



A puzzling anomaly in the 4-mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process

Olivier Poirot, Sandra Jeudy, Chantal Abergel, Jean-Michel Claverie

► To cite this version:

Olivier Poirot, Sandra Jeudy, Chantal Abergel, Jean-Michel Claverie. A puzzling anomaly in the 4-mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process. Journal of Virology, 2019, <10.1128/JVI.01206-19>. <hal-02314004>

HAL Id: hal-02314004

<https://amu.hal.science/hal-02314004v1>

Submitted on 11 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

A puzzling anomaly in the 4-mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process

Olivier Poirot, Sandra Jeudy, Chantal Abergel, Jean-Michel Claverie

► To cite this version:

Olivier Poirot, Sandra Jeudy, Chantal Abergel, Jean-Michel Claverie. A puzzling anomaly in the 4-mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process. Journal of Virology, American Society for Microbiology, 2019, 10.1128/JVI.01206-19 . hal-02314004

HAL Id: hal-02314004

<https://hal-amu.archives-ouvertes.fr/hal-02314004>

Submitted on 11 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1

2

3

4 A puzzling anomaly in the 4-mer composition of the giant pandoravirus genomes reveals a
5 stringent new evolutionary selection process

6

7 Running title: Unique compositional anomaly in pandoraviruses

8

9 Olivier Poirot^{a#}, Sandra Jeudy^a, Chantal Abergel^a, Jean-Michel Claverie^{a#}

10 ^a Aix Marseille Univ., CNRS, IGS, Information Génomique & Structurale (UMR7256),

11 Institut de Microbiologie de la Méditerranée (FR 3489), Marseille, France

12

13 [#] Correspondance to : jean-michel.claverie@univ-amu.fr ; olivier.poirot@igs.cnrs-mrs.fr

14

15

16 **Keywords:** Chaos Game Representation; Pandoravirus; Giant viruses; 4-mer statistics;

17 Genome composition; DNA editing; Host-virus relationship .

18

19 **Abstract**

20 The Pandoraviridae is a rapidly growing family of giant viruses, all of which have been
21 isolated using laboratory strains of Acanthamoeba. The genomes of ten distinct strains
22 have been fully characterized, reaching up to 2.5 Mb in size. These double-stranded DNA
23 genomes encode the largest of all known viral proteomes and are propagated in oblate
24 virions that are among the largest ever-described (1.2 μm long and 0.5 μm wide). The
25 evolutionary origin of these atypical viruses is the object of numerous speculations.
26 Applying the Chaos Game Representation to the pandoravirus genome sequences, we
27 discovered that the tetranucleotide (4-mer) “AGCT” is totally absent from the genomes of
28 2 strains (*P. dulcis* and *P. quercus*) and strongly underrepresented in others. Given the
29 amazingly low probability of such an observation in the corresponding randomized
30 sequences, we investigated its biological significance through a comprehensive study of
31 the 4-mer compositions of all viral genomes. Our results indicate that “AGCT” was
32 specifically eliminated during the evolution of the Pandoraviridae and that none of the
33 previously proposed host-virus antagonistic relationships could explain this phenomenon.
34 Unlike the three other families of giant viruses (Mimiviridae, Pithoviridae, Molliviridae)
35 infecting the same Acanthamoeba host, the pandoraviruses exhibit a puzzling genomic
36 anomaly suggesting a highly specific DNA editing in response to a new kind of strong
37 evolutionary pressure.

38 **Importance**

39 The recent years have seen the discovery of several families of giant DNA viruses all
40 infecting the ubiquitous amoebozoia of the genus Acanthamoeba. With dsDNA genomes
41 reaching 2.5 Mb in length packaged in oblate particles the size of a bacterium, the

42 pandoraviruses are the most complex and largest viruses known as of today. In addition
43 to their spectacular dimensions, the pandoraviruses encode the largest proportion of
44 proteins without homolog in other organisms which are thought to result from a *de novo*
45 gene creation process. While using comparative genomics to investigate the evolutionary
46 forces responsible for the emergence of such an unusual giant virus family, we discovered
47 a unique bias in the tetranucleotide composition of the pandoravirus genomes that can
48 only result from an undescribed evolutionary process not encountered in any other
49 microorganism.

50 **Introduction**

51 The Pandoraviruses are among the growing number of families of environmental
52 giant DNA viruses infecting protozoans and isolated using the laboratory host
53 *Acanthamoeba* (Protozoa/Lobosa/Ameobida/ Acanthamoebidae/ *Acanthamoeba*)¹⁻⁴. As
54 of today, they exhibit the largest fully characterized viral genomes, made of linear dsDNA
55 molecules from 1.9 to 2.5 Mb in size, predicted to encode up to 2500 proteins¹⁻³. After
56 their internalization by phagocytosis, these viruses multiply in their amoebal host
57 through a lytic cycle lasting about 12 hours, ending with the production of hundreds of
58 giant amphora-shaped particles (1.2 μm long and 0.5 μm wide)¹⁻³. The phylogenetic
59 structure of the Pandoraviridae family exhibits two separate clusters referred to as A-
60 and B- clades^{2,3} (Fig. 1). Despite this clear phylogenetic signal (computed using a core set
61 of 455 orthologous proteins), strains belonging to clade A or B did not exhibit noticeable
62 differences in terms of virion morphology, infectious cycle, host range, or global genome
63 structure and statistics (e.g. nucleotide composition, gene number, gene density)¹⁻³.

64 In addition to their unusual virion morphology and gigantic genomes, the pandoraviruses
65 exhibit other unique features such as an unmatched proportion (>90%) of genes coding
66 for proteins without any database homologs (ORFans) outside of the Pandoraviridae
67 family, and strain-specific genes contributing to an unlimited pan-genome¹⁻³. These
68 features, confirmed by the analysis of additional strains⁵, led us to suggest that a process
69 of *de novo* and *in situ* gene creation might be at work in pandoraviruses^{2,3}. Following this
70 history of unexpected findings, we thought that further analyses of the Pandoraviridae
71 might reveal additional surprises.

72 While searching for hidden genomic patterns eventually linked to evolutionary processes
73 unique to the pandoraviruses, we used a Chaos Game graphical representation of their
74 genome sequences⁶⁻⁷. This method converts long one-dimensional DNA sequence into a
75 fractal-like image, through which a human observer may detect specific patterns. This
76 representation illustrates in a holistic manner the frequencies of all oligonucleotides of
77 arbitrary length k (k-mers) in a given DNA sequence. Using this approach led us to
78 discover that the 4-mer “AGCT” was uniquely absent from the genome of *Pandoravirus*
79 *dulcis*, providing the starting point of the present study (Fig.2).

80

81 **Results**

82 ***The absence of any given 4-mer in a long random DNA sequence is highly improbable***

83 After detecting the absence of the “AGCT” word in the Chaos Game graphical
84 representation of the *P. dulcis* genome, we computed the number of occurrence of all 4-
85 mers in the ten available Pandoravirus genome sequences using direct counting⁸. This
86 revealed that “AGCT” was also absent from the genome of *P. quercus*. Notice that
87 although these strains belong to the same A-clade, their genome sequences are
88 nevertheless far from identical (their orthologous coding-regions share 72% nucleotide
89 identity on average), hence the common missing “AGCT” is not a mere consequence of
90 their sequence similarity.

91 Such a plain finding might not sound very interesting, until one realizes to what extent not
92 encountering a single occurrence of “AGCT” in DNA sequences respectively 1.908.524 bp

93 (*P. dulcis*) and 2.077.288 bp (*P. quercus*) is unlikely, as shown below, using increasingly
94 sophisticated computations.

95 In the simplest case, let us first consider a random DNA sequence with equal proportions
96 of the four nucleotides (%A=%T=%C=%G=25%). Since there are 256 distinct 4-mers, the
97 probability for each of them to occur at a given position in an increasingly long sequence
98 tends to $p_{AGCT} = 1/256$. In a random sequence of approximately 2 Mbp, one thus
99 expects an average of about 7800 occurrences for each distinct 4-mers. This already
100 suggests how unlikely it is for one of them to be absent.

101 To estimate the order of magnitude of such probability, the DNA sequence is seen as
102 consisting of 4 sets of non-overlapping 4-mers collected according to 4 different “reading
103 frames” (e.g. 4-mers 1-4, 5-8, 9-12, ..., etc, for frame 1). The different reading frames thus
104 correspond to approximately 500,000 positions each.

105 At each of these position, the probability for “AGCT” not to occur is $q_{AGCT} = 255/256$.
106 For one reading frame, this probability becomes approximately

$$107 \quad Q_{AGCT} = \left(255/256\right)^{500,000} \cong 1.2 \cdot 10^{-850} \quad (1)$$

108 and:

$$109 \quad 4 \times Q_{AGCT} \cong 5 \cdot 10^{-850} \quad (2)$$

110 for the 4 reading frames (assuming them to be independent for the sake of simplicity).

111 Such a value is smaller than any that could be computed in reference to a physical
112 process. For instance, one second approximately corresponds to $2 \cdot 10^{-18}$ of the age of the
113 universe.

114 The above probability should actually be corrected to account for the fact that we did not
 115 specifically search for “AGCT” while analyzing the viral genome. Any missing 4-mer would
 116 have raised the same interest. A Bonferroni correction should then be applied to
 117 compensate for the multiple testing of 256 different 4-mers. However, the probability of
 118 not finding any 4-mer, Q_{any} , remains an incommensurably small number.

$$119 \quad Q_{any} \cong 256 \times 5 \cdot 10^{-850} \cong 1.3 \cdot 10^{-847} \quad (3)$$

120 We may further argue that this event was bound to occur in at least one genome given
 121 the huge amount of DNA sequence that is now available, for instance in Genbank. The
 122 calculation runs as follows; The april 2019 release of Genbank contains about $3.2 \cdot 10^{11}$ bp.
 123 Assuming that all Genbank entries are 2 Mb-long sequences, this would correspond to 1.6
 124 10^5 theoretical pandoravirus genomes. The order of magnitude of the probability of
 125 observing one of them missing any of the 4-mers remains amazingly small at about

$$126 \quad Q_{any/Genbank} \cong 1.6 \cdot 10^5 \times Q_{any} \cong 2.1 \cdot 10^{-842} \quad (4)$$

127 Finally, one may want to make a final adjustment by taking into account that the *P. dulcis*
 128 genome is 64% G+C rich. This slightly changes the probability of random occurrence of
 129 “AGCT” from $p_{AGCT} = 1/256 = 0.00391$ to

$$130 \quad p_{AGCT} = (0.18)^2 \times (0.32)^2 = 3.31 \cdot 10^{-3} \quad (5)$$

131 then

$$132 \quad 4 \times Q_{AGCT} = (1 - p_{AGCT})^{500,000} \cong 8.9 \cdot 10^{-719} \quad (6)$$

133 Using the same Bonferroni correction as above lead to the final conservative estimate:

$$134 \quad Q_{any/Genbank} < 4 \cdot 10^{-711} \quad (7)$$

135 still an incommensurably small probability (e.g. the same as not getting a single head in
136 2360 tosses of a fair coin).

137 As the above computation remains an approximation (neglecting the overlap of
138 neighboring 4-mers), we estimated how unlikely it is that any 4-mer would be missing
139 from large DNA sequences by a different approach. We computer generated a large
140 number of random sequences of increasing sizes and recorded the threshold at which
141 point none of the 4-mers is missing. Fig. 3 displays the results of such computer
142 experiment. It shows how fast the probability of any 4-mer missing is decreasing with the
143 random sequence size. In this experiment, we found that the proportion of sequences
144 larger than 10,000 bp missing anyone of the 256 4-mers was less than 1/10,000.

145

146 ***Caveat: randomized sequences exhibit strongly unnatural 4-mer distributions***

147 The above results already suggested that it is impossible for the *P. dulcis* and *P.*
148 *quercus* genomes to be missing “AGCT” solely by chance without invoking a biological
149 constraint. However, this conclusion rests on the assumption that the randomization
150 process suitably modeled these genomes. However, the frequency distribution of the
151 various 4-mers found in the actual *P. dulcis* genome (and of other pandoraviruses) and
152 the one computed from its randomized sequence are strongly different (Fig. 4). While the
153 natural sequence consist of 4-mers occurring at frequencies distributed along a large and
154 rather continuous interval, the randomized sequence exhibits 4-mers occurring around 5
155 narrow peaks of frequencies with none in between. As expected from a good quality
156 randomization, these peaks correspond to the frequencies of the five types of 4-mers:
157 those consisting of only A or T at the lower end, those consisting of only G or C at the

158 higher end, and those consisting of (A or T)/(G or C) in proportions 1/3, 2/2, and 3,1 in
159 between. The more continuous and spread out natural distribution is the testimony of
160 multiple evolutionary constraints, most of them unknown, that have resulted in a distinct
161 4-mer usage, like a dialect or a language tic inherited from past generations⁹.

162 First, notice that the missing “AGCT” does not correspond to the 4-mer type with the
163 lowest expected frequency (but the middle one). Second, it is clear that the above
164 probability calculations based on such distorted model of the natural sequence, cannot
165 be used as a reliable estimate of statistical significance. This problem is similar to the one
166 encountered when trying to evaluate the quality of local sequence alignments in similarity
167 searches^{10, 11}.

168 We can mitigate the effect of the above stringent randomization (only preserving the
169 original nucleotide composition) by using the *P. dulcis* and *P. quercus* actual genome
170 sequences to evaluate to what extent the absence of “AGCT” might be the mere
171 statistical consequence of the frequency of its constituent 3-mers: AGC and GCT.

172 As shown in Table 1, AGC and GCT are not among the least frequent 3-mers found in the
173 *P. dulcis* or *P. quercus* genomes. As the theoretical average is 1/64 (≈ 0.0156), their
174 proportions range from 0.0156 to 0.0097 within the coding and non-coding regions of the
175 genomes. On one given strand, AGC and GCT also do not strongly segregate from each
176 other’s in coding *versus* intergenic regions (Table 1). By combining the AGC 3-mer
177 frequency with that of the single nucleotide T ($p_{(t)}=0.182$ for *P. dulcis*, $p_{(t)}=0.196$ for *P.*
178 *quercus*), the expected number of “AGCT” per strand is 4286 for *P. dulcis* and 4898 for *P.*
179 *quercus*, while none is observed. Such stark contrast between expected and observed
180 values is unique to the “AGCT” 4-mer. By comparison, the palindromic “ACGT” 4-mer

181 (with an identical composition) exhibits a statistical behavior (Table 1, bottom lines) much
182 closer to the 3-mer-dependent random sequence model.

183

184 ***No 4-mer is missing from the largest actual viral genomes***

185 As vividly illustrated in Fig. 4, the 4-mer distributions in randomized sequences
186 strongly depart from that in natural genomes. We thus analyzed all complete genome
187 sequences available in the viral section of Genbank¹², to investigate to what extent the
188 absence of a given 4-mer was exceptional for genomes in the size range corresponding to
189 Pandoraviruses.

190 We found that the next largest viral genomes missing a 4-mers were those of five phages
191 infecting enterobacteria, with unusual genome sizes in the 345kb-359kb range¹³⁻¹⁶. Except
192 for *P. dulcis* and *P. quercus*, none of the 26 largest publicly available viral genomes
193 (including 25 large/giant eukaryotic viruses, and phage G)¹² were missing a 4-mer (Fig. 5).
194 Thus, even by comparison with natural sequences, *P. dulcis* and *P. quercus* appear
195 exceptional.

196 We noticed that the five large enterobacteria-infecting phages pointed out by our
197 analysis, were all missing the same “GCGC” 4-mer although they exhibit divergent
198 genomic sequences and were isolated from different hosts¹³⁻¹⁶. This palindromic 4-mer
199 might be the target of isoschizomeric restriction endonucleases functionally homologous
200 to HhaI found in *Haemophilus haemolyticus*, a Gammaproteobacteria. Many of them
201 have been described (see <https://enzymefinder.neb.com>). We will return to the
202 hypothesis that some 4-mers might be missing in response to a host or viral defense
203 mechanism¹⁷ in the discussion section.

204

205 **The anomalous distribution of “AGCT” correlates with the Pandoraviridae phylogenetic**
206 **structure**

207 The absence of “AGCT” in *P. dulcis* and *P. quercus* genomes becomes even more
208 intriguing when put in the context of the phylogenetic structure of the whole
209 pandoravirus family. As shown in Fig. 1, the Pandoraviridae neatly cluster into two
210 separate clades. For well-conserved proteins (such as the DNA polB), the percentage of
211 identical residues between intra-clade orthologs is in the 82% to 90% range, and in the
212 72% to 76% range between the two clades. The corresponding genome sequences are
213 thus far from being identical (and only partially collinear) within each clade. It is thus
214 quite remarkable that the “AGCT” count exhibits a consistent trend to be very low in A-
215 clade members, and at least 10 times higher in B-clade strains. Such a contrast was strong
216 enough to pre-classify three unpublished isolates prior to complete genome assembly and
217 finishing (data not shown).

218 The large difference in “AGCT” counts could be due to the deletion of a genomic region
219 concentrating most of them, for instance within a repeated structure absent from the A-
220 clade isolates. However, Fig. 6 shows that this is not at all the case. In B- clade isolates,
221 the numerous occurrences of “AGCT” are rather uniformly distributed along the whole
222 genomes. However, we noticed that the “AGCT” distribution in the *P. neocaledonia*
223 genome exhibits a change of slope at one of its extremities, as if the corresponding
224 segment had been acquired from a A-clade strain. Such hypothesis was confirmed using a
225 dot-plot comparison with the *P. salinus* genome, to which this terminal segment is clearly
226 homologous (Fig. 7).

227

228 **“AGCT” was specifically deleted from A-clade pandoravirus genomes**

229 We have seen in the previous section that the extreme difference in the “AGCT”
230 count in *P. dulcis* (N=0) and *P. neocaledonia* (N=544) is not due to the local deletion of an
231 “AGCT”-rich segment. We then investigated if that difference was limited to “AGCT”, or if
232 other 4-mers exhibited large differences in counts. Fig. 8 shows that this was not the case.
233 If the frequencies of the various 4-mers within each genome exhibit tremendous
234 differences (very much at odd with their distribution in randomized sequences, see Fig.
235 4), the frequency for each 4-mer (low, average or high) was very similar across the two
236 different viral genomes (Spearman correlation, $r=0.9859$). The difference in “AGCT”
237 count is thus not the consequence of the use of globally distinct 4-mer vocabularies by
238 the two pandoravirus clades. It appears to be due to a selection specifically exerted
239 against the presence of “AGCT” in the genomes of A-clade pandoraviruses.

240 Another argument in favor of an active selection against the presence of “AGCT” is
241 provided by the following statistical computation. We first identified the orthologous
242 proteins in *P. dulcis* and *P. neocaledonia*, using the best-reciprocal Blastp match criterium.
243 We identified 585 orthologous ORFs. In *P. neocaledonia*, 180 of them were found to
244 contain one or several “AGCT” (for a total of 350 occurrences). We then computed the
245 average percentage of nucleotide identity in the alignments of these 180 *P. neocaledonia*
246 ORFs with their *P. dulcis* orthologous counterparts. The value was 69%.

247 According to a neutral scenario (and neglecting multiple hits), the probability is thus
248 $p = 0.69$ that any nucleotide remains the same along the evolutionary trajectory
249 separating the two pandoraviruses. For a given “AGCT”, the probability to remain intact

250 over the same evolutionary distance is $p_{intact} = 0.69^4 = 0.227$, such as none of the
251 four positions is changed. For the sake of simplicity, we will neglect the chance creation of
252 new “AGCT” during the process. As a result, we then expect *P. dulcis* orthologous ORFs to
253 exhibit 68 occurrences (i.e. 0.227×350) of “AGCT”.

254 This simple calculation already indicates that the “AGCT” 4-mer diverged much faster (at
255 least 80 times faster since $350 \times 0.227 / 80 < 1$) than the rest of the orthologous coding
256 regions. This result suggests that the absence of “AGCT” in *P. dulcis* and *P. quercus*, as
257 well as its distinctive low frequency in all A-clade strains is the consequence of an active
258 counter selection. We discuss possible molecular mechanisms in the following section.
259 The above calculation could not be extended to interORFs regions, due to their much
260 lower conservation and their unreliable pairwise alignments.

261

262 **Discussion**

263 **Which model for the counter selection of “AGCT”?**

264 Following our statistical computations on random sequences confirmed by the
265 analysis of actual genome sequences, we can safely assume that the genome of the
266 common ancestor of the A- and B-clade pandoraviruses was not missing any 4-mers. Our
267 discussion will thus take for granted that the difference in “AGCT” frequency between the
268 two Pandoraviridae clades is the consequence of a loss in the A-clade rather than a gain in
269 the B-clade. Such phenomenon probably predated the split of the two clades as the
270 number of “AGCT” found in B-clade Pandoravirus genomes (≈ 500) is already 15 times
271 lower than expected in the corresponding randomized sequences (≈ 7800).

272

273 Any model proposed to explain our results must take into account that the two types of
274 pandoraviruses replicate with the same efficiency in various laboratory strains of
275 Acanthamoeba. From this we can reasonably assume that both clades do not differ much
276 in their range of natural hosts (one of which is known to be an Acanthamoeba for A-clade
277 *Pandoravirus inopinatum*¹⁸). The cause of the marked difference in “AGCT” counts
278 between the two clades must thus reside within the viruses themselves. Such inference is
279 further supported by the fact that none of the other families of giant viruses¹⁹ infecting
280 the very same Acanthamoeba hosts exhibit a similar 4-mer anomaly in their genome
281 composition.

282 The first model that comes to mind is inspired from the well-documented restriction-
283 modification systems that many bacteria use to counteract bacteriophage infections. The
284 host bacterial cells express DNA sites (most often short palindromes) specific
285 endonucleases that cut the invading phage genome before it could replicate. Such
286 defense mechanism imposes the bacteria to protect the cognate motif in its own genome
287 using a specific methylase. According to the Red Queen evolutionary concept, the
288 bacteriophages could counteract the host’s defense by removing the targeted site from
289 their own genome¹⁷. The absence of the palindrome “GCGC” that we previously noticed
290 in several large enterobacterial phages¹³⁻¹⁶ could result from such evolutionary strategy.

291 Translating such a model in our system thus requires three distinct assumptions: 1) that
292 the Acanthamoeba cells express an antiviral endonuclease specific for “AGCT”; 2) that B-
293 clade pandoraviruses are immune from it (as other Acanthamoeba-infecting viruses); 3)

294 that A-clade pandoraviruses evolved a different strategy by removing the endonuclease
295 target from their genomes.

296 Such a model was readily invalidated by simply attempting to digest the B-clade *P.*
297 *neocaledonia* genomic DNA (extracted from infectious particles) with commercial
298 restriction enzymes (such as PvuII) targeting “cAGCTg” (212 occurrences) and AluI,
299 targeting “AGCT” (544 occurrences). The resulting Pulsed-field gel electrophoresis (PFGE)
300 pattern showed that these sites were not protected (Fig. 9). Accordingly, the PacBio data
301 used to sequence the *P. neocaledonia* genome² did not indicate the presence of modified
302 nucleotides at the “AGCT” sites²⁰.

303 We must point out that the above results simultaneously invalidate a symmetrical model
304 where the “AGCT”-specific endonuclease would have been encoded by the
305 pandoraviruses, together with the protective cognate methylase. Such a hijacked
306 restriction/modification system would have been attractive as it is found in
307 chloroviruses²¹, another family of large eukaryotic DNA viruses. Unfortunately, it does not
308 apply here. Accordingly, no homolog of the cognate DNA-methyl transferase was
309 detected among the *P. neocaledonia* or *P. macleodensis* protein-coding gene contents.
310 Further nailing the coffin of such restriction/modification hypothetical model, no
311 difference in terms of potentially relevant endonuclease or DNA methylase was found
312 between the gene contents of the A-clade *P. dulcis* and *P. quercus* and those of the B-
313 clade *P. neocaledonia* and *P. macleodensis*.

314 A more hypothetical model would assume that the “AGCT” motif is targeted at the
315 transcript level (i.e. “AGCU”) rather than at the DNA level. Classical endonucleases and

316 DNA methylases would thus not be involved in the host-virus confrontation. There are
317 several arguments against a mechanism directly targeting viral transcripts.

318 First, as B-clade pandoraviruses exhibit similar proportions of “AGCT” in ORFs and inter-
319 ORF regions, the A-clade strains would have had no incentive to eliminate the motif from
320 their intergenic regions, as *P. dulcis* and *P. quercus* have done totally in reaching zero
321 occurrences. “AGCT” is also still present in some protein-coding regions of *P. inopinatum*
322 (N=15), *P. salinus* (N=3), and *P. celtis* (N=1).

323 Second, very few motif-specific RNases are known, and to our knowledge, only one is
324 viral: a protein encoded in the bacteriophage T4 RegB gene²². We found no significant
325 homolog of this protein in the pandoraviruses or *Acanthamoeba*. We also looked for
326 mRNA methylases that could act as a protective mechanism for the viral transcript. A
327 single one was described in another family of eukaryotic DNA virus: the product of the
328 Megavirus Mg18 gene²³. Again, no significant homolog of this protein was detected in the
329 pandoraviruses.

330 In conclusion to this section, if the presence of “AGCT” decreases the virus fitness, we
331 found no evidence that it is due to a DNA or RNA nuclease-mediated defense mechanism
332 in *Acanthamoeba*. However, it could still be due to an unknown inhibitory mechanism
333 acting at the transcription regulation level to which B-clade pandoviruses would exhibit
334 some immunity. The corresponding proteins could be encoded among the numerous
335 ORFans found in pandoravirus genomes¹⁻³. Alternatively, the “AGCT” deficit could be due
336 to a restriction imposed by unknown additional hosts in nature, although quite an unlikely
337 scenario given the ubiquity and abundance of *Acanthamoeba* in the environment.

338 Finally, could “AGCT” be deleterious for some intrinsic reasons, for instance due to its
339 palindromic structure and composition? This is very unlikely, when one compare the
340 absent “AGCT” in *P. dulcis* and *P. quercus*, with other 4-mers with identical structures and
341 compositions. For instance “ACGT” occurs at 5822 and 6165 positions (in *P. dulcis* and *P.*
342 *quercus*, respectively), and “GATC” occurs at 8114 and 8567 times) in (*P. dulcis* and *P.*
343 *quercus*, respectively). The presence or absence of “AGCT” does not either exert a strong
344 constraint on protein sequences, as the amino-acids encoded by “AGC” or “GTC” (Serine
345 and Alanine, respectively) have many possible alternative codons and are easily
346 replaceable residues given their mild physicochemical properties. Finally, we found no
347 evidence that the removal of “AGCT” was due to a specific (for instance, enzyme-
348 mediated) process targeting then replacing the forbidden 4-mer by a constant alternative
349 word. Replacement patterns for 72 *P. dulcis* sites unambiguously mapped to their
350 homologous *P. neocaledonia* “AGCT” counterparts are indicated in Table 2. It suggests
351 that the complete loss of “AGCT” in the A-clade strains is due to a stringent, nevertheless
352 random (i.e. non-directed) evolutionary process.

353 The analysis of long nucleotide (and amino acid) sequences as overlapping k-mers
354 has a long history in bioinformatics. Initially proposed in the context of the RNA folding
355 problem²³, the concept was then quickly applied to many other areas including gene
356 parsing²⁴, the detection of regulatory motifs^{25, 26}, and has become central to the fast
357 implementation of large-scale similarity search^{27, 28}, sequence assembly²⁹, and the binning
358 of metagenomics data^{30, 31}. However, its popularity should not hide that most of the
359 observed frequency disparities (starting from the simplest mononucleotide composition)
360 between k-mers within a given organism, or across species have not yet received
361 convincing biological explanations^{32, 33}. This suggests that profound and unexpected

362 biological insights may one day come out from the analysis of k-mer frequencies, and in
363 particular from their most improbable fluctuations. In a daring parallel with the delayed
364 understanding of the CRISPR/CAS system from the initial spotting of intriguing repeats³⁴,
365 we would like to expect that the pandoraviridae “AGCT” distribution anomaly might lead
366 to the discovery of a novel defense mechanism against viral infection.

367

368 **Materials and Methods**

369 **Chaos game representation**

370 Chaos game representation (CGR) was introduced in 1990 by Jeffrey⁶ to visually detect
371 global patterns in large DNA sequences. It was inspired from a method generating fractals
372 within a polygon as a sequence of points, iteratively positioned according to a rule based
373 on their distance to one of the vertices of the polygon. To apply this method to DNA
374 sequences, one uses a square with corners labelled A, T, G and C. Starting from the
375 center of the square, the sequence is used to determine the position of the next point at
376 the center of the line connecting the previous point and the corner corresponding to the
377 current nucleotide. In addition to global patterns, the resulting graph also reveals the
378 differential frequencies of substrings (k-mers), for instance leaving a blank area at the
379 position corresponding to a missing substring (Fig. 2). CGR thus allows the rapid detection
380 of compositional anomaly of k-mers for increasing n values, instead of comparing large
381 statistical tables. Once the k-mer (4-mer) distributions of interest were determined by
382 CGR, they were further analyzed and compared using a standard counting package⁸.

383

384 **Pulse-field gel electrophoresis (PFGE)**

385 Approximately 5,000 pandoravirus particles were embedded in 1% low gelling agarose
386 and the plugs were incubated in lysis buffer (50mM Tris-HCl pH8.0, 50mM EDTA, 1% (v/v)
387 N-laurylsarcosine, 1mM DTT and 1mg/mL proteinase K) for 16h at 50°C. After lysis, the
388 plugs were washed once in sterile water and twice in TE buffer (10mM Tris-HCl pH8.0 and
389 1mM EDTA) with 1mM PMSF, for 15 min at 50°C. The plugs were then equilibrated in the
390 appropriate restriction buffer and digested with 20 units of PvuII or AluI at 37°C for 14
391 hours. Digested plugs were washed once in sterile water for 15 min, once in lysis buffer
392 for 2h and three times in TE buffer. Electrophoresis was carried out in 0.5X TAE for 18 h at
393 6V/cm, 120° included angle and 14°C constant temperature in a CHEF-MAPPER system
394 (Bio-Rad) with pulsed times ramped from 0.2s to 120s.

395

396 **Availability of data**

397 All virus genome sequences analyzed in this work are freely available from the public
398 GenBank repository (URL://www.ncbi.nlm.nih.gov/genbank/). The Pandoravirus
399 sequences used here correspond to the following accession numbers: *P. dulcis*
400 (NC_021858), *P. neocaledonia* (NC_037666), *P. macleodensis* (NC_037665), *P. salinus*
401 (NC_022098), *P. quercus* (NC_037667), *P. celtis* (NC_), *P. inopinatum* (NC_026440), *P.*
402 *pampulha* (LT972219.1), *P. massiliensis* (LT972215.1), *P. braziliensis* (LT972217).

403

404 **References**

405 1. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V,
406 Bertaux L, Bruley C, Garin J, Claverie JM, Abergel C. 2013. Pandoraviruses: amoeba
407 viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*
408 341:281-286.

409 2. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic JM, Beucher L, Philippe N,
410 Bertaux L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie JM. 2018. Diversity
411 and evolution of the emerging Pandoraviridae family. *Nat Commun* 9:2285.

412 3. Legendre M, Alempic JM, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S, Couté
413 Y, Abergel C, Claverie JM. 2019. Pandoravirus celtis illustrates the microevolution
414 processes at work in the giant Pandoraviridae genomes. *Front Microbiol* 10:430.

415 4. Abergel C, Legendre M, Claverie JM. 2015. The rapidly expanding universe of giant
416 viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev* 39:779-
417 796.

418 5. Aherfi S, Andreani J, Baptiste E, Oumessoum A, Dornas FP, Andrade ACDSP, Chabriere
419 E, Abrahao J, Levasseur A, Raoult D, La Scola B, Colson P. 2018. A large open pangenome
420 and a small core genome for giant pandoraviruses. *Front Microbiol* 9:1486.

421 6) Jeffrey HJ. 1990. Chaos game representation of gene structure. *Nucleic Acids Res*
422 18:2163-2170.

423 7) Hoang T, Yin C, Yau SS. 2016. Numerical encoding of DNA sequences by chaos game
424 representation with application in similarity comparison. *Genomics* 108:134-142.

425 8) Mullan LJ, Bleasby AJ. 2002. Short EMBOSS User Guide. *European Molecular Biology*
426 *Open Software Suite. Brief Bioinform.* 3:92-94.

- 427 9) Phillips GJ, Arnold J, Ivarie R. 1987. Mono- through hexanucleotide composition of the
428 *Escherichia coli* genome: a Markov chain analysis. *Nucleic Acids Res* 15:2611-2626.
- 429 10) Altschul SF, Erickson BW. 1985. Significance of nucleotide sequence alignments: a
430 method for random sequence permutation that preserves dinucleotide and codon usage.
431 *Mol Biol Evol* 2:526-538.
- 432 11) Pagni M, Jongeneel CV. 2001. Making sense of score statistics for sequence
433 alignments. *Brief Bioinform* 2:51-67.
- 434 12) Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015. NCBI viral genomes resource.
435 *Nucleic Acids Res* 43:D571-7.
- 436 13) Abbasifar R, Griffiths MW, Sabour PM, Ackermann HW, Vandersteegen K, Lavigne R,
437 Noben JP, Alanis Villa A, Abbasifar A, Nash JH, Kropinski AM. 2014. Supersize me:
438 *Cronobacter sakazakii* phage GAP32. *Virology* 460-461:138-146.
- 439 14) Kim MS, Hong SS, Park K, Myung H. 2013. Genomic analysis of bacteriophage
440 PBECO4 infecting *Escherichia coli* O157:H7. *Arch Virol* 158:2399-2403.
- 441 15) Šimoliūnas E, Kaliniene L, Truncaite L, Klaus V, Zajančauskaite A, Meškys R. 2012.
442 Genome of *Klebsiella* sp.-infecting bacteriophage vB_KleM_RaK2. *J Virol* 86:5406.
- 443 16) Pan YJ, Lin TL, Lin YT, Su PA, Chen CT, Hsieh PF, Hsu CR, Chen CC, Hsieh YC, Wang JT.
444 2015. Identification of capsular types in carbapenem-resistant *Klebsiella pneumoniae*
445 strains by *wzc* sequencing and implications for capsule depolymerase treatment.
446 *Antimicrob Agents Chemother* 59:1038-1047.
- 447 17) Sharp PM. 1986. Molecular evolution of bacteriophages: evidence of selection
448 against the recognition sites of host restriction enzymes. *Mol Biol Evol* 3:75-83.

449 18) Antwerpen MH, Georgi E, Zoeller L, Woelfel R, Stoecker K, Scheid P. 2015. Whole-
450 genome sequencing of a pandoravirus isolated from keratitis-inducing acanthamoeba.
451 *Genome Announc* 3(2): pii: e00136-15.

452 19) Abergel C, Legendre M, Claverie JM. 2015. The rapidly expanding universe of giant
453 viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev* 39: 779-
454 796.

455 20) Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner
456 SW. 2010. Direct detection of DNA methylation during single-molecule, real-time
457 sequencing. *Nat Methods* 7:461-465.

458 21) Agarkova IV, Dunigan DD, Van Etten JL. 2006. Virion-associated restriction
459 endonucleases of chloroviruses. *J Virol* 80:8114-8123.

460 22) Odaert B, Saïda F, Aliprandi P, Durand S, Créchet JB, Guerois R, Laalami S, Uzan M,
461 Bontems F. 2007. Structural and functional studies of RegB, a new member of a family of
462 sequence-specific ribonucleases involved in mRNA inactivation on the ribosome. *J Biol*
463 *Chem* 282:2019-2028.

464 23) Priet S, Lartigue A, Debart F, Claverie JM, Abergel C. 2015. mRNA maturation in giant
465 viruses: variation on a theme. *Nucleic Acids Res* 43:3776-3788.

466 24) Dumas JP, Ninio J. 1982. Efficient algorithms for folding and comparing nucleic acid
467 sequences. *Nucleic Acids Res* 10:197-206.

468 25) Claverie JM, Bougueleret L. 1986. Heuristic informational analysis of sequences.
469 *Nucleic Acids Res* 14:179-196.

470 26) Brendel V, Beckmann JS, Trifonov EN. 1986. Linguistics of nucleotide sequences:
471 morphology and comparison of vocabularies. *J Biomol Struct Dyn* 4:11-21.

- 472 27) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
473 search tool. *J Mol Biol* 215:403-410.
- 474 28) Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res.* 12:656-664.
- 475 29) Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G,
476 Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z,
477 Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: an empirically
478 improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.
- 479 30) Chan CK, Hsu AL, Halgamuge SK, Tang SL. 2008. Binning sequences using very sparse
480 labels within a metagenome. *BMC Bioinformatics* 9:215.
- 481 31) Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. 2004. Application of
482 tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*
483 6:938-947.
- 484 32) Karlin S, Mrázek J, Campbell AM. 1997. Compositional biases of bacterial genomes
485 and evolutionary implications. *J Bacteriol* 179:3899-3913.
- 486 33) Bohlin J, Pettersson JH. 2019. Evolution of genomic base composition: from single cell
487 microbes to multicellular animals. *Comput Struct Biotechnol J* 17:362-370.
- 488 34) Ishino Y, Krupovic M, Forterre P. 2018. History of CRISPR-Cas from encounter with a
489 mysterious repeated sequence to Genome editing technology. *J Bacteriol* 200:pii: e00580-
490 17.
- 491 35) Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for
492 creating dotplots on genome scale. *Bioinformatics* 23:1026-1028.

493

494 **Figure Legends**

495 **Figure 1. Phylogenetic structure of the Pandoraviridae.** Adapted from [ref. 3]. The
496 number of occurrences of the “AGCT” 4-mer is indicated for the genome of each strain.
497 The counts are given for one DNA strand and are identical for both strands (“AGCT” is
498 palindromic).

499

500 **Figure 2. Chaos game representation of the *P. dulcis* genome.** The largest square left
501 blank (circled in red) corresponds to “AGCT”, indicating the absence of this 4-mer in the
502 genome.

503

504 **Figure 3. Influence of random sequence length on the number of missing 4-mers.** 10.000
505 random sequences up to 10.000 bp in size were analyzed. Except for extremely rare
506 fluctuations, no sequence longer than 4000 bp exhibits a missing 4-mer. 4-mer overlaps
507 as well as nucleotide compositions are taken into account in this analysis.

508

509 **Figure 4. Distribution of 4-mer frequencies in natural and randomized genome**
510 **sequences.** Top: histogram of the number of distinct 4-mers occurring at various numbers
511 of occurrences in the *P. dulcis* genome; Bottom: same analysis after randomization.

512

513 **Figure 5. Missing 4-mers in the largest viral genomes.** Except for *P. dulcis* and *P. quercus*,
514 the largest viral genomes missing a 4-mers are those of 5 distinct bacteriophages
515 (accession numbers: NC_019401, NC_025447, NC_027364, NC_027399, NC_019526).

516

517 **Figure 6. Cumulative distribution of “AGCT” occurrences along the different**
518 **pandoravirus genomes.** The “AGCT” word appears uniformly spread throughout the B-
519 clade pandoravirus genomes, except for a clear rarefaction at the end of the *P.*
520 *neocaledonia* genome sequence.

521

522 **Figure 7. DNA sequence dot-plot comparison of *P. neocaledonia* (horizontal) and *P.***
523 ***salinus* (vertical).** The two genomes only exhibit remnants of collinearity except for the
524 terminal region of *P. neocaledonia* (red circle) coinciding with a low “AGCT” density
525 typical of A-clade strains (Fig. 6). Dot plot generated using GEPARD³⁵ with parameters:
526 word size=15, window size=0.

527

528 **Figure 8. Comparison of the proportion of all 4-mers in *P. dulcis* (A-clade) vs. *P.***
529 ***neocaledonia* (B-clade).** The 4 most frequent 4-mers are “GCGC”, “CGCG”, “CGCC”, and
530 “GGCG”.

531

532 **Figure 9. Digestion of *P. neocaledonia* DNA at “AGCT” sites.** Lane 1: undigested *P.*
533 *neocaledonia* DNA (2.2 Mb) migrating as expected. The bottom band (below 48.5 kb)
534 correspond to an episome not always present. Lane 2: *P. neocaledonia* DNA digested by
535 the PvuII restriction enzyme (cutting site: cAGCTg). Lane 3: *P. neocaledonia* DNA digested
536 by the AluI restriction enzyme (cutting site: AGCT). These results demonstrate that the
537 “AGCT” sites are not protected by modified nucleotides.

538

539 **Acknowledgements**

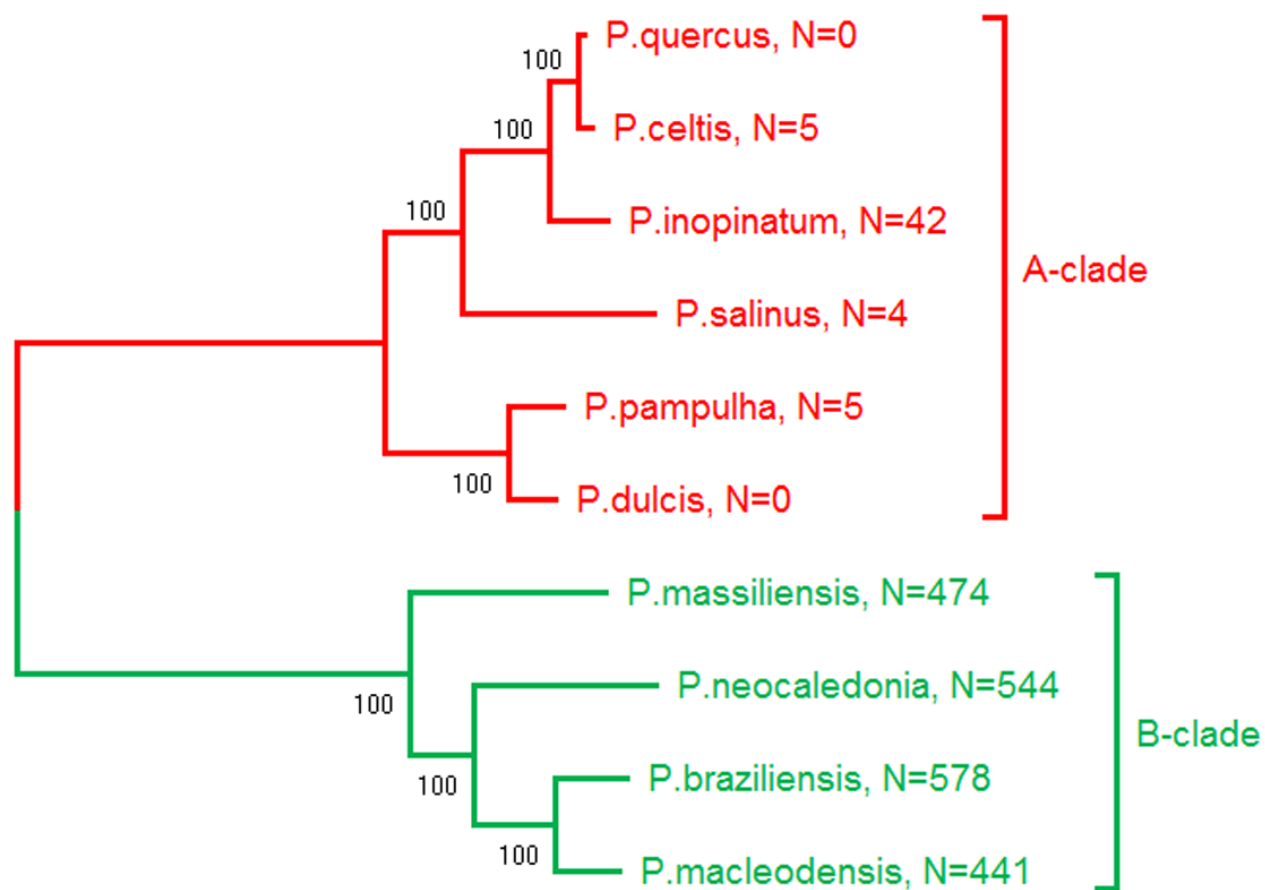
540 We thank Dr. Sacha Schutz for his inspiring blog (URL: <http://dridk.me/>) that initiated our
541 interest in the Chaos Game Representation technique. We thank Dr. Matthieu Legendre
542 for verifying the absence of modified nucleotides at “AGCT” sites using the PACBIO
543 sequence data. Our laboratory is supported by the French National Research Agency
544 (ANR-14-CE14-0023-01), France Genomique (ANR-10-INSB-01-01), Institut Français de
545 Bioinformatique (ANR-11-INSB-0013), the Fondation Bettencourt-Schueller (OTP51251),
546 and by the Provence-Alpes-Côte-d’Azur region (2010 12125). We acknowledge the
547 support of the PACA-Bioinfo platform. The funding bodies had no role in the design of the
548 study, analysis, and interpretation of data and in writing the manuscript.

549

550 **Competing interests**

551 The authors declare that they have no competing interests

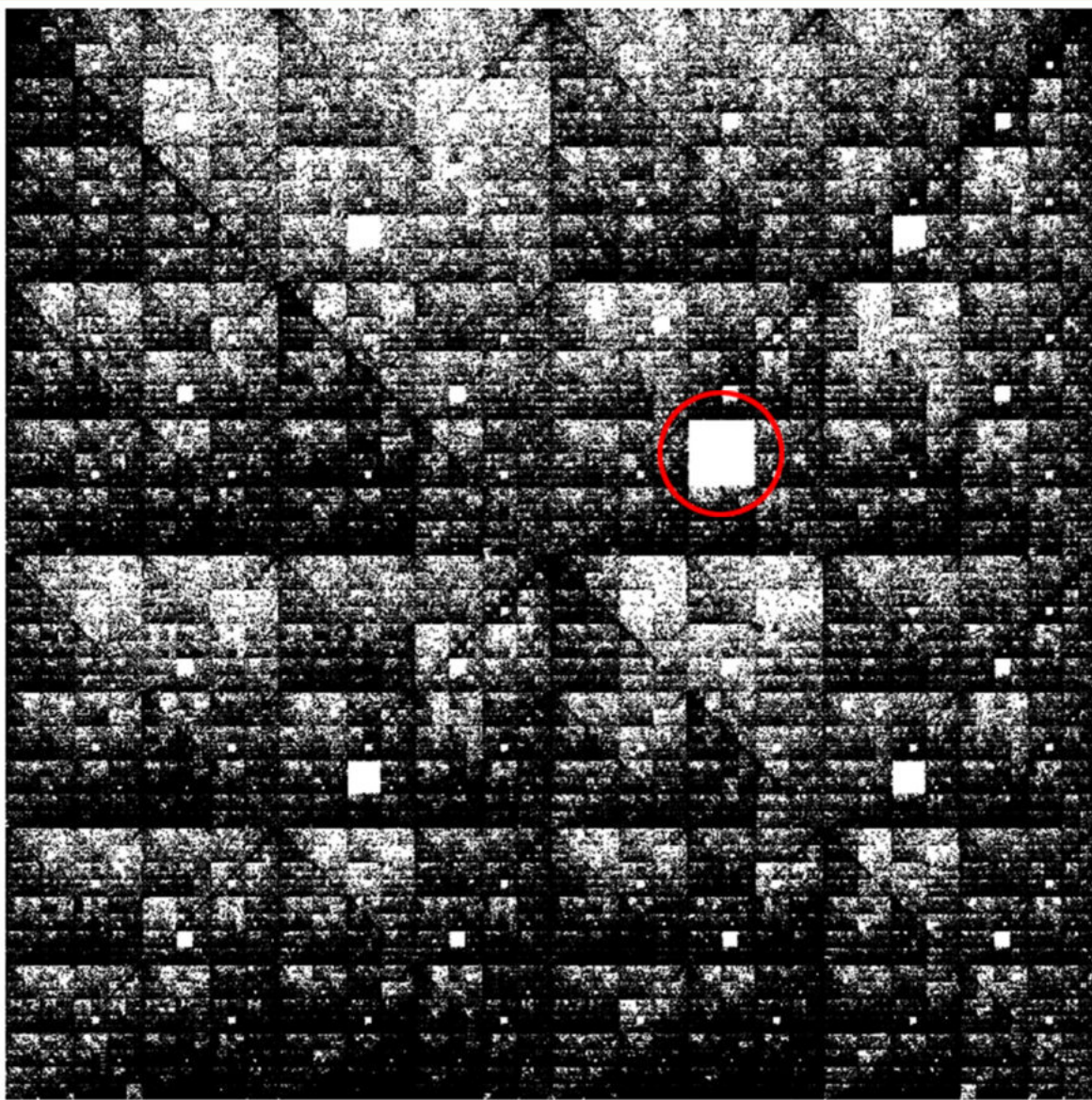
552



A

word points ☒ squares ☐ AGCT: 0 / 0

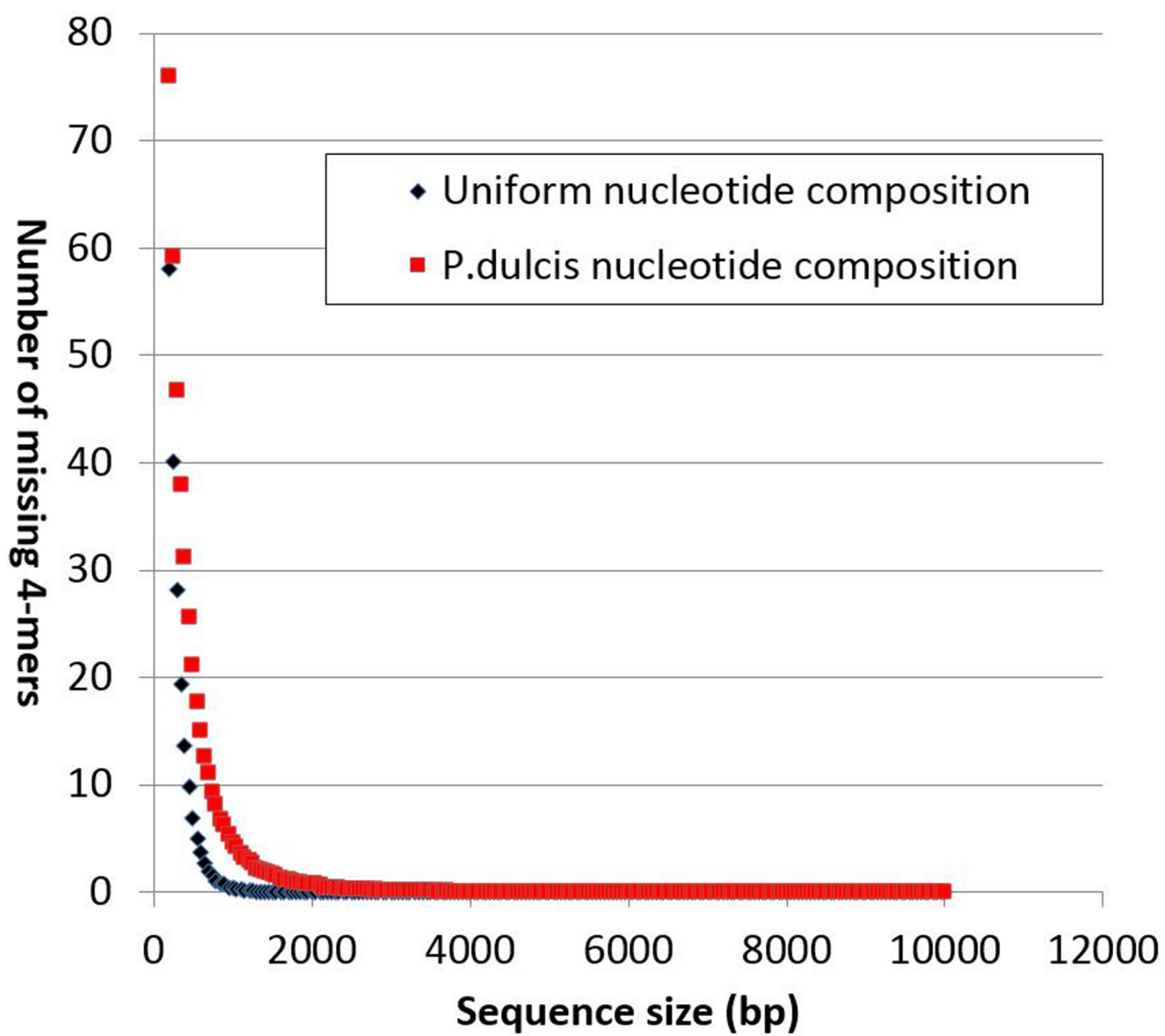
T



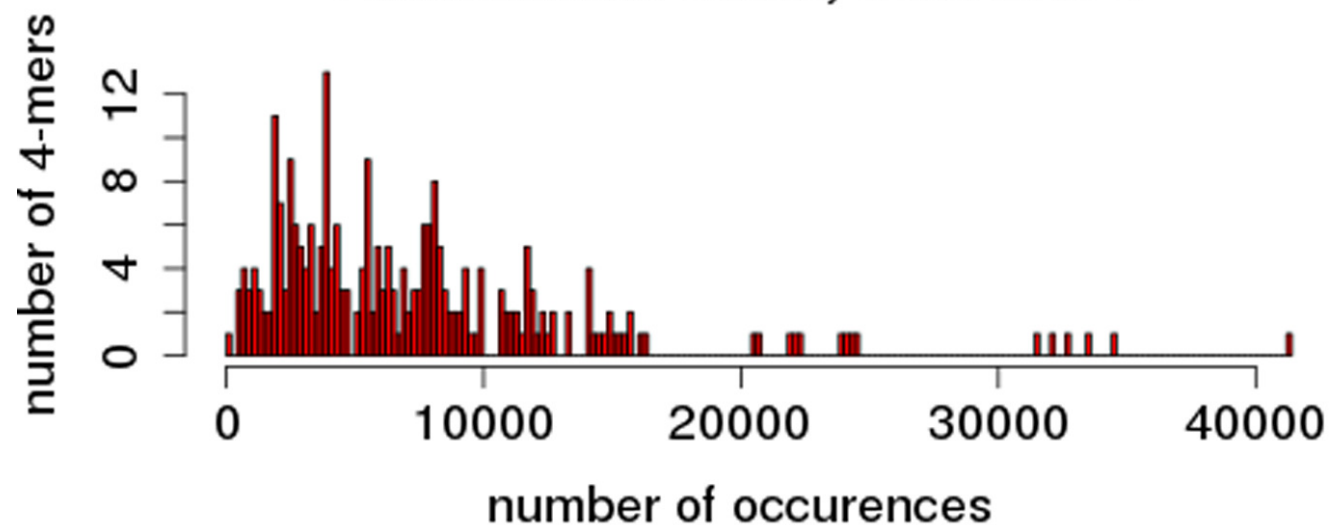
C

word size ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

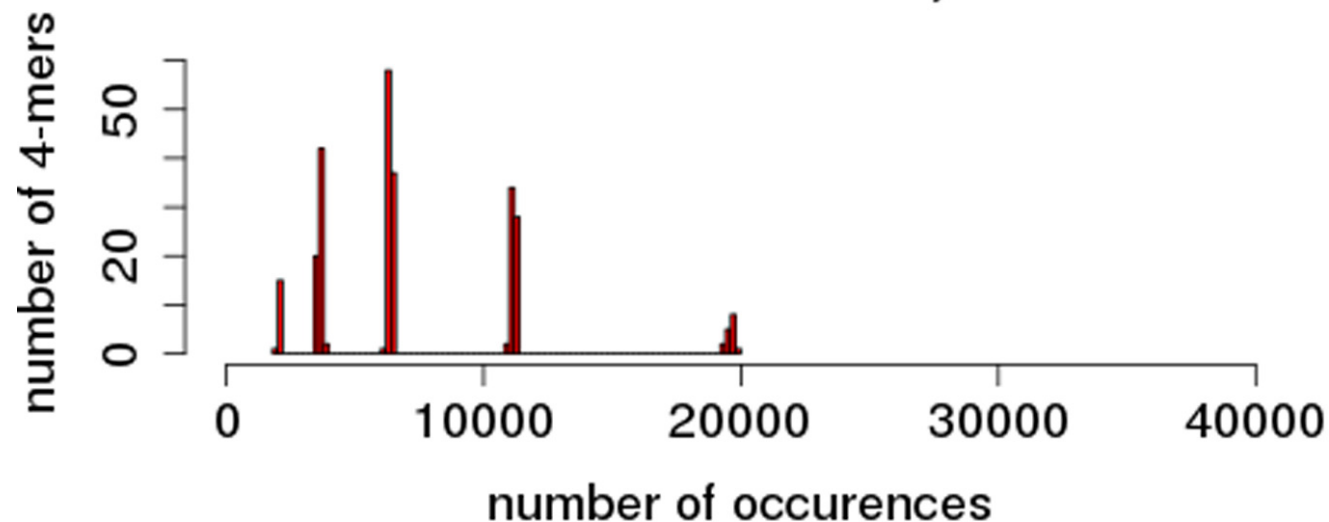
G

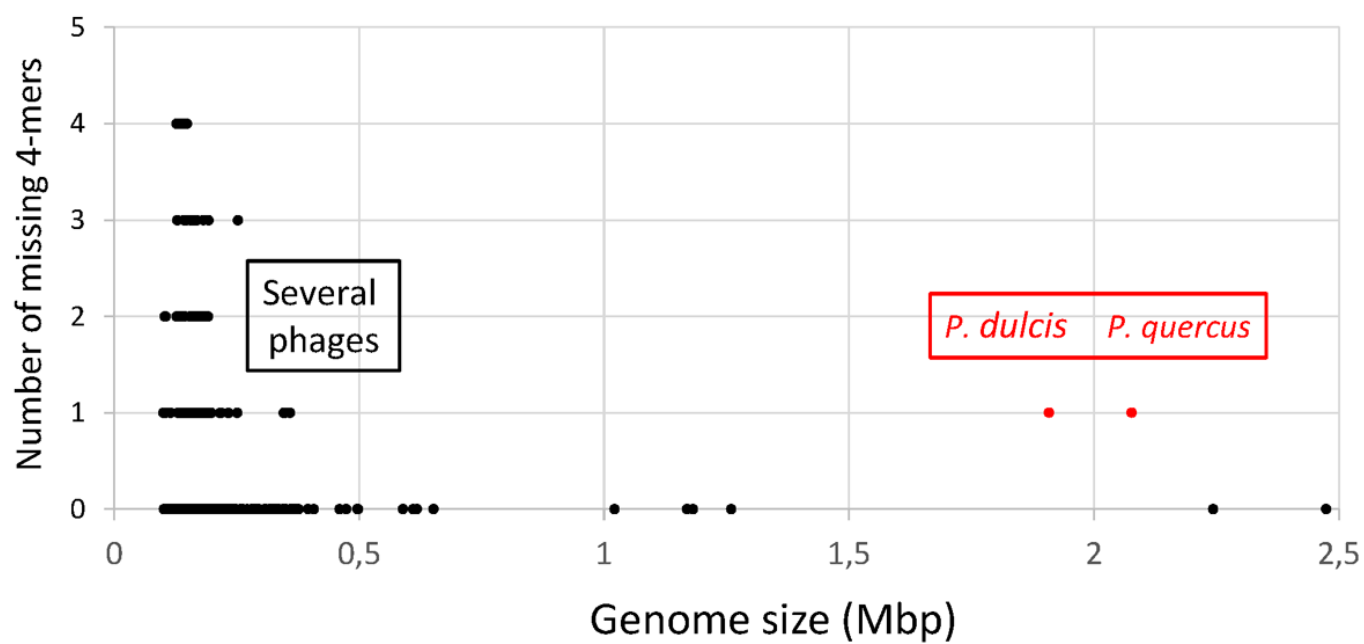


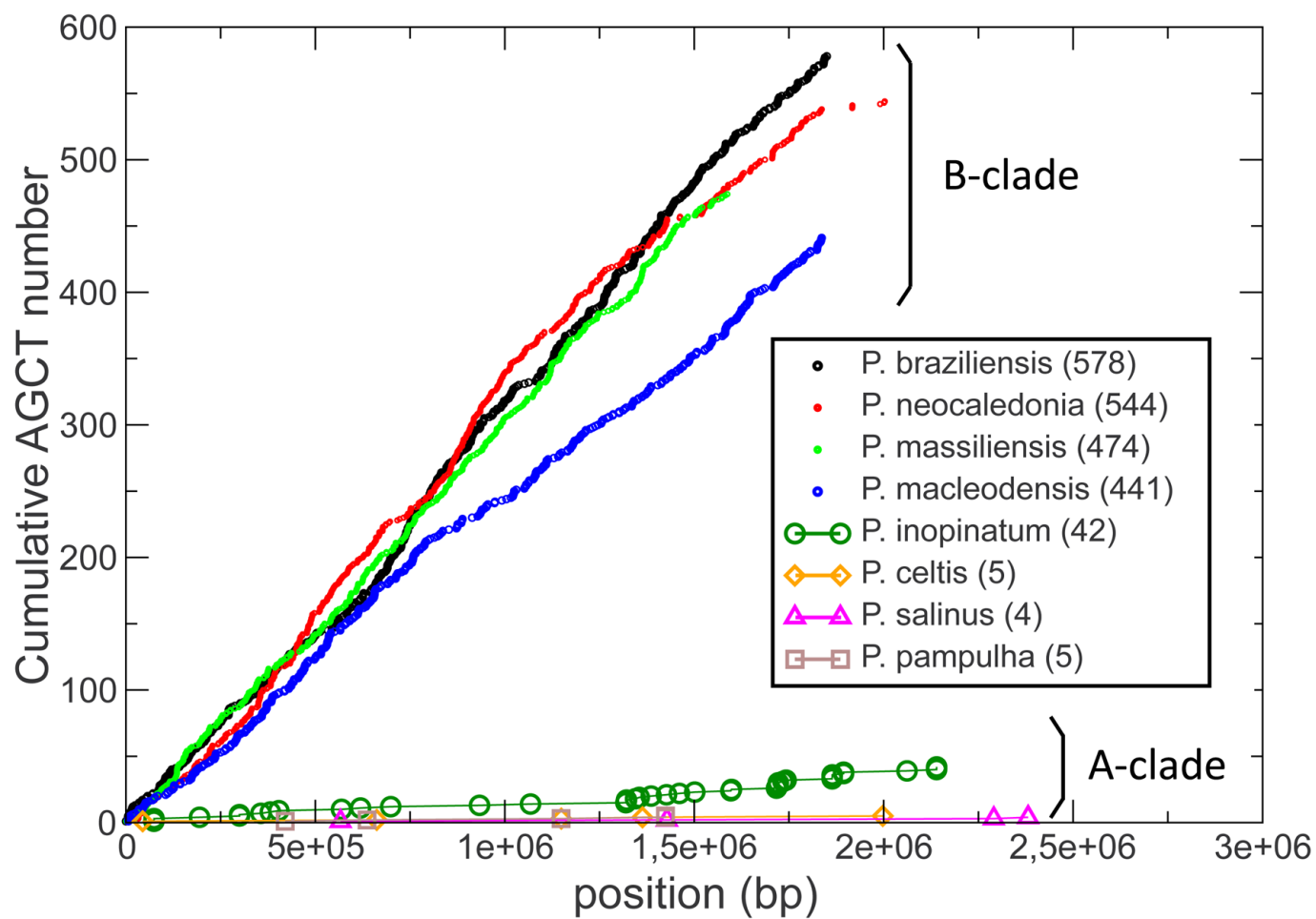
Pandoravirus dulcis, word size=4



Pandoravirus dulcis shuffled, word size=4







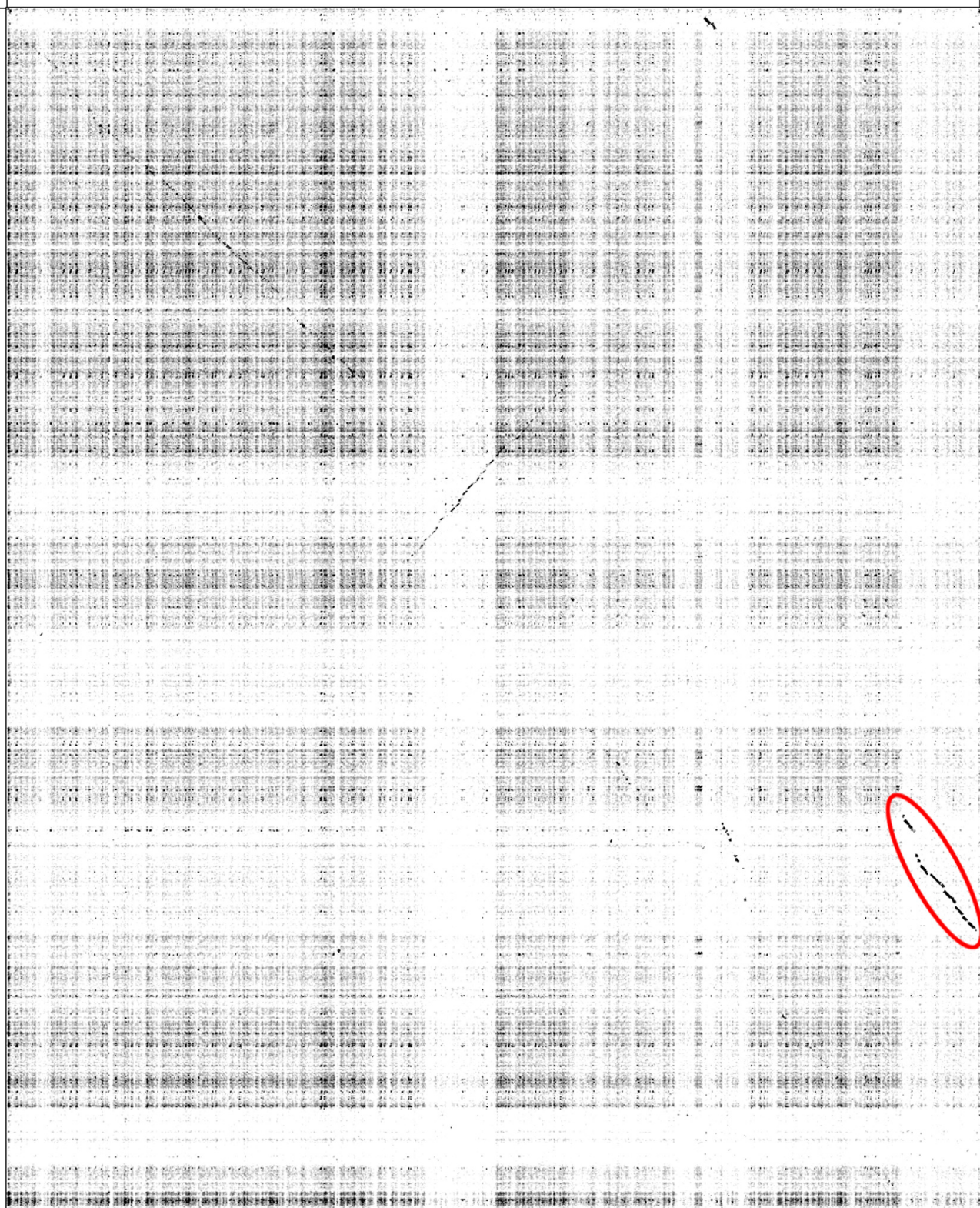
0 0

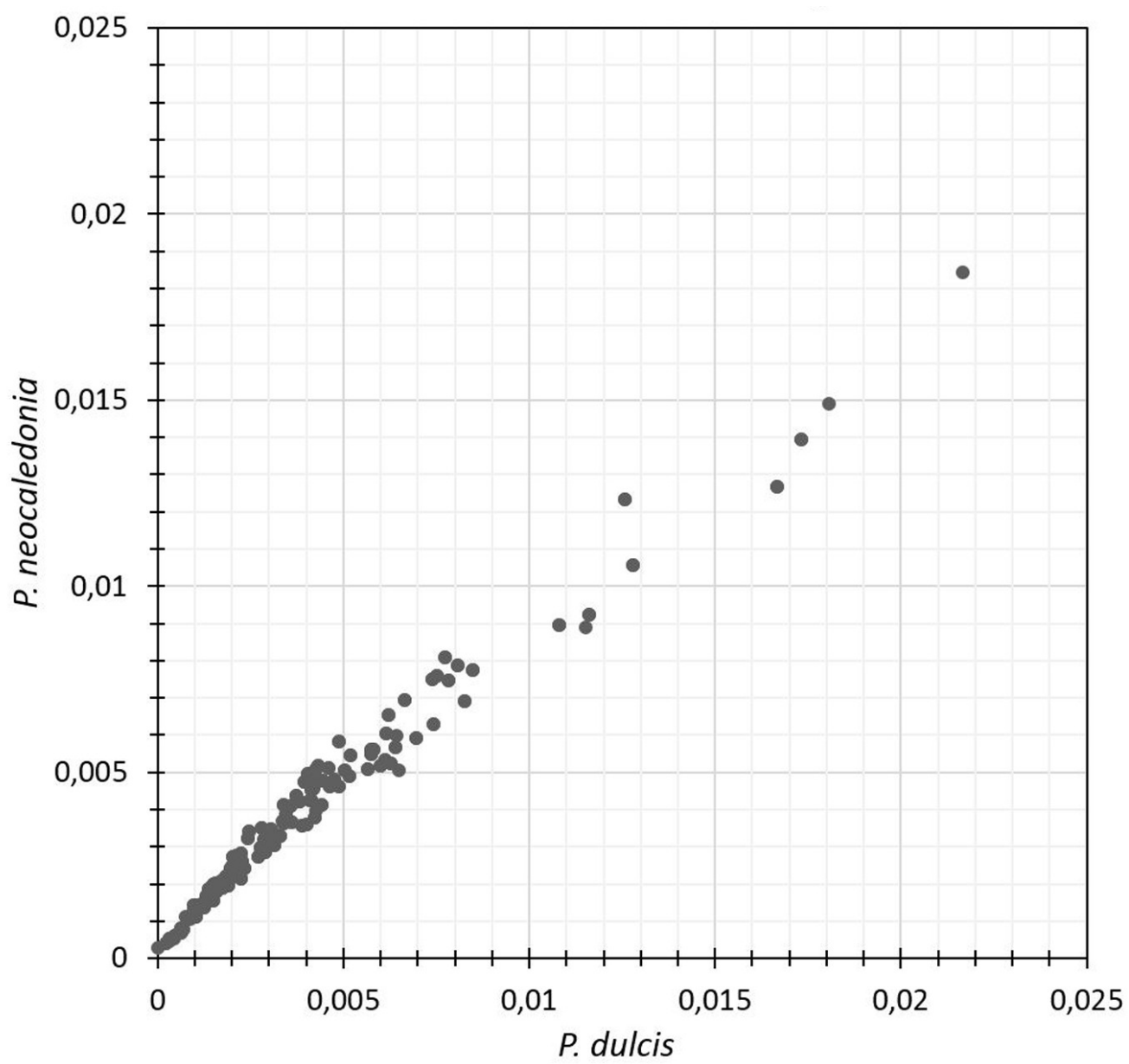
P. neocaledonia

2003190

P. salinus

2473869





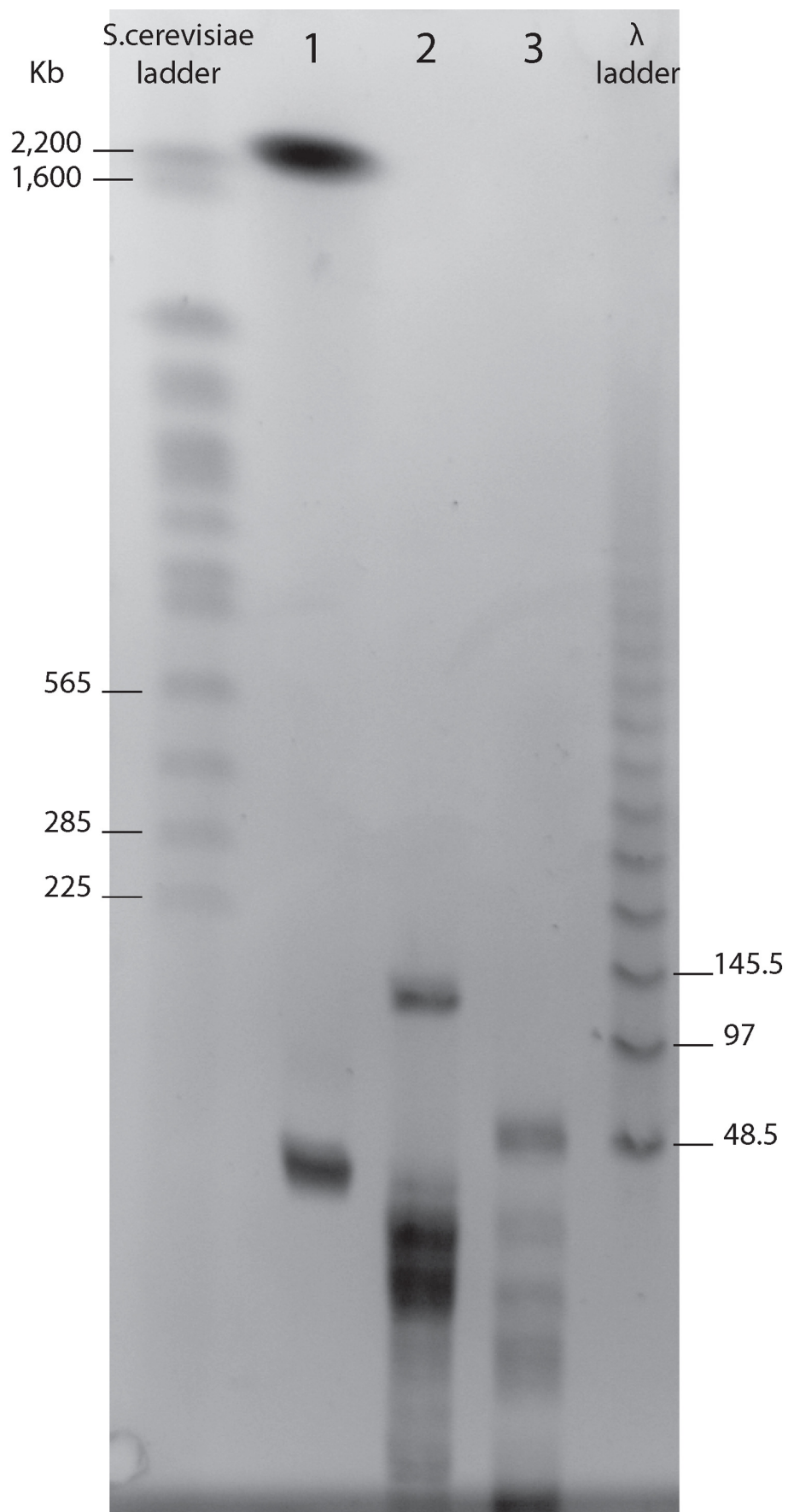


Table 1. Distribution of the AGC (and the complementary GCT) 3-mers

| Statistics | P. dulcis | | | P. quercus | | |
|--|------------------------------|------------------|------------------|------------------------------|------------------|------------------|
| Genome size (bp) | 1,908,524 | | | 2,077,288 | | |
| | interORF | ORF | global | interORF | ORF | global |
| AGC frequency (strand 1) | 0.0101 (1/99) | 0.0112 (1/89) | 0.0109 (1/92) | 0.0098 (1/102) | 0.0110 (1/90) | 0.0106 (1/94) |
| GCT frequency (strand 1) | 0.0102 (1/98) | 0.0156 (1/64) | 0.0138 (1/72) | 0.0097 (1/103) | 0.0145 (1/68) | 0.0129 (1/77) |
| AGC/GCT (2 strands, global) | 0.0123 (1/81) | | | 0.0118 (1/85) | | |
| AGC/GCT overall rank | 37/64 | | | 43/64 | | |
| p(AGC).p(T) | $2.24 \cdot 10^{-3}$ (1/446) | | | $2.31 \cdot 10^{-3}$ (1/432) | | |
| AGCT expected number (one strand x p(AGC).p(T)) | 4286 | | | 4898 | | |
| AGCT observed number | 0 | | | 0 | | |
| ACGT expected number (one strand x p(ACG).p(T)) | 7884 | | | 8387 | | |
| ACGT observed number | 5822 | | | 6165 | | |

Table 2. Homologous site replacements between *P. neocaledonia* and *P. dulcis*.

| <i>P. neocaledonia</i> → <i>P. dulcis</i> variant | Number |
|--|--------|
| AGCT → AGTT | 31 |
| AGCT → AACT | 18 |
| AGCT → GGCT | 4 |
| AGCT → AACC | 4 |
| AGCT → AATT | 3 |
| AGCT → GGCG | 2 |
| AGCT → [ACGA, ACTT, AGAT, AGCC, AGGC, CATT, GGCC, GGTT, GTCT, TGCC, TGGT, TGTC] | 1 |