Conflict RNA modification, host-parasite co-evolution, and the origins of DNA and DNA-binding proteins1

Enzymatic RNA modification, host-virus conflicts and the origins of DNA and DNA-binding proteins

Paul J. McLaughlin [1] and Liam P. Keegan [2,3,4]

[1] Institute of Structural and Molecular Biology, Michael Swann Building, School of Biological Sciences, Kings Buildings [2] Department of Molecular Biology and Functional Genomics Stockholm University S-106 91 Stockholm, [3] Centre for Integrative Physiology, Hugh Robson Building, George Square, Edinburgh, EH8 9XDThe University of Edinburgh, Edinburgh, UK.

[4] Corresponding author

Telephone number ++ 44 131 467 8417
Fax number ++ 44 131 467 8456
Email addresses: Liam.Keegan@igmm.ed.ac.uk,
Character count: 50,345
Running title: ADARs and immunity

Abstract

Nearly one hundred and fifty different enzymatically modified forms of the four canonical residues in RNA have been identified. For instance, enzymes of the ADAR (adenosine deaminase acting on RNA), family convert adenosine residues to inosines in cellular dsRNAs. Recent findings show that DNA Endonuclease V enzymes have undergone an evolutionary transition from cleaving 3' to deoxyinosine in DNA and ssDNA to cleaving 3' to inosine in dsRNA and ssRNA in humans. Recent work on dsRNA-binding domains of ADARs and other proteins also shows that a degree of sequence-specificity is achieved by direct readout in the minor groove. However the level of sequence specificity observed is much less than that of DNA major groove-binding helix-turn-helix proteins.

We suggest that the evolution of more sequence-specific binding proteins following the opening up of the major groove by the RNA to DNA genome transitions represents the major advantage that DNA genomes have over RNA genomes. We propose that a hypothetic RNA modification, a ribose reductase acting on genomic dsRNA (RRAR) must have existed. We discuss why this is the most satisfactory explanation for the origin of DNA. The evolution of this RNA modification and later steps to DNA genomes are likely to have been driven by cellular conflicts with viruses. RNA modifications continue to be involved in host-virus conflicts; in vertebrates edited cellular dsRNAs with inosine-uracil base pairs appear to be recognized as self RNA and to suppress activation of innate immune sensors that detect viral dsRNA.

Introduction

ADAR (adenosine deaminase acting on RNA) enzymes deaminate adenosine bases to inosines in cellular dsRNAs (Keegan et al., 2004). ADARs evolved from adenosine deaminases acting on tRNAs (ADATs), that convert adenosine bases to inosine in tRNA anticodons. Inosine at position 34 in tRNA anticodons permits a wider range of possible base-pairings and facilitates wobble decoding at the third position in eight tRNA types in Eukaryotes and one tRNA in *E. coli*. ADARs evolved from ADATs partly by addition of dsRNA-binding domains. Inosine behaves like guanosine in Watson-Crick base pairing during cDNA synthesis and translation so that ADAR editing of dsRNA is detected as A to G base changes between genomic and cDNA sequences. Pre-mRNA editing by ADARs often leads to changes codon meaning and production of edited isoforms of ion channel subunits and other CNS proteins in vertebrates and especially in *Drosophila*.

Recent findings show that inosine-uracil (I-U) base pairs introduced by ADAR editing in vertebrate cellular dsRNA (Fig. 1 A), also assist innate immune discrimination between self and non-self RNAs (Karikó et al., 2005; Vitali and Scadden, 2010). Animal cells use a range of Pattern Recognition Receptors (PRRs) to distinguish viral RNAs from cellular RNAs, activating interferon signalling in response to viral RNAs (Takeuchi and Akira, 2010). The mechanisms by which modified bases such as inosine in cellular RNAs modulate the activities of antiviral PRRs is an important area for future work and more understanding of modified base recognition by proteins is required. Inosine is only one of approximately one hundred and fifty known enzymatic modifications in RNA (Machnicka et al., 2013). Many of these were first discovered in tRNAs and rRNAs but 5-meC and 6-meA have now been found to be widely distributed and present at specific locations in many mRNAs.

An evolutionary shift from DNA cleavage to RNA cleavage by Endonuclease V enzymes recognizing inosines or deoxyinosines

Endonuclease V in bacteria is a DNA repair enzyme that cleaves phosphpdiester bond 3' to deoxyinosines arising from deamination of adenosines. Endo nuclease V flips out the

deoxyinosine base into a recognition pocket(Figure 1 B) (Dalhus et al., 2009). Endo V retains the cleaved DNA product after the reaction probably to allow ordered recruitment of repair the strand cleavage and remove doexyinosine. However the human homolog of Endo V does not cleave DNA 3' of deoxyinosine. In fact, surprisingly, human Endo V cleaves dsRNA and ssRNA 3' to inosines (Morita et al., 2013; Vik et al., 2013). Human Endo V does also bind at deoxyinosine in DNA but, requires a ribose 2' hydroxyl on the adjacent base which is involved in the cleavage reaction, the DNA (Vik et al., 2013). It will be interesting to establish when Endo V specificity switched between dsRNA and DNA. The switch seems to be less complete in mouse Endo V than it is in human Endo V. Possibly the increased level of editing in primates compared to mice that is associated with dsRNA formed by pairing of highly sequence similar Alu repeats has helped to recruit a new ribonuclease to help clear these edited RNA. Bacterial Endo V also recognizes a wide range of base mismatches and aberrant structures in DNA. Human Endo V might also cleave dsRNAs having other modified bases or structural features.

The switch between DNA and RNA substrates is rather simple for Endo V proteins because they use the minor groove to recognize base mismatches. In dsRNA the movement of the phosphate strands away from one another is stopped by a steric clash in the minor groove between the ribose 2' hydroxyl and the ribose of the next base on the same strand (Fig. 2 A,B) (Dickerson et al., 1982). This steric clash brings the phosphate strands closer together over the major groove and makes the major groove narrower and much deeper in A-helix RNA than in the major groove in the more familiar B-helix of DNA. The B-DNA strand is narrower, the base edges in the major groove are brought closer to the surface and the major groove is wider (Fig 2 A,B). DNA is still able to adopt the A-helix conformation of dsRNA during transcription and during replication by DNA polymerases. In the case of Endo V nuclease the transition between DNA and dsRNA substrates may involve a recent change in preference towards dsRNA, however, during the evolution of life many minor groove binding proteins may have undergone a switch form recognizing dsRNA to recognizing DNA.

It has been proposed that ribonucleotide reductases arose only at a late stage in molecular evolution, as genome-encoded proteins. Ribozymes are unlikely to have carried out the reduction of ribose earlier because the catalytic mechanism involves radicals that are difficult to control and damaging to the nucleic acid of a ribozyme (Poole et al., 2002). In now clear also that binding to the minor groove in DNA is still the mechanism used by a surprising number of key genomic processes. If these minor groove processes evolved in RNA genomes then the sizes and coding capacities achieved by ancient RNA genomes may have been much greater than we have anticipated. When the structure of a DNA target complex of the major DNA mismatch-recognition protein MutS was determined, it was found that the protein does not use the major groove to recognize the aberrant base, even though the capacity to distinguish between correct and incorrect bases should be higher in the major groove (Fig. 1 A) (Seeman et al., 1976). Instead MutS interacts with the minor groove of DNA (Lamers et al., 2000). This observation of minor groove interaction has since been extended to some other DNA mismatch recognition proteins.

Minor groove dsRNA sequence recognition by ADAR double strand RNA binding domains (dsRBDs), versus major groove sequence recognition by helix-turn-helix proteins

Sequence specificity of double-stranded RNA-binding domain sequences binding long A-helical dsRNA was expected to be negligible. Specific adenosine recognition by ADAR dsRBDs was thought to depend mainly on the unique molecular shapes of the range of different RNA hairpins in which edited adenosine residues occur in edited pre-mRNAs.

However, Fredrick Allain and colleagues in Zurich now call attention to the fact that there is some sequence specificity in binding of ADAR2 dsRBDs to regular dsRNA (Masliah et al., 2013; Stefl et al., 2010; Stefl et al., 2006). The ADAR2 dsRBD2 helix α1 lies along the dsRNA minor groove and makes sequence-specific contacts (Fig. 2 A,B). Other dsRBDs such as ADAR2 dsRBD1 similarly exhibit some sequence specificity. In the minor groove the guanine amino group is the most distinctive base-specific feature and is particularly important for sequence specific recognition as seen in the dsRNA complex of the *A. aeolicus* RNaseIII dsRBD (Masliah et al., 2013). Sequence specific activities of RNaseIII on different substrates are strongly affected by positioning of guanine bases as positive and negative determinants. Multiple dsRDBs are present in many dsRNA-binding proteins and the two dsRBDs in ADAR2 combine to allow increased sequence specificity. ADARs also obtain significant further sequence specificity from the binding of the catalytic deaminase domain (Goodman et al., 2012), positioned to adjacent to dsRBD2 on dsRNA (Rice et al., 2012), which contacts the target adenosine and probably also some preferred flanking the adenosine.

The sequence specificity of minor groove binding by dsRDBs must be contrasted however with the much more impressive sequence specificity achieved by DNA-binding proteins. In the dsRDB-dsRNA complex the protein domain steps over the deep and inaccessible major groove, making contacts only with the phosphate backbones. The lambda repressor DNA-binding domain contains a version of the ancient helix-turn-helix (HTH) fold that is one of the oldest structures used by proteins to recognize specific DNA sequences. The second, recognition alpha helix of the helix-turn-helix fold in lambda repressor penetrates the major groove of the DNA in the complex with a lambda phage operator DNA (Fig. 2A,B right side), (Jordan and Pabo, 1988). A much larger number of sequence-specific contacts are made by amino acid side chains of the lambda repressor recognition helix with the bases in the major groove than ADAR2 dsRBD makes in the minor groove of dsRNA. This difference is due to intrinsic features of dsRNA and DNA. The major groove is superior to the minor groove in allowing base sequences to be discriminated by proteins making hydrogen bonds and other contacts with exocyclic amino and methyl groups (Fig. 1) (Seeman et al., 1976). Different base pairs are much less distinguishable from the minor groove sides of the bases. HTH motifs are present in ancient promoter-proximal DNA binding gene-specific regulatory protein such as proteins MYB that are also conserved between eukaryotes and archaea. Although the HTH motifs and the details of how they interact with the major groove differs between eukaryotic and bacterial gene-specific regulators the derivation of the DNA binding domains from common ancestors is not in doubt.

A modified RNA genome intermediate in the transition from RNA genomes to DNA genomes.

The major groove in dsRNA can be accessed by proteins when it is widened by unusual base pairing or by an extensive supporting folded RNA structure (Hermann and Westhof, 1999). Protein access to the major groove in a long genomic dsRNA (Steitz, 1993), would require a significant structure alteration that would be difficult to maintain in a replicating molecule. The helix-turn-helix (HTH) major groove-binding proteins had evolved already when eukaryotes and bacteria separated (Aravind et al., 2005), suggesting that DNA must have evolved already also. However, evolutionary comparison studies on DNA polymerases show a surprising diversity in the origins of these proteins, suggesting that complete DNA genomes appeared after the separation of bacteria from eukaryotes or perhaps even several times independently in prokaryotes and eukaryotes (Forterre, 2006; Forterre and Grosjean, 2000, 2009).

We suggest that DNA bases first arose as modified RNA bases at key points in genomic dsRNA and that full DNA genomes arose only much later. A hypothetical ribose

reductase acting on genomic dsRNA (RRAR) could have converted RNA bases to DNA bases in a dsRNA genome. This would have allowed the evolution of major groove-binding HTH-motif proteins to evolve before complete DNA genomes were formed. The local modification process envisaged for the RNA genome would have operated rather similarly to current DNA methylation

Most of the difficulty in explaining how the RNA to DNA genome transition came about is due to the need to base the explanation on the current actions of ribonucleotide reductase producing free monodeoxynucleotides. If DNA arose at a late evolutionary time in cells that were already complex, then the evolutionary search process that discovered DNA will need to have been carried out directly on the RNA genome itself. It is likely that many enzymatic modifications of genomic dsRNA were tested over the course of evolution till the one modification was discovered that offered a significant advantage. New enzymatic RNA modifications are likely to have appeared over the whole history of life and many may have disappeared again if they found no situation in which they offered a benefit.

To understand why studying RNA modification leads to the conclusion that DNA must have first evolved as an enzymatic modification of RNA it is helpful to consider the example of the queueosine (Q) base modification in tRNAs. There are two general types of mechanisms involved in enzymatic modifications of RNA. Most commonly, particular bases in a specific RNA molecule are modified, sometimes in serial multistep processes in which the chemical structure of this single base becomes more and more elaborated from the original canonical genome-encoded base. In a much less common RNA modification process, a complicated modified base such as queueosine (Q) is partly presynthesised off the target RNA. In this case free guanosine is put through a series of modifications to form a pre-queueosine base which. This is then substituted for a guanosine at position 34 in some specific tRNAs by a specific transglycosylase enzyme. Further modifications then occur on the pre-Q base after insertion into the tRNA transcript. Based on our knowledge of the range of postsynthetic enzymatic RNA modifications occurring at position 34 in different tRNA it is easy to explain the evolution of the substitutional queueosine modification mechanism. The initial evolution of the queueosine base was by the usual type of postsynthetic modification of tRNA transcripts at position 34. Evolutionary selection for improved accuracy of decoding by the final enzymatically modified tRNAs defined the required chemical structure for the modified base. The subsequent evolution of a specific transglycosylase allowed the early part of the base modification process to become a separate presynthetic process carried out on a single nucleoside off the RNA.

Let us now attempt to carry out a thought experiment in which we imagine that we know nothing about post-synthetic modifications in tRNAs. All we know is that a particular enzymatically modified form of guanosine is made off the tRNA and then inserted into a tRNA where it plays a critical role. How would we explain the evolution of queueosine in this case? This would be very difficult. We could not account for how either the structure of the queueosine base nor how the insertion into tRNAs specifically had ever managed to evolve. This thought experiment on the evolution of the queueosine base illustrates the same difficulty that we face in trying to explain the origin of DNA. We cannot satisfactorily explain the origin of DNA by starting from the current production of DNA mononucelotides by ribonucleotide reductases. That starting point leaves far too many questions unanswered. A lost or still undiscovered RNA modification that reduced ribose directly in dsRNA genomes or caused the loss of the 2' hydroxyl in some way is a more likely explanation for production of the first DNA bases.

Different RNA modifications that have been conserved in different RNA classes are very finely tailored to the functional requirements of these RNAs. Many of the modifications conserved in rRNAs or tRNAs provide increased stability or increase the fidelity of

translation. Ribosomal RNAs use an alternative RNA modification on the ribose 2' hydroxyl. If introduced in dsRNA it might reduce any tendency to hydrolysis of the RNA genome that was caused by the presence of the 2' hydroxyl on ribose (Poole et al., 2000). Most importantly however ribose 2' methylation does not remove the 2' hydroxyl. The steric clash that forces dsRNA into the A-helix conformation would still occur and formation of the B-helix as it does with deoxyribose. Ribose 2' methylation is useful to stabilise rRNAs but a deoxyribose is required to give the wide major groove of the DNA helix. Cells will have discovered this by an evolutionary search process. Postsynthetic ribose reduction in stable, folded or catalytic RNAs is likely to have been harmful in ribozymes, rRNAs or tRNAs. In genomic RNA however it immediately conferred some benefit. Viruses may have been first to evolve the ribose reduction to protect their genomes from host defences. Continuing conflicts with viruses are likely to have driven subsequent steps in DNA evolution also.

As to the identity of the original ribose reductase acting on genomic dsRNA (RRAR), it is worth considering the ribonucleotide reductases (RNRs) first, even if this only serves to outline requirements for this RNA modification activity. Current RNRs act on free ribonucleotide di- or tri-phosphates. Taking the closest possible analogy to the queueosine base discussed earlier one possible RRAR could have been an enzyme ancestral to the current Class I and II ribonucleotide reductases (RNRs) that acted on a ribose within an RNA strand. A larger substrate binding pocket would have been needed in the hypothetical RRAR ancestor than in the current RNRs. The C3 carbon of the target ribose is in the deepest part of the active site in Class I and II ribonucleotide reductases and more space will have been required for a continuous RNA strand to exit. The core of the active site would have been conserved in the subsequent conversion to RNRs. Current Class I and II RNRs may have filled in an original larger RNA substrate cavity in different ways, possibly with some changes in ancillary aspects of the radical handling processes. Class II RNRs have a large cobalamin cofactor now involved in the RNR reaction that is not in Class I RNRs. The 5' phosphate is not buried in the active sites nor contacted by protein in current RNRs. It might be possible to demonstrate the principle of post synthetic ribose reduction in RNA by engineering an RNR to contact this phosphate in an RNA strand and modify the ribose on the 3' base of an ssRNA or dsRNA oligonucleotide. In fact we do not know what sort of enzyme the RRAR may have been. There may be some entirely different way to reduce ribose in RNA. A possible problem is that in RNRs the C3 carbon loses a hydrogen temporarily during the reaction so it would need to be established whether this reaction mechanism would cause breakage of an RNA chain. A reaction mechanism that does not break an RNA chain would be better but so long as the products could be rejoined within the reactions this could be acceptable. The nature of the chemical reaction is not central to the argument we are making here. What is central is that the reaction that first made DNA must have been targeted to the already existing genomic dsRNA.

If a dsRNA modifying ribose reductase made the first DNA bases at some positions in an RNA genome then further steps to current DNA genomes are much simpler to imagine. The cellular RNA-dependent RNA polymerase will have adapted to use a mixed RNA/DNA template, but newly synthesised strands would still be RNA. To maintain the DNA modification through successive replications the RNA-modifying enzyme might have evolved to maintain modifications on newly synthesised, hemi-modified substrates as current genomic DNA methyltransferases do. Free DNA nucleotides would eventually have become available through eating other cells or by assimilation from the environment or by salvage within cells. Maintenance of the DNA modification might come to be assisted by patch repair DNA polymerases that evolved from RNA dependent RNA polymerases to incorporate salvaged deoxynucleotides nucleotides. RNRs will have evolved to increase the supply of deoxynucleotides.

Evolutionary selection driving the entire process would be related to the important role of modified nucleotides in cellular conflicts mostly with viruses and parasites. New cellular enzymes involved are likely to have been recruited at steps during the process from viruses that may have been first to evolve them. Conflicts with viruses affecting cellular genome replication or gene expression are likely to have led to the cellular genome being modified first at the critical places. The evolution of helix-turn-helix motif proteins specifically recognizing critical regions of the genome began the evolution of much more sequence-specific protein-based genome control sites. Until DNA major groove binding proteins such high levels of specificity in targeting sites in the RNA genome would probably have had to be based on protein-assisted searches for homology by complexes containing specific guide RNAs.

Greater genome stability or larger genomes may not be the key advantages of DNA.

We have argued here that even the successful production of deoxynucleotides at a later stage in evolution still leaves too many unanswered questions about how genomic DNA was first formed. Ribonucleotide reductases could even have evolved to produce deoxynucleotides for metabolic purposes without thereby leading to formation of DNA. However, because of the difficult chemistry involved in ribose reduction the hypothetical RRAR that was involved in making the first DNA bases is likely to have been as complex a protein as the RNRs. It is not plausible that cells could have had a large enough RNA genome or that they could have evolved with an RNA genome for long enough to produce such a protein if cells had not substantially solved problems associated with genomic RNA instability. Much of the thinking about the transition from RNA to DNA genomes has been based on the idea that DNA genomes are more stable than RNA genomes and that they would have needed to form as early as possible in evolution (Freeland et al., 1999) (Forterre, 2006; Forterre and Grosjean, 2000, 2009). The idea that susceptibility of RNA to hydrolysis *in vitro* would have prevented formation of large RNA genomes in cells does not accord with our present understanding of the biological stabilization of RNA molecules inside cells. Stabilities of RNA molecules are fully under cellular control. The RNA genome and especially separated RNA strands will also have been carefully associated with protective proteins.

The earliest cells are likely to have been thermophiles. Contrary to simple views, dsRNA is more robust in some ways than DNA. RNA has a higher melting temperature than DNA and this could be raised further by interactions with poly-cationic molecules such as polyamines and with proteins. Conversion of ribose to deoxyribose also severely weakens the base-sugar bond so that depurination is a major type of lesion in DNA that would not have posed a difficulty for RNA genomes (Lindahl, 1993). Current thermophilic archaebacteria with DNA genomes cope with depurination rates elevated more than 1000-fold over ambient temperature. However the damage due to higher temeperatures, plus potential further large increases in depurination due to mutagens, may have helped to made RNA genomes more suitable for early cells. Some of the DNA mismatch recognition proteins using minor groove binding may have been directly inherited from RNA genomes and they may have helped to sustain RNA genomes. Proofreading by DNA polymerases is another minor groove process. Their high-fidelity, replication capacity may have been inherited from ancestral RNA-dependent RNA polymerases (RdRPs) even though no currently known RdRP has a proofreading capacity. Interestingly, if repair of damage due to depurination evolved mainly after a later transition to DNA genomes then proteins involved in this type of repair may not have arisen directly from proteins previously involved in RNA genome repair. We might find that major groove interactions and different protein domains among DNA mismatch recognition proteins involved specifically in repairing depurination and base loss.

How could a large RNA genome have functioned?

Based on our argument about the importance of major groove recognition in DNA current archaeal genome organization is not similar to that in the last cells with RNA genomes. Operon formation by genes in DNA may only have become possible when many different sequence-specific DNA binding proteins had evolved to create adaptive patterns of transcriptional control for many genes. Genes encoding functionally-related proteins could then come together in the genome to be within operon transcription units having suitable expression patterns. On the other hand our argument implies that cells showing very limited use of major groove DNA-binding proteins are closer to the last cellular RNA genome, even if they currently have large genomes.

Trypansosmes are primitive protozoans (Cavalier-Smith, 2010), and they may have retained minimalist transcriptional control of gene expression as a primitive feature from an RNA genome ancestor (Berriman et al., 2005; Iyer et al., 2008). *T. brucei* has a remarkably low number of sequence-specific DNA-binding proteins compared to Metazoans, with not very many more DNA-binding domain types than the basic set shared between eukaryotes and archaebacteria. Some other protozoans that are not parasites also have low numbers of sequence-specific DNA-binding proteins, although the numbers fluctuate greatly due to expansions of particular protein families (Iyer et al., 2008). Some Protozoans, such as *Trypanosoma brucei*, have remarkable genome organizations in which coding sequences of functionally unrelated proteins align along one strand and are expressed as a single transcript. This crude grouping of unrelated genes in *T. brucei* is very different from the assembly of functionally-related genes into operons in bacteria and archaea. After transcription the long transcripts are split up into monocistronic mRNAs by trans-splicing to a common spliced leader sequence and polyadenylation. Genome sequences of other primitive protists indicate that use of the trans-splicing mechanism and strand-switch gene arrangements are also not related to parasitism. Genes in *T. brucei* are defined not so much at the genome level but at the single-stranded RNA level, by the ribosome, with the help of the trans-splicing apparatus.

It is the importance of major groove DNA binding proteins that defines trypanosome genomes as being closer to the last cellular RNA genomes. Trypanosomes have large genomes. It is surprising to realise that the last RNA genome might have been unexpectedly large also. It also could have served to produce a cell similar to a contemporary eukaryotic cell of a primitive type. Most current bacterial and eukaryotic cells and organisms have evolved many more DNA-binding proteins and more complex transcriptional regulatory patterns. Bacterial and archaeal transcriptional repressors and activators are at the pinnacles of specific genome sequence recognition by individual proteins. They use major groove interactions to define large, usually symmetrical DNA-binding sites that are the operators controlling bacterial gene expression (Garvie and Wolberger, 2001). Eukaryotic major groove DNA-binding proteins have instead developed mainly towards more complex cooperative binding to smaller individual sites. The significant proportion of human genes encoding sequence-specific DNA-binding proteins is the basis for the complicated differential regulation of transcription of individual genes in cells and tissues. The sheer number of human genes involved emphasises the importance of DNA major groove recognition (Jolma et al., 2013; Ravasi et al., 2010).

Do any complete or partial cellular RNA genomes remain?

The recent discoveries of RNA-mediated trans-generational inheritance will lead to an exploration of whether genetic information may be sustained or inherited further as RNA alone, with more independence from DNA genomes. The diversity of non-Mendelian inheritance processes in the natural world may be quite substantial and we can now

investigate them with greater confidence (Jablonka and Raz, 2009). Transitions to DNA genomes will have taken some time while the many ssRNA- and dsRNA-binding proteins already in existence adjusted to loss of the 2' hydroxyl on ribose to become DNA-binding proteins. This transition process might be still incomplete in some organisms.

Trypanosome biology also offers hints that some cells may have found advantages in not fully completing the transition to DNA. An RNA chromosome might be more resistant to the insertion of the types of selfish elements that have bombarded DNA genomes since the evolution of viruses with reverse transcriptases. The RNA genome could organize backup protection mechanisms for the DNA genome; kinetoplastid macronuclei appear to operate a system in which RNA sequences that direct genomic DNA correction are inherited at least into the next cell generation. Associating these with an RNA chromosome would ensure inheritance permanently. Some DNA genomes also capture copies of transposons and selfish DNA elements at particular locations to provide targeting sequences for RNAi-based defence; the targeted elements are cognate to DNA though; they enter the DNA genome. An RNA chromosome might be targeted by RNA viruses or could have been adopted by an RNAi mechanism to retain targeting sequences for defense against that do not enter DNA. DNA genomes may have tended to drive RNA genomes out through genomic conflicts but this might have reached a sustained stalemate in some cases.

In another process, any intact cell with an RNA genome still in existence might, losing out to DNA genome cells in evolution to greater complexity, have adopted a parasitic lifestyle. We will need to be able to recognize the potential presence of complete or partial RNA genomes in cells that have DNA as well as in environmental sequences. Clearly, sequences from an RNA genome should be found only by RT-PCR and not PCR amplification. However identifying sequences encoding very distinctive members of ancient minor groove-binding protein or polymerase families might provide the first evidence. Studies on Endo V and other minor groove binding proteins may identify patterns of amino acid sequence change associated with the switch from RNA binding to DNA binding or to dual RNA and DNA binding.

Inosines in bacterial conflicts

A deep search for adenosine deaminase domains has found new examples fused at the carboxy termini of bacterial toxin proteins (Iyer et al., 2011). The bacterium transfers the carboxy-terminal toxin domain from structures on its cell surface into another cell. Toxin proteins have various N-terminal domains involved in delivering the protein to the cell surface and into the target cell. A variety of different domains are found fused at the carboxy termini of toxin proteins and many of the toxin domains damage nucleic acids. In these bacterial toxin-antitoxin systems the toxin genes are located in the chromosome beside an antitoxin gene which prevents damage to the cell. In one type of toxin-antitoxin system the antitoxin encodes a ubiquitination-targeting protein directed at the toxin (Zhang et al., 2012). If the second bacterium receiving the toxin protein does not have the antitoxin gene then its nucleic acids will be attacked. In some bacteria toxin-antitoxin mechanisms cause strain-specific contact inhibition effects likely to affect resource sharing for food supply or in the formation of biofilms. The toxin-antitoxin system acts as a system of self or kin recognition.

The biochemical activities of the predicted adenosine deaminase domain toxins have not been characterised yet but some are similar to bacterial adenosine deaminases acting on tRNAs (ADATs). Eukaryotes use adenosine deamination to inosine to facilitate wobble decoding in eight tRNAs but bacteria use adenosine deamination in only one tRNA. This lack of universality in the use of inosine for wobble decoding has been an enigma. The involvement of adenosine deaminase domains in bacterial genomic conflict may now account for this difference between bacterial and eukaryotes. Bacteria may have reduced the number

of edited tRNAs as these might be targets for the deaminase toxins. The findings are a further reminder of the importance of genome conflicts and host parasite conflicts (Hamilton and Hamilton, 2001), in the evolution of RNA and DNA modifications.

Inosines in vertebrate antiviral responses and inosine recognition in dsRNA by RIG-I-like sensors

In vertebrates inosine-uracil (I-U) base pairs introduced by ADAR editing or adenosines in dsRNA assist innate immune discrimination between self and non-self RNAs (Karikó et al., 2005; Vitali and Scadden, 2010), and may even allow endogenous cellular RNAs to suppress undesireable immune responses. Mutations in human ADAR1 cause Aicardi-Goutieres #syndrome, a childhood encephalopathy with interferon expression and symptoms resembling those caused by congenital virus infections (Rice et al., 2012).

Animal cells use a range of Pattern Recognition Receptors (PRRs) to distinguish viral RNAs from cellular RNAs, activating interferon signalling in response to viral RNAs (Takeuchi and Akira, 2010). Immune responses to exogenous RNAs are unwanted in cell transfections with RNAs such as mRNAs encoding pluripotency factors that are transfected for iPS cell induction. Many of the RNA modification systems in vertebrates, including ADAR RNA editing, operate in the nucleus. Viruses replicating in the cytoplasm are expected to have limited access to these modifications and therefore naked RNAs or transfected unmodified mRNAs induce innate immune responses because PRRs recognize them as viruses. Some viruses do gain protection from the innate immune system by recruiting host RNA modifications or encoding their own (Nallagatla et al., 2008). Since cellular RNAs do not normally induce interferon signalling it seemed likely that naturally-occurring modified bases in cellular RNAs should be neutral or inhibitory to innate immune responses against exogenous RNA. Kariko et al showed that incorporating moderate levels of some modified bases that naturally occur in eukaryotic tRNAs and rRNAs during *in vitro* mRNA synthesis prevents immune induction by the transfected RNA (Kariko et al., 2005). Most modified bases do not change the encoded protein sequence and synthesising mRNAs with modified nucleotides blocks the immune induction during iPS cell induction.

The mechanisms by which modified bases in cellular RNAs modulate the activities of antiviral sensors represent an important area for future work. In the case of inosine within dsRNA, Vitali et al have shown that I-U base pairs in dsRNA inhibit activation of RIG-I-like cytoplasmic antiviral RNA sensors in particular (Vitali and Scadden, 2010). It is possible to suggest the likely basis for this discrimination by RIG-I-like receptors. Crystal structures of dsRNA oligonucleotides containing I-U base pairs show that an I-U base pairs forms two hydrogen bonds as a wobble base pair. The uracil or inosine base is moved inwards from the minor groove (Figure 1 A), and the dsRNA bends way from the minor groove slightly (Pan et al., 1998). The effect of multiple sequential I-U base pairs is more severe, bending is likely to be greater. There may also be significantly weakened stacking of bases since RNAs with multiple I-U base pairs have very unstable pairing (Serra et al., 2004). We suggest that RIG-I has a minor-groove scanning mechanism that detects the presence of the I-U base pair through the altered shape of the minor group and through effects on the wobble base pair on base stacking.

The RIG-I innate immune sensor is much more recently evolved than Endo V and Mut S and only found in Metazoans. However the mechanisms by which proteins such as Endo V and RIG-I recognize modified or non-canonical bases in dsRNA are likely to be similar to DNA mismatch recognition. The I-U wobble base pair in dsRNA is isosteric with the G-U wobble base pair that occurs more commonly in structures of stable RNA. Interestingly though, G-U wobble base pairs do not inhibit RIG-I as effectively as I-U base pairs do. Since guanosine has an amino group in the minor groove that inosine lacks this

discrimination is consistent with a minor groove mechanism for I-U base pair recognition by RIG-I activation. Whether RIG-I is inhibited by other modified bases remains to be determined. The crucial step in RIG-I activation by dsRNA is the release of signalling N-terminal CARD domains that interact with ubiquitin to mediate antiviral signalling. The CARD domains are bound to the helicase domain in the inactive RIG-I structure at a position slightly overlapping where dsRNA is bound in the active, signalling conformation of RIG-I. dsRNA containing I-U base pairs may be able to bind to the helicase domain without causing displacement of the CARD domains or a different inactive RIG-I conformation may be generated.

Conclusion

Decades of research have now brought us to the point where an even wider range of important roles of RNA modifications are coming into focus. We have appreciated for a long time that RNA modifications in tRNAs and rRNAs are critical for translation. Faithful operation of the genetic code for proteins largely depends on RNA modifications. The earlier identifications of modified bases in stable RNAs were laborious and further modifications are likely to be found. A wider range of modified bases are also now being found at specific positions in mRNAs. The tolerance of reverse transcriptases in reading past modified bases had previously erased the evidence for these modified bases from cDNA sequences; only the RNA editing events that affect base pairing were readily detected. Reverse transcriptases may have good reasons to be so tolerant of modified bases; new sequencing technologies for finding RNA modifications and nanopore sequencing technologies that sequence RNAs directly may lead to a goldrush in the identification of new RNA and DNA modifications. Further important roles of modifications in genome-like dsRNA have also become clear. Many types of evolutionary conflict have involved arms races of nucleic acid attack and defence and modified nucleic acids are critical in these processes. It is not really surprising therefore that vertebrates have now been found to use modified nucleic acids to distinguish their own nucleic acids from those of viruses and parasites. This will be a major arena for further work.

Studies on ADAR dsRNA editing can also be brought together with evolving ideas on the origins of DNA genomes. We argue that understanding RNA modifications helps to illuminate the origin of DNA. We suggest a search for an existing or ancient ribose reductase acting on genomic dsRNA (RRAR). We suggest that the transition to DNA genomes could have occurred later than previously suggested and present a significantly revised view of the evolutionary benefits of DNA genomes. Indeed, DNA would be an almost trivial modification of dsRNA if it were not for the fact that DNA structure, for the first time, exposed the entire genome to proteins searching for specific sequences. We have not discussed the twenty five modifications known to occur in DNA but similar principles with regard to genome conflicts apply even more strongly to these. Some, such as 5-meC and 6-mA are the same modifications that occur in RNA. All enzymatic modifications of DNA may one day be recognized as further modification steps evolutionarily appended to the reduction of ribose in dsRNA.

References.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The many faces of the helix-turn-helix domain: Transcription regulation and beyond⋆. FEMS microbiology reviews *29*, 231-262.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., and Haas, B. (2005). The genome of the African trypanosome Trypanosoma brucei. Science *309*, 416-422.

Cavalier-Smith, T. (2010). Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biology letters *6*, 342-345.

Dalhus, B., Arvai, A.S., Rosnes, I., Olsen, Ø.E., Backe, P.H., Alseth, I., Gao, H., Cao, W., Tainer, J.A., and Bjørås, M. (2009). Structures of endonuclease V with DNA reveal initiation of deaminated adenine repair. Nature structural & molecular biology *16*, 138-143.

Dickerson, R.E., Drew, H.R., Conner, B.N., Wing, R.M., Fratini, A.V., and Kopka, M.L. (1982). The anatomy of a-, b-, and z-dna. Science *216*, 475-485.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. Virus research *117*, 5-16.

Forterre, P., and Grosjean, H. (2000). The interplay between RNA and DNA modifications: back to the RNA world.

Forterre, P., and Grosjean, H. (2009). The Interplay between RNA and DNA Modifications. DNA and RNA Modification Enzymes, 259.

Freeland, S.J., Knight, R.D., and Landweber, L.F. (1999). Do proteins predate DNA? Science *286*, 690-692.

Garvie, C.W., and Wolberger, C. (2001). Recognition of specific DNA sequences. Molecular cell *8*, 937-946.

Goodman, R.A., Macbeth, M.R., and Beal, P.A. (2012). ADAR proteins: Structure and catalytic mechanism. In Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing (Springer), pp. 1-33.

Hamilton, W.D., and Hamilton, W.D. (2001). Narrow roads of gene land: Volume 2: evolution of sex, Vol 2 (Oxford University Press).

Hermann, T., and Westhof, E. (1999). Non-Watson-Crick base pairs in RNA-protein recognition. Chemistry & biology *6*, R335-R343.

Iyer, L.M., Anantharaman, V., Wolf, M.Y., and Aravind, L. (2008). Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. International journal for parasitology *38*, 1-31.

Iyer, L.M., Zhang, D., Rogozin, I.B., and Aravind, L. (2011). Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. Nucleic acids research *39*, 9473-9497.

Jablonka, E., and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. The Quarterly review of biology *84*, 131-176.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., and Wei, G. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327-339.

Jordan, S.R., and Pabo, C.O. (1988). Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions. Science *242*, 893-899.

Kariko, K., Buckstein, M., Ni, H., and Weissman, D. (2005). Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. Immunity *23*, 165-175.

Karikó, K., Buckstein, M., Ni, H., and Weissman, D. (2005). Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. Immunity *23*, 165-175.

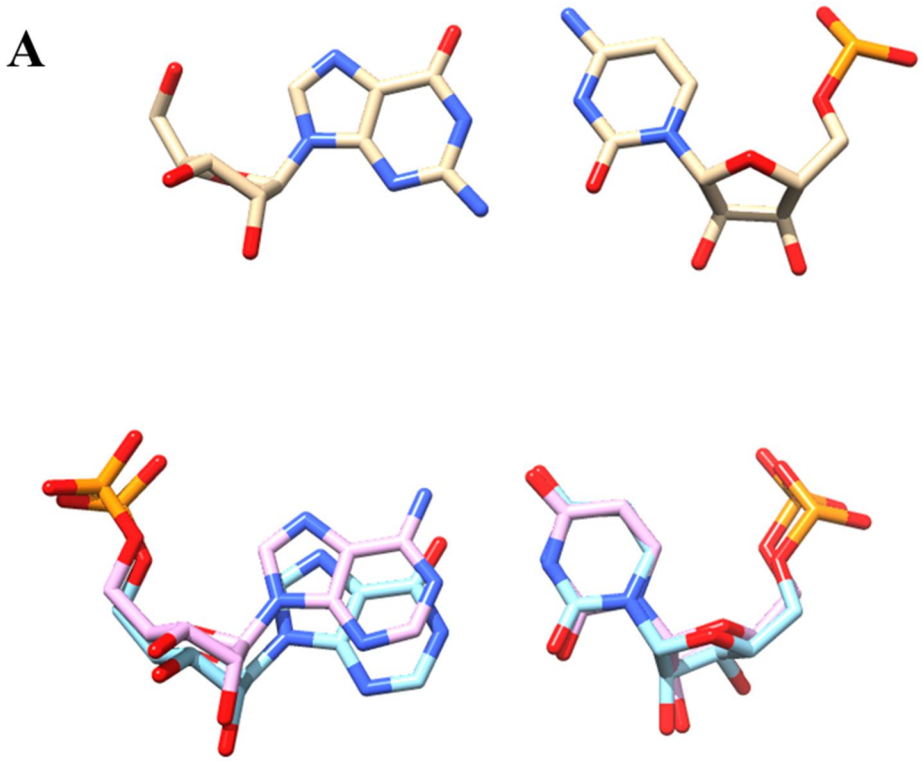
Keegan, L.P., Leroy, A., Sproul, D., and O'Connell, M.A. (2004). Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. Genome Biol *5*, 209.

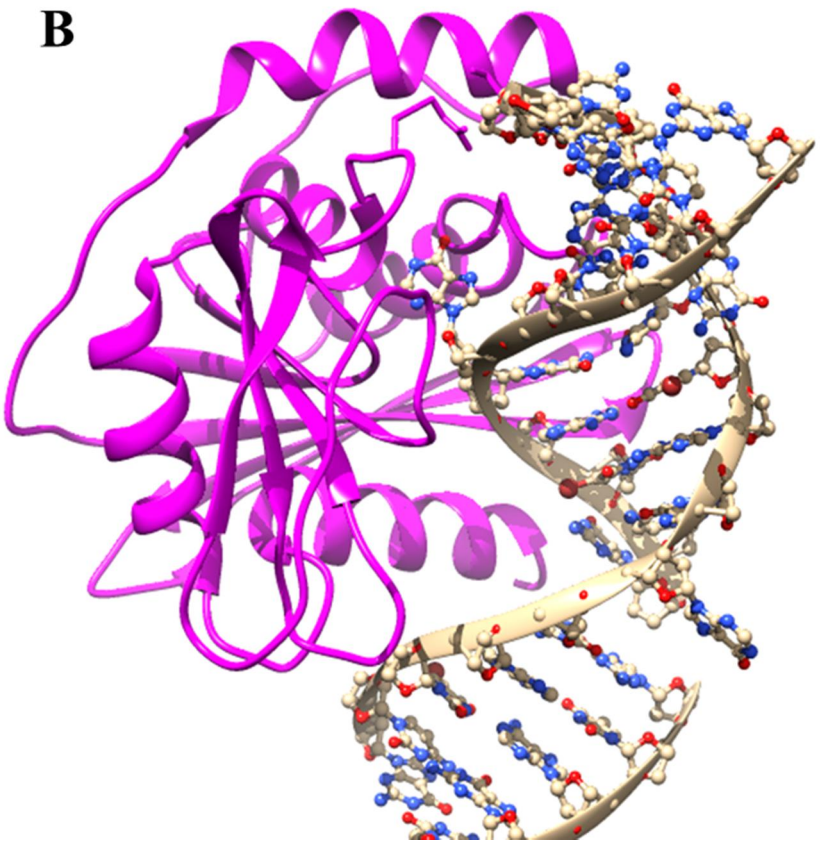
Lamers, M.H., Perrakis, A., Enzlin, J.H., Winterwerp, H.H., de Wind, N., and Sixma, T.K. (2000). The crystal structure of DNA mismatch repair protein MutS binding to a G· T mismatch. Nature *407*, 711-717.
Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709-715.
Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K.M.*, et al.* (2013). MODOMICS: a database of RNA modification pathways--2013 update. Nucleic Acids Res *41*, D262-267.
Masliah, G., Barraud, P., and Allain, F.H.-T. (2013). RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cellular and Molecular Life Sciences *70*, 1875-1895.
Morita, Y., Shibutani, T., Nakanishi, N., Nishikura, K., Iwai, S., and Kuraoka, I. (2013). Human endonuclease V is a ribonuclease specific for inosine-containing RNA. Nature communications *4*.
Nallagatla, S.R., Toroney, R., and Bevilacqua, P.C. (2008). A brilliant disguise for self RNA. RNA biology *5*, 140-144.
Pan, B., Nath, S., Sun, L., Hart, D., and Sundaralingam, M. (1998). Crystal structure of an RNA octamer duplex r (CCCIUGGG) 2 incorporating tandem I· U wobbles. Nucleic acids research *26*, 5699-5706.
Poole, A., Penny, D., and Sjöberg, B.-M. (2000). Methyl-RNA: an evolutionary bridge between RNA and DNA? Chemistry & biology *7*, R207-R216.
Poole, A.M., Logan, D.T., and Sjöberg, B.-M. (2002). The evolution of the ribonucleotide reductases: much ado about oxygen. Journal of molecular evolution *55*, 180-196.
Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N.*, et al.* An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744-752.
Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N.*, et al.* (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744-752.
Rice, G.I., Kasher, P.R., Forte, G.M., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A.*, et al.* (2012). Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. Nat Genet *44*, 1243-1248.
Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proceedings of the National Academy of Sciences *73*, 804-808.
Serra, M.J., Smolter, P.E., and Westhof, E. (2004). Pronounced instability of tandem IU base pairs in RNA. Nucleic Acids Res *32*, 1824-1828.
Stefl, R., Oberstrass, F.C., Hood, J.L., Jourdan, M., Zimmermann, M., Skrisovska, L., Maris, C., Peng, L., Hofr, C., and Emeson, R.B. (2010). The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. Cell *143*, 225-237.
Stefl, R., Xu, M., Skrisovska, L., Emeson, R.B., and Allain, F.H. (2006). Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. Structure *14*, 345-355.
Steitz, T.A. (1993). Similarities and differences between RNA and DNA recognition by proteins. In "The RNA World", R.F. Gesteland, and J.F. Atkins, eds. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press), pp. 219-237.
Takeuchi, O., and Akira, S. (2010). Pattern recognition receptors and inflammation. Cell *140*, 805-820.
Vik, E.S., Nawaz, M.S., Andersen, P.S., Fladeby, C., Bjørås, M., Dalhus, B., and Alseth, I. (2013). Endonuclease V cleaves at inosines in RNA. Nature communications *4*.
Vitali, P., and Scadden, A.D. (2010). Double-stranded RNAs containing multiple IU pairs are sufficient to suppress interferon induction and apoptosis. Nat Struct Mol Biol *17*, 1043-1050.
Zhang, D., de Souza, R.F., Anantharaman, V., Iyer, L.M., and Aravind, L. (2012). Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol Direct *7*, 18.

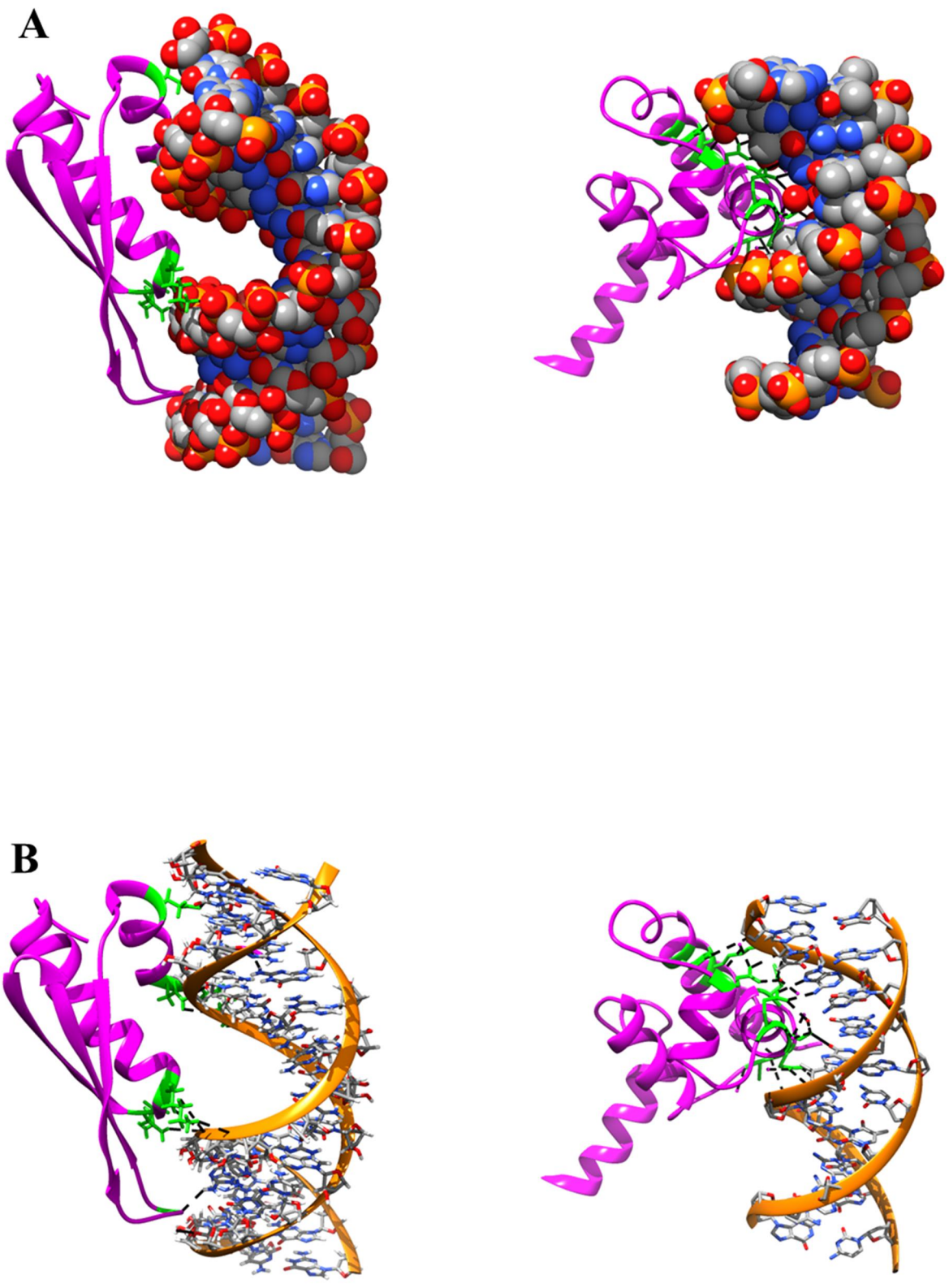
Figure legends

Figure 1. The canonical base pairs and the inosine-uracil wobble base pair in dsRNA and the recognition of deoxyinosine by bacterial Endonuclease V. **A.** A G-C base pair in dsRNA is shown in the upper half and an A-U base pair in dsRNA is in the lower half. The major groove is above the bases and the minor groove is below the bases in all cases. In the lower image the inosine-uracil wobble base pair is superimposed on the A-U base pair to show the shift inwards towards the major groove. **B**. Ribbon diagram of bacterial Endonuclease V complexed with DNA. The protein recognizes the minor groove and deoxyinosine is flipped out of the minor groove.

Figure 2. Minor groove sequence recognition in dsRNA versus major groove sequence recognition in DNA. **A.** The left hand image shows a ribbon diagram of human ADAR2 dsRBD 2 bound to the *GluR B Q/R* site RNA hairpin. The dsRBD domain passes over the major groove to interact with two successive minor grooves and some sequence-specific base contacts are made by helix1 in the upper minor groove. The right hand image shows one of the symmetrical DNA-binding domains of lambda phage repressor in a complex with an operator half-site. Note the wider and shallower major groove than in dsRNA. The recognition alpha helix penetrates the major groove and a much larger number of specific contacts are formed between amino acid side chains and the bases in the presented major groove surface. Amino acids making specific contacts are marked in green with hydrogen bonds extending from them in black. **B** The same structures shown using stick and ribbon models of the nucleic acids to see the hydrogen protein-nucleic acid contacts more clearly.


A


B


A
B