

OPINION

Confounded cytosine! Tinkering and the evolution of DNA

Anthony Poole, David Penny and Britt-Marie Sjöberg

Early in the history of DNA, thymine replaced uracil, thus solving a short-term problem for storing genetic information — mutation of cytosine to uracil through deamination. Any engineer would have replaced cytosine, but evolution is a tinkerer not an engineer. By keeping cytosine and replacing uracil the problem was never eliminated, returning once again with the advent of DNA methylation.

The origin of DNA is a fundamental question in evolution. Early on, DNA replaced RNA, reflecting the superior information-storage capacity of DNA^{1,2}. Modern biochemical pathways provide an insight into this transition, as do RNA and uracil-DNA (U-DNA) viruses^{2,3}, suggesting that the replacement took place in two steps (FIGS 1, 2a): replacement of ribose with deoxyribose, then replacement of uracil (U) with thymine (T)⁴. The first step was probably very complex, and has recently been reviewed elsewhere^{2,5}. Here we look at the second (U→T) replacement, which is emerging as another example of why evolution is best viewed as a tinkerer, not as an engineer with an eye for ‘good’ design (BOX 1).

Central to the story is cytosine (C), which readily deaminates to form U. This turns C•G pairs into U•G mispairs, and is an ongoing process in DNA^{6,7} (FIG. 2b). Without repair, replication of a U•G mispair would give one U•A pair (which is read as a T•A pair) and one C•G pair. All organisms carry the machinery for repair of C deaminations — uracil-*N*-glycosylase (UNG), which recognizes and removes any U that it detects, leaving an aban-

sic site. This is patched up by base-excision repair^{8,9} (FIG. 3), which creates a gap in the DNA opposite G. DNA polymerase then fills the gap with dC, thus repairing the mutation. Occasionally, U (from dUTP) is incorporated opposite A, so both U•G and U•A pairs turn up in DNA. The UNG recognizes and removes U arising from either C deamination or misincorporation, allowing DNA to be faithfully repaired^{10–12}.

Before T was a constituent of DNA, it would have been harder to detect C deamina-

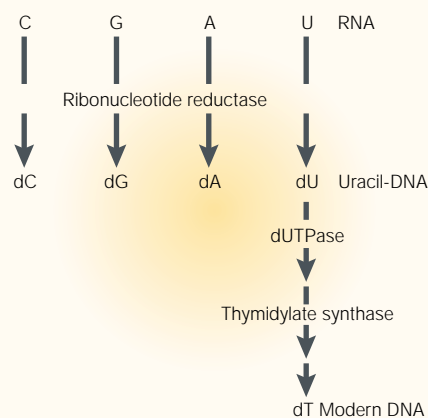


Figure 1 | **Stepwise evolution of DNA on the basis of what is inferred from modern biochemical pathways.** Ribonucleotide reductases catalyse synthesis of deoxyribonucleotides from ribonucleotides, with the exception of dT, which is synthesized from dU by thymidylate synthase (TS). dUTP is acted on by dUTPase to produce dUMP, the substrate for dTMP synthesis by TS. This is then brought up to the triphosphate level (indicated by the last arrow) before incorporation into DNA.

tions, because U was a bona fide constituent of early DNA at U•A pairs. It is widely accepted that U→T replacement solved this problem because it allowed any U arising by C deamination to be detected unambiguously¹³ (FIG. 4). However, replacing U with T would not by itself eliminate mutations arising from C deamination — it simply allows C→U mutations to be recognized, because U would otherwise be absent from DNA⁹. In the absence of repair (assuming the only role for T was to provide a means of recognizing C→U deaminations), there is no obvious selection pressure for the U→T replacement.

François Jacob¹⁴ has likened evolution to “tinkering”. In contrast to an engineer, who works by design, obtaining all the necessary materials needed for construction of a prototype and finally testing it before putting it to work, a tinkerer makes use of whatever is at hand. This means that the result, although functional, is often far from perfect. A consequence of this *modus operandi* is that if something works in the short term it will be used, even if a better alternative is conceivable. New innovations cannot arise only to become useful when a subsequent function evolves, because there is no selection to maintain such innovations before they become useful.

Recent progress on the biochemistry of U removal reveals an unexpected diversity of reactions catalysed by members of the uracil-DNA glycosylase family (even though they all share a common origin), and allows the U→T conundrum to be resolved. New data¹⁵ on a closely related phenomenon — the repair of deaminated 5-methylcytosine (5-meC, which deaminates to T, resulting in a T•G mismatch; FIG. 2b) — highlights the usefulness of the tinkering analogy for evolution. The problems solved by replacing U with T resurfaced once again when C methylation became a feature of the genome, with a member of the U-DNA glycosylase family being recruited to repair 5-meC→T deaminations.

Driving thymidylate synthase evolution
If we view the U→T replacement in terms of

Box 1 | Tinkering versus engineering

Is cytosine a component of the genetic material simply because there is no possible alternative to cytosine? If this were so, then cytosine deamination would be unavoidable. Jacob's point, however, is that in evolution solutions are found using what is readily available¹⁴. An engineer can choose to scrap an entire project and start again from scratch. In the current context, if C is no good because of deamination, an engineer can replace it. If G is no good — because there is no alternative to C — then it too can be replaced. Indeed, if DNA therefore turns out to be no good, because there is no alternative to C•G pairs, this can be replaced.

The problem of C deamination is recognized by prebiotic chemists³⁶ as causing headaches for solving the origin-of-life problem. There is nevertheless no expectation that nucleic acids (with C) are the only possible medium for biological information storage — they are merely the only medium we are aware of.

So there are two positions that can be held: that DNA containing C is the only possible genetic material, so the tinkering analogy does not hold; or that C and DNA came to be used for storage of genetic material, in spite of the tendency for C to deaminate. Our point is that if C were available, and were selected for in evolution, it would not make any difference if it later turned out to have an Achilles' heel — once it became central to the genetic material, it was effectively impossible to replace it, at any level. A helpful analogy is Francis Crick's³⁷ concept of a "frozen accident" — a feature on which other features are built becomes so central to the working of the "machine" that it cannot be replaced in evolution, even if there is a conceivably better alternative. All those features that are built on the central feature would also have to be changed, so the central feature is effectively "frozen". Frozen accidents are therefore a consequence of evolution through tinkering.

tinkering, repair of C→U deaminations probably evolved before — or perhaps concurrent with — thymidylate synthase (which synthesizes dTMP from dUMP for the incorporation of T (from dTTP) into DNA; BOX 2). T could not have arisen in the absence of UNG as this implies forethought, not tinkering. It makes more sense if some form of repair arose first. This could have been by removal of all U, removal and repair only at U•G mismatches or preferential (but imperfect) U•G repair. We will consider each of these possibilities in turn.

If repair were by removal of all U, as per modern UNG, U would be removed where it had arisen by C deamination (which would be beneficial), but also where it was correctly paired with A (FIG. 4). Removal of U opposite A would be energetically wasteful, because the base-excision repair pathway would be stuck in a futile cycle. Nevertheless, if the cost (in terms of energy) of removing and reincorporating U is less than the benefit (in terms of improved genetic stability) of repairing C deamination events, then UNG function might have been selected for (FIG. 4). The cost of the futile cycle would have driven selection towards replacement of U by T (that is, driven the evolution of thymidylate synthase), and subsequently favoured evolution of mechanisms (such as dUTPase) for reducing the misincorporation of U into DNA.

However, extensive removal of U at U•A pairs was probably a greater cost than the benefit of repairing U•G. On average, 25% of the genome would be U, and the risk of mutation by subsequent misincorporation

would be ever present, as futile cycling would be continuous.

At the other end of the spectrum is the second possibility — removal and repair only at U•G mismatches, as per the UNG-related enzyme, mismatch-specific uracil DNA-glycosylase (MUG)^{16,17}. This would leave U•A pairs untouched (FIG. 4).

However, if U•G mismatches were recognized and replaced unambiguously from the outset, what selection would there have been for U→T replacement?

Leaky MUG — a possible solution
The suggestion that either UNG or MUG predate thymidylate synthase does not seem to solve the problem of the U→T transition. The

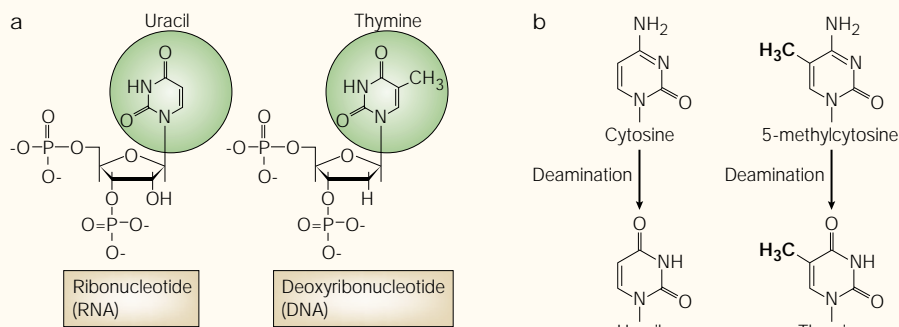


Figure 2 | **DNA structure and the chemistry of cytosine deamination.** **a** | Differences between RNA and DNA. RNA and DNA differ at the 2' position of the ribose — RNA has a hydroxyl group, whereas in DNA this is reduced to hydrogen. The second difference is at the level of the base — RNA contains uracil, which is equivalent to thymine in DNA. The difference between these two bases is the presence, in thymine, of a methyl group at the 5' position of the pyrimidine ring. **b** | Cytosine deamination. Spontaneous deamination of cytosine and 5-methylcytosine is shown. The former deaminates to uracil, whereas the latter deaminates to give thymine. Both deamination events are mutagenic if unrepaired, leading to C•G→T•A changes in one daughter DNA strand after replication.

ideal solution would be a trade-off between UNG and MUG activities — preferential (but imperfect) U•G repair (FIG. 4). Selection for U replacement could then have been driven by initial 'leakiness' of repair (that is, occasional U-excision repair at U•A pairs). Such a 'leaky' MUG (FIG. 4), with properties common to both modern enzymes, would have later evolved by duplication and divergence into the two specific enzymes seen in modern repair.

Increasing knowledge of the modern enzymes from sequence, structural, mutagenesis and biochemical data provides a powerful way to explore this leaky MUG idea. Extant MUG specifically flips out U binding to G opposite, thus precluding N-glycoside cleavage at U•A pairs. 'Flipping-out' of U is common to both MUG and UNG, but MUG recognizes G opposite U, with little affinity for U itself¹⁸, whereas UNG has high affinity for U, ignoring the complementary strand and even excising U from single-stranded DNA¹⁶.

In MUG, interaction with G is mediated largely by two residues (glycine 143 and serine 145), which hydrogen-bond with groups on G (REF. 16) that normally pair with C. Mutating these residues should therefore weaken discrimination between G and A, producing a leaky MUG. More significantly, substrate-affinity studies using catalytically inactive mutants of UNG show that UNG naturally has a tenfold faster association rate for U•G pairs than for U•A pairs¹⁹. This results from the greater ease with which U•G pairs can be prised apart, allowing the enzyme to bind U (REF. 19). So, even for a rudimentary enzymatic reaction of U-specific recognition and cleavage, U•G mismatches might have been preferred by the prototype UNG/MUG.

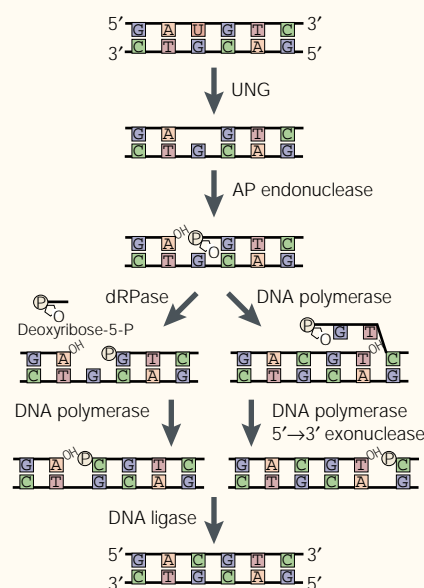


Figure 3 | Base-excision repair pathways. The two alternative pathways for base-excision repair of C→U deaminations are shown. All steps downstream of uracil-*N*-glycosylase (UNG) are part of general base-excision repair, which acts on abasic lesions from a number of sources, including those produced by action of DNA glycosylases. It can be argued that the right-hand pathway pre-dates that on the left, given that deoxyribose-5-phosphatase (dRPase) would further reduce the possibility of errors introduced during repair — removal of only the abasic site eliminates the possibility of downstream errors introduced by resynthesis of a short patch by DNA polymerase. (AP endonuclease, apurinic/apryrimidic endonuclease.) (Adapted from REF. 8.)

Stopping the cycle: eliminating uracil
If the proto-UNG/MUG preferentially excised U at U•G pairs, but occasionally excised U opposite A (leaky MUG; FIG. 4), evolution of thymidylate synthase and consequent U→T replacement would have been advantageous. It would result in improved fidelity — and possibly energy conservation — by eliminating occasional futile removal at U•A pairs. Reduction in base-excision repair activity would have reduced errors resulting from incorrect replacement of adjacent downstream nucleotides during re-synthesis^{8,9} (FIG. 3). The U•G preference would have become relaxed, with speed of recognition and removal being selected for as genomes increased in size. We therefore propose that the advent of thymidylate synthase permitted the elimination of occasional U excision opposite A. Hence the role of T in DNA was simply to allow the specific repair of U•G mispairs. Other than this function, there is no clear advantage in having T in DNA rather than U.

The advent of dUTPase would have pro-

vided further improvement, turning the dUTP pool into dUMP (FIG. 5), which thymidylate synthase could convert to dTMP. This would have reduced dUTP misincorporation and enhanced dTTP incorporation (FIG. 5). In this model, then, the base-excision repair pathway pre-dated leaky MUG activity (which excises only the base, leaving an abasic site). Abasic lesions arise non-enzymatically in DNA and are a major source of mutation^{9,20}. So base-excision repair probably evolved to repair abasic lesions before its role in excision of U.

Why MUG?

The above picture describes the probable nature of the proto-UNG/MUG and how it could have driven U→T replacement, as well as how UNG could subsequently evolve. But what selection pressure gave rise to MUG? Given the much higher rate at which UNG removes U from DNA, MUG is probably not important in U-base-excision repair *in vivo*^{21,22}, and the function of discriminate U•G removal in bacteria is unclear. There is evidence, however, that MUG might provide a means of maintaining genetic material during viral infection.

Two bacteriophage that infect *Bacillus subtilis* — PBS1 and PBS2 — are unusual among DNA viruses as their DNA contains U not T (REF. 3). These viruses make a protein that inhibits UNG but not MUG¹⁷. So, in an infected bacterium, UNG-mediated repair is

inhibited, allowing the production of phage U-DNA. But any U•G mispairs arising from C→U deamination in the host cannot be repaired. Although U•A pairs are tolerated, lack of repair during infection could leave a bacterium with serious lesions, so the ability to specifically repair U•G mispairs would be advantageous. Both PBS1 and PBS2 are lytic phage (where the host cell is killed upon phage release), although given predicted phage diversity^{23,24} some lysogenic phage (where the host cell survives phage release) might also have U-DNA genomes.

The hypothesis that MUG maintains genome fidelity during lysogenic infection by U-containing DNA phage would be borne out if MUG were found only in prokaryotes infected by U-DNA phage. Whereas host-mediated repair of U•G mispairs would be beneficial for both lytic and lysogenic phage, the selective advantage of MUG to the host would be in genome integrity during lysogenic infection. The existence of MUG in bacteria opens up the exciting possibility that, as per the RNA world, the U-DNA world is alive and well in viruses.

The trouble with T

As eukaryotic genomes became more complex, additional mechanisms of gene regulation developed. One such mechanism is DNA methylation, where a methyl group is added to position 5 on the cytosine ring, forming 5-mC (REF. 15). The ability to regu-

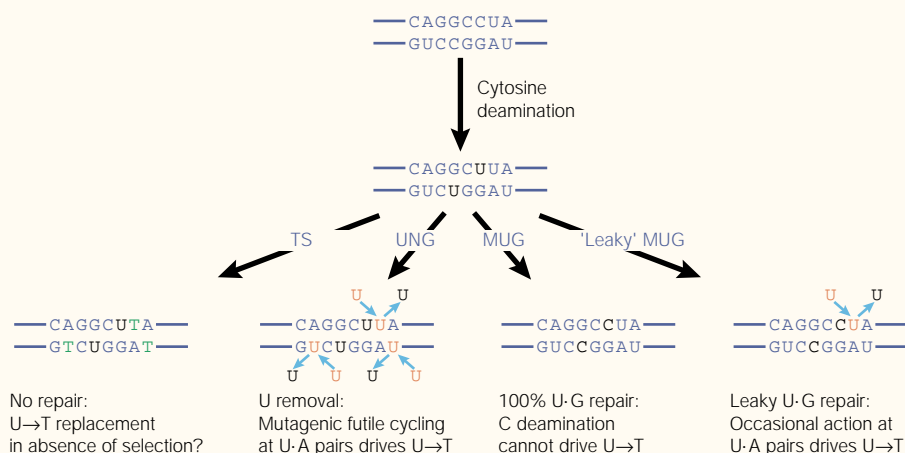


Figure 4 | 'Leaky' mismatch-specific uracil-DNA glycosylase as a driving force for evolution of U→T replacement in DNA. Replacement of dU with dT occurred after the origin of DNA, probably in response to the problem of C deamination to U. This leads to loss of information as C is incorrectly read as U. Four evolutionary hypotheses are shown. The standard explanation for the origin of thymidylate synthase (TS) is that it solves this problem. However, replacing U with T without pre-existing uracil-*N*-glycosylase (UNG) activity does not eliminate the problem. An alternative explanation is that UNG arose first, allowing repair of deaminated cytosines, but creating an additional problem — extensive futile repair at undamaged sites. Unlike UNG, mismatch-specific uracil-DNA glycosylase (MUG) acts specifically on U•G mismatches, so if this evolved before TS, it is hard to see what selection pressure would drive U→T replacement. We propose that a 'leaky' MUG arose in response to C→U deamination. The enzyme would have had a preference for U•G repair, but occasionally act on U•A pairs. Occasional futile repair might have favoured evolution of TS and replacement of U by T.

tions, but the problem of C deamination was never eliminated — it re-emerged in the form of 5-meC deamination. Tinkering also makes sense of the evolution of the 5-meC apparatus, which subsequently drove the recruitment of the U-excision apparatus into T excision because of the ‘unforeseen’ side effect of 5-meC→T deamination. All this could have been avoided simply by eliminating C early in the evolution of the genetic material — but how boring life would be if evolution worked by engineering.

Anthony Poole and David Penny are at the Institute of Molecular BioSciences, PO Box 11222, Massey University, Palmerston North, New Zealand. Britt-Marie Sjöberg is in the Department of Molecular Biology, Stockholm University, 106 91 Stockholm, Sweden.

e-mails: a.m.poole@massey.ac.nz; d.penny@massey.ac.nz; britt-marie.sjoberg@molbio.su.se

Links

ENCYCLOPEDIA OF LIFE SCIENCES **Codon usage in molecular evolution**

- Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
- Poole, A., Penny, D. & Sjöberg, B.-M. Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chem. Biol.* **7**, R207–R216 (2000).
- Hitzeman, R. A. & Price, A. R. Relationship of *Bacillus subtilis* DNA polymerase III to bacteriophage PBS2-induced DNA polymerase and to the replication of uracil-containing DNA. *J. Virol.* **28**, 697–709 (1978).
- Lazcano, A., Guerrero, R., Margulis, L. & Oró, J. The evolutionary transition from RNA to DNA in early cells. *J. Mol. Evol.* **27**, 283–290 (1988).
- Sutherland, J. D. & Whitfield, J. N. Prebiotic chemistry: a biorganic perspective. *Tetrahedron* **53**, 11493–11527 (1997).
- Frederico, L. A., Kunkel, T. A. & Ramsay Shaw, B. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**, 2532–2537 (1990).
- Frederico, L. A., Kunkel, T. A. & Ramsay Shaw, B. Cytosine deamination in mismatched base pairs. *Biochemistry* **32**, 6523–6530 (1993).
- Dianov, G. & Lindahl, T. Reconstitution of the DNA base excision-repair pathway. *Curr. Biol.* **4**, 1069–1076 (1994).
- Lindahl, T. & Wood, R. D. Quality control by DNA repair. *Science* **286**, 1897–1905 (1999).
- Mosbaugh, D. W. & Bennett, S. E. Uracil-excision DNA repair. *Prog. Nucleic Acid Res. Mol. Biol.* **48**, 315–370 (1994).
- Krokan, H. E., Standal, R. & Slupphaug, G. DNA glycosylases in the base excision repair of DNA. *Biochem. J.* **325**, 1–16 (1997).
- Cunningham, R. P. DNA glycosylases. *Mutat. Res.* **383**, 189–196 (1997).
- Kornberg, A. & Baker, T. A. *DNA Replication* 2nd edn Ch. 2 (W. H. Freeman, New York, 1992).
- Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
- Jones, P. A. The DNA methylation paradox. *Trends Genet.* **15**, 34–37 (1999).
- Barrett, T. E. *et al.* Crystal structure of a G:T/U mismatch-specific DNA glycosylase: mismatch recognition by complementary-strand interactions. *Cell* **92**, 117–129 (1998).
- Gallinari, P. & Jiricny, J. A new class of uracil-DNA glycosylases related to human thymine-DNA glycosylase. *Nature* **383**, 735–738 (1996).
- Barrett, T. E. *et al.* Crystal structure of a thwarted mismatch glycosylase DNA repair complex. *EMBO J.* **18**, 6599–6609 (1999).
- Panayotou, G., Brown, T., Barlow, T., Pearl, L. H. & Savva, R. Direct measurement of the substrate preference of uracil-DNA glycosylase. *J. Biol. Chem.* **273**,

- 45–50 (1998).
- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- Lutsenko, E. & Bhagwat, A. S. Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. *Mutat. Res.* **437**, 11–20 (1999).
- Lutsenko, E. & Bhagwat, A. S. The role of the *Escherichia coli* MUG protein in the removal of uracil and 3,*N*¹-Ethenocytosine from DNA. *J. Biol. Chem.* **274**, 31034–31038 (1999).
- Maniloff, J. & Ackermann, H.-W. Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. *Arch. Virol.* **143**, 2051–2063 (1998).
- Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA* **96**, 2192–2197 (1999).
- Hendrich, B., Hardeland, U., Ng, H. -H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301–304 (1999).
- Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
- O'Neill, R. J. W., O'Neill, M. J. & Graves, J. A. M. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998).
- Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.* **20**, 116–117 (1998).
- Garrick, D., Fiering, S., Martin, D. I. K. & Whitelaw, E. Repeat induced gene silencing in mammals. *Nature Genet.* **18**, 56–59 (1998).

OPINION

Retroviral recombination: what drives the switch?

Matteo Negroni and Henri Buc

The high rate of recombination in retroviruses is due to the frequent template switching that occurs during reverse transcription. Although the mechanism that leads to this switch is still a matter of debate, there is increasing evidence that specific RNA structures are involved. And the implications might go beyond retroviral genetic variability.

Diploidy is great! We, eukaryotes, have developed a specialized system to exchange genetic information between homologous chromosomes at meiosis. By passing only half our genetic complement to the next generation, and leaving a mating partner to provide the other half, genetic information is reshuffled. Diploidy also ensures that if the DNA is damaged on one chromosome, its homologous counterpart can be copied by the cell's repair machinery.

Retroviruses seem to share similar concerns. They have developed a sophisticated

- Regev, A., Lamb, M. J. & Jablonka, E. The role of DNA methylation in invertebrates: developmental regulation or genome defence? *Mol. Biol. Evol.* **15**, 880–891 (1998).
- Simmen, M. W. *et al.* Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* **283**, 1164–1167 (1999).
- Lieb, M. & Bhagwat, A. S. Very short patch repair: reducing the cost of cytosine methylation. *Mol. Microbiol.* **20**, 467–473 (1996).
- Tsutakawa, S. E., Jingami, H. & Morikawa, K. Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell* **99**, 615–623 (1999).
- Adams, M. A. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
- Tweedie, S. *et al.* Vestiges of a DNA methylation system in *Drosophila melanogaster*? *Nature Genet.* **23**, 389–390 (1999).
- Shapiro, R. Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. *Proc. Natl Acad. Sci. USA* **96**, 4396–4401 (1999).
- Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
- El-Hajj, H. H., Wang, L. & Weiss, B. Multiple mutant of *Escherichia coli* synthesizing virtually thymineless DNA during limited growth. *J. Bacteriol.* **174**, 4450–4456 (1992).
- Garcia, G. A. & Goodenough-Lashua, D. M. In *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) 135–168 (ASM, Washington DC, 1998).
- El-Hajj, H. H., Zhang, H. & Weiss, B. Lethality of a dut (deoxyuridine triphosphatase) mutation in *Escherichia coli*. *J. Bacteriol.* **170**, 1069–1075 (1988).

Acknowledgements.

This work was supported by the Swedish Natural Science Research Council (to B.M.S) and by the New Zealand Marsden Fund.

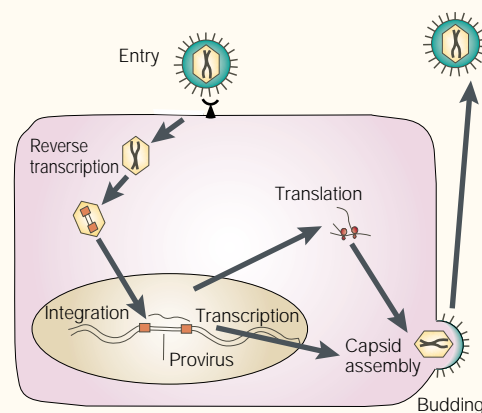


Figure 1 | Reverse transcription in the life cycle of retroviruses. The retroviral infectious cycle. Reverse transcription takes place in the cytoplasm of the host cell within the viral structure called the capsid (hexagon). The reverse transcriptase converts the viral RNA into double-stranded DNA, which integrates in the host genome to form what is called a ‘provirus’. Transcription by the cellular RNA polymerase II generates the viral messenger RNAs, as well as the new genomic RNAs that will be packaged into the budding virions. (Adapted from REF. 34.)