commun. 5, 4835 (2014). 20. Ramaswami, G. & Li, J. B. Identification of human RNA editing sites: A historical perspective. Methods 107, 42–47 (2016). 21. Kleinman, C. L. & Li, J. B. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome” Science 335, 1302 (2012). author reply 1302. 22. Lin, W., Piskol, R., Tan, M. H. & Li, J. B. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science 335, 1302–1302 (2012). 23. Pickrell, J. K., Gilad, Y. & Pritchard, J. K. Comment on Widespread RNA and DNA sequence differences in the human transcriptome”. Science 335, 1302–1302 (2012). 24. Teleni, A. et al. Deep sequencing of 10,000 human genomes. Proc. Natl Acad. Sci. USA 113, 11901–11906 (2016). 25. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature 583, 83–89 (2020). 26. Zaranek, A. W., Levanon, E. Y., Zecharia, T., Clegg, T. & Church, G. M. A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing. PLoS Genet. 6, e1000954 (2010). 27. Ramaswami, G. et al. Identifying RNA editing sites using RNA sequencing data alone. Nat. Methods 10, 128–132 (2013). 28. Ramaswami, G. et al. Accurate identification of human Alu and non-Alu RNA editing sites. Nat. Methods 9, 579–581 (2012). 29. Mele, M. et al. The human transcriptome across tissues and individuals. Science 348, 660–665 (2015). 30. Zhang, F., Lu, Y., Yan, S., Xing, Q. & Tian, W. SPINT: An SNP-free toolkit for identifying RNA editing sites. Bioinformatics 33, 3538–3548 (2017). 31. Giudice, C. L., Tangaro, M. A., Pesole, G. & Picardi, E. Investigating RNA editing in deep transcriptome datasets with REDitools and REDiPortal. Nat. Protoc. 15, 1098–1131 (2020). 32. John, D., Weirick, T., Dimmeler, S. & Uchida, S. RNAEditor: Easy detection of RNA editing events and the introduction of editing islands. Brief. Bioinform. 18, 993–1001 (2017). 33. Picardi, E. et al. Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci. Rep. 5, 14941 (2015). 34. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013). 35. Ewing, J., Green, T. & Bass, B. L. Predicting sites of ADAR editing in human genomes. Nature 583, 83–89 (2020). 36. D’Erchia, A. M. et al. Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass, and respiratory activity. Mitochondrion 20, 13–21 (2015). 37. Rosenberg, B. R., Hamilton, C. E., Mwangi, M. M., Dewell, S. & Papavasiliou, F. N. Transcriptome-wide RNA editing reveals numerous APOBEC1 mRNA editing targets in transcript 3' UTRs. Nat. Struct. Mol. Biol. 18, 230 (2011). 38. Powell, L. M. et al. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell 50, 831–840 (1987). 39. Polson, A. G., Crain, P. F., Pomerantz, S. C., McCloskey, J. A. & Bass, B. L. The mechanism of adenosine to inosine conversion by the double-stranded RNA-activating/modifying activity: A high-performance liquid chromatography-mass spectrometry analysis. Biochemistry 30, 11507–11514 (1991). 40. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Res. 47, D442–D450 (2019). 41. Nithalik, A. et al. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol. Cell. Proteomics 11, O11013698 (2012).
Acknowledgements

We thank the GTEx consortium for making their RNA sequencing data publicly available. We thank Ariel Feiglin and Ilana Buchumenski for their help. This research has been supported by the International Collaboration Grant from the Jacki and Bruce Barron Cancer Research Scholars Program, a partnership of the Israel Cancer Research Fund and City of Hope, as supported by The Harvey L. Miller Family Foundation [205467 to E.Y.L.], and the Israel Science Foundation (grant numbers 1945/18 to E.E. and 2039/20, 231/21 to E.Y.L.).

Author contributions

O.G. and E.E. performed most of the bioinformatics data analyses and wrote the software. Y.S. performed the proteomics analysis, E.K. the ADAR assignment, U.B.Z. the RNA structure analysis, T.D.M. analyzed the clinical data, N.B. worked with the TCGA data, R.C.F. performed t-SNE and S.H.R. has contributed to specific data analyses. A.A.S. carried out the sequencing tasks. S.H.R. and Z.T. curated the software. E.Y.L., O.G., and E.E. conceived the study, designed the analyses, and wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare competing financial interests. E.E. is a consultant to Korro Bio, a company that develops RNA editors. E.Y.L was a consultant to ADARx. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28841-4.

Correspondence and requests for materials should be addressed to Erez Y. Levanon or Eli Eisenberg.

Peer review information Nature Communications thanks Rui Zhang, Graziano Pesole, and Meng How Tan for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022