DNA repair genes in the Megavirales pangenome

Romain Blanc-Mathieu[1] and Hiroyuki Ogata[1,*]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan (R.B.M.: roblanc@kuicr.kyoto-u.ac.jp; H.O.: ogata@kuicr.kyoto-u.ac.jp)

* Corresponding author: ogata@kuicr.kyoto-u.ac.jp

Short title: DNA repair genes in Megavirales

***Highlights (for review)**

Megavirales pangenome encodes key enzymes involved in five major DNA repair pathways.

Larger Megavirales genomes tend to encode more genes for DNA repair enzymes than smaller Megavirales genomes in consistence with Drake's rule. This suggests that these enzymes play a crucial role in maintaining the integrity of their large genomes.

A few DNA repair enzymes are inferred to be encoded in the last common ancestors of individual families in the order Megavirales, suggesting that they already had the potential to achieve a relatively large genome compared to extant DNA viruses lacking DNA repair genes.

Individual families of Megavirales gradually increased their genome size by acquiring new repair functions from their hosts through horizontal gene transfer.

Abstract

The order “Megavirales” represent a group of eukaryotic viruses with a large genome encoding a few hundred up to two thousand five hundred genes. Several members of Megavirales possess genes involved in major DNA repair pathways. Some of these genes were likely inherited from an ancient virus world and some others were derived from the genomes of their hosts. Here we examine molecular phylogenies of key DNA repair enzymes in light of recent hypotheses on the origin of Megavirales, and propose that the last common ancestors of the individual families of the order Megavirales already possessed DNA repair functions to achieve and maintain a moderately large genome and that this repair capacity gradually increased, in a family-dependent manner, during their recent evolution.

Introduction

Nucleo-cytoplasmic large DNA viruses (NCLDVs) are a group of viruses that infect diverse members of eukaryotes and possess a large double-stranded DNA genome varying in size from 100 kb to 2.5 Mb [1,2]. They are considered to form a monophyletic group based on the conservation of several genes [3], which led to the recent proposal of the order "Megavirales" to refer to this viral group [4].

Microbiologists and evolutionary biologists have investigated the evolution of Megavirales genes by sequence and molecular phylogenetic analyses and proposed theories on the origin of Megavirales and its association with the emergence of eukaryotes [5-9]. Depending on hypotheses, Megavirales genes can have different origins. A recent hypothesis postulates that Megavirales evolved from ancient DNA transposons of the Polintons family that are themselves the remnant of Tectoviridae-like bacteriophages that entered the protoeukaryotic cell along with the Alphaproteobacterial endosymbiont [6,10]. Under this scenario, Megavirales are the product of a "melting pot" of early viral evolution and part of their genes were directly derived from an ancient virus world as "hallmark viral genes" [11]. Alternatively, Megavirales genes can be acquired from known cellular organisms including viral hosts ("the accretion hypothesis") [12-14], or could have been vertically inherited from unknown and ancestral cellular organisms ("the reductive evolution from the fourth (or more) domain(s) hypothesis") [5,15]. These scenarios are not totally exclusive of each other. For instance, an ancient cellular organism might have

evolved to an ancestor of Megavirales by reductive genome evolution, followed by later re-accumulation of cellular genes in the course of viral evolution.

One of the notable features uncovered from the genomics of Megavirales is that some members abundantly encode DNA repair genes in their genomes [16,17]. DNA repair functions are essential for cellular organisms (i.e., "large genomes") but are infrequent or absent in smaller viral genomes. Certain DNA repair enzymes are known to be specific to a subset of Megavirales members having a particularly large genome [18]. Therefore, Megavirales DNA repair genes are of interest to understand the evolution of their large genomes. In this manuscript, we review the homologs of key DNA repair genes in the pangenome of Megavirales with a special focus on their phylogenetic relationships with eukaryotic homologs.

Classification of Megavirales

The order Megavirales is currently composed of seven families: *Mimiviridae*, *Marseilleviridae*, *Phycodnaviridae*, *Poxviridae*, *Asfarviridae*, *Iridoviridae* and *Ascoviridae* [4]. Following Santini and colleagues [19], instead of *Mimiviridae*, we used the term "Megaviridae" family which includes viruses classified in *Mimiviridae* plus *Phaeocystis globosa virus* (PgV), *Cafeteria roenbergensi virus* (CroV), Organic lake phycodnaviruses (OLPV1 and 2), *Aureococcus anophagefferens virus* (AaV) [20] and *Chrysochromulina ericina virus* (CeV) [21]. The recent discoveries of pandoraviruses (1.9 Mb ~ 2.5 Mb; [22,23]), *Mollivirus sibericum* (650 kb, [24]), *Pithovirus sibericum* (610 kb, [25]), substantially expanded the known diversity of Megavirales lineages [26,27]. These viruses

wait for their official taxonomic classifications; a phylogenetic analysis suggests that pandoraviruses are a group of viruses belonging to the *Phycodnaviridae* family [2].

Why do Megavirales genomes encode many DNA repair genes?

There are five known DNA repair pathways: the base excision repair (BER), nucleotide excision repair (NER), double-strand break (DSB) repair, mismatch repair (MMR), and direct damage reversal (DDR) pathways (**Box 1**). Most of the key enzymes in these DNA repair pathways can be found in the genomes of Megavirales (**Table S1**). Transcriptomic ([24,28-33]) and proteomic ([24,34-37]) studies revealed that many of these DNA repair genes are transcriptionally active during infection and that gene products (i.e., enzymes) are packaged into Megavirales virions (**Table S1**). Individual Megavirales genomes encode 0 up to 15 of those genes (3.6 genes on average). In contrast, other smaller double-stranded DNA viruses encode 0 to 5 of those genes (0.4 genes on average). Redrejo-Rodríguez and Salas provided a comprehensive review on the BER pathway genes in the pangenome of Megavirales [38]. Interestingly, members of the Megaviridae family, which possess a relatively large genome compared to the members of other families, encode the full or almost full set of genes needed to accomplish a viral BER pathway.

Microorganisms (viruses, prokaryotes and unicellular eukaryotes) with a larger genome tend to encode a larger number of genes than those with a smaller genome [39]. As a consequence, the target-size of deleterious mutation is greater for microorganisms with a larger genome. We thus expect a lower mutation rate per base per generation (i.e., a higher fidelity in replication and repair) for microorganisms with a larger genome than with a

smaller genome. Drake postulated that the mutation rate per nucleotide site per generation scales inversely with genome size of DNA-based microorganisms and proposed that the microbial mutation rate per genome may have evolved towards a nearly invariant value across taxa [40]. Subsequent analyses demonstrated that most microorganisms with an estimated mutation rate per nucleotide site per generation conform to the inverse scaling observed by Drake [41,42]. There are currently no data available for the mutation rate for the viruses of Megavirales. However, in line with Drake's rule, the genome size positively correlates with the number of DNA repair genes in Megavirales (Spearman's ρ = 0.4, p = $1.7x10^{-4}$) (**Figure 1**). Exceptions to this trend are the three pandoraviruses encoding only two of the known repair genes that we analyzed. Pandoraviruses may rely on host's DNA repair machinery or they may encode either highly divergent or unidentified DNA repair genes.

DNA repair genes from an ancient virus world

Previous phylogenetic analyses (see references in **Table 1**) suggested an old origin for some of the DNA repair genes encoded in the Megavirales pangenome, which may predate the radiation of major eukaryotic lineages. Most of DNA processing genes are not conserved across the three domains of life. Mre11/Rad50 are among the few DNA processing enzymes that are universally conserved in the three domains [43]. Based on this observation, some authors suggested that the last universal common ancestor (LUCA) was capable of processing DNA [44]. However, Forterre argues against this hypothesis based on the fact that majority of the core enzymes for DNA replication are not homologous between

bacteria and eukaryote/archaea [45], and proposes that LUCA still had an RNA genome and that the descendant cellular lineages (leading to the three domains) acquired DNA processing enzymes including Mre11/Rad50 not from cellular organisms but from viruses [46]. Forterre's scenario assumes the existence of viral ancestors that possessed Mre11/Rad50 genes. Interestingly, some Megavirales members possess Mre11 and/or Rad50 homologs (**Table S1**). Among these, *Acanthamoeba polyphaga mimivirus* encodes a fused version of Mre11 and Rad50 genes, which was shown to be phylogenetically distinct from the canonical version of Mre11/Rad50 that are universally conserved in the three domains of life [47]. The non-canonical gene version was also found in other viruses, plasmids and a few unrelated cellular organisms including diatoms, *Dictyostelium*, *Micromonas*, *Deferribacter* and *Clostridium*. The scattered phylogenetic distribution of the non-canonical type of genes in cellular organisms suggests that the mimivirus Mre11/Rad50 gene was not derived from its host. It is more parsimonious to assume that the ancestral non-canonical gene was encoded in an ancient virus and an ancient virus-cell gene transfer gave birth to the two Mre11/Rad50 lineages (i.e., the canonical cellular and the non-canonical viral lineages) than to hypothesize an ancient gene duplication in a cellular organism followed by independent gene losses in different cellular lineages (**Figure 2**). Although one cannot determine the direction of this ancient gene transfer (either virus-to-cell or cell-to-virus) from the tree topology, the inferred existence of Mre11/Rad50 in an ancestral virus (prior to LUCA) is compatible with Forterre's scenario.

NAD-dependent DNA ligases, found in a minority of Megavirales, form a monophyletic group outside cellular homologs, and may be derived from an ancient virus

world through bacteriophages [48]. In several Megavirales lineages, the ancestral NAD-dependent ligase was displaced by an ATP-dependent ligase from cellular organisms [48,49]. The NAD-dependent DNA ligase genes in PgV and CeV are fused with the coding region of a DNA polymerase of family X (PolX) [21,38].

Apurinic/apyrimidinic (AP) endonucleases of Megaviridae, Marseilleviridae and Entomopoxvirinae also form a monophyletic group, thus being unlikely to be recently emerged in Megavirales [50] (**Figure S1A**). The Entomopoxvirinae AP endonuclease genes are fused with the coding region of PolX [50].

All the currently sequenced Megaviridae genomes encode MutS homologs (MutS7 or MutS8 subfamily) [18]. MutS7 was also found in *Heterocapsa circularisquama DNA virus* (HcDNAV), the mitochondria of octocorals and several Epsilonproteobacteria. HcDNAV (356 kb) is distantly related to *African swine fever virus* based on DNA polymerase sequences [51]. MutS7 shows an atypical domain architecture with an additional C-terminal HNH endonuclease domain. The HNH domain is suggested to have the role of MutH, an endonuclease that introduces a nick in the newly synthesized DNA strand. Outside Megaviridae, MutS8 was only found in '*Candidatus* Amoebophilus asiaticus', an amoeba-associated bacterium. Both of the two MutS subfamilies are only distantly related to other subfamilies such as bacterial MutS1/MutS2 and eukaryotic MutS homologs, suggesting an early divergence of the MutS subfamilies that likely predates the radiation of major eukaryotic lineages.

Horizontal acquisition from eukaryotes

In contrast to the above documented cases illustrating old origins of Megavirales DNA repair genes, there are cases of recent transfers of DNA repair genes from eukaryotes to Megavirales (**Table 1**). For example, the uracil DNA glycosylase (UDG) of pandoraviruses was clearly acquired from their host (**Figure S1B**). The same phylogenetic tree of UDGs also indicates (albeit less clearly) the horizontal acquisition of UDGs in Megaviridae and Entomopoxvirinae. Other similar cases include the 3-methyladenine (3-MeA) DNA glycosylases of *Megavirus lba* and *Megavirus chiliensis* (**Figure S1C**). The xeroderma pigmentosum complementation group G (XPG) protein of *Emiliania huxleyi virus* 86 (EhV-86) is also another case of gene transfer from its host (**Figure S1D**). Photolyases of Poxviridae were also acquired from eukaryotes (**Figure S1E**). Megavirales PolX involved multiple events of acquisition from different cellular organisms including eukaryotes [50] (**Figure S1F**). These observations clearly indicate that Megavirales genomes accumulated numerous DNA repair genes in the course of their evolution after the establishment of the lineages corresponding to the currently recognized viral families. Nonetheless, it should be noted that some of the Megavirales homologs for the above mentioned enzymes do not show evidence of recent gene transfer, including the UDGs of Poxviridae (but Entomopoxvirinae) and Marseilleviridae, 3-MeA DNA glycosylases of pandoraviruses and *Mollivirus*, and XPGs of Poxviridae. Their origins may be resolved in the future with the accumulation of more genomic data. Alternatively, the origin of these genes may be too old to be resolved by current phylogenetic methods.

Conclusions

The pangenome of the order Megavirales encodes most of the key enzymes involved in DNA repair pathways. The number of DNA repair enzymes encoded in individual genomes is positively correlated with the genome size. It should also be noted that all genes in tested Megavirales genomes show a nonsynonymous to synonymous substitution ratio below one, suggesting that most of the genes in the Megavirales pangenome are functional and contribute to virus fitness [52,53]. The repair enzymes are thus needed to maintain a large genome encoding the large number of functional genes as predicted from Drake's model. The last common ancestor of Megavirales encoded a few DNA repair enzymes including the NAD-dependent DNA ligase and the AP endonuclease, which are probably inherited from an ancient virus world. The Megavirales DSB repair genes (Mre11/Rad50) are also likely derived from an ancient virus world predating LUCA. For several other DNA repair genes, their origin can be traced back up to the last common ancestor of viral families (such as MutS7 of Megaviridae and XPG of Poxviridae). These suggest that the Megavirales lineages leading to the last common ancestors of viral families already possessed a relatively large genome encoding a number of genes with important functions that require DNA repair genes for maintenance. In contrast to these genes of ancient origins, many of the other key enzymes in the BER, NER and DDR pathways appear to originate from recent horizontal gene transfers from hosts. Megavirales thus gradually reinforced their DNA repair machinery, which further augments their coding capacity, in the course of more recent evolution of individual viral families. Given the long evolutionary time required to achieve the current distribution of DNA repair genes in Megavirales from its ancestor, the large genome of an extant Megavirales should not be seen as an entity that is

rapidly expanding through rampant acquisition of genes and junk DNA, but may be seen as a functionally constrained entity like the genomes of many other extant living organisms.

Acknowledgement

We are grateful to Dr. Susumu Goto for his valuable comments on an earlier version of our manuscript. This work was in part supported by JSPS KAKENHI (Grant number 26430184), Canon Foundation (Project number 203143100025) and the Collaborative Research Program of Institute for Chemical Research, Kyoto University (Grant number 2015-125). Computation time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University.

References

1. Iyer LM, Aravind L, Koonin EV: **Common origin of four diverse families of large eukaryotic DNA viruses**. *J Virol* 2001, **75**:11720-11734.
2. Yutin N, Koonin EV: **Pandoraviruses are highly derived phycodnaviruses**. *Biol Direct* 2013, **8**:25.
3. Iyer LM, Balaji S, Koonin EV, Aravind L: **Evolutionary genomics of nucleo-cytoplasmic large DNA viruses**. *Virus Res* 2006, **117**:156-184.
4. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, Cheng XW, Federici BA, Van Etten JL, Koonin EV, et al.: **"Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses**. *Arch Virol* 2013, **158**:2517-2521.
5. Claverie JM: **Viruses take center stage in cellular evolution**. *Genome Biol* 2006, **7**:110.
6. Krupovic M, Koonin EV: **Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution**. *Nat Rev Microbiol* 2015, **13**:105-115.

**

An excellent opinion article in which the authors describe the evolutionary relationships among bacterial tectovirus and selfish double-stranded DNA elements of eukaryotes. Authors then described a coherent evolutionary scenario where Polintons were the first group of eukaryotic double-stranded DNA viruses that evolved from bacteriophages. Megavirales, other large eukaryotic DNA viruses and various other selfish genetic elements derived from Polintons.

7. Filee J: **Route of NCLDV evolution: the genomic accordion**. *Curr Opin Virol* 2013, **3**:595-599.
8. Moreira D, Brochier-Armanet C: **Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes**. *BMC Evol Biol* 2008, **8**:12.
9. Takemura M, Yokobori S, Ogata H: **Evolution of Eukaryotic DNA Polymerases via Interaction Between Cells and Large DNA Viruses**. *J Mol Evol* 2015, **81**:24-33.

*

An original research article presenting molecular evidence that suggests DNA polymerase genes evolved through horizontal gene transfer between the viral and archaeal-eukaryotic lineages. This result revives hypotheses regarding the putative role of NCLDVs in the genesis of eukaryotic nucleus.

10. Koonin EV, Krupovic M, Yutin N: **Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses**. *Ann N Y Acad Sci* 2015, **1341**:10-24.

*

This review article described the diversity of double stranded DNA viruses of eukaryotes and assert the hypothesis stating that Megavirales along with transposons and plasmids were derived from Polintonviruses.

11. Koonin EV, Senkevich TG, Dolja VV: **The ancient Virus World and evolution of cells**. *Biol Direct* 2006, **1**:29.
12. Filee J, Siguier P, Chandler M: **I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses**. *Trends Genet* 2007, **23**:10-15.
13. Yutin N, Wolf YI, Koonin EV: **Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life**. *Virology* 2014, **466-467**:38-52.

*

A comprehensive phylogenomic analysis of giant viruses (Mimiviruses, Pithoviruses and Pandoraviruses) in which authors systematically trace the origin of some universal cellular genes present in these viruses to test the fourth domain hypothesis.

14. Moreira D, Lopez-Garcia P: **Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes?** *Philos Trans R Soc Lond B Biol Sci* 2015, **370**:20140327.

*

In this review article authors explain and demonstrate how the high evolutionary rate of viruses can lead to artefactual trees supporting the fourth domain hypothesis.

15. Abergel C, Legendre M, Claverie JM: **The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus**. *FEMS Microbiol Rev* 2015, **39**:779-796.

**

An excellent review article in which authors portray the four families of giant viruses and extend the fourth domain hypothesis to postulate that each giant viral lineage evolved by reduction from unknown domains of protocellular life.

16. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM: **The 1.2-megabase genome sequence of Mimivirus**. *Science* 2004, **306**:1344-1350.
17. Fischer MG, Allen MJ, Wilson WH, Suttle CA: **Giant virus with a remarkable complement of genes infects marine zooplankton**. *Proc Natl Acad Sci U S A* 2010, **107**:19508-19513.
18. Ogata H, Ray J, Toyoda K, Sandaa RA, Nagasaki K, Bratbak G, Claverie JM: **Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment**. *ISME J* 2011, **5**:1143-1151.
19. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, Barbe V, Wommack KE, Noordeloos AA, Brussaard CP, et al.: **Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes**. *Proc Natl Acad Sci U S A* 2013, **110**:10800-10805.
20. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW: **Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution**. *Virology* 2014, **466-467**:60-70.
21. Gallot-Lavallee L, Pagarete A, Legendre M, Santini S, Sandaa RA, Himmelbauer H, Ogata H, Bratbak G, Claverie JM: **The 474-Kilobase-Pair Complete Genome Sequence of CeV-01B, a Virus Infecting Haptolina (Chrysochromulina) ericina (Prymnesiophyceae)**. *Genome Announc* 2015, **3**.
22. Philippe N, Legendre M, Doutre G, Coute Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, et al.: **Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes**. *Science* 2013, **341**:281-286.
23. Antwerpen MH, Georgi E, Zoeller L, Woelfel R, Stoecker K, Scheid P: **Whole-genome sequencing of a pandoravirus isolated from keratitis-inducing acanthamoeba**. *Genome Announc* 2015, **3**.
24. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic JM, Ramus C, Bruley C, Labadie K, et al.: **In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba**. *Proc Natl Acad Sci U S A* 2015, **112**:E5327-5335.
25. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, et al.: **Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology**. *Proc Natl Acad Sci U S A* 2014, **111**:4274-4279.

26. Sharma V, Colson P, Chabrol O, Pontarotti P, Raoult D: **Pithovirus sibericum, a new bona fide member of the "Fourth TRUC" club**. *Front Microbiol* 2015, **6**:722.
27. Sharma V, Colson P, Chabrol O, Scheid P, Pontarotti P, Raoult D: **Welcome to pandoraviruses at the 'Fourth TRUC' club**. *Front Microbiol* 2015, **6**:423.
28. Majji S, Thodima V, Sample R, Whitley D, Deng Y, Mao J, Chinchar VG: **Transcriptome analysis of Frog virus 3, the type species of the genus Ranavirus, family Iridoviridae**. *Virology* 2009, **391**:293-303.
29. Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, Lartigue A, Lescot M, Bernadac A, Poulain J, et al.: **mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus**. *Genome Res* 2010, **20**:664-674.
30. Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova IV, Dunigan DD, Van Etten JL: **Microarray analysis of Paramecium bursaria chlorella virus 1 transcription**. *J Virol* 2010, **84**:532-542.
31. Rubins KH, Hensley LE, Bell GW, Wang C, Lefkowitz EJ, Brown PO, Relman DA: **Comparative analysis of viral gene expression programs during poxvirus infection: a transcriptional map of the vaccinia and monkeypox genomes**. *PLoS One* 2008, **3**:e2628.
32. Assarsson E, Greenbaum JA, Sundstrom M, Schaffer L, Hammond JA, Pasquetto V, Oseroff C, Hendrickson RC, Lefkowitz EJ, Tscharke DC, et al.: **Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes**. *Proc Natl Acad Sci U S A* 2008, **105**:2140-2145.
33. Yang Z, Bruno DP, Martens CA, Porcella SF, Moss B: **Genome-wide analysis of the 5' and 3' ends of vaccinia virus early mRNAs delineates regulatory sequences of annotated and anomalous transcripts**. *J Virol* 2011, **85**:5897-5909.
34. Ince IA, Boeren S, van Oers MM, Vlak JM: **Temporal proteomic analysis and label-free quantification of viral proteins of an invertebrate iridovirus**. *J Gen Virol* 2015, **96**:196-205.
35. Wong CK, Young VL, Kleffmann T, Ward VK: **Genomic and proteomic analysis of invertebrate iridovirus type 9**. *J Virol* 2011, **85**:7900-7911.
36. Renesto P, Abergel C, Decloquement P, Moinier D, Azza S, Ogata H, Fourquet P, Gorvel JP, Claverie JM: **Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products**. *J Virol* 2006, **80**:11678-11685.
37. Fischer MG, Kelly I, Foster LJ, Suttle CA: **The virion of Cafeteria roenbergensis virus (CroV) contains a complex suite of proteins for transcription and DNA repair**. *Virology* 2014, **466-467**:82-94.
38. Redrejo-Rodriguez M, Salas ML: **Repair of base damage and genome maintenance in the nucleo-cytoplasmic large DNA viruses**. *Virus Res* 2014, **179**:12-25.
39. Hou Y, Lin S: **Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes**. *PLoS One* 2009, **4**:e6978.
40. Drake JW: **A constant rate of spontaneous mutation in DNA-based microbes**. *Proc Natl Acad Sci U S A* 1991, **88**:7160-7164.
41. Lynch M: **Evolution of the mutation rate**. *Trends Genet* 2010, **26**:345-352.
42. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M: **Drift-barrier hypothesis and mutation-rate evolution**. *Proc Natl Acad Sci U S A* 2012, **109**:18488-18492.
43. de Souza RF, Iyer LM, Aravind L: **Diversity and evolution of chromatin proteins encoded by DNA viruses**. *Biochim Biophys Acta* 2010, **1799**:302-318.

44. Leipe DD, Aravind L, Koonin EV: **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27**:3389-3401.
45. Forterre P: **Why are there so many diverse replication machineries?** *J Mol Biol* 2013, **425**:4714-4726.
46. Forterre P: **Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain**. *Proc Natl Acad Sci U S A* 2006, **103**:3669-3674.
47. Yoshida T, Claverie JM, Ogata H: **Mimivirus reveals Mre11/Rad50 fusion proteins with a sporadic distribution in eukaryotes, bacteria, viruses and plasmids**. *Virol J* 2011, **8**:427.
48. Yutin N, Koonin EV: **Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity**. *Biol Direct* 2009, **4**:51.
49. Filee J, Pouget N, Chandler M: **Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses**. *BMC Evol Biol* 2008, **8**:320.
50. Yutin N, Koonin EV: **Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes**. *Virol J* 2012, **9**:161.
51. Ogata H, Toyoda K, Tomaru Y, Nakayama N, Shirai Y, Claverie JM, Nagasaki K: **Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus**. *Virol J* 2009, **6**:178.
52. Ogata H, Claverie JM: **Unique genes in giant viruses: regular substitution pattern and anomalously short size**. *Genome Res* 2007, **17**:1353-1361.
53. Doutre G, Philippe N, Abergel C, Claverie JM: **Genome analysis of the first Marseilleviridae representative from Australia indicates that most of its genes contribute to virus fitness**. *J Virol* 2014, **88**:14340-14349.
54. Fromme JC, Verdine GL: **Base excision repair**. *Adv Protein Chem* 2004, **69**:1-41.
55. Kim YJ, Wilson DM, 3rd: **Overview of base excision repair biochemistry**. *Curr Mol Pharmacol* 2012, **5**:3-13.
56. Setlow RB, Carrier WL: **The Disappearance of Thymine Dimers from DNA: An Error-Correcting Mechanism**. *Proc Natl Acad Sci U S A* 1964, **51**:226-231.
57. Costa RM, Chigancas V, Galhardo Rda S, Carvalho H, Menck CF: **The eukaryotic nucleotide excision repair pathway**. *Biochimie* 2003, **85**:1083-1099.
58. Kamarthapu V, Nudler E: **Rethinking transcription coupled DNA repair**. *Curr Opin Microbiol* 2015, **24**:15-20.
59. Habraken Y, Sung P, Prakash L, Prakash S: **Yeast excision repair gene RAD2 encodes a single-stranded DNA endonuclease**. *Nature* 1993, **366**:365-368.
60. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH: **Understanding nucleotide excision repair and its roles in cancer and ageing**. *Nat Rev Mol Cell Biol* 2014, **15**:465-481.
61. Lammens K, Bemeleit DJ, Mockel C, Clausing E, Schele A, Hartung S, Schiller CB, Lucas M, Angermuller C, Soding J, et al.: **The Mre11:Rad50 Structure Shows an ATP-Dependent Molecular Clamp in DNA Double-Strand Break Repair**. *Cell* 2011, **145**:54-66.
62. Kreuzer KN: **Recombination-dependent DNA replication in phage T4**. *Trends Biochem Sci* 2000, **25**:165-173.
63. Storvik KA, Foster PL: **The SMC-like protein complex SbcCD enhances DNA polymerase IV-dependent spontaneous mutation in *Escherichia coli***. *J Bacteriol* 2011, **193**:660-669.
64. Schofield MJ, Hsieh P: **DNA mismatch repair: molecular mechanisms and biological function**. *Annu Rev Microbiol* 2003, **57**:579-608.

65. Iyer RR, Pluciennik A, Burdett V, Modrich PL: **DNA mismatch repair: functions and mechanisms**. *Chem Rev* 2006, **106**:302-323.
66. Eker AP, Quayle C, Chaves I, van der Horst GT: **DNA repair in mammalian cells: Direct DNA damage reversal: elegant solutions for nasty problems**. *Cell Mol Life Sci* 2009, **66**:968-980.
67. Yi C, He C: **DNA repair by reversal of DNA damage**. *Cold Spring Harb Perspect Biol* 2013, **5**:a012575.
68. Eddy SR: **Hidden Markov models**. *Curr Opin Struct Biol* 1996, **6**:361-365.

Box 1: DNA repair pathways

Base Excision Repair (BER)

Most damage to bases in DNA is repaired by the BER pathway [54]. There are several variations in BER, resulting in the replacement of either a single nucleotide or a slightly longer stretch of DNA strand. A typical BER is initiated by any of several DNA glycosylases (uracil DNA glycosylase (UDG), Fpg or 3-methyladenine (3-MeA) DNA glycosylase) to remove an incorrect or damaged substrate base. This creates an abasic site, which is incised by an apurinic/apyrimidinic (AP) endonuclease. A lyase or phosphodiesterase then removes the remaining sugar, leaving a gap that will be filled by a family X DNA polymerase before a DNA ligase seals the nicked DNA strand [55].

Nucleotide Excision Repair (NER)

NER, initially described in bacteria [56] and later shown to be present in all domains of life, is the most versatile DNA repair pathway [57,58]. One of the key enzymes recruited in the NER pathway is the endonuclease of the eukaryotic XPG family, a single-stranded structure-specific DNA endonuclease, which cleaves single-stranded DNA during NER to excise damaged DNA [59,60].

Double-Strand Break (DSB) Repair

DSBs are a major cause of genomic instability [61]. DSBs occur upon exposure of the genome to DNA-damaging agents but also emerge during normal DNA metabolic processes including replication, recombination and viral recombination-dependent DNA

replication [62,63]. Eukaryotic/archaeal Mre11/Rad50 (and the orthologous bacterial SbcD/SbcC) constitute the key machinery that recognizes DSBs and bridges two DNA ends to initiate DSB repair.

Mismatch Repair (MMR)

MMR recognizes and corrects base-base mismatches and small indels introduced during normal replication, leading to 50- to 1000-folds enhancement of replication fidelity [64,65]. In bacteria, MutS (MutS1) recognizes mismatches/indels and MutH, an endonuclease, introduces a nick in the newly synthesized DNA strand to start MMR. MutS homologs are ubiquitous in cellular organisms, and exhibit ancient paralogs with different functions in MMR, recombination or chromosomal stability [64].

Direct Damage Reversal (DDR)

In contrast to the sophisticated BER, NER, DSB and MMR pathways, DDR makes use of a single protein to remove lesions without incision of the sugar-phosphate backbone or base excision [66]. The reversing of UV light-induced photolesions by photolyases is one of the three major DNA repair mechanisms by DDR that have been identified to date [67].

Figure Legends

Figure 1. Relationship between the size of genome and the number of DNA repair genes in Megavirales. Only key enzymes involved in major DNA repair pathways are considered. DNA repair proteins were first identified in annotated Megavirales genomes through keyword searches. The resulting sequences were used to build hidden Markov model (HMM) profiles. These profiles, plus Pfam profiles for DNA repair domains, were searched against Megavirales proteomes using hmmsearch [68] to identify homologs (E-value < 0.01, score > 100, and score > 5 times the bias composition correction score).

Figure 2. A simplified phylogenetic tree showing two lineages of Mre11/Rad50 homologs and possible scenarios for the origin of the two lineages. (A) A scenario that assumes an ancestral gene duplication in a cellular organism (prior to the last universal common ancestor, LUCA). In this scenario, the two canonical and non-canonical gene lineages emerged through a gene duplication in an ancient cellular organism. The presence of the non-canonical gene in viral lineages and plasmids are explained by gene transfer from cells to viruses. This scenario is not parsimonious, because it has to assume numerous and independent losses of non-canonical genes in different cellular lineages. (B) A scenario that assumes the presence of Mre11/Rad50 in an ancient viral lineage. The ancient viral lineage may be the origin of the canonical gene lineages in cellular organisms (i.e., transfers of the viral genes to cells in agreement with Forterre's evolutionary model). In this scenario, the presence of the non-canonical genes in several cellular organisms can be readily explained by additional and later gene transfers from viruses to cells.

Table 1. Family distribution and inferred evolutionary origin of selected Megavirales DNA repair genes.

Repair pathway	Proteins	Viral families	Recent acquisition from hosts	Non-eukaryotic origin	References
BER	UDG	Mars, Mega, Pando, Pox	Ento, Pando,	Mars, Mega*, Pox (but Ento),	Supplementary figure S1B
	3-MeA DNA glycosylases	Mega, Molli, Pando,	Mega	Molli, Pando	Supplementary figure S1C
	AP endonuclease	Mars, Mega, Ento	-	Mars, Mega, Ento	[50]; Supplementary figure S1A
	PolX	Mega, MpV1, Ento	Mega (but AaV, CeV, PgV)	Ento, MpV1, AaV, CeV, PgV	[50]; Supplementary figure S1F
	NAD-dependent ligase	Ento, Irido, Mega	-	Yes	[50]
	ATP-dependent ligase	APMVB, Asf, Mars, Phyco, Pitho, Pox	Yes	-	[50]
NER	XPG	Asco, EhV, Irido, Mars, Mega, Pox	EhV, Mega,	Asco, Irido, Mars, Pox	Supplementary figure S1D
DSB	Mre11/Rad50	Asco, Irido, Mega	-	Asco, Irido, Mega	[47]
MMR	MutS7 and 8	Mega	-	Mega	[18]
DDR	Photolyases	Mega, Pox	Mega*, Pox	-	[50]; Supplementary figure S1E

Names with an asterisk correspond to cases where the evolutionary origins of genes were not clear from phylogeny. Abbreviations: Asco, *Ascoviridae*; Asf, *Asfarviridae*; Irido, *Iridoviridae*; Mars, *Marseilleviridae*; Mega, Megaviridae; Phyco, *Phycodnaviridae*; Pox, *Poxviridae*; AaV, *Aureococcus anophagefferens virus*; APMVB, *Acanthamoeba polyphaga moumouvirus*; CeV, *Chrysochromulina ericina virus*; EhV, *Emiliania huxleyi virus*; Ento, *Entomopoxvirinae*; Molli, *Mollivirus sibericum*; MpV1, *Micromonas sp. RCC1109 virus* MpV1; Pando, pandoraviruses; PgV, *Phaeocystis globosa virus*; Pitho, *Pithovirus sibericum*.

Figure 1

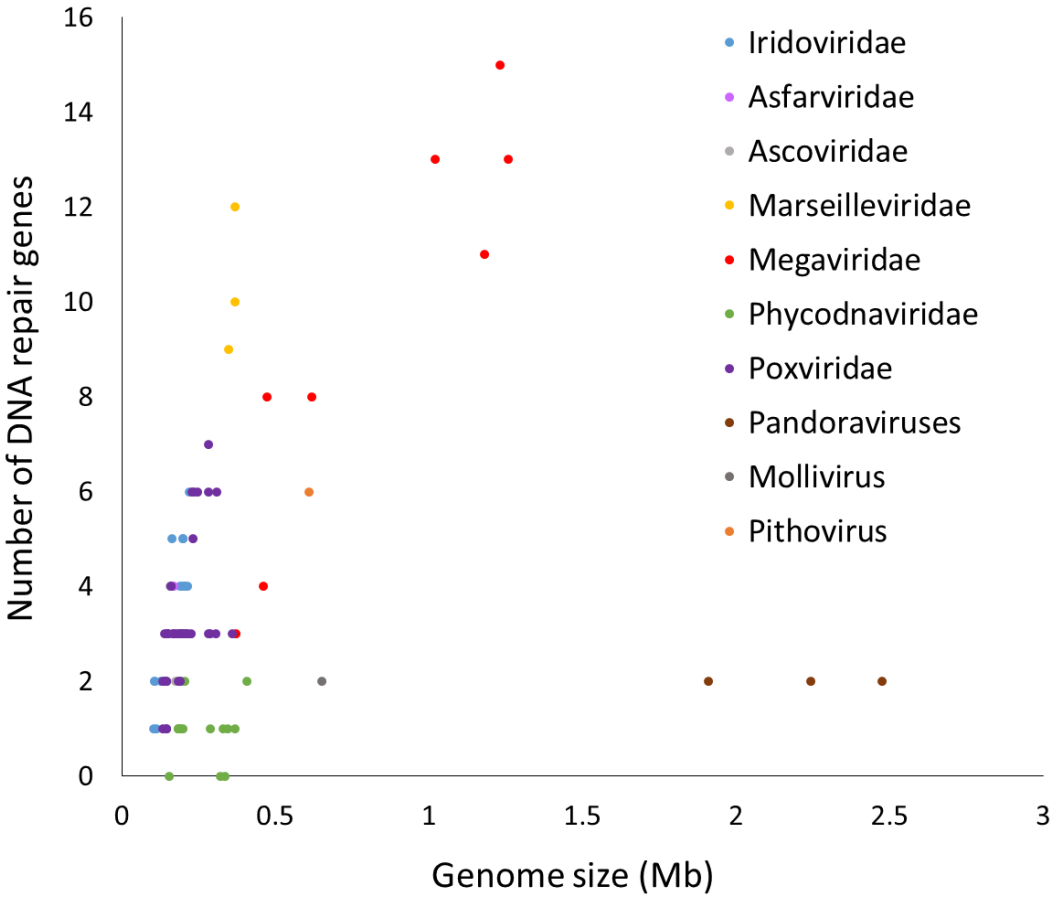

16
14
12
10
8
6
4
2
0
Number of DNA repair genes
0
0.5
1
1.5
2
2.5
3
Genome size (Mb)
Iridoviridae
Asfarviridae
Ascoviridae
Marseilleviridae
Megaviridae
Phycodnaviridae
Poxviridae
Pandoraviruses
Mollivirus
Pithovirus

Figure 2

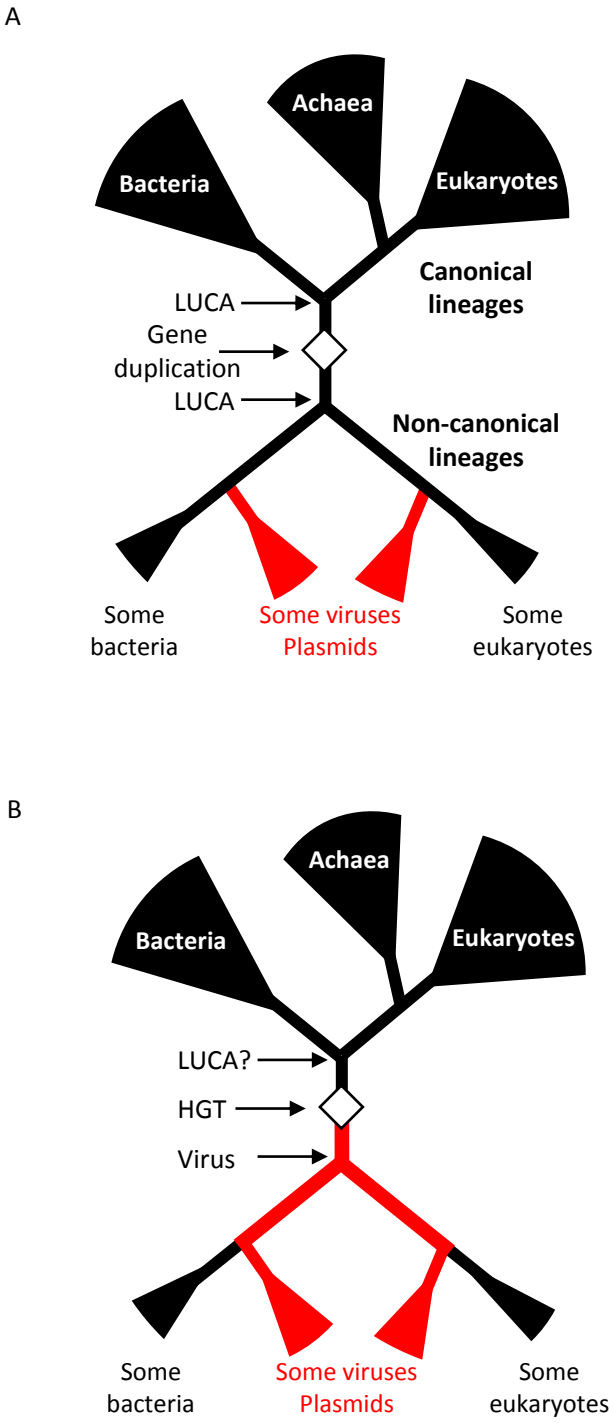

A
Achaea
Bacteria
Eukaryotes
Canonical lineages
LUCA
Gene duplication
LUCA
Non-canonical lineages
Some bacteria
Some viruses Plasmids
Some eukaryotes
B
Achaea
Bacteria
Eukaryotes
LUCA?
HGT
Virus
Some bacteria
Some viruses Plasmids
Some eukaryotes

Table S1

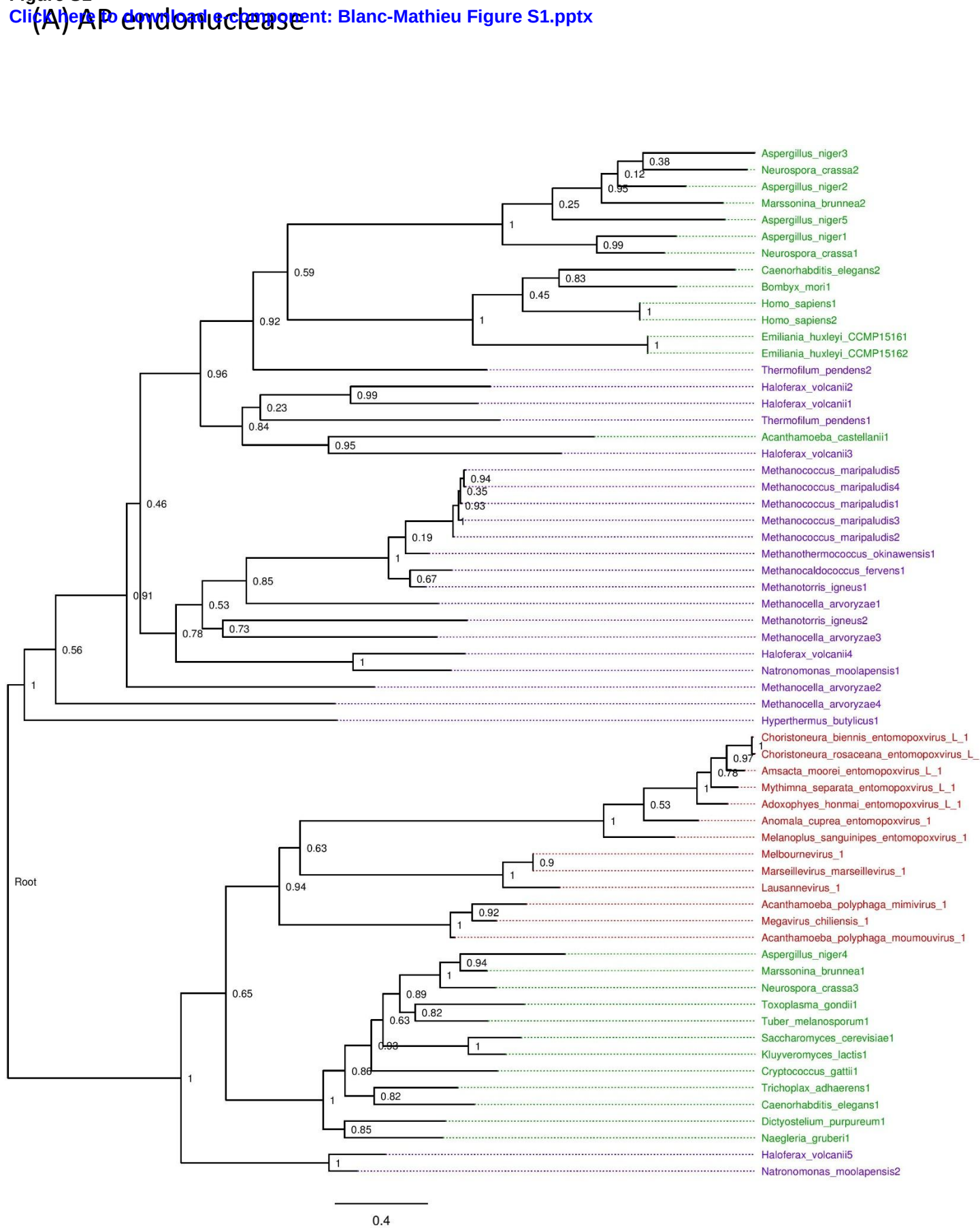
Click here to download e-component: Blanc-Mathieu Table S1 REVISED.xlsx

Table S1. Number of DNA repair genes in Megavirales.

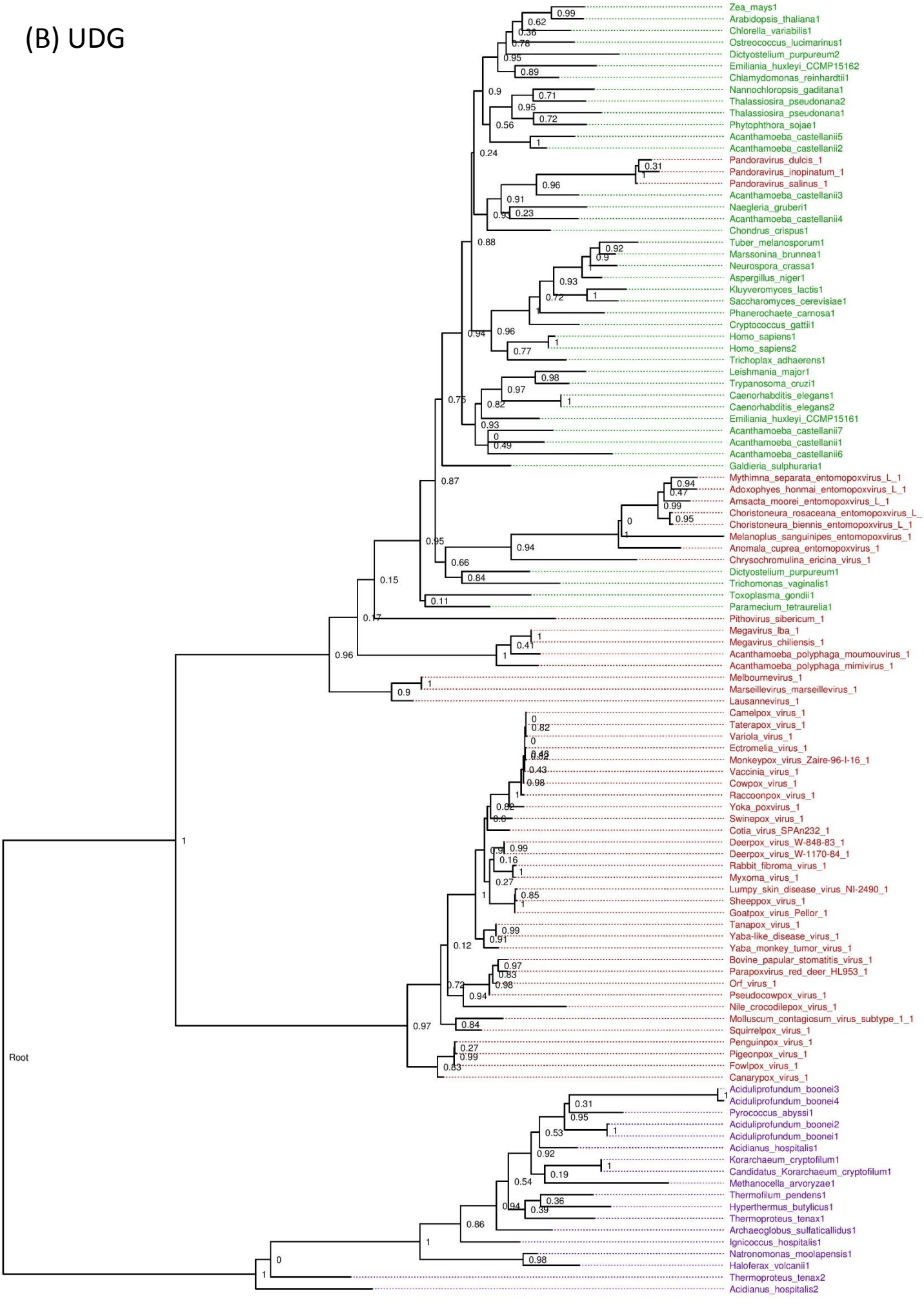
Virus name	Family	ID
African swine fever virus	Asfarviridae	NC_001659
Heliothis virescens ascovirus 3e	Ascoviridae	NC_009233
Spodoptera frugiperda ascovirus 1a	Ascoviridae	NC_008361
Trichoplusia ni ascovirus 2c	Ascoviridae	NC_008518
Ambystoma tigrinum virus	Iridoviridae	NC_005832
Anopheles minimus irodovirus	Iridoviridae	NC_023848
Armadillidium vulgare iridescent virus	Iridoviridae	NC_024451
European catfish virus	Iridoviridae	NC_017940
Frog virus 3	Iridoviridae	NC_005946
Infectious spleen and kidney necrosis virus	Iridoviridae	NC_003494
Invertebrate iridescent virus 22	Iridoviridae	NC_023615
Invertebrate iridescent virus 3	Iridoviridae	NC_008187
Invertebrate iridescent virus 30	Iridoviridae	NC_023611
Invertebrate iridescent virus 6	Iridoviridae	NC_003038
Invertebrate iridovirus 22	Iridoviridae	NC_021901
Invertebrate iridovirus 25	Iridoviridae	NC_023613
Lymphocystis disease virus - isolate China	Iridoviridae	NC_005902
Lymphocystis disease virus 1	Iridoviridae	NC_001824
Singapore grouper iridovirus	Iridoviridae	NC_006549
Wiseana iridescent virus	Iridoviridae	NC_015780
Lausannevirus	Marseilleviridae	NC_015326
Marseillevirus marseillevirus	Marseilleviridae	NC_013756
Melbournevirus	Marseilleviridae	NC_025412
Acanthamoeba polyphaga mimivirus	Megaviridae	NC_014649
Acanthamoeba polyphaga moumouvirus	Megaviridae	NC_020104
Aureococcus anophagefferens virus	Megaviridae	NC_024697
Cafeteria roenbergensis virus BV-PW1	Megaviridae	NC_014637
Chrysochromulina ericina virus	Megaviridae	NC_028094
Megavirus chiliensis	Megaviridae	NC_016072
Megavirus lba	Megaviridae	NC_020232
Phaeocystis globosa virus	Megaviridae	NC_021312
Acanthocystis turfacea Chlorella virus 1	Phycodnaviridae	NC_008724
Bathycoccus sp. RCC1105 virus BpV1	Phycodnaviridae	NC_014765
Ectocarpus siliculosus virus 1	Phycodnaviridae	NC_002687
Emiliania huxleyi virus 86	Phycodnaviridae	NC_007346
Feldmannia species virus	Phycodnaviridae	NC_011183
Micromonas pusilla virus 12T	Phycodnaviridae	NC_020864
Micromonas sp. RCC1109 virus MpV1	Phycodnaviridae	NC_014767
Ostreococcus lucimarinus virus 1	Phycodnaviridae	NC_014766
Ostreococcus lucimarinus virus OlV5	Phycodnaviridae	NC_020852
Ostreococcus tauri virus 1	Phycodnaviridae	NC_013288
Ostreococcus tauri virus 2	Phycodnaviridae	NC_014789
Ostreococcus virus OsV5	Phycodnaviridae	NC_010191
Paramecium bursaria Chlorella virus 1	Phycodnaviridae	NC_000852

Figure S1

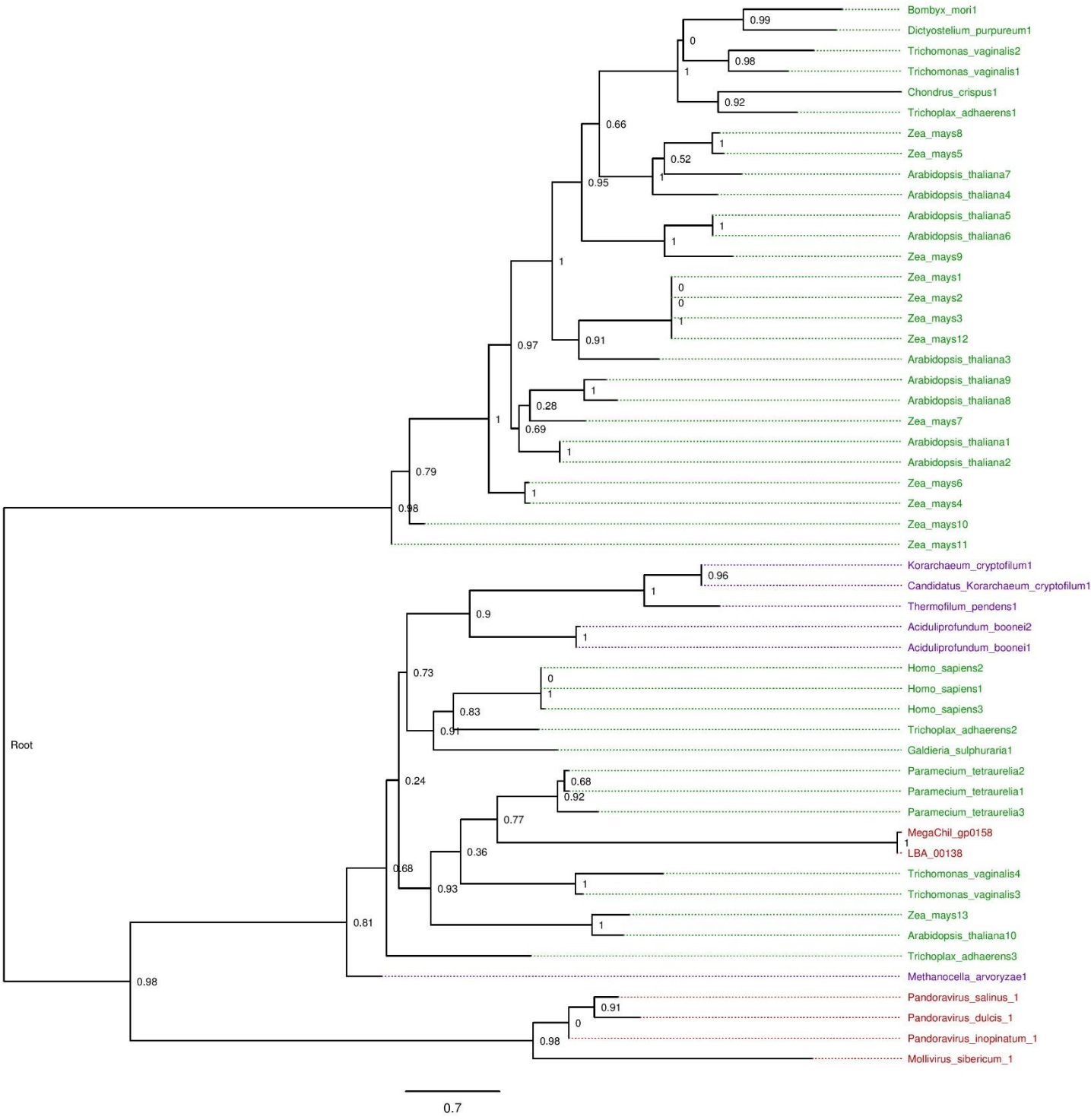
Click here to download e component: Blanc-Mathieu Figure S1.pptx



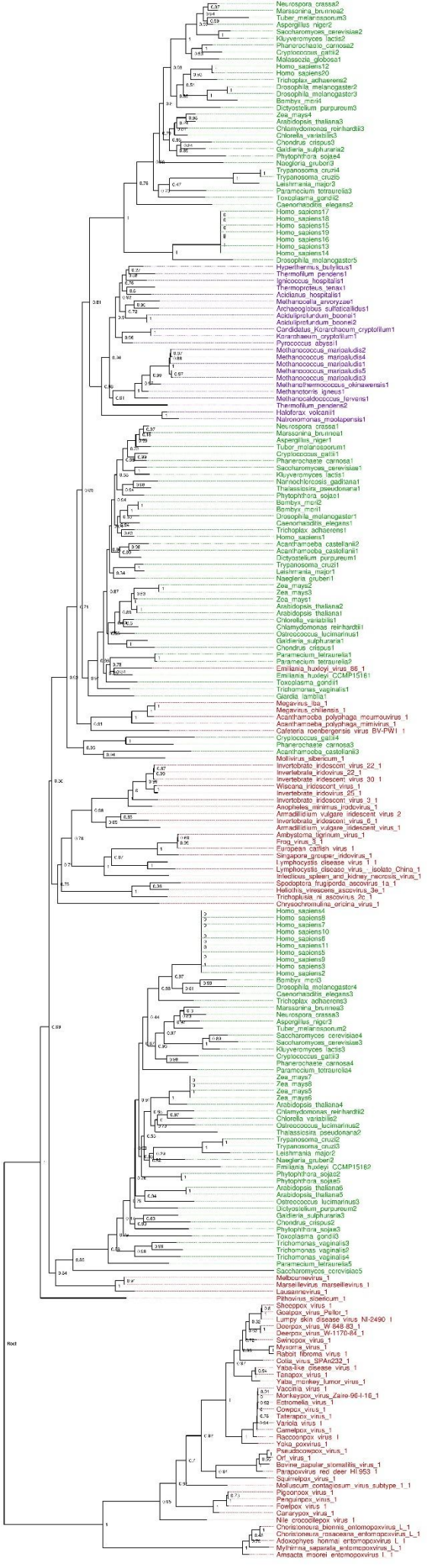
(B) UDG


Zea_mays1
Arabidopsis_thaliana1
Chlorella_variabilis1
Ostreococcus_lucimarinus1
Dictyostelium_purpureum2
Emiliania_huxleyi_CCMP15162
Chlamydomonas_reinhardtii1
Nannochloropsis_gaditana1
Thalassiosira_pseudonana2
Thalassiosira_pseudonana1
Phytophthora_sojae1
Acanthamoeba_castellanii5
Acanthamoeba_castellanii2
Pandoravirus_dulcis_1
Pandoravirus_inopinatum_1
Pandoravirus_salinus_1
Acanthamoeba_castellanii3
Naegleria_gruberi1
Acanthamoeba_castellanii4
Chondrus_crispus1
Tuber_melanosporum1
Marssonina_brunnea1
Neurospora_crassa1
Aspergillus_niger1
Kluyveromyces_lactis1
Saccharomyces_cerevisiae1
Phanerochaete_carnosa1
Cryptococcus_gattii1
Homo_sapiens1
Homo_sapiens2
Trichoplax_adhaerens1
Leishmania_major1
Trypanosoma_cruzi1
Caenorhabditis_elegans1
Caenorhabditis_elegans2
Emiliania_huxleyi_CCMP15161
Acanthamoeba_castellanii7
Acanthamoeba_castellanii1
Acanthamoeba_castellanii6
Galdieria_sulphuraria1
Mythimna_separata_entomopoxvirus_L_1
Adoxophyes_honmai_entomopoxvirus_L_1
Amsacta_moorei_entomopoxvirus_L_1
Choristoneura_rosaceana_entomopoxvirus_L_
Choristoneura_biennis_entomopoxvirus_L_1
Melanoplus_sanguinipes_entomopoxvirus_1
Anomala_cuprea_entomopoxvirus_1
Chrysochromulina_ericina_virus_1
Dictyostelium_purpureum1
Trichomonas_vaginalis1
Toxoplasma_gondii1
Paramecium_tetraurelia1
Pithovirus_sibericum_1
Megavirus_lba_1
Megavirus_chiliensis_1
Acanthamoeba_polyphaga_moumouvirus_1
Acanthamoeba_polyphaga_mimivirus_1
Melbournevirus_1
Marseillevirus_marseillevirus_1
Lausannevirus_1
Camelpox_virus_1
Taterapox_virus_1
Variola_virus_1
Ectromelia_virus_1
Monkeypox_virus_Zaire-96-I-16_1
Vaccinia_virus_1
Cowpox_virus_1
Raccoonpox_virus_1
Yoka_poxvirus_1
Swinepox_virus_1
Cotia_virus_SPAn232_1
Deerpox_virus_W-848-83_1
Deerpox_virus_W-1170-84_1
Rabbit_fibroma_virus_1
Myxoma_virus_1
Lumpy_skin_disease_virus_NI-2490_1
Sheeppox_virus_1
Goatpox_virus_Pellor_1
Tanapox_virus_1
Yaba-like_disease_virus_1
Yaba_monkey_tumor_virus_1
Bovine_papular_stomatitis_virus_1
Parapoxvirus_red_deer_HL953_1
Orf_virus_1
Pseudocowpox_virus_1
Nile_crocodilepox_virus_1
Molluscum_contagiosum_virus_subtype_1_1
Squirrelpox_virus_1
Penguinpox_virus_1
Pigeonpox_virus_1
Fowlpox_virus_1
Canarypox_virus_1
Aciduliprofundum_boonei3
Aciduliprofundum_boonei4
Pyrococcus_abyssi1
Aciduliprofundum_boonei2
Aciduliprofundum_boonei1
Acidianus_hospitalis1
Korarchaeum_cryptofilum1
Candidatus_Korarchaeum_cryptofilum1
Methanocella_arvoryzae1
Thermofilum_pendens1
Hyperthermus_butylicus1
Thermoproteus_tenax1
Archaeoglobus_sulfaticallidus1
Ignicoccus_hospitalis1
Natronomonas_moolapensis1
Haloferax_volcanii1
Thermoproteus_tenax2
Acidianus_hospitalis2
Root
0.9

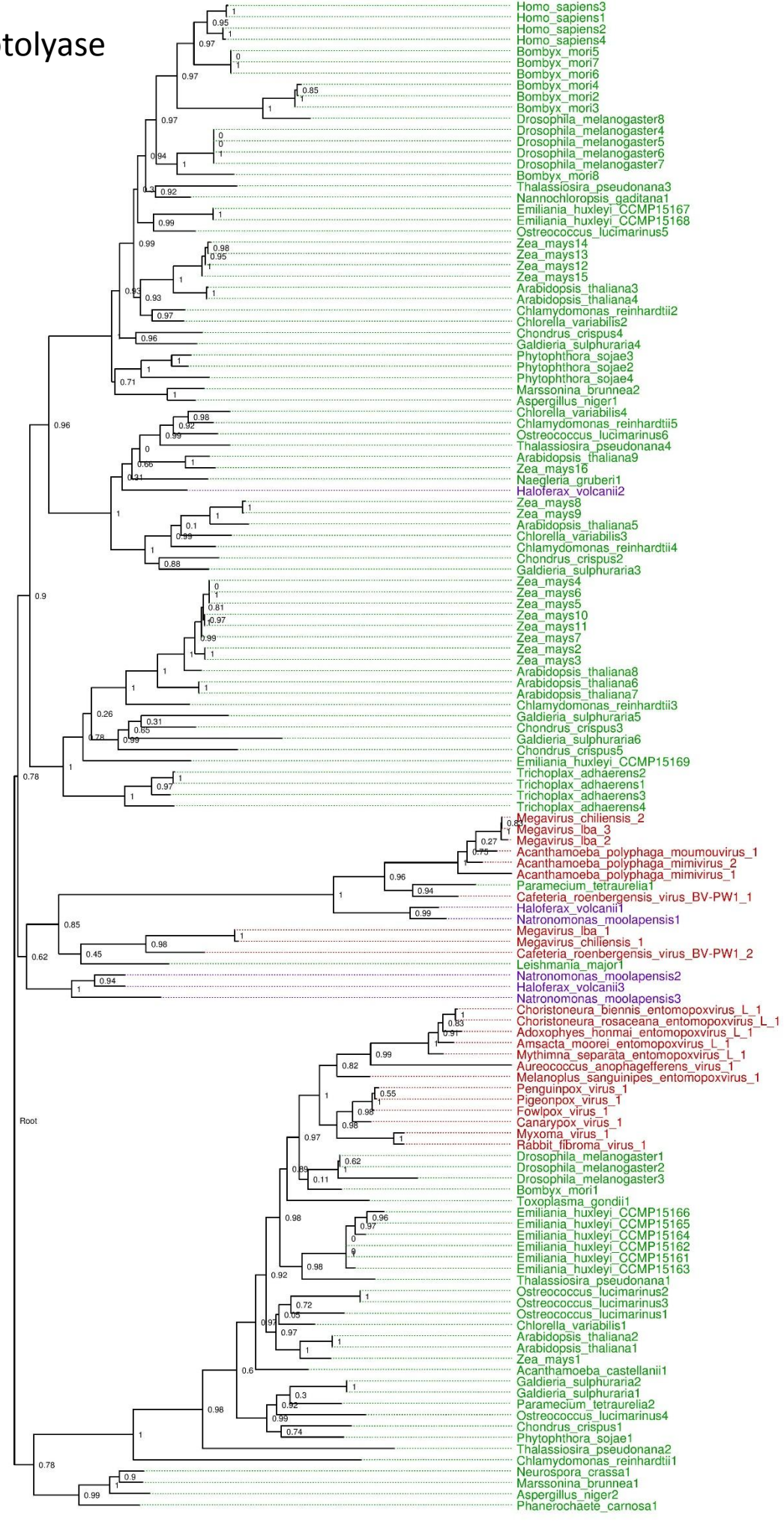
(C) 3-MeA DNA glycosylase



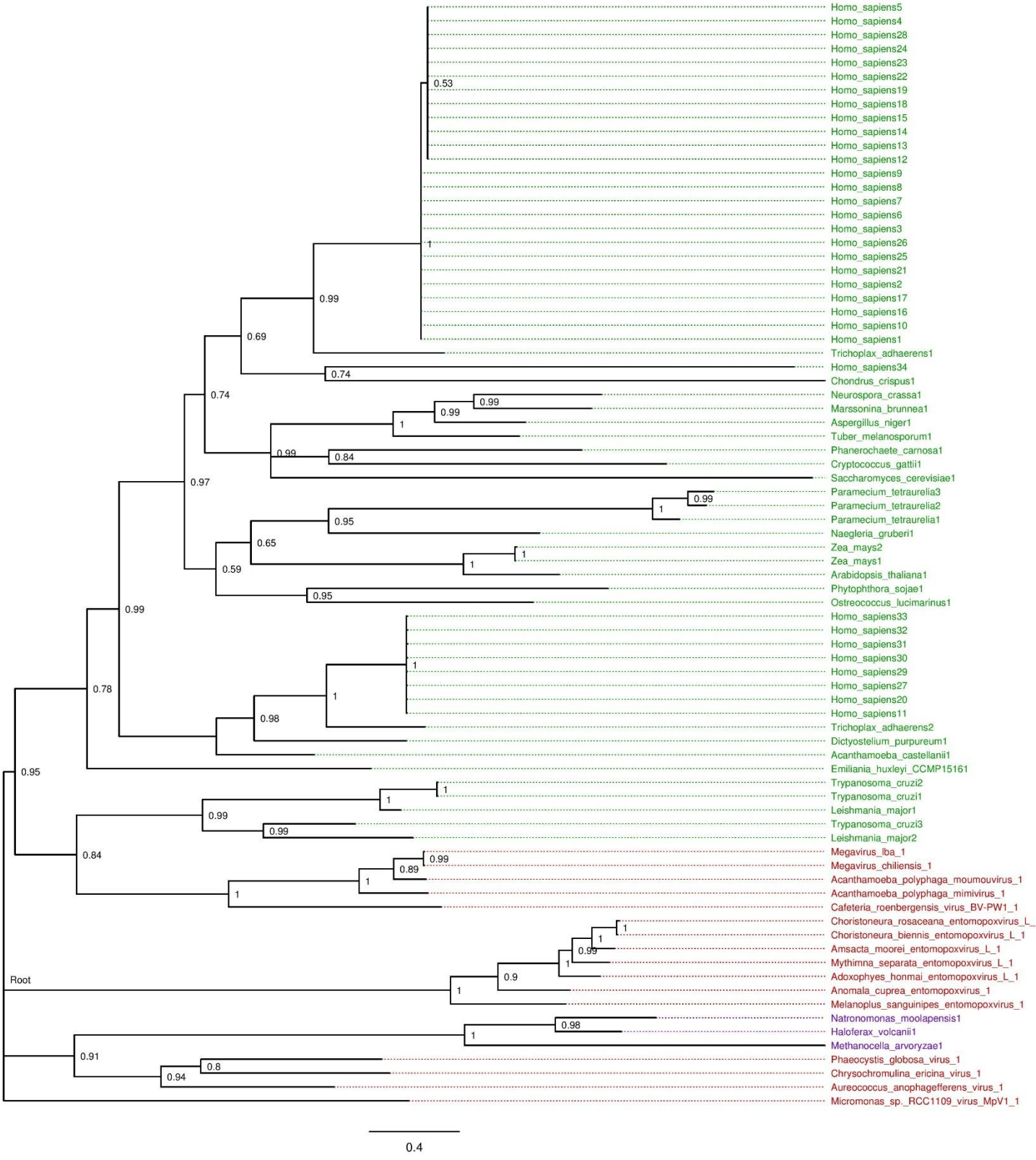
(D) XPG


Neurospora_crassa2
Marssonina_brunnea2
Tuber_melanosporum3
Aspergillus_niger2
Saccharomyces_cerevisiae2
Kluyveromyces_lactis2
Phanerochaete_carnosa2
Cryptococcus_gattii2
Malassezia_globosa1
Homo_sapiens12
Homo_sapiens20
Trichoplax_adhaerens2
Drosophila_melanogaster2
Drosophila_melanogaster3
Bombyx_mori4
Dictyostelium_purpureum3
Zea_mays4
Arabidopsis_thaliana3
Chlamydomonas_reinhardtii3
Chlorella_variabilis3
Chondrus_crispus3
Galdieria_sulphuraria2
Phytophthora_sojae4
Naegleria_gruberi3
Trypanosoma_cruzi4
Trypanosoma_cruzi5
Leishmania_major3
Paramecium_tetraurelia3
Toxoplasma_gondii2
Caenorhaditis_elegans2
Homo_sapiens17
Homo_sapiens18
Homo_sapiens15
Homo_sapiens19
Homo_sapiens16
Homo_sapiens13
Homo_sapiens14
Drosophila_melanogaster5
Hyperthermus_butylicus1
Thermofilum_pendens1
Ignicoccus_hospitalis1
Thermoproteus_tenax1
Acidianus_hospitalis1
Methanocella_arvoryzae1
Archaeoglobus_sulfaticallidus1
Acidulprofundum_boonei1
Acidulprofundum_boonei2
Candidatus_Korarchaeum_cryptofilum1
Korarchaeum_cryptofilum1
Pyrococcus_abyssi1
Methanococcus_maripaludis2
Methanococcus_maripaludis4
Methanococcus_maripaludis1
Methanococcus_maripaludis5
Methanococcus_maripaludis3
Methanothermococcus_okinawensis1
Methanotorris_igneus1
Methanocaldococcus_fervens1
Thermofilum_pendens2
Haloferax_volcanii1
Natronomonas_moolapensis1
Neurospora_crassa1
Marssonina_brunnea1
Aspergillus_niger1
Tuber_melanosporum1
Cryptococcus_gattii1
Phanerochaete_carnosa1
Saccharomyces_cerevisiae1
Kluyveromyces_lactis1
Nannochloropsis_gaditana1
Thalassiosira_pseudonana1
Phytophthora_sojae1
Bombyx_mori2
Bombyx_mori1
Drosophila_melanogaster1
Caenorhaditis_elegans1
Trichoplax_adhaerens1
Homo_sapiens1
Acanthamoeba_castellanii2
Acanthamoeba_castellanii1
Dictyostelium_purpureum1
Trypanosoma_cruzi1
Leishmania_major1
Naegleria_gruberi1
Zea_mays2
Zea_mays3
Zea_mays1
Arabidopsis_thaliana2
Arabidopsis_thaliana1
Chlorella_variabilis1
Chlamydomonas_reinhardtii1
Ostreococcus_lucimarinus1
Galdieria_sulphuraria1
Chondrus_crispus1
Paramecium_tetraurelia1
Paramecium_tetraurelia2
Emiliania_huxleyi_virus_86_1
Emiliania_huxleyi_CCMP1516_1
Toxoplasma_gondii1
Trichomonas_vaginalis1
Giardia_lamblia1
Megavirus_lba_1
Megavirus_chilensis_1
Acanthamoeba_polyphaga_moumouvirus_1
Acanthamoeba_polyphaga_mimivirus_1
Cafeteria_roenbergensis_virus_BV-PW1_1
Cryptococcus_gattii4
Phanerochaete_carnosa3
Acanthamoeba_castellanii3
Mollivirus_sibericum_1
Invertebrate_iridescent_virus_22_1
Invertebrate_iridovirus_22_1
Invertebrate_iridescent_virus_30_1
Wiseana_iridescent_virus_1
Invertebrate_iridovirus_25_1
Invertebrate_iridescent_virus_3_1
Anopheles_minimus_iridovirus_1
Armadillidium_vulgare_iridescent_virus_2
Invertebrate_iridescent_virus_6_1
Armadillidium_vulgare_iridescent_virus_1
Ambystoma_tigrinum_virus_1
Frog_virus_3_1
European_catfish_virus_1
Singapore_grouper_iridovirus_1
Lymphocystis_disease_virus_1_1
Lymphocystis_disease_virus__isolate_China_1
Infectious_spleen_and_kidney_necrosis_virus_1
Spodoptera_frugiperda_ascovirus_1a_1
Heliothis_virescens_ascovirus_3e_1
Trichoplusia_ni_ascovirus_2c_1
Chrysochromulina_ericina_virus_1
Homo_sapiens4
Homo_sapiens8
Homo_sapiens7
Homo_sapiens10
Homo_sapiens6
Homo_sapiens11
Homo_sapiens5
Homo_sapiens9
Homo_sapiens3
Homo_sapiens2
Bombyx_mori3
Drosophila_melanogaster4
Caenorhaditis_elegans3
Trichoplax_adhaerens3
Marssonina_brunnea3
Neurospora_crassa3
Aspergillus_niger3
Tuber_melanosporum2
Saccharomyces_cerevisiae4
Saccharomyces_cerevisiae3
Kluyveromyces_lactis3
Cryptococcus_gattii3
Phanerochaete_carnosa4
Paramecium_tetraurelia4
Zea_mays7
Zea_mays8
Zea_mays5
Zea_mays6
Arabidopsis_thaliana4
Chlamydomonas_reinhardtii2
Chlorella_variabilis2
Ostreococcus_lucimarinus2
Thalassiosira_pseudonana2
Trypanosoma_cruzi2
Trypanosoma_cruzi3
Leishmania_major2
Naegleria_gruberi2
Emiliania_huxleyi_CCMP1516_2
Phytophthora_sojae2
Phytophthora_sojae5
Arabidopsis_thaliana6
Arabidopsis_thaliana5
Ostreococcus_lucimarinus3
Dictyostelium_purpureum2
Galdieria_sulphuraria3
Chondrus_crispus2
Phytophthora_sojae3
Toxoplasma_gondii3
Trichomonas_vaginalis3
Trichomonas_vaginalis2
Trichomonas_vaginalis4
Paramecium_tetraurelia5
Saccharomyces_cerevisiae5
Melbournevirus_1
Marseillevirus_marseillevirus_1
Lausannevirus_1
Pithovirus_sibericum_1
Sheeppox_virus_1
Goatpox_virus_Pellor_1
Lumpy_skin_disease_virus_NI-2490_1
Deerpox_virus_W-848-83_1
Deerpox_virus_W-1170-84_1
Swinepox_virus_1
Myxoma_virus_1
Rabbit_fibroma_virus_1
Cotia_virus_SPAn232_1
Yaba-like_disease_virus_1
Tanapox_virus_1
Yaba_monkey_tumor_virus_1
Vaccinia_virus_1
Monkeypox_virus_Zaire-96-I-16_1
Ectromelia_virus_1
Cowpox_virus_1
Taterapox_virus_1
Variola_virus_1
Camelpox_virus_1
Raccoonpox_virus_1
Yoka_poxvirus_1
Pseudocowpox_virus_1
Orf_virus_1
Bovine_papular_stomatitis_virus_1
Parapoxvirus_red_deer_NZ1953_1
Squirrelpox_virus_1
Molluscum_contagiosum_virus_subtype_1_1
Pigeonpox_virus_1
Penguinpox_virus_1
Fowlpox_virus_1
Canarypox_virus_1
Nile_crocodilepox_virus_1
Chorioboura_blomnis_entomopoxvirus_L_1
Choristoneura_rosaceana_entomopoxvirus_L_1
Adoxophyes_honmai_entomopoxvirus_L_1
Mythimna_separata_entomopoxvirus_L_1
Amsacta_moorei_entomopoxvirus_L_1
Root
0.7

(E) Photolyase



(F) PolX


Homo_sapiens5
Homo_sapiens4
Homo_sapiens28
Homo_sapiens24
Homo_sapiens23
Homo_sapiens22
Homo_sapiens19
Homo_sapiens18
Homo_sapiens15
Homo_sapiens14
Homo_sapiens13
Homo_sapiens12
Homo_sapiens9
Homo_sapiens8
Homo_sapiens7
Homo_sapiens6
Homo_sapiens3
Homo_sapiens26
Homo_sapiens25
Homo_sapiens21
Homo_sapiens2
Homo_sapiens17
Homo_sapiens16
Homo_sapiens10
Homo_sapiens1
Trichoplax_adhaerens1
Homo_sapiens34
Chondrus_crispus1
Neurospora_crassa1
Marssonina_brunnea1
Aspergillus_niger1
Tuber_melanosporum1
Phanerochaete_carnosa1
Cryptococcus_gattii1
Saccharomyces_cerevisiae1
Paramecium_tetraurelia3
Paramecium_tetraurelia2
Paramecium_tetraurelia1
Naegleria_gruberi1
Zea_mays2
Zea_mays1
Arabidopsis_thaliana1
Phytophthora_sojae1
Ostreococcus_lucimarinus1
Homo_sapiens33
Homo_sapiens32
Homo_sapiens31
Homo_sapiens30
Homo_sapiens29
Homo_sapiens27
Homo_sapiens20
Homo_sapiens11
Trichoplax_adhaerens2
Dictyostelium_purpureum1
Acanthamoeba_castellanii1
Emiliania_huxleyi_CCMP15161
Trypanosoma_cruzi2
Trypanosoma_cruzi1
Leishmania_major1
Trypanosoma_cruzi3
Leishmania_major2
Megavirus_lba_1
Megavirus_chiliensis_1
Acanthamoeba_polyphaga_moumouvirus_1
Acanthamoeba_polyphaga_mimivirus_1
Cafeteria_roenbergensis_virus_BV-PW1_1
Choristoneura_rosaceana_entomopoxvirus_L_
Choristoneura_biennis_entomopoxvirus_L_1
Amsacta_moorei_entomopoxvirus_L_1
Mythimna_separata_entomopoxvirus_L_1
Adoxophyes_honmai_entomopoxvirus_L_1
Anomala_cuprea_entomopoxvirus_1
Melanoplus_sanguinipes_entomopoxvirus_1
Natronomonas_moolapensis1
Haloferax_volcanii1
Methanocella_arvoryzae1
Phaeocystis_globosa_virus_1
Chrysochromulina_ericina_virus_1
Aureococcus_anophagefferens_virus_1
Micromonas_sp._RCC1109_virus_MpV1_1
Root
0.4

Supplementary Figure S1. Phylogenetic trees of amino acid sequences corresponding to (**A**) AP endonuclease, (**B**) UDG, (**C**) 3-MeA DNA glycosylase, (**D**) XPG, (**E**) photolyase and (**F**) PolX. Sequences from Megavirales, eukaryotes and archaea are included. B,C,D and E are maximum likelihood phylogenies reconstructed using the *ete toolkit* package (Huerta-Cepas et al. 2010) with the workflow "*mafft_ensi-trimal02-prottest_default-raxml*". The models selected were PROTGAMMALG (B,D) and PROTGAMMAVT (C,E). A and F are calculated by Bayesian inference with the CAT-GTR infinite mixture model (Lartillot and Philippe 2004) (4 chains were used and stopped when the "*maxdiff*" was less than 0.1; 1000 first trees were burned and the consensus phylogeny was build using every remaining trees). Statistical support at node is given as SH-like values (B,C,D and E) or posterior probabilities (A and F). Number at scale bar indicate the number of substitutions per site. These phylogenetic trees are midpoint rooted. Red: Megavirales, Green: Eukaryotes, Purple: Archaea.

References:

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:1.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.