bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-wide characterization of RNA editing in chicken: lack of evidence for non-A-to-I events

Laure Frésard[1,2,3], Sophie Leroux[1,2,3], Pierre-François Roux[4,5] Christophe Klopp[6], Stéphane Fabre[1,2,3], Diane Esquerré[1,2,3,7], Patrice Dehais[1,2,3,6], Anis Djari[1,2,3,6], David Gourichon[8], Sandrine Lagarrigue[4,5§], Frédérique Pitel[1,2,3§]

[1]UMR INRA / [2]INPT ENSAT / [3]INPT ENVT, Génétique, Physiologie et Systèmes d'élevage, INRA, F-31326 Castanet Tolosan, France

[4]Agrocampus Ouest, [5]INRA, UMR1348 Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Élevage, F-35000 Rennes, France

[6]INRA, Sigenae UR875 Biométrie et Intelligence Artificielle, F-31326 Castanet-Tolosan, France

[7]INRA, GeT-PlaGe Genotoul, F-31326 Castanet-Tolosan, France

[8]INRA, PEAT Pôle d'Expérimentation Avicole de Tours, F- 37380 Nouzilly, France

[§]Corresponding authors and equal contributors

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

RNA editing corresponds to a post-transcriptional nucleotide change in the RNA sequence, creating an alternative nucleotide, not present in the DNA sequence. This leads to a diversification of transcription products with potential functional consequences. Two nucleotide substitutions are mainly described in animals, from adenosine to inosine (A-to-I) and from cytidine to uridine (C-to-U). This phenomenon is more and more described in mammals, notably since the availability of next generation sequencing technologies allowing a whole genome screening of RNA-DNA differences. The number of studies recording RNA editing in other vertebrates like chicken are still limited. We chose to use high throughput sequencing technologies to search for RNA editing in chicken, to understand to what extent this phenomenon is conserved in vertebrates.

We performed RNA and DNA sequencing from 8 embryos. Being aware of common pitfalls inherent to sequence analyses leading to false positive discovery, we stringently filtered our datasets and found less than 40 reliable candidates. Conservation of particular sites of RNA editing was attested by the presence of 3 edited sites previously detected in mammals. We then characterized editing levels for selected candidates in several tissues and at different time points, from 4.5 days of embryonic development to adults, and observed a clear tissue-specificity and a gradual editing level increase with time.

By characterizing the RNA editing landscape in chicken, our results highlight the extent of evolutionary conservation of this phenomenon within vertebrates, and provide support of an absence of non A-to-I events from the chicken transcriptome.

KEYWORDS

RNA editing - Chicken - RNA-DNA difference - RNA-Seq - DNA-Seq

BACKGROUND

A fascinating reality of the genome, receiving more and more empirical evidences, is that its biology is far more complex than previously thought. The rule "one gene has one DNA sequence leading to one mRNA translated into one protein", even if not (yet) an exception, is now well-known to be transgressed in a vast field of possibilities. Taking the example of the human genome, the number of

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes, the percentage of the genome that is transcribed, the alternative transcripts count per gene, or the way their expression is regulated, are all characteristics for which knowledge is moving with an extraordinary pace. The ENCODE project brought a lot of data and analyses in this line [1]. Among transformations that RNA transcripts undergo during maturation, RNA editing is a phenomenon leading to differences between the final RNA sequence and the DNA region it was transcribed from. The term was first used by Benne *et al* in 1986 [2], and can now be defined, in a broad sense, as a nucleotide insertion, deletion or substitution in the RNA sequence, occurring in various types of RNA, from tRNA to mRNA, either coding or not [3]. Substitutions comprise several types of modifications, the most common in vertebrates being the A-to-I conversion, catalyzed by the ADAR family enzymes (Adenosine Desaminase that Acts on RNA) [4] and leading to an A-to-G reading of the cDNA molecule [5, 6] and C-to-U conversion, catalyzed by the APOBEC enzyme [7, 8].

RNA editing is limited to eukaryotes, with a few exceptions (see [9] for review). It is observed in chloroplasts, widespread in mitochondria, and also found as a nuclear phenomenon in animals. It seems to have arisen through different mechanisms in different lineages, rather than being inherited from a common ancestor, and whether natural selection was involved in its evolution is still debated [9-11]. While RNA editing is more and more characterized in mammals, especially in human, mouse and rat [12-18], only a few studies have been performed in birds and were targeting specific genes. The apolipoprotein B (APOB) RNA editing, well-known in mammals, seems to be absent from chicken [19] and zebra finch [20]. In chicken, the CYFIP2 (cytoplasmic FMR1 interacting protein 2) and FLNA (filamin A) genes are edited in brain and liver [21], the splicing regulator NOVA1 (Neuro-Oncological Ventral Antigen 1) is edited in the brain [22], the $GABA_A$ (gamma-Aminobutyric Acid Type A) Receptor, alpha3 subunit (GABRA3) is edited in the brain and the retina [23, 24]. But no genome-wide study is available to really assess the extent of RNA editing in this species. High-throughput RNA sequencing actually allows performing a deeper transcriptome analysis than previous technologies, including RNA editing through a genome-wide approach [25]. This has been performed on several species, including human and mouse [12, 13, 15, 26, 27] but never in avian species. The number of editing sites (or detected as RDD : RNA-DNA Differences) observed in mammals strongly varies between studies, even on the same tissues of the same species, and an increasing number of analyses point the requirement of very careful bioinformatics procedures to limit technical artifacts [14, 15, 28-32].

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To improve the available knowledge about the extent of RNA editing in chicken, we chose an approach without *a priori* by using DNA and RNA sequencing on the same samples through Next Generation Sequencing (NGS) technology of whole embryos. Our results support the fact that RNA editing is not a frequent event in chicken, is mostly limited to the canonical A-to-I conversions, and shows strong tissue- and developmental-specificities.

RESULTS

Sequences analysis

DNA and RNA sequences were obtained from the same samples of chicken embryos. In average, 141,534,451 DNA reads and 65,302,559 RNA reads were aligned and analyzed for each embryo. The genome coverage reaches 93% for DNA reads and 22% for the RNA reads. A summary on DNA and RNA sequences aligned on Galgal4 chicken assembly is presented in Table 1.

Data filtering – biases detection

The first step was to detect RDD sites, *i.e.* positions homozygous in DNA and presenting an alternative sequence in RNA. To consider a position as potentially candidate, we fixed a minimum read-depth threshold of 15 both in DNA and RNA alignments for each embryo. We only kept candidates for which the alternative nucleotide frequency in DNA was null (Figure 1A). A total of 1,327 RDD sites met this criterion. The next filtering steps are aiming to avoid common pitfalls in sequences analysis, in order to decrease the number of putative false positive RDD candidates (Figure 1). To increase the robustness of the results and avoid putative false positive due to an artifact present only in one sample, we only considered RDD sites detected in at least 2 biological replicates. We ended up with 324 RDD sites (Figure1B).

It has previously been shown that polymorphisms overrepresented in read extremities are likely to be false positives [33-35]. In order to avoid this bias, we only considered RDD sites in which the RDD allele was, in median, not in the 10% extremities of reads overlapping them (Figure 1C). Two additional filters related to sequencing were applied: we removed candidates with an over-representation of one allele on one strand and discarded positions where more than one alternative nucleotide was found in proportions superior to 5% (Figure 1C). A total of 112 RDD sites passed all

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

filters. We then removed candidates in splicing sites from non-coding regions (Figure 1D), and filtered for regions containing homopolymers (Figure 1E). A total of 84 candidates remained. We applied a last filter by removing candidates harboring the “edited” pattern in the genomic DNA reads (Figure 1F). The goal was here to take into account putative candidate regions for which the corresponding DNA reads were present, but unmapped or not mapped to the same position as the "edited" RNA reads.
At the end of the analysis, we found 36 reliable RDD candidates (Table 2). A total of 17 chicken genes are potentially impacted by these RDD sites, knowing that one site can be associated with several genes and that we are probably missing non-annotated genes for candidates highlighted in intergenic regions. Interestingly, many of these candidates were organized in clusters, the 36 positions corresponding to 20 different genomic regions (Table 2). A total of 7 clusters, in 5 annotated genes and in 2 intergenic regions, could be counted up, encompassing 12 to 1439 bp. The distance between 2 clustered RDDs ranged from 3 to 807 bp, for a number of detected sites comprised between 2 and 5.

RDD types

We distinguished canonical RDD (A-to-G and C-to-T) from non-canonical RDD (other base changes). As the sequencing process was not strand-specific, the complement bases of canonical changes were also considered as canonical (*i.e.* T-to-C and G-to-A).
When comparing our datasets before and after filtering, we observed a clear enrichment in canonical changes throughout successive filters, which was quite reassuring in terms of results accuracy (Figure 2). Before filtering, all possible base changes were represented, at a frequency ranging from 5 % to 20% (Figure 2). Altogether, canonical base changes represented 50% of RDD candidates. After filtering, canonical base changes represented all modifications except one, at position chr6: 29787642. This non-canonical A-to-C position seemed to be the result of a misalignment involving an alternative splice-site. This position was selected for pyrosequencing validations.
Among the canonical modifications, we found only A-to-G or its complement T-to-C modifications, and no C-to-T conversion.
We then characterized the RDD candidates with regards to their putative functional features.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Functional RDD and tissue expression

Three RDD sites, located on CYFIP2, GRIA2 and COG3, were potentially functional, because leading to a non-synonymous change, and thus potentially having deleterious effects on the encoded protein. Most of the remaining candidates are located in gene introns, upstream or downstream regions of genes (Table 2, Figure 3).

By using 5 different *in silico* predictors of the amino-acids substitutions putative effects, we showed that none of the 3 non-synonymous substitutions was likely to be deleterious (Table 3). These substitutions were localized in highly conserved regions of the proteins (Additional file 2). A striking observation is that the K/E editing site affecting the CYFIP2 gene changes an amino-acid conserved between all examined Vertebrate species into an amino-acid which is coded without editing by the genomic sequence of Ray-finned fishes.

Characterization of candidates

We designed primers for 14 RDD candidates corresponding to 9 genomic regions, comprising missense variants, intron, upstream or downstream regions, intergenic position, and the remaining non-canonical modification. We first confirmed the homozygous status of the 14 selected RDD sites on DNA by Sanger sequencing. Their RDD status was then tested by pyrosequencing, and 13 RDD candidates were confirmed as edited loci (Figure 4). It is interesting to note that the unique site not validated by pyrosequencing corresponds to the non-canonical RDD candidate. A subset of 7 validated candidates was then tested in the other available tissues: individual heart, brain and liver tissues from three developmental times, comprising the same stage as the original HiSeq samples (day 4.5), an older embryonic stage (day 15), and an adult stage (11 months of age). Among these candidates, three are clustered on chromosome 13 (Figure5B.abc) and two are clustered on chromosome 2 (Figure 5C.ab). These positions were tested for tissue and stage effects on editing levels (Table 4). Tissue effect and stage effect were significant for all candidates (p-value≤0.05), and an interaction between tissue and stage was also observed for all but one candidates. There was a clear effect of both tissue and stage on the editing level. Interestingly, for 5 candidates out of 7, there was a continuous increase in editing level with age, from about 50% to more than 80%, independently of the tested tissue (Figure 5). In both clustered regions (Figure 5BC), all candidates harbored the same profile and only differed by their editing level. For one candidate, chr1: 167109833 (Figure 5Aa),

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the editing level was increasing during embryonic development and was less important in adult stage. On chr13: 10717577 (Figure5A.b), editing was mainly present in brain, with a level increase with time, and really low in other tissues. Interestingly, the editing level was tissue-specific, at every developmental time point, increasing for most of the candidates from liver to brain (Figure 5).

DISCUSSION

Among animals, RNA editing is well described in mammals, but similar studies were lacking in other vertebrates like chicken. The goals of this study were to screen the entire chicken transcriptome for editing sites, and to characterize this phenomenon at different stages of development and tissues to extend the analysis of its conservation among vertebrates.

To do so, we used DNA-Seq and RNA-Seq technologies, allowing us to screen the whole chicken genome for such events. This approach was used recently in several species to detect RDD [12-15, 26, 27, 36-38]. While a large number of new RDD sites was first described using this approach, in particular in humans with more than 10,000 sites observed [37], these results were then contested [28, 29], showing that RNA editing would be a limited process when taking into account possible high-throughput sequencing technologies biases. Later studies confirmed this questioning by finding much less RDD sites when stringently filtering the dataset [15, 29]. For example, Pickrell and colleagues demonstrated that up to 94% of the 10,210 edited sites highlighted by Li and collaborators [37] were likely to be false positive.

We carefully looked at common analysis pitfalls when detecting RDD sites in our dataset. First, we applied a stringent filter by taking into account only RDD sites observed in at least 2 biological replicates. This filter ensures to keep true biological phenomena, and to remove candidates due to individual-dependent artefacts (as specific sequencing errors or somatic mutations putatively not seen in DNA). As there is an over-representation of mis-called SNP in read extremities [33-35], each RDD site with a biased distribution of the alternative nucleotide towards the extremities of the reads was discarded from the analyses. In accordance with previous studies [28], we chose to consider only the distribution of the “edited” nucleotide position, to increase the stringency of the method: for candidates with small RDD levels, when considering both nucleotides in the filter, the candidate can be falsely declared as unbiased and kept.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Only sites with a balanced proportion of alternative nucleotides in forward and reverse strands were kept. At this step of analysis, we found 112 RDD candidates (Figure 1ABC). Another study filtering its datasets for the same biases, with a more stringent filter concerning the number of biological replicates (at least two thirds of replicates detected as edited) [15], found between 128 (in mouse liver) and 447 (in mouse adipose) RDD candidates at the same filtering step, *i.e.* more candidates than our results while their study was limited to the exome. It could constitute an argument in favor of the scarcity of RNA editing in chicken. We chose to be really stringent by keeping only RDD positions with a total absence of the alternative nucleotide on DNA. It appears that the SAMtools mpileup SNP detection software declared homozygous DNA positions where we could find the alternative nucleotide harbored by several reads. We are aware of the possible loss of real candidates, but the aim was here to maximize the reliability of the results. The final step, eliminating candidates for which the edited pattern was found in the DNA reads, removed 48 candidates, even if they passed our stringent filters. At the end of the filtering steps, we kept less than 10% of putative RDD sites detected at the beginning of our study, which is similar to results obtained in recent studies, taking into account biases linked to high-throughtput sequencing [15, 17].

Compared with previous studies, a distinguishable feature of our analysis is the search of the "edited" pattern in the DNA reads of candidates highlighted through RNA-Seq. In several cases, while the “edited” RNA reads map to a candidate region, the corresponding DNA reads do not map to the chicken genome: the RNA read can be mapped to a paralogous region due to the splitting of introns, while the original region is either absent from the genome or is carrying too many mismatches between our individuals and the reference sequence. This can be explained by an incomplete genome assembly and / or several regions with assembly errors. Indeed, the chicken genome assembly is still incomplete, especially regarding microchromosomes [39, 40]. The false RDD status of many candidates due to DNA polymorphism in paralogous regions has already been highlighted [14]. A similar observation has been made by Piskol *et al* [30], leading to the conclusion that non-canonical editing site are likely to be false positive RDDs. Previous studies successfully detected a high number of edited sites in human [13, 36], mainly in Alu sequences. These human repeated sequences clearly show a high propensity to harbor A-to-I editing sites, which is a strong argument in favor of their true existence, and the human genome assembly is more complete than the chicken one (but with much more repeated regions). But given the results obtained in our study, we confirm that a part of the

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

putative edited sites obtained from RNA-Seq data may be removed when taking into account not only the DNA sequence from the same sample or the biases previously underlined as the position in the reads or the strand bias, but also the putative alignment problems due to the genome assembly or to the discontinuous nature of RNA-Seq data.

At the end of the filtering steps, the number of RDD candidates was considerably reduced. Our filters were quite stringent, and we may have missed a few real positions. But as we still detected a false positive candidate through an experimental validation, this high stringency was surely appropriate.

We could also have missed true candidates because of the alignment stringency: if some regions are extensively edited, the resulting RNA sequence becomes really different from its DNA matrix. As a consequence, reads sequenced from edited RNA carry too many mismatches to be kept in the alignment [9, 41], except when using appropriate computational methods [42]. But given the small extent of RDD in chicken, this hypothesis is not likely.

Nevertheless, this low number of candidates tends to suggest that, as in other non-primate animals, RNA editing is a limited event in chicken [43].

The proportion of canonical RDD changes increased across filters, which is reassuring about the reliability of the pipeline: only one non-canonical change could be observed, shown not to be a true conversion, due to misalignments along an alternative splice site. The status of this false candidate was confirmed by pyrosequencing.

Interestingly, no C-to-T conversion was observed in our dataset. In particular, confirming previous studies on RNA editing in chicken, we could not find any editing in APOB transcripts [19]. This is in accordance with the missing of APOBEC1 from the chicken genome, as this enzyme seems to be required for C-to-U APOB RNA editing in vertebrates [5].

After the detection of RDD sites in the chicken transcriptome, our aim was to further characterize several interesting edited positions.

All the tested candidates were shared between the studied tissues in our analysis, with one candidate presenting a very low level of editing in the liver, whatever the stage (chr13: 10717577, Figure 5Ab). But the edition level varies between the analyzed tissues and ages, confirming in our chicken model that RNA editing varies across times and tissues.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Interestingly, RNA editing levels change over time. As generally observed in mammals, with few exceptions [44-47], the A-to-I editing level increased during development, as it is the case for 6 out of 7 tested candidates. Even if the tissue and stage specificity of edition is clear in candidates tested from cluster regions (Figure5BC), it is even more pronounced for tested candidates highlighted separately (Figure5A). These time- and tissue-specific phenomena are not only due to the level of expression of ADARs [46, 48] and more work is needed to decipher the spatio-temporal regulation of RNA editing. The low level of edition at embryonic stages in almost all the tested candidates could be explained by a putative importance for adequate embryologic development, as it was hypothesized for the GRIA2 Q/R site in mammals [46], even if it has to be confirmed.

As previously highlighted, our results confirm that editing at a particular position often comes with editing sites nearby, but no clear functional explanation has been proposed yet [46]. The regional sequence composition and RNA molecule tertiary structure seem to be involved in these clustered editing sites [49, 50].

One interesting result is that only a few candidates were directly affecting the protein sequence by changing an amino-acid. It has been shown that RNA editing can impact protein function, like modifying ions channels in some tissues [51, 52], or impacting the ligand-binding affinity [53]. Nevertheless, our results show that RNA editing in chicken is more frequently silent, as already observed [54]. More studies should be performed to confirm these results. But a significant number of candidates are located in non-coding parts of the chicken genome, at least given the current state of the annotation. As in a previous study in human Alu regions [49], we observed a high number of edited sites in introns. Even if our data come from polyA+ RNAs, these sites may correspond to editing in pre-mRNAs. But they may be part of non-coding RNAs too, where editing has been discovered in several species, and the biological significance of which is still largely unknown [54, 55].

We highlighted 3 candidates that were previously described as edited in mammals, one K/E substitution already observed in the CYPIF2 gene [21], one I/V conversion located in the COG3 gene [16, 17, 56] - previously described in human, mouse and rat - and the R/G site in the GRIA2 gene [57], which means that these editing events are not restricted to mammals and appeared before the Sauropsid-Synapsid divergence. Possible implications of an altered editing efficiency at the R/G site in GRIA2 in mental disorders in human and mouse were recently observed [58]. Concerning the COG3 editing site, no functional implication is documented at this time, but as underlined in another study

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[16], the conservation of this site both in mammals and in a broader way in vertebrates implies a putative functional role. Similarly, the functional significance of the conserved CYFIP2 K/E editing, which is higher in brain than in other tissues in human, is not known, but may be implicated in apoptosis [45].

The editing sites are located in highly conserved regions between vertebrates (Additional file 2). Interestingly, the modification observed in the CYFIP2 gene results in a conversion from a Glutamic acid to a Lysine. This amino-acid is only present in the Ray-finned fishes, and shared by all of them (http://www.ensembl.org/index.html). The other species for which the homologous CYFIP2 sequence is available are all K-coding at this position, which asks the question of the functionality of this E residue, only present in fishes as a chromosomal codon, but resulting from an editing phenomenon in several Vertebrate species, including chicken.

The very small number of conserved edited sites between species has already been underlined [59] and may be the signature of their functional importance.

CONCLUSIONS

This study constitutes, to our knowledge, the first whole genome screening of RNA editing in chicken. By using a stringent pipeline, we focused on really reliable RNA editing events and thus removed most putative false positives, a big pitfall in RNA editing discovery through high-throughput sequencing. Our pipeline predicts reliable RNA editing site; most of the tested sites are confirmed through an independent validation method, avoiding biases encountered when using NGS data. RNA editing seems to be a very limited phenomenon in chicken, at least in whole embryo at 4.5 days of age, as attested by a whole genome screening through RNA-Seq. This whole genome analysis shows that the A-to-I editing mechanism may be the only one present in chicken. Several edited loci are conserved between chicken and other vertebrate species, including human, which indicates that, while RNA editing arose long ago in the evolution, some particular nucleotides from a few genes are subject to RNA editing. This conservation is probably linked to the molecular mechanisms involved, but more deeply questions the functionality of editing at these specific loci. Even if the spreading of RNA editing is more and more characterized, a huge effort to discover the putative functionality of this phenomenon is still needed.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Tissues dataset

The material used in this study for the embryo sequences dataset was previously described [60] [SRA study accession number: SRP033603]. Briefly, two chicken lines were crossed, Line 6 [61] and Line R^- [62]. Chickens were bred at INRA, UE1295 Pôle d'Expérimentation Avicole de Tours, F-37380 Nouzilly in accordance with European Union Guidelines for animal care. Twelve F1 were produced from 2 families: 8 embryos (embryonic day 4.5) and 4 adults from the same batch. Embryos were kept as a whole, while 3 adult tissues were harvested: brain, heart, and liver. Additional embryos were produced at embryonic days 4.5 (n=8) and 15 (n=8), from a cross between the same lines, and 3 embryonic tissues were harvested: brain, heart and liver. Genomic DNA and total RNA were concurrently extracted from the same samples of crushed whole embryos or individual tissues (AllPrep DNA/RNA Mini Kit, Qiagen). RNA quality was measured by a BioAnalyzer (Agilent); all samples had a RIN (RNA Integrity Number) ≥ 9.9.

Sequencing

RNA sequencing

Libraries with a mean insert size of 200bp were prepared following Illumina instruction for RNA-Seq analysis, by selecting polyA+ fragments (TruSeq RNA Sample Prep Kit) from each sample. Samples were tagged to allow subsequent identification, amplified by PCR and quantified by qPCR (Agilent QPCR Library Quantification Kit).

A total of 8 embyo libraries were sequenced (paired-ends, 100 bp) in triplicate on an Illumina HiSeq 2000 sequencer (Illumina, TruSeq PE Cluster Kit v3, cBot and TruSeq SBS Kit v3) by randomizing their position in 6 different sequencing lanes.

DNA sequencing

DNA from 8 embryos was sequenced on 5 lanes of Illumina HiSeq 2000. Library preparation (mean insert size 328bp), DNA quantification and sequencing (paired-ends, 100bp) were performed according to the manufacturer instructions (TruSeq DNA Sample Prep Kit Illumina, Agilent QPCR Library Quantification Kit, TruSeq PE Cluster Kit v3 cBot TruSeq SBS Kit v3).

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Computational analyses

When not specified, analyses were performed with homemade Perl and R scripts.

Genomic sequences analyses

Sequences were aligned to the current chicken genome assembly (Gallus gallus 4) using the BWA program version 0.7.0, option aln [63]. Sequences were then filtered on mapping quality (MAPQ≥30). SAMtools rmdup command was used to remove possible PCR duplicates.

PolyA RNA sequences analysis

Sequences were aligned with Tophat software version 2.0.5 on the chicken reference genome Galgal4 as described in [60].

Sequences mapping uniquely on the reference genome, without PCR duplicates and with a minimum mapping quality of 30, were selected.

Identification of RNA/DNA differences

Sequences were locally realigned and recalibrated before SNP detection with GATK software version 1.6.11 and BamUtil (bam recab command).

SAMtools software version 0.1.19 was used with mpileup utility to detect SNPs between DNA and RNA samples from each individual. We set a maximum coverage of 10,000 for each calling to take into account as many reads as possible in the calling. SNPs were detected independently on each biological replicate.

Editing detection

SNPs were analyzed from VCF files obtained from SAMtools mpileup detection. For each biological replicate, only variations where DNA was homozygous either for the reference allele or for the alternative allele, and where RNA was heterozygous, were kept.

Several successive filters were applied to consider a position as a putatively RDD site. We first only considered positions with a sufficient depth, keeping only candidates presenting a minimum of 15 reads both in DNA and RNA alignments. To increase the likelihood of a site to be a true RDD position

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

by avoiding a sample artefact, we set to 2 the number of biological replicates that must carry the same modification.

We then applied several filters inherent to technical bias due to high-throughput sequencing. Positional bias was checked and all RDD candidates for which the median position of the "edited" allele among reads overlapping them was in the 10 first or 10 last bases were discarded. The strand bias was also considered; to be kept, a RDD candidate must present a proportion of edited allele on the forward strand really close to its proportion on reverse strand (delta≤0.5).

We checked the biallelical status of each selected candidate, a third allele being detected in less than 5% of cases being considered as a sequencing error.

An additional filter was applied to ensure that the alternative nucleotide frequency on DNA was null.

The functional consequence of each RDD in each transcript was predicted using the Ensembl Variant Effect Predictor (VEP) version 71 [64]. Non-coding splicing site regions were removed to take into account putative misaligned reads at these sites [31]. Then, positions belonging to homopolymers (n≥5) were removed because they may generate false positive candidates [13].

The chicken genome assembly still lacks several assembled regions, due to sequence assembly errors or missing fragments. A fragment detected as uniquely mapped may thus be present, with several polymorphisms, at genomic regions absent from the reference sequence, but present in the DNA reads from our samples. Therefore, a last filter was performed by searching the “editing site” (40 bp surrounding the candidate locus) in the DNA reads from samples thought to be edited. This pattern was searched with fuzznuc [65].

Validation assays and editing characterization

Sanger sequencing

We first checked the homozygous status of RDD sites by Sanger sequencing on DNA. The 8 biological replicates were tested. Primers were designed using PyroMark Assay Design software to allow further cDNA pyrosequencing (Additional file 1).

Pyrosequencing

RDD sites were tested on a Qiagen PyroMark Q24 sequencer. Primers were designed with PyroMark Assay Design software (Additional file 1). PCR products were made using PyroMark PCR Kit

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Qiagen). We performed the analyses through the PyroMark Q24 1.0.10 software with default analysis parameters.

Tissue and stage effects on the editing level were tested through an analysis of variance in a model taking into account tissues, stages and the interaction between tissues and stages for each tested candidate.

In Silico prediction of protein structure and function

To predict the putative effect of the editing conversions on protein structure and function, we used several bioinformatic tools: SIFT is based on sequence homology and the physical properties of amino acids (http://sift.bii.a-star.edu.sg/); PolyPhen2 uses physical and comparative considerations (http://genetics.bwh.harvard.edu/pph2/); MutationAssessor is based on evolutionary conservation of the affected amino acid in protein homologs (http://mutationassessor.org/); CHASM (computed using CRAVAT 3.0: http://www.cravat.us/) is based on the probability that a modification gives the cells a selective survival advantage; ProSMS predicts protein stability changes due to single amino acid modifications (http://babel.ucmp.umu.se/prosms/).

ACKNOWLEDGMENTS

We thank the entire staff of the PEAT experimental unit for their excellent animal care, and Juliette Riquet, Julie Demars, Gwenola Tosser, Annie Robic and Bertrand Servin for helpful discussions about the results presented in this study. Sequencing was performed at GeT-PlaGe Genotoul platform.

This paper is dedicated to the memory of André Bordas.

REFERENCES

1. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
2. Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC: **Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA.** *Cell* 1986, **46:**819-826.
3. Knoop V: **When you can't trust the DNA: RNA editing changes transcript sequences.** *Cell Mol Life Sci* 2011, **68:**567-586.
4. Bass BL: **RNA editing by adenosine deaminases that act on RNA.** *Annu Rev Biochem* 2002, **71:**817-846.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Gott JM, Emeson RB: **Functions and mechanisms of RNA editing.** *Annu Rev Genet* 2000, **34:**499-531.
6. Lee JH, Ang JK, Xiao X: **Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants.** *Rna* 2013, **19:**725-732.
7. Blanc V, Davidson NO: **C-to-U RNA Editing: Mechanisms Leading to Genetic Diversity.** *J Biol Chem* 2003, **278:**1395-1398.
8. Lau PP, Xiong WJ, Zhu HJ, Chen SH, Chan L: **Apolipoprotein B mRNA editing is an intranuclear event that occurs posttranscriptionally coincident with splicing and polyadenylation.** *J Biol Chem* 1991, **266:**20550-20554.
9. Gray MW: **Evolutionary origin of RNA editing.** *Biochemistry* 2012, **51:**5235-5242.
10. Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF: **Cell biology. Irremediable complexity?** *Science* 2010, **330:**920-921.
11. Speijer D: **Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity.** *Bioessays* 2011, **33:**344-349.
12. Park E, Williams B, Wold BJ, Mortazavi A: **RNA editing in the human ENCODE RNA-seq data.** *Genome Res* 2012, **22:**1626-1633.
13. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB: **Accurate identification of human Alu and non-Alu RNA editing sites.** *Nat Methods* 2012, **9:**579-581.
14. Schrider DR, Gout JF, Hahn MW: **Very few RNA and DNA sequence differences in the human transcriptome.** *PLoS One* 2011, **6:**e25842.
15. Lagarrigue S, Hormozdiari F, Martin LJ, Lecerf F, Hasin Y, Rau C, Hagopian R, Xiao Y, Yan J, Drake TA, Ghazalpour A, Eskin E, Lusis AJ: **Limited RNA editing in exons of mouse liver and adipose.** *Genetics* 2013, **193:**1107-1115.
16. Holmes AP, Wood SH, Merry BJ, de Magalhaes JP: **A-to-I RNA editing does not change with age in the healthy male rat brain.** *Biogerontology* 2013, **14:**395-400.
17. Danecek P, Nellaker C, McIntyre RE, Buendia-Buendia JE, Bumpstead S, Ponting CP, Flint J, Durbin R, Keane TM, Adams DJ: **High levels of RNA-editing site conservation amongst 15 laboratory mouse strains.** *Genome Biol* 2012, **13:**26.
18. Maas S, Godfried Sie CP, Stoev I, Dupuis DE, Latona J, Porman AM, Evans B, Rekawek P, Kluempers V, Mutter M, Gommans WM, Lopresti D: **Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing sites.** *Biochem Biophys Res Commun* 2011, **412:**407-412.
19. Teng B, Davidson NO: **Evolution of intestinal apolipoprotein B mRNA editing. Chicken apolipoprotein B mRNA is not edited, but chicken enterocytes contain in vitro editing enhancement factor(s).** *J Biol Chem* 1992, **267:**21265-21272.
20. Severi F, Chicca A, Conticello SG: **Analysis of reptilian APOBEC1 suggests that RNA editing may not be its ancestral function.** *Mol Biol Evol* 2011, **28:**1125-1129.
21. Levanon EY, Hallegger M, Kinar Y, Shemesh R, Djinovic-Carugo K, Rechavi G, Jantsch MF, Eisenberg E: **Evolutionarily conserved human targets of adenosine to inosine RNA editing.** *Nucleic Acids Res* 2005, **33:**1162-1168.
22. Irimia M, Denuc A, Ferran JL, Pernaute B, Puelles L, Roy SW, Garcia-Fernandez J, Marfany G: **Evolutionarily conserved A-to-I editing increases protein stability of the alternative splicing factor Nova1.** *RNA Biol* 2012, **9:**12-21.
23. Daniel C, Wahlstedt H, Ohlson J, Björk P, Öhman M: **Adenosine-to-Inosine RNA Editing Affects Trafficking of the γ-Aminobutyric Acid Type A (GABAA) Receptor.** *J Biol Chem* 2011, **286:**2031-2040.
24. Ring H, Boije H, Daniel C, Ohlson J, Ohman M, Hallbook F: **Increased A-to-I RNA editing of the transcript for GABAA receptor subunit alpha3 during chick retinal development.** *Vis Neurosci* 2010, **27:**149-157.
25. Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, Zhou H: **Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens.** *BMC Bioinformatics* 2011, **12 Suppl 10:**S5.
26. Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, Guo J, Dong Z, Bao L, Wang J: **Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome.** *Nat Biotechnol* 2012, **30:**253-260.
27. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X: **Accurate identification of A-to-I RNA editing in human by transcriptome sequencing.** *Genome Res* 2012, **22:**142-150.
28. Kleinman CL, Majewski J: **Comment on "Widespread RNA and DNA sequence differences in the human transcriptome".** *Science* 2012, **335:**1302; author reply 1302.
29. Pickrell JK, Gilad Y, Pritchard JK: **Comment on "Widespread RNA and DNA sequence differences in the human transcriptome".** *Science* 2012, **335:**1302; author reply 1302.
30. Piskol R, Peng Z, Wang J, Li JB: **Lack of evidence for existence of noncanonical RNA editing.** *Nat Biotechnol* 2013, **31:**19-20.
31. Lin W, Piskol R, Tan MH, Li JB: **Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome".** *Science* 2012, **335:**1302.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

32. Chen JY, Peng Z, Zhang R, Yang XZ, Tan BC, Fang H, Liu CJ, Shi M, Ye ZQ, Zhang YE, Deng M, Zhang X, Li CY: **RNA editome in rhesus macaque shaped by purifying selection.** *PLoS Genet* 2014, **10:**e1004274.
33. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming.** *Nucleic Acids Res* 2010, **38:**e131.
34. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: **Improving RNA-Seq expression estimates by correcting for fragment bias.** *Genome Biol* 2011, **12:**R22.
35. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36:**e105.
36. Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, Levanon EY: **A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes.** *Genome Res* 2014, **24:**365-376.
37. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG: **Widespread RNA and DNA sequence differences in the human transcriptome.** *Science* 2011, **333:**53-58.
38. Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G: **Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing.** *Nucleic Acids Res* 2010, **38:**4755-4767.
39. Groenen MA, Megens HJ, Zare Y, Warren WC, Hillier LW, Crooijmans RP, Vereijken A, Okimoto R, Muir WM, Cheng HH: **The development and characterization of a 60K SNP chip for chicken.** *BMC Genomics* 2011, **12:**274.
40. International Chicken Genome Sequencing C: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:**695-716.
41. Feagin JE, Abraham JM, Stuart K: **Extensive editing of the cytochrome c oxidase III transcript in Trypanosoma brucei.** *Cell* 1988, **53:**413-422.
42. Carmi S, Borukhov I, Levanon EY: **Identification of widespread ultra-edited human RNAs.** *PLoS Genet* 2011, **7:**e1002317.
43. Eisenberg E, Nemzer S, Kinar Y, Sorek R, Rechavi G, Levanon EY: **Is abundant A-to-I RNA editing primate-specific?** *Trends Genet* 2005, **21:**77-81.
44. Enstero M, Daniel C, Wahlstedt H, Major F, Ohman M: **Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA.** *Nucleic Acids Res* 2009, **37:**6916-6926.
45. Shtrichman R, Germanguz I, Mandel R, Ziskind A, Nahor I, Safran M, Osenberg S, Sherf O, Rechavi G, Itskovitz-Eldor J: **Altered A-to-I RNA editing in human embryogenesis.** *PLoS One* 2012, **7:**e41576.
46. Veno MT, Bramsen JB, Bendixen C, Panitz F, Holm IE, Ohman M, Kjems J: **Spatio-temporal regulation of ADAR editing during development in porcine neural tissues.** *RNA Biol* 2012, **9:**1054-1065.
47. Wahlstedt H, Daniel C, Enstero M, Ohman M: **Large-scale mRNA sequencing determines global regulation of RNA editing during brain development.** *Genome Res* 2009, **19:**978-986.
48. Garncarz W, Tariq A, Handl C, Pusch O, Jantsch MF: **A high-throughput screen to identify enhancers of ADAR-mediated RNA-editing.** *RNA Biol* 2013, **10:**192-204.
49. Athanasiadis A, Rich A, Maas S: **Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome.** *PLoS Biol* 2004, **2:**e391.
50. Tian N, Yang Y, Sachsenmaier N, Muggenhumer D, Bi J, Waldsich C, Jantsch MF, Jin Y: **A structural determinant required for RNA editing.** *Nucleic Acids Res* 2011, **39:**5669-5681.
51. Seeburg PH, Hartner J: **Regulation of ion channel/neurotransmitter receptor function by RNA editing.** *Curr Opin Neurobiol* 2003, **13:**279-283.
52. Song W, Liu Z, Tan J, Nomura Y, Dong K: **RNA Editing Generates Tissue-specific Sodium Channels with Distinct Gating Properties.** *J Biol Chem* 2004, **279:**32554-32561.
53. Wang Q, O'Brien PJ, Chen C-X, Cho D-SC, Murray JM, Nishikura K: **Altered G Protein-Coupling Functions of RNA Editing Isoform and Splicing Variant Serotonin2C Receptors.** *J Neurochem* 2000, **74:**1290-1300.
54. Nishikura K: **Functions and regulation of RNA editing by ADAR deaminases.** *Annu Rev Biochem* 2010, **79:**321-349.
55. Savva YA, Reenan RA: **Identification of evolutionarily meaningful information within the mammalian RNA editing landscape.** *Genome Biol* 2014, **15:**103.
56. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S: **Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution.** *Nature* 2009, **461:**809-813.
57. Yang JH, Sklar P, Axel R, Maniatis T: **Purification and characterization of a human RNA adenosine deaminase for glutamate receptor B pre-mRNA editing.** *Proc Natl Acad Sci U S A* 1997, **94:**4354-4359.
58. Kubota-Sakashita M, Iwamoto K, Bundo M, Kato T: **A role of ADAR2 and RNA editing of glutamate receptors in mood disorders and schizophrenia.** *Mol Brain* 2014, **7:**5.
59. Pinto Y, Cohen HY, Levanon EY: **Mammalian conserved ADAR targets comprise only a small fragment of the human editosome.** *Genome Biol* 2014, **15:**R5.
60. Frésard L, Leroux S, Servin B, Gourichon D, Dehais P, Cristobal MS, Marsaud N, Vignoles F, Bed'hom B, Coville J-L, Hormozdiari F, Beaumont C, Zerjal T, Vignal A, Morisson M, Lagarrigue S, Pitel F:

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transcriptome-wide investigation of genomic imprinting in chicken. *Nucleic Acids Res* 2014, **42:**3768-3782.

61. Bumstead N, Barrow PA: **Genetics of resistance to Salmonella typhimurium in newly hatched chicks.** *Br Poult Sci* 1988, **29:**521-529.
62. Bordas A, Tixier-Boichard M, Merat P: **Direct and correlated responses to divergent selection for residual food intake in Rhode Island Red laying hens.** *Br Poult Sci* 1992, **33:**741-754.
63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
64. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26:**2069-2070.
65. Olson SA: **EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite.** *Brief Bioinform* 2002, **3:**87-91.

FIGURE LEGENDS

Figure 1 Number of RDD candidates obtained after each filter

Figure 2 Proportion of base changes of RDD candidates before/after filters

Figure 3 Distribution of RDD sites in genomic features

Figure 4 Validation of candidates by Sanger sequencing (DNA) (red arrow) and pyrosequencing (cDNA) (grey). A: Example of a canonical RDD (T-to-C) at position chr2: 86000926. The sequence is in reverse-complement. The RDD status is confirmed by pyrosequencing (A: 53% - G: 47%)

B : Example of a non-canonical RDD (A-to-C) at position chr6: 29787642. The alternative nucleotide is not detected (A:100% - C: 0%).

Figure 5 Editing levels observed across tissues and time.

A: 2 selected candidates (a: chr1: 167109833; b: chr13: 10717577). B: Cluster 1 candidates (a: chr13: 931 843; b: chr13: 931855; c: chr13: 931888). C: Cluster 2 candidates (b: chr2: 86000926; c: chr2: 86001370). On abscissa axis: 1: Embryo stage 4.5 days – Brain, 2: Embryo stage 15 days – Brain, 3: Adult 11 months – Brain, 4: Embryo stage 4.5 days – Heart, 5: Embryo stage 15 days – Heart, 6: Adult

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11 months – Heart, 7: Embryo stage 4.5 days – Liver, 8: Embryo stage 15 days – Liver, 9: Adult 11 months – Liver.

TABLES

Table 1 Number of analyzed RNA and DNA sequences in the study (after alignment on Galgal4)

	Embryo (n=8)
Mean total number of reads (DNA)	141 534 451
Mean total number of reads (RNA)	65 302 559
Total number of reads (DNA)	1 132 275 604
Total number of reads (RNA)	522 420 469
Mean coverage (DNA) - min coverage 5 reads (% genome)	93.1 ± 1.0
Mean coverage (RNA) - min coverage 5 reads (% genome)	22.3 ± 2.1
Mean coverage (RNA) - min coverage 15 reads (% genome)	16.4 ± 2.6

Table 2 RDD candidates after filtering steps. In bold, candidates tested for validation

Chromosome	Position	DNA Nucleotide	RDD Nucleotide	Gene Name	Gene Short Name	Consequence (VEP analysis)	Number Of Edited Replicates	Mean Depth	Mean Frequency Of Edited Nucleotide
1	36367200	A	G	ENSGALG00000027799	*TMEM19*	3_prime_UTR_variant	2	25±2.83	0.27±0.11
1	74991790	T	C	ENSGALG00000014342	*TEAD4*	downstream_gene_variant, intron_variant	7	27.14±9.44	0.38±0.1
1	74992334	T	C	ENSGALG00000014342	*TEAD4*	downstream_gene_variant, intron_variant	3	38.33±12.34	0.46±0.07
1	74992422	T	C	ENSGALG00000014342	*TEAD4*	downstream_gene_variant, intron_variant	3	30.33±6.66	0.3±0.06
1	74993229	T	C	ENSGALG00000014342	*TEAD4*	downstream_gene_variant, intron_variant	3	25.33±7.51	0.39±0.06
1	**167109833**	**A**	**G**	**ENSGALG00000016980**	***COG3***	**missense_variant**	**4**	**100.25±32.8**	**0.47±0.08**
2	**86000822**	**T**	**C**	**ENSGALG00000013191**	***NDUFS6***	**upstream_gene_variant**	**5**	**73.2±19.64**	**0.42±0.06**
2	86000881	T	C	ENSGALG00000013191	*NDUFS6*	upstream_gene_variant	2	114.5±36.06	0.21±0
2	**86000926**	**T**	**C**	**ENSGALG00000013191**	***NDUFS6***	**upstream_gene_variant**	**5**	**113.2±64.58**	**0.48±0.06**
2	**86001360**	**T**	**C**	**ENSGALG00000013191**	***NDUFS6***	**upstream_gene_variant**	**4**	**61.75±24.92**	**0.26±0.06**
2	**86001370**	**T**	**C**	**ENSGALG00000013191**	***NDUFS6***	**upstream_gene_variant**	**5**	**55.8±26.13**	**0.43±0.09**
2	110994632	A	G	-		intergenic_variant	3	29.33±3.79	0.67±0.11
3	**2384093**	**A**	**G**	**-**	**-**	**intergenic_variant**	**5**	**21.8±6.26**	**0.28±0.13**
3	**2384105**	**A**	**G**	**-**		**intergenic_variant**	**2**	**23±9.9**	**0.34±0.06**
3	38271681	T	C	ENSGALG00000011013	*PCNXL2*	splice_region_variant, synonymous_variant	2	40±16.97	0.87±0.01
4	17996546	T	C	-		intergenic_variant	3	24.33±7.09	0.27±0.04
4	17999411	T	C	ENSGALG00000009128	*novel gene*	downstream_gene_variant	3	36±14.42	0.3±0.07
4	17999509	T	C	ENSGALG00000009128	*novel gene*	downstream_gene_variant	5	33±14.82	0.26±0.05
4	**20956439**	**A**	**G**	**ENSGALG00000009405**	***GRIA2***	**missense_variant, splice_region_variant**	**5**	**48±53.30**	**0.42±0.09**
4	**73361456**	**A**	**G**	**ENSGALG00000014395**	***DHX15***	**downstream_gene_variant**	**4**	**25.5±3.87**	**0.39±0.16**
6	15730446	A	G	ENSGALG00000005243	*PPP3CB*	intron_variant	2	18.5±4.95	0.38±0.25
6	**29787642**	**A**	**C**	**ENSGALG00000009427**	***TIAR***	**intron_variant**	**4**	**49.5±14.84**	**0.29±0.12**
6	34848568	A	G	ENSGALG00000010517	*TCTN3*	downstream_gene_variant	3	34±7	0.2±0.04
12	2800528	T	C	ENSGALG00000003738 ENSGALG00000003799	*EMC3* *USP4*	downstream_gene_variant	3	29.33±10.97	0.55±0.19
12	2800601	T	C	ENSGALG00000003738 ENSGALG00000003799	*EMC3* *USP4*	downstream_gene_variant	2	32.5±2.12	0.22±0.01
13	**931843**	**T**	**C**	**ENSGALG00000000946**	***PFDN1***	**intron_variant**	**2**	**19.5±0.71**	**0.28±0.1**
13	**931855**	**T**	**C**	**ENSGALG00000000946**	***PFDN1***	**intron_variant**	**3**	**22.67±4.73**	**0.65±0.09**
13	**931888**	**T**	**C**	**ENSGALG00000000946**	***PFDN1***	**intron_variant**	**2**	**23.5±2.12**	**0.28±0.06**
13	**10717577**	**T**	**C**	**ENSGALG00000003818**	***CYFIP2***	**missense_variant**	**2**	**40.5±17.68**	**0.27±0.08**
17	4705	T	C	ENSGALG00000014171	*novel gene*	intron_variant	3	45±6.56	0.18±0.03
22	5274	T	C	-	-	intergenic_variant	2	16±0	0.7±0.18
Z	27752047	A	G	-	-	intergenic_variant	3	29±4.58	0.23±0.06
Z	27752050	A	G	-	-	intergenic_variant	3	28±4	0.84±0.05
Z	27752217	A	G	-	-	intergenic_variant	2	31±4.21	0.57±0.01
Z	30809802	T	C	-	-	intergenic_variant	2	23±7.1	0.25±0.05
Z	30815599	T	C	ENSGALG00000005846	*MPDZ*	downstream_gene_variant	2	21.5±4.95	0.33±0.16

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3 *In silico* prediction of functional consequence on edited variants

	Software				
Variant	SIFT	PolyPhen2	Mut. Ass.	CHASM*	ProSMS
COG3 I632V	tolerated	benign	neutral	0.40 (0.22)	could destabilize
CYFIP2 K320E	tolerated	benign	neutral	0.32 (0.29)	no effect
GRIA2 R764G	tolerated	benign	medium	0.17 (0.49)	no effect

* Computed using Cravat 3.0, functional score close to 1 means functional effect (score p-value)

Table 4 P-values from an analysis of variance for tissue and stage effect on editing frequency.

Positions of candidates tested on multiple tissues by pyrosequencing are shown.

Chromosome	Position	Tissue effect	Stage effect	Tissus/Stage Interaction
1	167 109 833	1.29E-19	1.63E-09	1.35E-07
2	86 000 926	1.63E-14	1.13E-21	4.79E-03
2	86 001 370	8.17E-05	2.39E-20	9.94E-02
13	10 717 577	2.08E-17	1.11E-04	2.32E-02
13	931 843	1.44E-16	1.67E-25	1.45E-06
13	931 855	4.02E-15	1.67E-17	0.14
13	931 888	4.57E-08	4.77E-24	0.02

bioRxiv preprint doi: https://doi.org/10.1101/008912; this version posted September 9, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ADDITIONAL FILES

Additional file 1 (Fresard_Table S1.xls):

Sequencing primers. [Btn] Biotin on 5' end for pyrosequencing only.

Additional file 2 (Fresard_Figure S1.tiff):

Alignment of protein sequence from different species.

Multi-species alignments were performed through the Muscle program in the PhyleasProg pipeline (phyleasprog.inra.fr), from reference protein sequences of fully sequenced genomes from Ensembl (www.ensembl.org).

The red arrows show the amino acid affected by the editing conversion. The overall conservation between all species is depicted under each multi-alignment. A. COG3 (I-->V) B. GRIA2 (R-->G) C. CYFIP2 (K-->E).

Figure 1

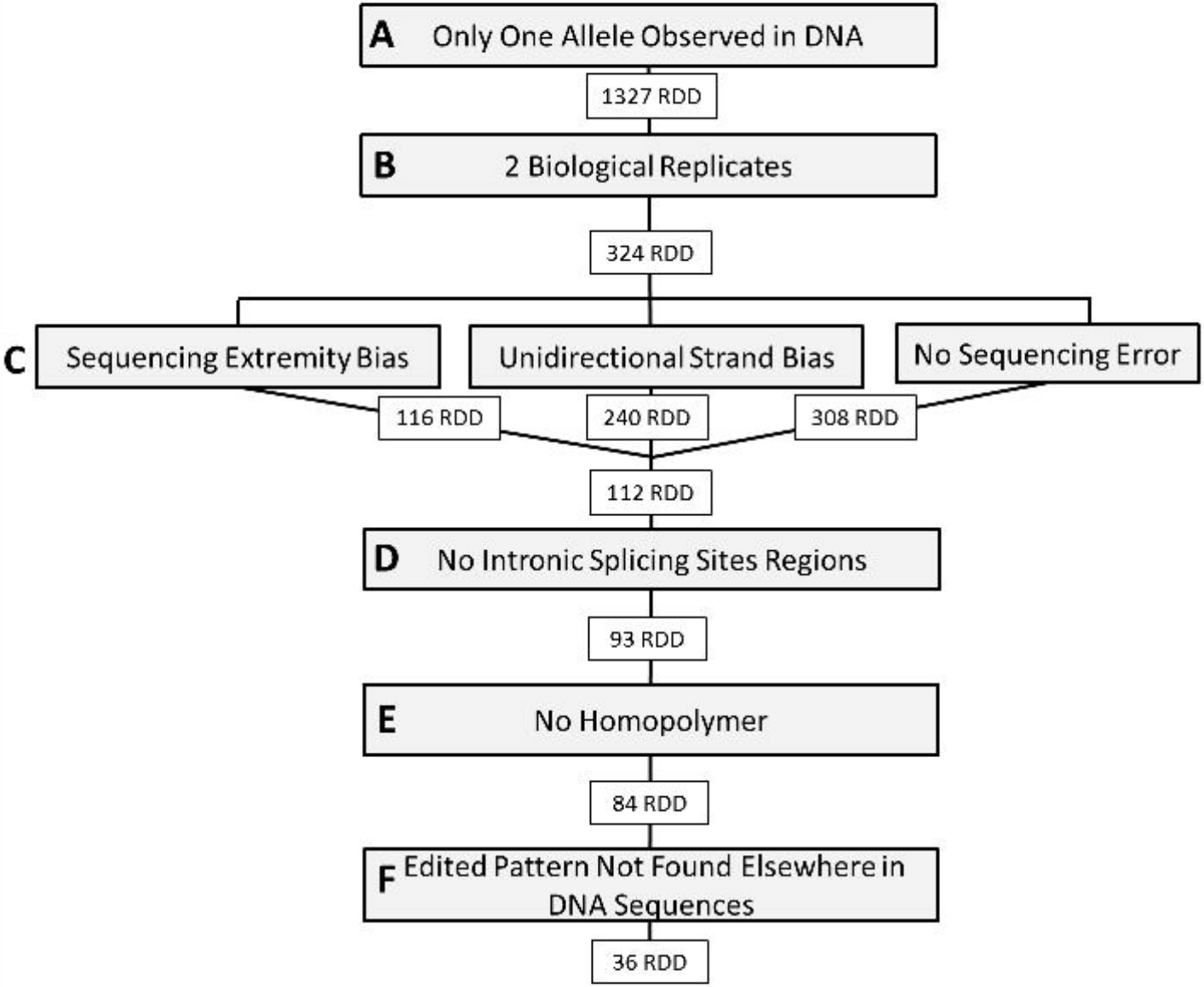

A Only One Allele Observed in DNA
1327 RDD
B 2 Biological Replicates
324 RDD
C Sequencing Extremity Bias
Unidirectional Strand Bias
No Sequencing Error
116 RDD
240 RDD
308 RDD
112 RDD
D No Intronic Splicing Sites Regions
93 RDD
E No Homopolymer
84 RDD
F Edited Pattern Not Found Elsewhere in DNA Sequences
36 RDD

Figure 2

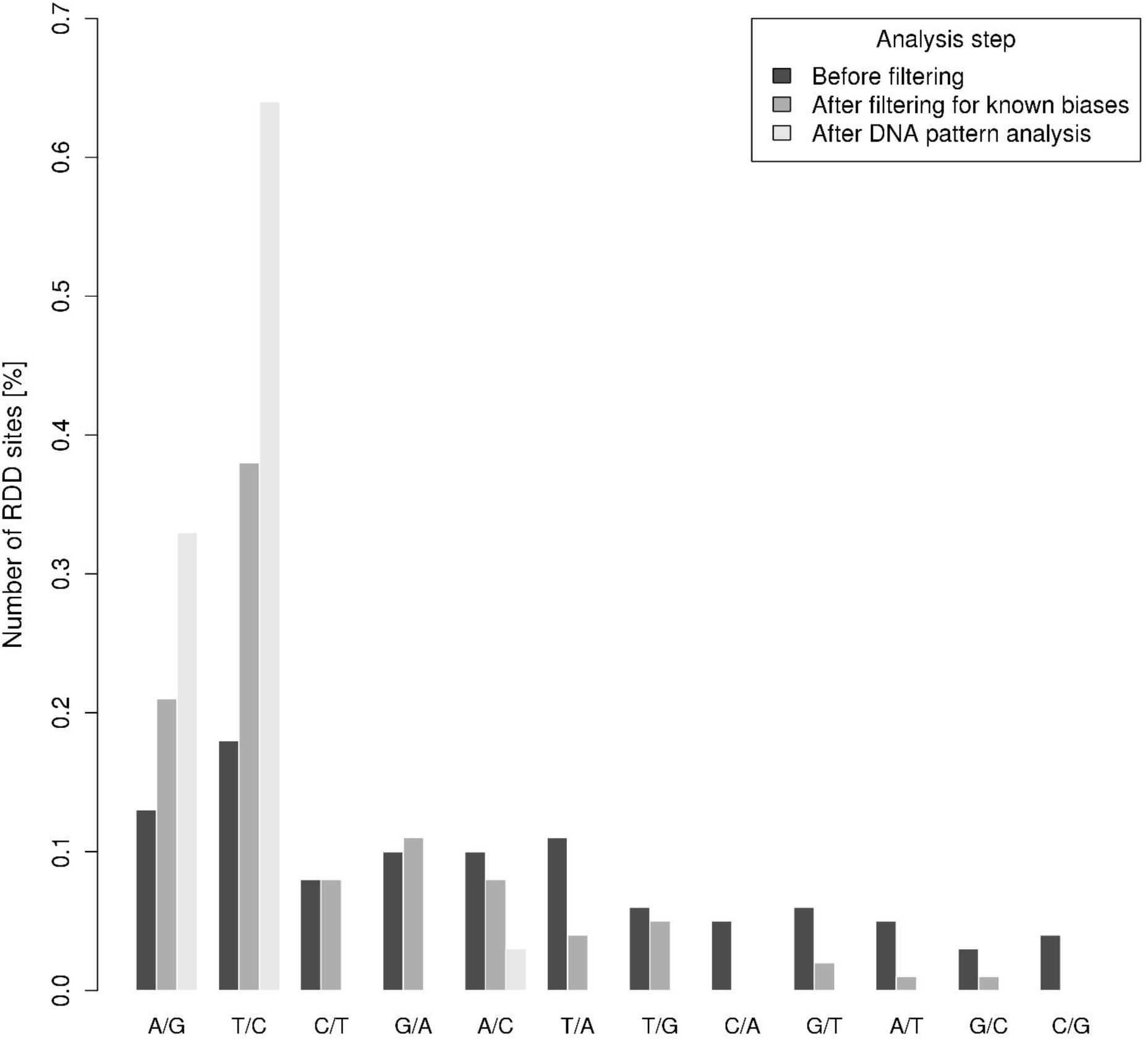

Analysis step
Before filtering
After filtering for known biases
After DNA pattern analysis
Number of RDD sites [%]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
A/G
T/C
C/T
G/A
A/C
T/A
T/G
C/A
G/T
A/T
G/C
C/G

Figure3

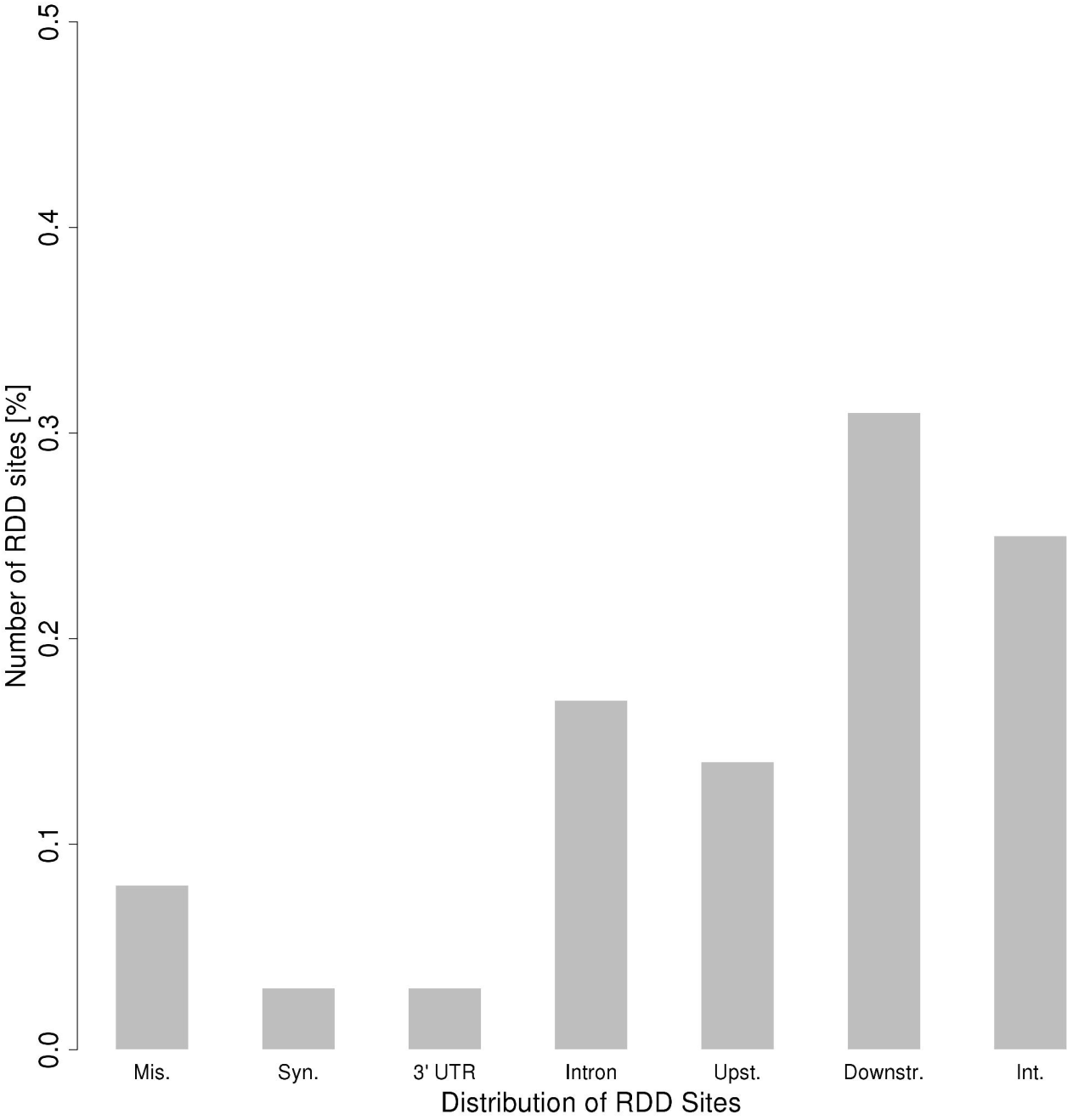
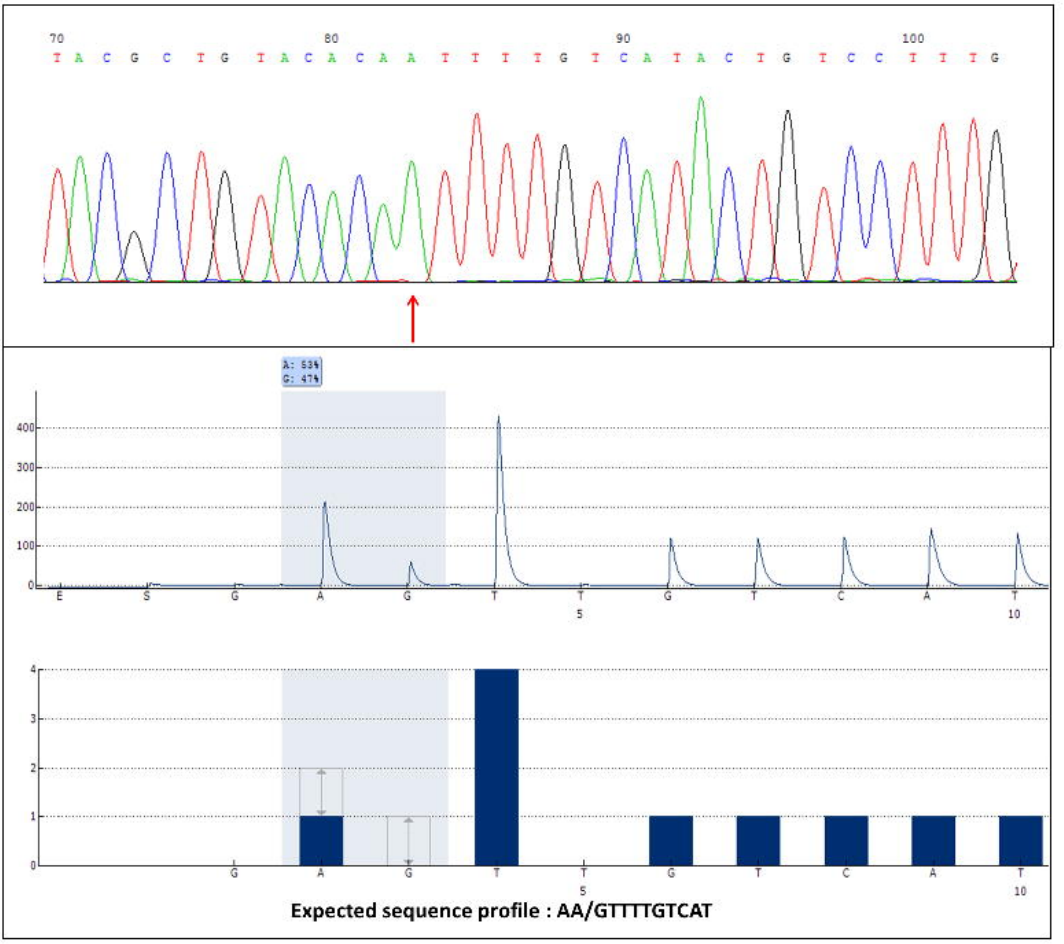
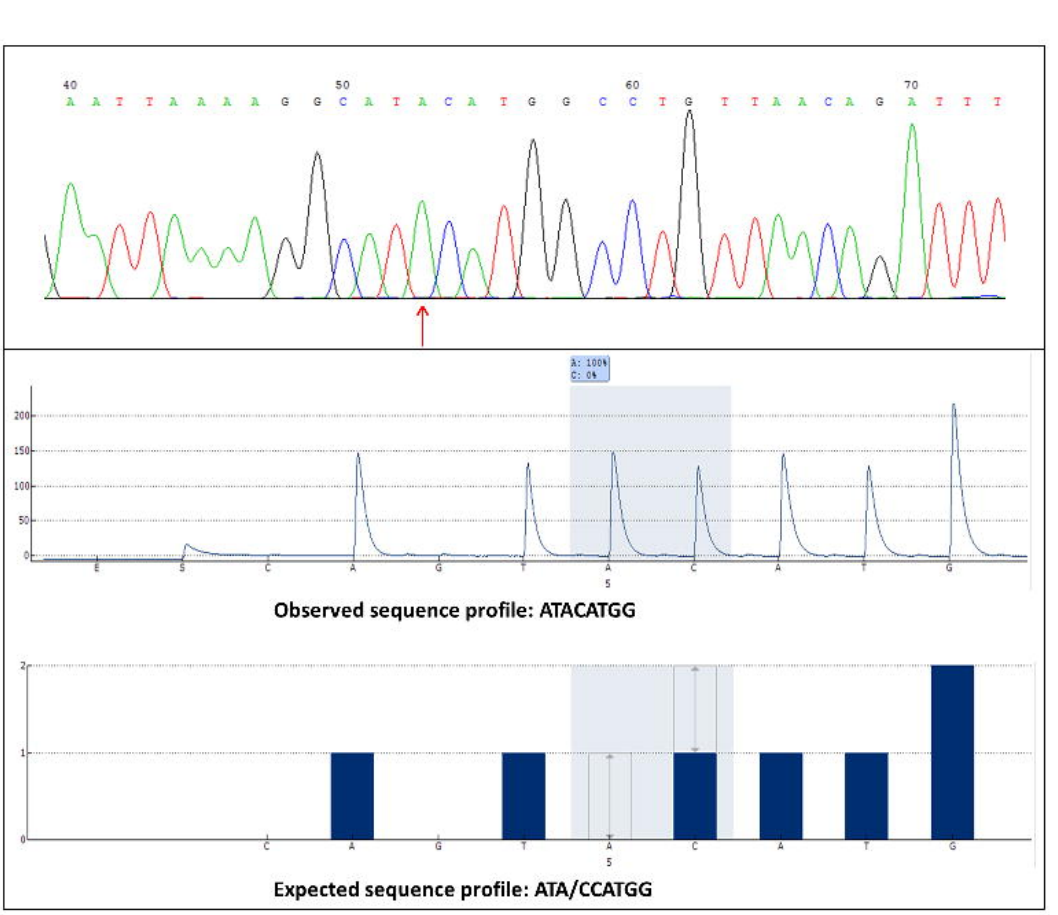

Number of RDD sites [%]
0.0
0.1
0.2
0.3
0.4
0.5
Mis.
Syn.
3' UTR
Intron
Upst.
Downstr.
Int.
Distribution of RDD Sites

Figure 4


A
DNA
cDNA
Expected sequence profile : AA/GTTTTGTCAT


B
DNA
cDNA
Observed sequence profile: ATACATGG
Expected sequence profile: ATA/CCATGG

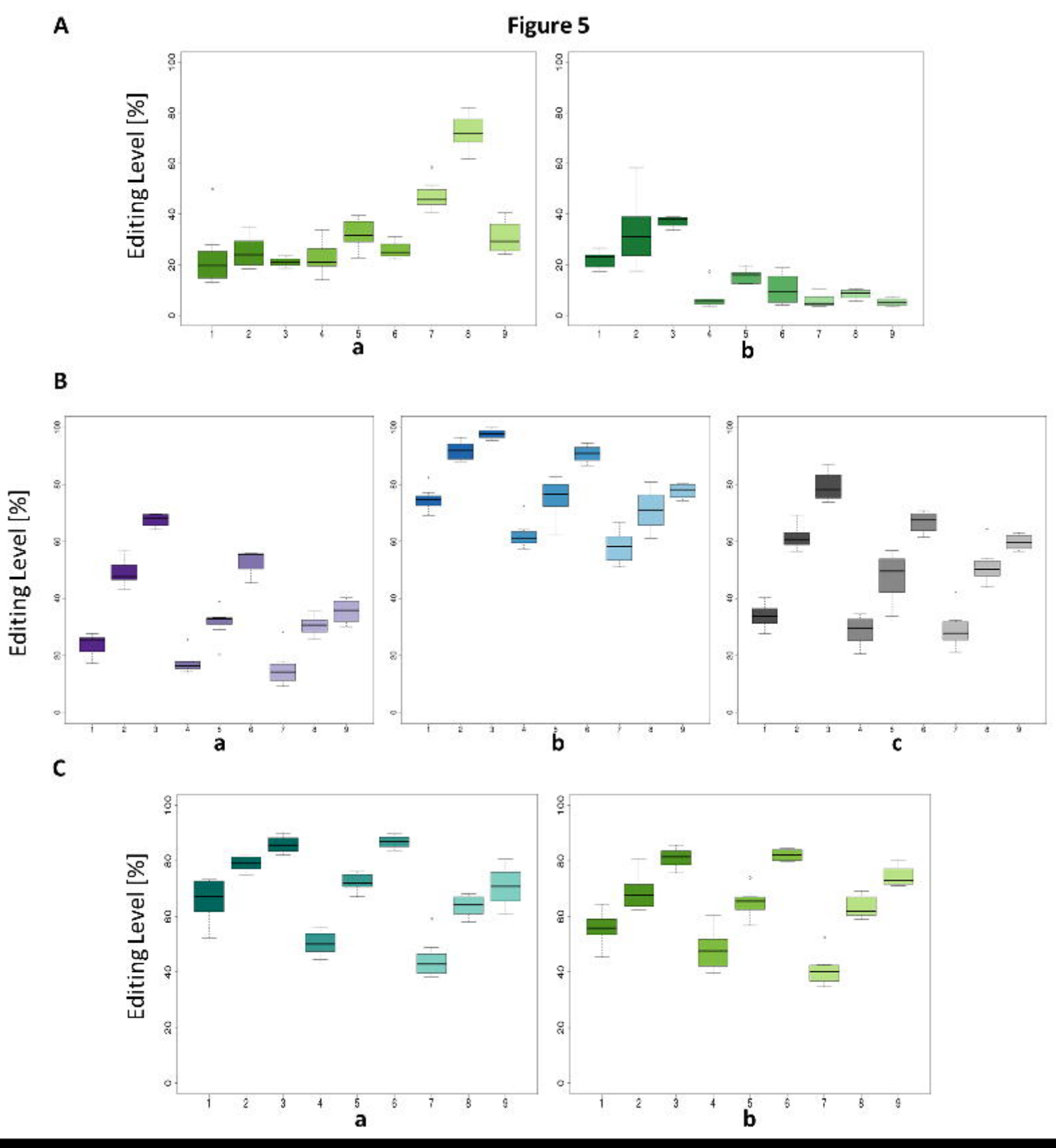



Figure 5