1   Title:

2   Retroviral origins of the *Caenorhabditis elegans* orphan gene F58H7.5.

3

4   Author:

5   Wadim J Kapulkin MRCVS, DVM, PhD *#

6   #corresponding author

7   e-mail: <wadim_kapulkin@yahoo.co.uk>

8

9   Phone: (+44)7466449215

10

11

12   Present Affiliation:

13

14   * Veterinary Consultancy

15   Conrada 16/65

16   01-922 Warsaw

17   Poland

18

19

20

21

22

23

24   Abstract:

25

26   This work describes the results of the genome-scale analysis of

27   endogenous retrovirus insertions in two *C. elegans* isolates: the prototype N2

28   (Bristol) and CB4856 (Hawaii). In total thirteen, identification of potentially

29   replication competent, endogenous retroviral elements is described. Ten

30   elements were identified as conserved between N2 and CB4856 by the

31   reciprocal match of paired LTRs. The description focuses on the particular

32   endogenous retrovirus insertion wich is identified on the proximal arm of the

33   chromosome IV (located at positions IV: 912,948 – 921,658 and IV: 899,767 – 908,485

34   of the N2 and CB4856 respectively). In both isolates the inserted provirus is

35   flanked by the predicted long terminal repeats (LTR)s of the length of 415 bp and

36   of identical sequence. Provided the absolute LTR sequence identity this

37   particular provirus represents insertion acquired prior to split from the common

38   ancestor, suggesting this insertion event is evolutionary recent. The identified

39   insertion of the endogenous retrovirus embeds the orphan gene F58H7.5,

40   specific to *C. elegans* lineage. This unprecedented example establishes that in

41   the evolutionary past *C. elegans*, had acquired the gene of the retroviral origins

42   presumably via mechanisms involving the RNA intermediate.

43

44

45

46

47   Importance:

48

49   This work describes the retroviral origins of *C. elegans* orphan gene F58H7.5.

50   Presented work implies that in the evolutionary past the *C. elegans* have

51   acquired new gene as a result of the infection event. *C. elegans* is presently

52   regarded as genetic model organism widely used in genetic research. The

53   genome of *C. elegans* have been sequenced nearly 20 years ago. This

54   unprecedented example establishes that in the evolutionary past C. elegans

55   genome, had acquired the gene of the retroviral origins presumably via

56   mechanisms involving the RNA intermediate.

57

58

59

60

61

62

63

64

65

66

67

68

69

70   Introduction:

71

72   Endogenous retroviruses and LTR retrotransposons are ubiquitously

73   present in eukaryotic genomes. Endogenous retroviruses and LTR

74   retrotransposons are widely assumed to replicate via RNA intermediates.

75   Endogenous retrovirus insertions present in the genomes are thought to

76   represent the remnants of the infectious events of the past evolutionary history.

77   Examples of an active endogenous retroviruses however are found almost

78   invariably among metazoan genomes, and contribute to the reverse flow of the

79   genetic information. Benign endogenous retroviruses that are found inserted into

80   genomes of animals and are known to contribute to phenotypic traits and overall

81   genetic variation. Importantly, certain endogenous retroviruses are known for the

82   potential to convert into active retroviruses (i.e. Moloney Leukemia Virus (Stoye

83   and Coffin 1987)) and Mouse Mammary Tumor Virus (Bittner 1936)) and those

84   are grouped together with circulating retroviruses (i.e. Rous Sarcoma Virus

85   (Rous 1911)) often with significant pathogenic potential an therefore pose an

86   active and significant threat to the health of the animal populations. Infectious

87   properties of the retroviral particles led to development of the retroviral vectors as

88   means to efficiently insert proviral DNA into genomes. In contrast, particularly

89   well studied genetic model organism, *C. elegans* appears to harbor considerably

90   fewer LTR elements than other animal models. Further, sparse endogenous

91   retroviruses inserted into *C. elegans* appear mostly dormant and do not appear

92   to significantly contribute to observable traits. One notable exception is the

93   RETR-1 element (Britten 1995) inserted into *C. elegans* plg-1 locus, (Papoli et al.

94   2007) contributing to natural variation in the copulatory plug polymorphism

95   (Doniach and Hodgkin 1995). The active transposition and replication of the

96   RETR-1 however has not been demonstrated (Preiss 2007), neither any of the

97   other endogenous (nor exogenous) retroviral elements present in *C. elegans*

98   genome. This work describes the identification of the 'first –pass set' of the

99   potentially replication competent proviruses inserted into *C. elegans* genome

100   present in both N2 (Bristol) and CB4856 (Hawaii) lines. The description is

101   focused on the distinct and unique proviral insertion on the proximal arm of the

102   chromosome IV identified by the reciprocal LTR match. This particular prediction

103   distinctly embeds the orphan gene of *C.elegans* F58H7.5. Given the recent

104   improvement in the genome engineering i.e. with the advancement in the

105   CRISPR-Cas9 methods (Jinek et al. 2012) applicable to *C.elegans* (Friedland et

106   al. 2013), I suggest the identified set might serve the practical and experimental

107   purpose: the proviruses once inserted into the *C.elegans* genome in the

108   evolutionary past could now be ablated (Yang et al. 2015).

109

110

111

112

113

114

115

116   Materials and methods:

117

118   Genebank-NCBI reference sequence of the prototype N2 (Bristol) *C.*

119   *elegans* genome (C. elegans Sequencing Consortium 1998): Chr I

120   NC_003279.8, Chr II NC_003280.10, Chr III NC_003281.10, Chr IV

121   NC_003282.8, Chr V NC_003283.11, Chr X NC_003284.9 and Hawaiian isolate

122   CB4856 (Thompson et al. 2015): I Chromosome CM003206.1, II Chromosome

123   CM003207.1, III Chromosome CM003208.1, IV Chromosome CM003209.1, V

124   Chromosome CM003210.1, X Chromosome CM003211.1.

125

126   *C. elegans* chromosome sequence assemblies were analyzed with LTR

127   finder program (Zhao Xu, Hao Wang 2007) at default parameters with tRNA

128   primer binding site predicted using *C. elegans* tRNA dataset, to detect individual

129   flanking LTR pairs. The first-pass screening, focused on the predictions were at

130   least one internal candidate reading frame were suggested by internally enabled

131   ScanProsite. DNA sequences predicted by the LTR finder were conceptually

132   translated into reading frames (<molbiol.ru/eng/scripts/01_13.html>) resulting in

133   strings of the one-letter amino-acid codes. Individual reading frames of coded

134   amino-acid strings were masked for in-frame stop codons. Stop-masked amino-

135   acid strings were analyzed with the external InterProScan search engine

136   (<www.ebi.ac.uk/Tools/pfa/iprscan/ > Zdobnov et al. 2001) for the retroviral

137   domain detection and output was cross-verified with the output of the LTR finder.

138   LTR identity in an isolate matched candidate predictions of the LTR pairs were

139 verified by the reciprocal sequence alignment (Altschul et al. 1990). Predicted

140 reading frames were verified with analysis by taxon restricted homology search

141 (either with the set of the NCBI reference sequence of the retro-transcribing

142 viruses or endogenous retroviruses). Identified proviruses were displayed on

143 Ensembl *C. elegans* N2(Bristol) database by BLAT (Kent 2002). The protein

144 domains architectures ideograms and the pairwise LOGO alignments were

145 drawn with MyDomains prosite (Hulo et al. 2008)

146 (<prosite.expasy.org/mydomains/>) and Weblogo3 (Crooks ey al. 2004,

147 Schneider and Stephens 1990) (<weblogo.threeplusone.com>) respectively. The

148 genomic landscape surrounding the confirmed retroviral insertions was GBrowse

149 (<gbrowse.org>) displayed at the WormBase website. Phylogenetic analysis on

150 predicted retroviral domains was conducted with pipeline in ETE3 at

151 (<www.genome.jp >). Alignment and phylogenetic reconstructions were

152 performed using the function "build" of ETE3 v3.0.0b32 (Huerta-Cepas et al.,

153 2016) as implemented on the GenomeNet (<www.genome.jp/tools/ete/>). The

154 compiled input files for reverse-transciptase polymerase, RNaseH and integrase

155 catalytic domain are given in the supplement files. Alignment was performed with

156 MAFFT v6.861b with the default options (Katoh and Standley, 2013). The initial

157 tree was constructed using FastTree v2.1.8 with default parameters (Price et al.,

158 2009). The ML bootstrapped trees were inferred either using RAxML v8.1.20 ran

159 with model PROTGAMMAJTT and default parameters (Stamatakis, 2014) and

160 PhyML v20160115 ran with model JTT and parameters: -f m --pinv e -o tlr --

161 nclasses 4 --bootstrap 100 --alpha e (Guindon et al., 2010). Branch supports

162 were computed out of 100 iterations.

163

164

165

166 Results:

167

168 The details of the genome-scale analysis are included in materials and

169 methods section and the results of the first pass screening are listed with (Table

170 1. and Table 4.) and graphically outlined (Fig. 5), however, will be described

171 elsewhere. Here, I describe the identification and analysis of the unusual provirus

172 inserted in sense (+) orientation on Chromosome IV of both prototype N2(Bristol)

173 strain and Hawaiian isolate CB4856. The provirus is found inserted at the

174 proximal arm of IV at positions IV:912948..921658 and IV:899767..908485 of an N2

175 and CB4856 respectively by the reciprocal match of the LTR pairs (Fig. 1.A. and

176 B). The two extracted sequences (8711 bp and 8719 bp respectively) align

177 precisely to ensembl N2(Bristol) assembly, with duplicate matches corresponding

178 to long terminal repeats (Fig. 1. C.). This suggests the two elements identified at

179 the proximal arm of the chromosome IV, might represent provirus inserted in the

180 ancestral *C. elegans* lineage, which have independently diverged since the

181 insertion time. To prove the identified provirus share a common ancestor, 1kb of

182 immediately flanking the insertion sequence was extracted and aligned,

183 confirming the 5' and 3' flanks were almost identical and have diverged

184   independently by four base substitution per thousand nucleotides (0.4%)

185   (Supplement file.1. Alignment of the insertions with 1kb flanking sequence). The

186   striking sequence conservation detected in the provirus flanking sequence

187   between N2(Bristol) and CB4856, indicates the element once inserted into the

188   genome of the ancestral *C. elegans* line, was inherited as a haplotypic block and

189   presumably maintained in same location on chromosome IV in both isolates.

190

191   **LTRs - Long Terminal Repeats** (Fig 1.A. and B. and supplement files). The

192   LTRs are identical in length in both *C.elegans* isolates. N2(Bristol) 5'-LTR is

193   located at IV:912,948 – 913,362 and 3'-LTR IV:921,244 – 921,658 and both

194   LTRs are of same length 415 bp. Hawaii CB4856 5'-LTR is located on

195   IV:899,767 – 900,181 and 3'-LTR IV:908071 – 908485 and both LTRs are of

196   same length 415 bp (Supplement file 2. LTR alignment). Comparison between 5'

197   LTRs and 3' LTRs in both isolates confirms the lack of divergent bases in the

198   LTR alignment. Provided the accepted model of the retroviral replication

199   (Telesnitsky A and Goff SP 1998) assumes the LTRs are identical at the insertion

200   time, therefore the lack of divergence in the LTR sequences in both geographical

201   isolates suggests that the identified insertion represent the recent evolutionary

202   event.

203

204   **PBS.** Primer Binding Sites (Fig 1.A. and B. and supplement files) are predicted

205   by LTR finder as sequences immediately downstream (12nt) of 5' LTR matching

206   the tRNA primer for the reverse transcription of viral RNA (Telesnitsky A and Goff

207   SP 1998). The LTR finder program identified the same iso-accepting methionyl

208   (CAT) tRNA to bind conserved PBS [5'-TAGCTAGCGAGTGAACCGAATTTCG]

209   (IV:913,374 – 913,398) in the C.elegans N2(Bristol) insertion and Hawaiian

210   CB4856 isolate (IV:900,193 – 900,217).

211

212   **PPT.** Poly Purine Tract sequence (Fig 1.A. and B. and supplement files)

213   immediately proceeding the 3' LTR are identified by the LTR finder in both

214   proviruses and are identical in two analyzed isolates [PPT: 5'-

215   TCAAAAGGGGGGAGG] located at (IV:921,229 – 921,243) in the C.elegans

216   N2(Bristol) insertion and (IV:908,056 – 908,070) Hawaiian CB4856 isolate.

217

218

219   The insertions of the provirus at the proximal arm of the chromosome IV,

220   are almost identical in the overall length. The provirus in the reference N2(Bristol)

221   assembly is 8,711 nt long and CB4856(Hawaii) is minimally longer 8, 719 nt

222   (Table 1. and supplement file). Given the LTRs, are of identical length in both

223   isolates and essentially lack any aberrant bases, the region between LTRs

224   (coding for retroviral proteins) is expected to harbor divergent nucleotides

225   accounting for the difference in the overall length. BLAT analysis indicates the

226   two regions (Fig.1 C. indicated by the black arrowheads) interrupting the

227   concordance of the alignment of the provirus variant found in chromosome IV in

228   CB4856 and the reference N2 assembly. Those two regions are displayed in

229   pairwise alignment (Supplement file 1.) and occur in region of provirus at bases

230   in range 5,00-5,57kb (Fig. 2) in a region connecting segments encoding for

231   Reverse Transcriptase-RNaseH and catalytic integrase proteins. Compared

232   regions in the reference assembly of an N2 and CB4856 assembly, identifies the

233   following segments as inconsistent between isolates: 1. Four base deletion in N2

234   (deletion flank GAC TTC----CGA CGC) or (ATGT) insertion in CB4856. 2. Three

235   base insertion in N2 (insertion flank CTG GGT – TCG CCG) or (TGC) deletion in

236   CB4856. 3. Seven base pair deletion in N2 (deletion flank AGT AGT------- ACA

237   TGG) or (ATCATGA) insertion in CB4856, with G to T single base substitution in

238   the left flank region. Collectively across the above 5,00-5,57kb region cumulative

239   aberrant bases (including two substitutions not altering the length) contribute to

240   the overall ~2,8% of divergence between two isolates.

241

242   As indicated in (Fig.1.A and 1.B), the provirus inserted into N2(Bristol) is

243   assigned two retroviral domains (Reverse transcriptase RT- domain: at

244   IV:916,714 – 917,250 and Integrase catalytic domain: IV:918,494 – 918,889) by

245   the LTR finder internally enabled PrositeScan. In contrast, CB4856 inserted

246   element is attributed with only one retroviral domain (Integrase catalytic domain:

247   IV:905,294 – 905,716). Provided high level of the overall sequence conservation

248   over the entire length of the insertion present in both isolates, that asymmetric

249   display is somewhat unusual for the elements shared between N2 and CB4856.

250   To address this discrepancy, the proviruses identified in N2 and CB4856, were

251   subjected to analysis with external protein domain search the InterProScan.

252

253 **Predicted proviral proteins and protein domains.** Provirus IV-8711 in

254 N2(Bristol) and IV-8719 in CB4856(Hawaii) exhibits organization of protein

255 domains typical for endogenous retroviruses/retrotransposons (Fig.1.C). The

256 protein domains predicted by the external InterProScan are presented as

257 pairwise consensus LOGO alignment of particular N2/CB4856 protein domain

258 pairs in (Fig.3.). The protein domains predicted by the external InterProScan

259 sequence LOGO's are presented in 5'->3' order as they occur encoded in the

260 provirus (Fig.3.A, gag; B. protease; C. RT polymerase; D. RNase H; E. integrase)

261 with respect to predicted reading frames. The summarized InterProScan search

262 results (below) demonstrate that the provirus identified in IV-8711 in N2(Bristol)

263 and IV-8719 in CB4856(Hawaii) in overall genetic organization appears typical

264 for the endogenous retroviruses. However an additional open reading frame is

265 found in the IV-8711/IV-8719 provirus, encoded by the ORF present in the region

266 between gene encoding the catalytic integrase protein and the 3'-LTR. This sub-

267 3'-LTR localization is typical for the retroviral envelope protein (env), but provided

268 the lack of the homology with retroviral env genes is termed below 3'-ORF and

269 represent the uncharacterized *C. elegans* protein F58H7.5 (Fig 1.C).

270

271 (i.) gag. Protein domain specifying the gag gene is defined by the Pfam protein

272 domain signature match (`PF03732/InterPro:IPR005162`). Gag is encoded in the

273 reading frame (+1) and is identical between IV-8711 in N2(Bristol) and IV-8719 in

274 CB4856(Hawaii) proviruses (Fig.3.A). Protein domain encoding for gag is 95

275 amino-acid residues long and encoded in positions: (1951-2235) in both isolates.

276

277   (ii.) pro. Protein domain specifying the Retroviral-type Aspartyl Protease gene is

278   defined by the Pfam protein domain signature match (`PF13650`). Aspartyl protease

279   is encoded in the reading frame (+1) and is identical between IV-8711 in

280   N2(Bristol) and IV-8719 in CB4856(Hawaii) proviruses (Fig.3.B). Protein domain

281   encoding for aspartyl protease is 95 amino-acid residues long and encoded in

282   positions: (2791-3075) in both isolates.

283

284   (iii.) pol. Protein domain specifying the Reverse-transcriptase polymerase gene

285   is defined by the ProSite protein domain signature match

286   (`PS50878/InterPro:IPR000477`). Reverse-transcriptase polymerase is encoded in

287   the reading frame (+2) and is not identical between IV-8711 in N2(Bristol) and IV-

288   8719 in CB4856(Hawaii) proviruses (Fig.3.C). Protein domain encoding for

289   aspartyl protease is 179 amino-acid residues long and encoded in positions:

290   (3767-4303) in both isolates. Across the Reverse-transcriptase polymerase

291   sequence there are three non-synonymous substitutions detected (positions 20

292   and 54 altering I->V and position 126 M->T) diverged IV-8711 in N2(Bristol) and

293   IV-8719 in CB4856(Hawaii) proviruses. In addition, the reference N2 sequence

294   contains in-frame stop codon predicted in position 119, absent in CB4856,

295   substituting for the serine in this position (Fig.3.C).

296

297   (iv.) RNaseH. Protein domain specifying the Retroviral RNaseH gene is defined

298   by the CDD protein domain signature match (`cd09274`). RNaseH is encoded in

299   the reading frame (+2) and is identical between IV-8711 in N2(Bristol) and IV-

300   8719 in CB4856(Hawaii) proviruses (Fig.3.D). RNaseH protein domain is 121

301   amino-acid residues long and is encoded in positions: (4583-4954) in both

302   isolates.

303

304   (v.) int. Integrase catalytic domain. Protein domain specifying the catalytic

305   integrase gene is defined by the ProSite protein domain signature match

306   (`PS50994/InterPro:IPR001584`). Markedly the Integrase is encoded in the

307   different reading frames: (+3) and (+2) frames are used in IV-8711 in N2(Bristol)

308   and IV-8719 in CB4856(Hawaii) proviruses respectively (Fig.3.E) and (Table.2).

309   This frame-shift is due to mutations listed in (Fig.2.3) and explains the lack of an

310   alignment in N-terminal part of predicted integrase sequence, until the metionine

311   residue in the position 32. Remaining part of the protein domain encoding for

312   integrase is 130 amino-acid residues long and is identical between IV-8711 in

313   N2(Bristol) and IV-8719 in CB4856(Hawaii) proviruses. Integrase catalytic

314   domain is encoded in positions: (5457-5942) and (5465-5950) in IV-8711 in

315   N2(Bristol) and in IV-8719 in CB4856(Hawaii) respectively.

316

317   (vi.) 3'-ORF. 3'-ORF embeds an uncharacterized protein F58H7.5 predicted in

318   chromosome IV of the reference sequence assembly of the *C. elegans* genome

319   (Fig. 1. C). In the InterProScan search I used to improve the retroviral domain

320   predictions, this unusual prediction is specified by the Phobius

321   (<phobius.sbc.su.se>) protein topology match. 3'-ORF is encoded in the reading

322   frame (+2) and (+1) in IV-8711 in N2(Bristol) and IV-8719 in CB4856(Hawaii)

323   proviruses (Fig.3.F) and (Table.2.) respectively. The 3'-ORF is defined by the

324   F58H7.5 prediction. Predicted 3'-ORF embeds the F58H7.5 starting from position

325   53 until 324, followed by the in-frame stop codon conserved in both isolates.

326   Provided the F58H7.5 coding sequence is continuous (intronless gene) it is

327   estimated it constitutes approximately 48,5% of the 3'-ORF. Within the F58H7.5

328   sequence there are two non-synonymous substitutions detected (3'-ORF position

329   87 H->P and 171 E->K) as divergent between IV-8711 in N2(Bristol) and IV-8719

330   in CB4856(Hawaii) proviruses (Fig.3.F). 3'-ORF is encoded in positions: (6887-

331   8563) and (6895-8571) in IV-8711 in N2(Bristol) and IV-8719 in CB4856(Hawaii)

332   respectively. Remarkably F58H7.5 is an orphan gene, taxonomically restricted to

333   *C. elegans* lineage. Conducted homology based search analysis indicates

334   F58H7.5 is uniquely present within *C. elegans* species, and therefore absent

335   from other caenorhabdids (including the sister species). This narrow taxonomic

336   distribution of F58H7.5 homology is in the stark contrast with the other proviral

337   frames encoding for the typical retroviral domains described above.

338

339

340   **Comparison with RETR-1 (III: 8852597 - 8861482).** Provided we demonstrated IV-

341   8711 in N2(Bristol) and IV-8719 in CB4856(Hawaii) proviruses, are 'identical-by-

342   descent' and diverged only in non-LTR regions, we compared the isolate specific

343   sequences to RETR-1 retrovirus (Britten 95). RETR-1 is distinctly detected by

344   LTR finder (Table I), specifically in the N2(Bristol) Chromosome III sequence.

345   This asymmetric display of the element is different from observed for IV-8711

346   and IV-8719 insertion, apparently present in both isolates. While the description

347   and detailed analysis of the RETR-1 is behind the scope of this work (will be

348   described eswhere), I compared IV-8711 in N2(Bristol) and IV-8719 in

349   CB4856(Hawaii) proviruses to asymmetric insertion of the RETR-1 inactivating

350   the plg-1 gene. RETR-1 (Table. 1) as displayed in N2(Bristol) chromosome III, is

351   8.886 kb in length (with LTR finder predicted 5' and 3' LTR sequences 518 bp

352   and 513 bp respectively). The particular reason of choosing the RETR-1 for the

353   comparative analysis is that in contrast with most other LTR finder predicted

354   retroviral insertions listed with (Table 1.), insertion into chromosome IV shared

355   between N2 and CB4856 aligns significantly with BLAST analysis with relatively

356   well characterized nematode endogenous retrovirus RETR-1. The overall domain

357   architecture of the RETR-1 used for comparison is slightly different to what is

358   described in the previous section concerning the IV-8711 in N2(Bristol) and IV-

359   8719 in CB4856(Hawaii) proviruses. The InterProScan search on the coded, stop

360   masked RETR-1 conceptual translation (described in materials and methods),

361   revealed the all relevant features of the protein domain organization recognized

362   other retroviruses, are grouped into frame (+1). This single (+1) frame of the

363   RETR-1, is fairly long and apparently continuous (uninterrupted by the in-frame

364   stop codons). InterProScan on the RETR-1 single frame (+1) identified regions

365   encoding for conserved protein domains corresponding to reverse-transcriptase

366   polymerase (PS50878/ InterPro:IPR000477), RNaseH (CDD: cd09274), and

367   catalytic Integrase (PS50994/InterPro:IPR001584). In contrast, two additional

368   protein domains typical for the retrovirus genome architecture, identified in the

369 provirus inserted into N2/CB4856 ancestral locus on chromosome IV, was not

370 identified in the RETR-1 using the same search criteria. Notably Pfam protein

371 domain signature matches (`PF03732/InterPro:IPR005162`) specifying gag gene

372 and (`PF13650`) specifying retroviral-type Aspartyl Protease gene respectively, are

373 apparently missing from (+1) long reading frame encoded by the RETR-1 (and

374 remaining reading frames). Positively identified protein domain signature

375 matches are compared in the Table.3 (Individual alignments are included in

376 supplementary materials). This analysis demonstrates the 3'-ORF embedding

377 F58H7.5 prediction is a unique and distinct part of the proviral insert in IV-8711 in

378 N2(Bristol) and IV-8719 in CB4856(Hawaii), but consistently absent from the

379 RETR-1 (and any other presently known reverse-transcribing agents (not

380 shown)).

381

382

383 **Comparison with other reverse-transcribing viruses.** Given the original

384 description and sequence analysis described in (Britten 1995) indicated the

385 remarkable homology in the RT polymerase region of RETR-1 to Cauliflower

386 Mosaic Virus (CaMV) RT (Table 1. in (Britten 1995)), we inferred that the

387 insertion detected by LTR finder at the *C. elegans* chromosome IV (IV-8711 in

388 N2 and IV-8719 in CB4856) represent related clan of endogenous retroviruses.

389 The reason for that is the reverse transcriptase polymerase region predicted by

390 the InterProScan in both IV-8711 in N2 and IV-8719 in CB4856 (Fig.3.C), aligns

391 with the CaMV RT and reverse transcriptases of other Caulimoviruses (not

392 shown). Therefore, the reverse transcriptase-polymerase domain phylogenetic

393 tree was build, using the predicted RT protein domains of retroviruses and other

394 reverse-transcribing agents (Fig.4). Both reverse-transcriptase polymerase

395 domain and RNaseH domain fast-tree cladograms of the reverse-transcribing

396 agents, the Caulimoviruses are nested within IV-8711-IV-8719/RETR-1 group.

397 Consistently, with two types of Maximum-Likelihood analysis (where RaxML and

398 PhyML nodes are supported by a bootstrap (Efron et al. 1996)), Caulimoviruses

399 are nested within the IV-8711-IV-8719/RETR-1 group. In particular, the reverse-

400 transcriptase polymerase domain cladograms include the RT-polymerase domain

401 predicted on HBV (Hepatitis B virus), which appear as an outgroup. The reverse-

402 transcriptase polymerase domain cladograms place the retrovirus clade and IV-

403 8711-IV-8719/RETR-1/Caulimovirus branch as sister taxa. The same relation is

404 observed in the cladorograms build based on RNaseH domains, however here

405 the outgroup root (HBV) is removed. Together in all examples provided (catalytic

406 integrase domain tree is not relevant here, as Caulimoviruses in most cases

407 maintain their DNA episomaly (Hull et al. 1987) and thereof do not encode

408 retroviral type integrase) IV-8711-IV-8719/RETR-1 group is basal to

409 Caulimovirus clade and therefore IV-8711-IV-8719/RETR-1 group appears as

410 paraphyletic taxon with respect to pararetroviruses. This result seems suprising

411 as Pararetroviuses existing in plants are currently grouped with Hepadnaviruses

412 (Coffin et al. 1997).

413

414

415   Discussion:

416

417   Retroviral insertions into animal genomes are established metrics of

418   dating the evolutionary time based on the LTR divergence. According to the

419   retrovirus replication model (Coffin et al. 1997) accepted for the endogenous

420   retroviruses and the LTR-retrotransposons, the LTRs are identical at the time of

421   an insertion, than subsequently diverge, acquiring independent mutations at

422   neutral rate. Examples of dating of the retroviral insertions and estimates of the

423   evolutionary splits are available in the literature i.e. (Johnson and Coffin 1999)

424   (Yohn et al. 2005) (Arnaud et al. 2007) (Jo et al. 2012) and are based on the

425   calculations which in principle could be used and applied towards dating of

426   retroviral insertions in *C. elegans* isolates. Those estimates could be useful in

427   dating of the evolutionary split between the reference N2(Bristol) and Hawaiian

428   isolates. In a given example, the conserved insertion pair of the IV-8711 in N2

429   and IV-8719 in CB4856, diverged only minimally. Presented evidence concerning

430   the divergence in the LTR region indicates the long terminal repeats in both

431   isolates did not acquired any new mutations since a split from the prototype

432   ancestral *C. elegans*. Therefore, I regard IV-8711 insertion represents the

433   evolutionary recent event. According to the accepted model this suggests the

434   split between N2(Bristol) and Hawaiian isolates could be estimated as

435   remarkably recent, essentially behind the resolution power of the dating. On the

436   other hand however, within the presented set of the retroviral insertions

437   conserved between N2(Bristol) and Hawaiian isolates, there are examples which

438 contradict the above evolutionary scaling (data not shown). In few examples

439 there is an evidence for LTR acquired mutations prior to split of N2(Bristol) and

440 Hawaiian isolates. In those examples, the 5'- and 3'-LTRs diverged but the

441 particular substitution is same in both isolates. The presence of this class of

442 substitutions conserved in both isolates, suggest some of the insertions into

443 prototype *C. elegans* genome have persisted for sufficient time in the ancestral

444 *C. elegans* line to permit for the LTR divergence prior to split of an N2 and

445 CB4856. Provided the alternative estimates for the split of N2 and Hawaii have

446 been proposed (Thomas et al. 2015), I suggest those might need to be revised or

447 rescaled using the metrics based on the LTR divergence. While detailed follow

448 up analysis on those particular examples is emerging, it however remains behind

449 the scope of this description.

450

451 The architecture of the unusual 3'-ORF embedding the F58H7.5 is

452 described in the results section. F58H7.5 is an uncharacterized *C. elegans* gene

453 of presently unknown function encoding for an novel protein. Based on the

454 homology searches it seems the F58H7.5 is a distinct and unique, taxon

455 restricted, species specific, orphan gene. The above orphan gene assignment is

456 for the following rationale: F58H7.5 predicted protein is not associated with any

457 of the known protein families represented in *C. elegans* neither elsewhere in the

458 taxonomy tree. Homology based search on predicted protein F58H7.5 indicates

459 no informative homologies in other caenorhabdids neither the sister species *C.*

460 *briggsae*[1]. Moreover, no additional homology is found in searches excluding the

461 *Caenorhabditis* genus. The F58H7.5 gene is clearly embedded into provirus

462 prediction on chromosome IV-8711 in N2(Bristol) and IV-8719 in CB4856(Hawaii)

463 and is predicted to lay in the region between the catalytic integrase and 3' LTR

464 (result section and Fig.1.C). The F58H7.5 gene is encoded in single frame in

465 continuous, uninterrupted manner and is apparently expressed (by virtue of the

466 annotated mRNA (reference sequence # NM_067613.1 || GI:17540863)). Lack of

467 introns spanning the coding sequence appears markedly relevant, as is

468 consistent with the retroviral origins of F58H7.5. Retroviruses and other reverse-

469 transcribing agents, do replicate via RNA intermediate (Baltimore 1970) and

470 (Temin and Mizutani 1970) and it is widely accepted that introns inserted in the

471 sense (+) orientation will be removed from genomic transcripts (Boeke et al.

472 1985) prior to insertion into chromosomal loci. Provided the above rationale, as

473 main conclusion of the described work, I suggest the annotated gene F58H7.5 in

474 the reference N2 assembly, is a retroviral-borne gene acquired via RNA

475 intermediate, specifically into *C. elegans* linage prior to the split of the N2(Bristol)

476 and CB4856(Hawaii) isolates. I note further, that while it has long been

477 established the retroviruses could embed the non-viral genes (or its portions) and

478 insert those into chromosomal loci (Spector et al. 1978), the acquisition of the

479 taxon restricted, species specific gene however has not been previously

480 postulated.

---

[1] The WormBase homology information specifies the weak (BLAST e-value 5.8e-06 over the 31.2% of the protein length) match with CBG12079 gene. CBG12079 (119 amino-acid residues, were 69 residues is a repeat of continued stretch of homo-polymeric charged glutamates, below) is considerably smaller gene, compared to F58H7.5 (272 amino-acid residues), and is interrupted by single intron. CBG12079 gene itself is found in ~3.5kb intron of Cbr-unc-73 on Cb-Chr-I, and therefore collectively unlikely orthologous with F58H7.5. [available via <http://www.wormbase.org/species/c_briggsae/protein/BP:CBP22639#06--10>

```
>CBG12079
MISHRGIVKFDTINSQFLWLSSTSKRACQSISSTDEKEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEQKMRIRWMWITK)]
```

481

482   The insertion of the retrovirus presently located on chromosome IV-8711

483   in N2 and IV-8719 in CB4856, represent the example of the ancestral retroviral

484   infection contrasting the RETR-1 insertion on chromosome III. While both IV-

485   8711/IV-8719 and RETR-1 share the protein domains typical for the organization

486   of endogenous retrovirus genome, the RETR-1 display in the LTR finder analysis

487   is polar i.e. RETR-1 is detected explicitly in N2(Bristol) but missing from

488   CB4856(Hawaii) (Table.1. and author's analysis. Not shown.). RETR-1 was

489   identified by *C. elegans* genome sequence analysis (Sulson and Waterston

490   1998) and is inserted into chromosome III of the reference N2 (spanning the

491   clone border F44E2/PAR3 Alan Coulson pers communication 1997). In the N2

492   reference sequence the RETR-1 is regarded transcriptionaly active in germline

493   and embryonic tissues, as judged by multiple ESTs sequences and *in situ*

494   hybridization (Yuji Kohara, personal communication 1997, and NEXTdb pattern

495   <nematode.lab.nig.ac.jp/db2/ShowCloneInfo.php?clone=88f8>). Others (Maydan

496   et al. 2007 and Maydan et al. 2010) have found by comparative genomic

497   hybridization that N2 RETR-1 locus is polymorphic (with large deletion spanning

498   the RETR-1 insertion site in N2) when compared to CB4856. Consistently,

499   (Papoli et al. 2008) demonstrated that RETR-1 is inserted into F44E2.11 in N2

500   and concluded is absent from plg-1(+) CB4856. Those findings are in the

501   substantial agreement with the results presented here (Table 1.). First RETR-1 is

502   accurately predicted by the LTR finder and served as positive control in the

503   genome-scale search. Second major retroviral protein domains predicted in the

504 stop-masked frames of the ancestral retroviral element inserted on chromosome

505 IV, appear conserved with domains predicted on the RETR-1, including reverse-

506 transcriptase, RNaseH and catalytic integrase. The presented results however

507 demonstrate, that comparing the gag gene region predicted consistently on the

508 stop-masked frame (+1) the IV-8711 in N2 and IV-8719 in CB4856 is absent from

509 RETR-1 (by direct comparison and by *de novo* InterProScan analysis). This rise

510 the possibility that gag gene in the RETR-1 (III) element is either rearranged

511 (deleted) or diverged significantly from other members of the clade, to sufficient

512 degree to prevent detection. Indeed the rearrangements appear common among

513 endogenous retroviral insertions (authors unpublished observation). In this

514 perspective the RETR-1 often regarded as 'active' (the original communication of

515 (Britten 1995) but also by (Dennis et al. 2012) who tackled the RETR-1

516 experimentally), might in the fact represent immobile (albeit transcriptionaly

517 active) insertion into plg-1 gene. Third, considering the evolutionary scale since

518 the insertion of into the locus on chromosome IV present in both N2 and Hawaii

519 isolates and RETR-1, it appears the later element was inserted into the N2

520 genome after the split of the Hawaii from the ancestral *C. elegans* line. While this

521 model would possibly explain the polar distribution of the RETR-1 (and possibly

522 some other elements detected discretely in the N2 (Table.1.) but not in CB4856),

523 the results of the LTR alignments appear to contradict this perspective. As

524 mentioned above the 5' and 3' LTRs in the elements present in the IV-8711 in N2

525 and IV-8719 in CB4856, are identical. In contrast the 5' and 3' LTRs predicted by

526 the LTR finder on the RETR-1 are different in length (518 and 513 bp

527   respectively (Table 1.), suggesting the insertion or deletion must have occurred)

528   and clearly acquired aberrant bases (>10 aberrant bases between 5' and 3' LTR)

529   at the 3' end, since the time insertion. Importantly, the LTR finder predicts so

530   called TSR (Target Site Repeats, short duplicated sequences of the host at the

531   integration site), implicated to occur by the accepted retroviral integration model

532   (Brown 1997) adjacent to RETR-1 insertion site [TSR: CGACTA]. The presence

533   of the TSRs flanking the RETR-1 insertion site might rule-out the inaccurate

534   prediction of LTRs. Considering, the observed divergence rate of the LTRs in the

535   RETR-1 and the divergence rate of the LTRs present in IV-8711 in N2 and in IV-

536   8719 in CB4856, those findings appear contradictory, if the neutral substitution

537   model of the LTRs divergence is assumed. This inconsistency might be due to

538   selective pressure on some endogenous retroviral insertions loci (i.e. serving as

539   the host restriction factors) or the effects of the negative selection (i.e. if the

540   insertion results in the trait affecting the overall fitness of the carrier) or perhaps

541   for some other reasons. This note however should serve as the word of caution,

542   considering the long distance evolutionary extrapolations based on the single

543   isolated example (given the examples supporting other evolutionary scenarios

544   could be drawn out of the examples given in (Table.1.). Those examples will be

545   discussed elsewhere).

546

547   Phylogenetic reconstructions of the predicted retroviral protein domains. The

548   intimate kinship established for the ancestral *C. elegans* provirus IV-8711 in N2

549   and in IV-8719 in CB4856 and RETR-1(III) is further supported by the

24

550   phylogenetic reconstructions. Based on the reverse-trascriptase polymerase

551   domain cladograms of the retro-transcribing viral agents, IV-8711-IV-8719/RETR-

552   1 like elements appear as paraphyletic group with Caulimovirus clade.

553   Cladograms establish the sister relationship between retrovirus clade and

554   caulimovirus/IV-8711-IV-8719/RETR-1. The above grouping is supported by

555   placement of the RT-polymerase domain predicted independently in genomes of

556   seven different Caulimoviruses and seven different retroviruses (representing

557   major branches in animal retrovirus taxonomy) respectively, in addition to five

558   members of the IV-8711-IV-8719/RETR-1 group. It appears even more striking

559   as the same phylogenetic kinship is supported by the RNaseH domain

560   cladograms. [In contrast to cladograms build based on RT-polymerase domain

561   (when single defined domain PS50878/ InterPro:IPR000477 is predicted on all

562   genomes included into tree) RNaseH domain is predicted either as RNaseH

563   (PS50879   RNASE_H domain, found typically in the vertebrate retroviruses) or

564   RNaseHI (cd09274 RNase_HI_RT_Ty3 found in the invertebrate retroviruses

565   including the group of IV-8711-IV-8719/RETR-1), while both types of the RNaseH

566   domain are identified in Caulimoviruses. Provided the IV-8711-IV-8719/RETR-1

567   group contains elements infecting nematodes and insects (and other retroviruses

568   that are associated with invertebrates other than insects and nematodes,

569   unpublished observation) the grouping with Caulimoviruses is surprising given

570   the later are found exclusively infecting plants. Curiously, the association

571   between described provirus inserted into chromosome IV and Caulimoviruses,

572   extends beyond the homology found in the polymerase domain. Insertion into

573 chromosome IV-8711 in N2 and IV-8719 is predicted by the LTR finder to utilize

574 the methionyl-tRNA as a primer for reverse transcription reaction. This coincides

575 with the tRNA preference described for Caulimoviruses (Hull et al. 1987), despite

576 the overall differences in the genome replication via reverse transcription

577 (however this Methionyl-tRNA bias remains speculative, as is based solely on the

578 computational prediction and thereof requires the experimental verification).

579 Regarding the above findings in the perspective concerning the scope of this

580 description, it needs to be emphasized, that while the above phylogenetic

581 reconstructions robustly confirm the identified region of the *C.elegans* genome

582 represents the inserted endogenous retrovirus, the phylogenetic tools can not be

583 applied efficiently towards the identified 3'-ORF embedding the F58H7.5 gene.

584 This is because (as noted above) the F58H7.5 is an orphan gene, taxonomically

585 restricted to *C. elegans* lineage, precluding the cross-taxa comparisons. The

586 variant of the F58H7.5 identified in CB4856 (result section Fig. 3.F) is than the

587 first known and constitutes the only prediction available for the phylogenetic

588 comparisons.

589

590 Considering the presented results in the genome-scale perspective, the

591 work focuses on the single example of the past infection of the *C. elegans* by the

592 retrovirus, represented as an insertion on the proximal arm of the chromosome

593 IV present in two modern geographical isolates. As mentioned in the initial

594 sentence of the results section, the detailed results of the genome scale analysis

595 are out of the scope of this description and will be described elsewhere. *C.*

596 *elegans* genome was analyzed with the LTR finder program and the combined

597 results are outlined in (Fig.5) and summarized in (Table 4.). In total LTR finder

598 predicted 567 and 588 candidate loci in the N2(Bristol) and CB4856(Hawaii)

599 respectively. Of above only fourteen predictions are listed in (Table 1.) where

600 thirteen was verified as retroviral insertions. Would there be many more

601 endogenous retrovirus insertions demanding the description ? While genome-

602 scale searches in *C. elegans* for the endogenous retroviruses, where attempted

603 by others (reviewed in (Bessereau 2006)) for various reasons, the entire picture

604 remains blurred. There is an instructive example included into (Table 1.) of

605 another provirus inherited from the ancestral *C. elegans* : the X inserted element

606 X-12435. This particular element is predicted on both N2 and CB4856 genomes

607 by the LTR finder by the reciprocal LTR match. However, in given case LTR

608 finder internally enabled ScanProsite detects the retroviral domain only on the

609 proviral locus in N2(Bristol) chromosome X but ignores the domain encoded by

610 the orthologous locus on CB4856(Hawaii) X. While the reasons for this behavior

611 remains largely enigmatic (i.e. it is formally possible that the domain encoding

612 sequences deteriorated enough to prevent the detection by the internally enabled

613 ScanProsite) this example suggests, there might be other proviral insertions

614 where support by the internally enabled ScanProsite is insufficient to detect the

615 protein domains supporting the search. To compensate for that insufficiency, I

616 propose, the frame-wise protein domain search would be enabled externally on

617 the conceptually translated, coded and stop-masked amino-acid strings. The

618 presented evidence implies, the implemented stop-masking offers the

619 improvement detecting the endogenous retroviruses insertions with InterProScan

620 search coupled with the LTR finder output. I further suggest, the stop-masking

621 approach implemented here, could be easily generalized into genome scale

622 analysis, with the following assumption: i. sequence of the entire chromosomes is

623 frame-wise translated into amino-acid coded strings, ii. coded amino-acid strings

624 are than stop-masked and searched for the co-occurring adjacent retroviral

625 domains indicative for the insertion of the provirus iii. the possible LTR pairs are

626 verified by the LTR finder.

627

628

629

630 Literature cited:

631

632 Arnaud, F., Caporale, M., Varela, M., Biek, R., Chessa, B., Alberti, A., …

633 Palmarini, M. (2007). A Paradigm for Virus–Host Coevolution: Sequential

634 Counter-Adaptations between Endogenous and Exogenous Retroviruses. PLoS

635 Pathogens, 3(11), e170. doi.org/10.1371/journal.ppat.0030170

636

637 Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses.

638 Nature. 1970 Jun 27;226(5252):1209-11.

639

640 Bittner JJ. SOME POSSIBLE EFFECTS OF NURSING ON THE MAMMARY

641 GLAND TUMOR INCIDENCE IN MICE. Science. 1936 Aug 14;84(2172):162.

642

643 Britten, R. J. (1995). Active gypsy/Ty3 retrotransposons or retroviruses in

644 Caenorhabditis elegans. Proceedings of the National Academy of Sciences of

645 the United States of America, 92(2), 599–601.

646

647 Bessereau JL. Transposons in C. elegans. 2006 Jan 18. In: WormBook: The

648 Online Review of C. elegans Biology [Internet]. Pasadena (CA): WormBook;

649 2005-.

650

651 Brown PO. Integration. In: Coffin JM, Hughes SH, Varmus HE, editors.

652 Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press;

653 1997.

654

655 Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor

656 (NY): Cold Spring Harbor Laboratory Press; 1997.

657

658 Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for

659 phylogenetic trees. Proceedings of the National Academy of Sciences of the

660 United States of America, 93(23), 13429.

661

662 Friedland, A. E., Tzur, Y. B., Esvelt, K. M., Colaiácovo, M. P., Church, G. M., &

663 Calarco, J. A. (2013). Heritable genome editing in C. elegans via a CRISPR-

664 Cas9 system. Nature Methods, 10(8), 741–743. doi.org/10.1038/nmeth.2532

665

666 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.

667 New algorithms and methods to estimate maximum-likelihood phylogenies:

668 assessing the performance of PhyML 3.0. Syst Biol. 2010 May;59(3):307-21. doi:

669 10.1093/sysbio/syq010.

670

671 Hodgkin J and Doniach T. Natural Variation and Copulatory Plug Formation in

672 Caenorhabditis elegans. Genetics 1995 May; 146: 149-164

673

674

675 Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis,

676 and Visualization of Phylogenomic Data. Molecular Biology and Evolution, 33(6),

677 1635–1638. http://doi.org/10.1093/molbev/msw046

678

679 Hull R, Covey SN, Maule AJ. Structure and replication of caulimovirus genomes.

680 J Cell Sci Suppl. 1987;7:213-29.

681

682 Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A

683 programmable dual-RNA-guided DNA endonuclease in adaptive bacterial

684 immunity. Science. 2012 Aug 17;337(6096):816-21. doi:

685 10.1126/science.1225829.

686

687 Jo H, Choi H, Choi MK, Song N, Kim JH, Oh JW, Seo K, Seo HG, Chun T, Kim

688 TH, Park C. Identification and classification of endogenous retroviruses in the

689 canine genome using degenerative PCR and in-silico data analysis.

690 Virology. 2012 Jan 20;422(2):195-204. doi: 10.1016/j.virol.2011.10.010.

691

692 Johnson, W. E., & Coffin, J. M. (1999). Constructing primate phylogenies from

693 ancient retrovirus sequences. Proceedings of the National Academy of Sciences

694 of the United States of America, 96(18), 10254–10260.

695

696 Maydan, J. S., Flibotte, S., Edgley, M. L., Lau, J., Selzer, R. R., Richmond, T. A.,

697 Pofahl, N. J., Thomas, J. H., & Moerman, D. G. (2007). Efficient high-resolution

698 deletion discovery in Caenorhabditis elegans by array comparative genomic

699 hybridization. Genome Res, 17, 337-47. doi:10.1101/gr.5690307

700

701 Jason S Maydan, Adam Lorch, Mark L Edgley, Stephane Flibotte, Donald G

702 Moerman. Copy number variation in the genomes of twelve natural isolates of

703 Caenorhabditis elegans. BMC Genomics. 2010; 11: 62. Published online 2010

704 Jan 25. doi: 10.1186/1471-2164-11-62

705

706 Palopoli MF, Rockman MV, TinMaung A, Ramsay C, Curwen S, Aduna A, Laurita

707 J, Kruglyak L. Molecular basis of the copulatory plug polymorphism in

708 Caenorhabditis elegans. Nature. 2008 Aug 21;454(7207):1019-22. doi:

709 10.1038/nature07171. Epub 2008 Jul 16.

710

711   Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing Large

712   Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular

713   Biology and Evolution, 26(7), 1641–1650. http://doi.org/10.1093/molbev/msp077

714

715   Rous, P. (1911). A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT

716   SEPARABLE FROM THE TUMOR CELLS. The Journal of Experimental

717   Medicine, 13(4), 397–411.

718

719   Spector, D. H., Varmus, H. E., & Bishop, J. M. (1978). Nucleotide sequences

720   related to the transforming gene of avian sarcoma virus are present in DNA of

721   uninfected vertebrates. Proceedings of the National Academy of Sciences of the

722   United States of America, 75(9), 4102–4106.

723

724   Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and

725   post-analysis of large phylogenies. Bioinformatics, 30(9), 1312–1313.

726   http://doi.org/10.1093/bioinformatics/btu033

727

728   Stoye, J. P., & Coffin, J. M. (1987). The four classes of endogenous murine

729   leukemia virus: structural relationships and potential for recombination. Journal of

730   Virology, 61(9), 2659–2669.

731

Sulson,J.E. and Waterston,R. et al. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998 Dec 11;282(5396):2012-8.

Telesnitsky A, Goff SP. Reverse Transcriptase and the Generation of Retroviral DNA. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. ('Fig.2. Process of reverse transcription of the retroviral genome'. The accepted model of the retroviral reverse transcription and replication).

Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature. 1970 Jun 27;226(5252):1211-3.

Thomas, C. G., Wang, W., Jovelin, R., Ghosh, R., Lomasko, T., Trinh, Q., … Cutter, A. D. (2015). Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. Genome Research, 25(5), 667–678. doi.org/10.1101/gr.187237.114

Thompson OA, Snoek LB, Nijveen H, Sterken MG, Volkers RJ, Brenchley R, Van't Hof A, Bevers RP, Cossins AR, Yanai I, Hajnal A, Schmid T, Perkins JD, Spencer D, Kruglyak L, Andersen EC, Moerman DG, Hillier LW, Kammenga JE, Waterston RH. Remarkably Divergent Regions Punctuate the Genome Assembly

754 of the Caenorhabditis elegans Hawaiian Strain CB4856. Genetics. 2015

755 Jul;200(3):975-89. doi: 10.1534

756

757 Yang L, Güell M, Niu D, George H, Lesha E, Grishin D, Aach J, Shrock E, Xu W,

758 Poci J, Cortazio R, Wilkinson RA, Fishman JA, Church G. Genome-wide

759 inactivation of porcine endogenous retroviruses (PERVs). Science. 2015 Nov

760 27;350(6264):1101-4. doi: 10.1126

761

762 Yohn, C. T., Jiang, Z., McGrath, S. D., Hayden, K. E., Khaitovich, P., Johnson,

763 M. E., … Eichler, E. E. (2005). Lineage-Specific Expansions of Retroviral

764 Insertions within the Genomes of African Great Apes but Not Humans and

765 Orangutans. PLoS Biology, 3(4), e110. doi.org/10.1371/journal.pbio.0030110

766

767 Zdobnov EM, Apweiler R. Bioinformatics, 2001 Sep;17(9):847-8. InterProScan--

768 an integration platform for the signature-recognition methods in InterPro.

769

770 Zhao Xu, Hao Wang, "LTR_FINDER: an efficient tool for the prediction of full-

771 length LTR retrotransposons", Nucleic Acids Res. 2007 July; 35(Web Server

772 issue): W265-W268.

773

774 Other references for the sequence search tools:

775

776

ScanProsite (stand-alone tool to scan PROSITE)	De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.* Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W362-5.
MyDomains	Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJA. *The 20 years of PROSITE.* Nucleic Acids Res. 2008 Jan;36(Database issue):D245-9.
BLAT	Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.
BLAST	Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
WebLogo3	Crooks GE, Hon G, Chandonia JM, Brenner SE

WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004) [Full Text]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. 18:6097-6100

MAFFT   Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability.
Mol Biol Evol. 2013 Apr;30(4):772-80. doi: 10.1093/molbev/mst010. Epub 2013 Jan 16.

777

778

779

780

781

782

783

784

785

786

787

788 Tables and Figures:

789

790

791 Figure 1. LTR finder graphics representing predicted provirus on the proximal arm of the

792 chromosome IV in N2 (A.), and CB4658 (B.) Blue oval represents conserved PBS and yellow oval

793 represents conserved PPT. Note RT and Integrase domains are predicted in N2 (A.) but only

794 integrase domain is detected in CB4856 (B.). Ensembl representation (C.) of BLAT alignment of

795 provirus sequences N2 (upper panel), and CB4658 (lower panel) with imposed Prosite Domain

796 architectures (415 bp LTRs are drawn as blue rectangles) scaled onto Ensembl BLAT images

797 (above the track with %GC content). The red horizontal bars represent BLAT alignments,

798 expectedly duplicated at the 5' and 3' LTR regions (upper panel). Additional BLAT match is

799 apparent with IV-8711 provirus of an N2 aligned with chromosome IV assembly of CB4856 (lower

800 panel). The extra bar represents non-continuous alignments in the RNaseH – integrase

801 connecting region (indicated by black arrowheads). Note the 3' region of the non-continuous

802 alignment overlaps with N-terminal region of the predicted integrase (see Fig. 2 for details).

803 F58H7.5 is apparent in both panels in the sector describing protein coding genes annotated in the

804 reference sequence assembly. Provided the ideograms are drawn up to scale F58H7.5 projects

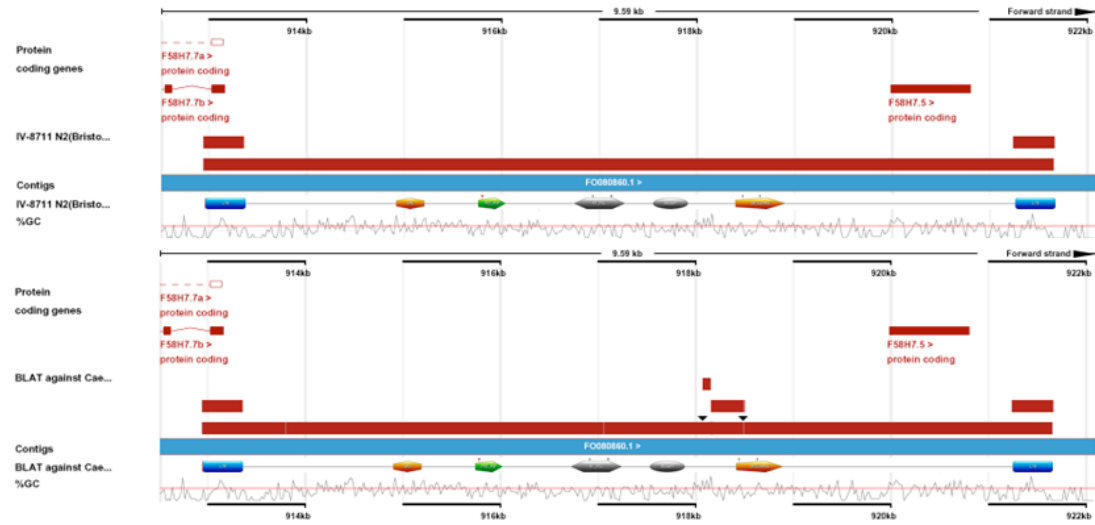805 on the 3' sub-LTR region of the predicted provirus.

806

807

A.

B.
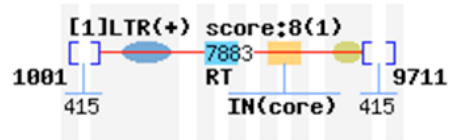
C.

808

809

810

811

812

813

814

815

816

817

818

819

38

820 Figure 2. Insertions and / or deletions in the region connecting the RNaseH and catalytic

821 integrase domain encoding region of IV-8711 in N2(Bristol) and IV-8719 in CB4856(Hawaii). The

822 frame altering (3.A.) lesion overlaps with region encoding for the predicted N-terminal catalytic

823 integrase domain (outlined in Fig.1.D.) . Residues corresponding to N-terminal catalytic integrase

824 domain alignment start are underlined (3.B). Note the non-synonymous substitution (T->G) in N2

825 sequence introduces the amber stop codon.

826

827

828

829   `1.`

830 `N2 5101   GATCCGATCCTGAAATGCATCAAGGACTTC----CGACGCCAGCAACGCCGATCGACATC  5156`

831 `Hii5101   ...............................ATGT............................  5160`

832

833   `2.`

834 `N2 5157   GTTCCTTCGACATGGGCAGGTGTGCTGGAGCACATCAAGCTTACTGAGTCTGGGTTGCTC  5216`

835 `Hii5161   ......................................................---..  5217`

836

837   `3.A.`

838 `N2 5517   CTGAGAATGACAGCTTCAGGTAACAAGTAGT-------ACATGGTTTGTTGGTTCACAAA  5569`

839 `Hii5518   ...............................T.ATCATGA......................  5577`

840

841 `3.B. Codon alignment over the N-terminal integrase: N2 (top) CB4856 (bottom)`

842 `N2 +3fr: ...........·A··S··G··N··K··*·········Y·`**`·M··V··C··W··F··T··K`**

843 `N2 5517   CTGAGAATGACAGCTTCAGGTAACAAGTAGT-------ACATGGTTTGTTGGTTCACAAA  5569`

844 `Hii5518   ...............................T.ATCATGA......................  5577`

845 `Hii+2fr: ...........·L··Q··V··T··S··I··I··M··N·`<u>`·M··V··C··W··F··T··K`</u>

846

847

848

849

850
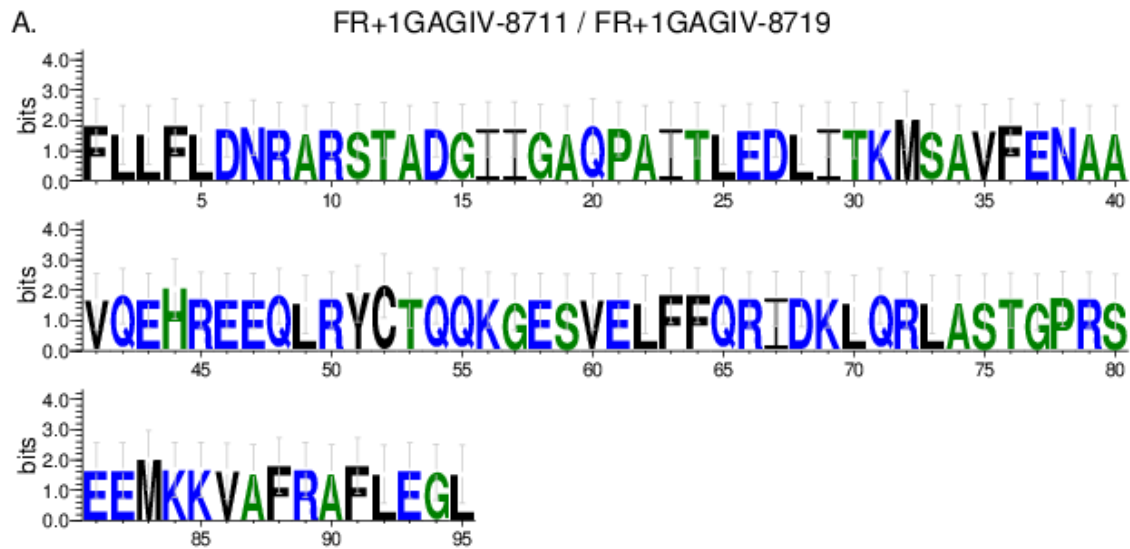
851

852

853

854

Figure. 3. Frame-wise LOGO alignments of the predicted protein domains on proviral insertions

on the proximal arm of the chromosome IV present in two *C. elegans* isolates: IV-8711 N2(Bristol)

and IV-8719(Hawaii). Group specific antigen gene (gag) (A.); Aspartyl protease (pro) (B.);

Reverse-transcriptase polymerase (pol) (C.); RnaseH (D.); Catalytic integrase (int)(E.); 3'-ORF

{F58H7.5} (F.). Stop codons are indicated by (*). Black arrows indicate the start and end of the

F58H7.5 coding region embedded into the 3'-ORF sequence.

861

862

863

864

865

866

867

B. FR+1PROIV-8711 / FR+1PROIV-8719

WebLogo 3.5.0

868

869

C. FR+2RT_POLIV-8711 / FR+2RT_POLIV-8719

WebLogo 3.5.0

870

41

871

872

873

874

875

876

D. FR+2RNaseH-IV-8711 / FR+2RNaseH-IV-8719

877

E.

FR+3INT-IV-8711 / FR+2INT-IV-8719

878

879

WebLogo 3.5.0

F.
FR+2-F58H7.5-IV-8711 / FR+1-F58H7.5-IV-8719

880

WebLogo 3.5.0

881   Figure 4. Retroviral protein domain cladograms. Reverse-transcriptase polymerase A. RNaseH B.
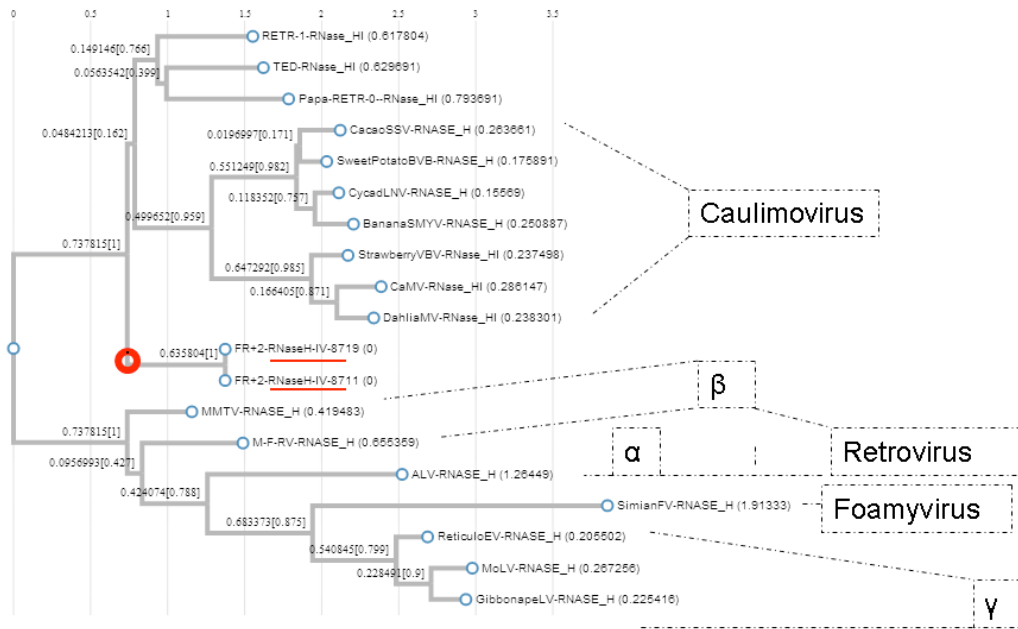
882   (description in the text)

883   A.
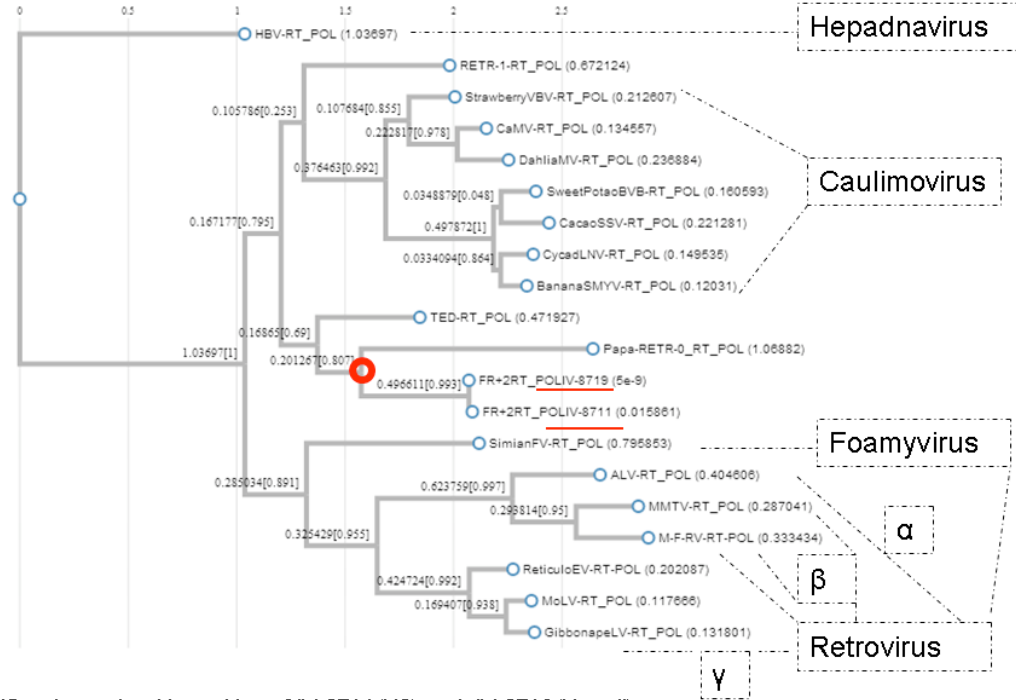
895   B.

○ Specifies the node with position of IV-8711(N2) and IV-8719(Hawaii)
   in the RNaseH domain tree of retro-transcribing agents.

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910   Fig.5. Genome-scale LTR finder predictions of the candidate LTR pairs. A. Chromosomal

911   representation of the LTR finder output pn N2(Bristol) genome, B. Chromosomal representation

912   of the LTR finder output on CB4856(Hawaii). The description is included into Tables 1. and 4.

Chr I
NC_003279.8
15,072434 MB

Chr II
NC_003280.10
15,279421 MB

Chr III
NC_003281.10
13,783801 MB

Chr IV
NC_003282.8
17,493829 MB

Chr V
NC_003283.11
20,924180MB

Chr X
NC_003284.9
17,718942MB

913
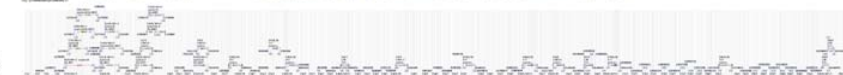
Chr I
CM003206
14,890789 MB

Chr II
CM003207
14,885952 MB

Chr III
CM003208
13,596826 MB

Chr IV
CM003209
17,183857 MB

Chr V
CM003210
20,182852 MB

Chr X
CM003211
17,537347 MB

914

915

916 Tables:

917 Table 1. Results of the first-pass genome-scale LTR finder predictions of the endogenous

918 retroviral elements inserted in chromosomal assemblies of two C. elegans isolates: reference

919 N2(Bristol) and CB4856(Hawaii).

Chr	Retroviral insertion LTR name	N2 insert length	Hawaii insert length	N2 chromosome position	N2 LTR 5'/3'	Hawaii chromosome position	Hawaii LTR 5'/3'
I	I-CeN2ii-1879 (*)	1879	1879	13147285 -13149163	245/245	12993371 – 12995249	245/245
	I-CeN2ii-5088	5088	5088	3847365 - 3852452	217/221	3776964 - 3782051	217/221
II	II-CeN2-16388	16388	-	13243641 - 13260028	161/161	-	-
III	III-CeN2-8886 (RETR-1)	8886	-	8852597 - 8861482	518/513	-	-
	III-CeN2ii-13588	13588	13506	13229612 - 13243199	517/517	13044879 - 13058384	517/517
IV	IV-CeN2ii-8711	8711	8719	912948 - 921658	415/415	899767 - 908485	415/415
V	V-CeN2ii-10045	10045	10048	4300964 - 4311008	318/318	4041012 - 4051059	316/320
	V-CeN2-16124	16124	-	8473016 - 8489139	180/193	-	-
	V-CeN2ii-19860	19860	19794	8825452 - 8845311	514/514	8520474 - 8540267	514/514
	V-CeN2-11685	11685	-	17146405 - 17158089	444/457	-	-
	V-CeN2ii-19417	19417	19417	18435912 - 18455328	251/244	18435912 - 18455328	251/244
	V-CeN2ii-12067	12067	12540	19270045 - 19282111	537/537	18647618 - 18660157	537/537
X	X-CeN2ii-12453 (**)	12453	12453	9191187 - 9203639	648/648	9033873 - 9046325	648/648

| | X-CeN2ii-5078 | 5078 | 5078 | 17644528 - 17649605 | 162/ 162 | 17462927 - 17468004 | 162/ 162 |

920   (*) - denotes unusually short element. (**) – denotes insertion identified on homologous CB4856

921   chromosome by the reciprocal LTR finder match only.

922

923

924

925

926

927

928   Table 2. Open Reading Frames used by predicted protein domains, presented in 5'->3' order in

929   IV-8711 in N2(Bristol) and IV-8719 in CB4856(Hawaii) proviruses. Discordant frames in the

930   catalytic integrase and 3'-ORF are highlighted by doubly lined frame.* Integrase reading frames

931   were trimmed at N-terminus to limit the mis-alignment due to the frame-shifting mutations (Fig.2.).

932

ORF	gag	Pro	Pol	RNaseH	Integrase (162nt)	3'-ORF	5'->3'
Reading frame used	+1	+1	+2	+2	+3	+2	IV-8711 (N2)
	+1	+1	+2	+2	+2	+1	IV-8719 (Hawaii)
Identity(%) similarity(%)	95/95 (100%) 95/95 (100%)	95/95 (100%) 95/95 (100%)	175/179 (98%) 177/179 (99%)	121/121 (100%), 121/121 (100%)	131/131 (100%)* 131/131 (100%)*	557/559 (99%) 558/559 (99%)	

933

934

935

936

937 Table 3. InterProScan identifiable retroviral protein domain signature matches of the retrovirus

938 inserted into chromosome IV-8711 of N2 (reference assembly) and assembled IV-8719 of

939 CB4856 (Thompson et al. 2015) compared to RETR-1 (III).

940

Protein Domain	RT_POL	RNaseH	Integrase	*C. elegans* isolate
RETR-1 (FR+1)	66/175(38%) 115/175(66%)	51/120(43%) 74/120(62%)	51/156(33%) 82/156(53%)	IV-8711 (N2)
Identity(%) similarity(%)	67/175(38%) 115/175(66%)	51/120(43%) 74/120(62%)	47/141(33%) 76/141(54%)	IV-8719 (Hawaii)

941

942

943

944

945

946

947

948

949

950

951

952

953

954

51

955 Table 4. LTR finder calculations on the number of the genome scale predictions per

956 chromosome. Values do vary, ranging from 2,31 - 1,77 (for chromosome X) to 8,05 and 8,52 (for

957 chromosome III and chromosome V) LTR finder predictions per MB in N2(Bristol) and

958 CB4856(Hawaii) respectively. While the lowest number of LTR finder predictions per MB is

959 observed consistently for sex chromosomes in both isolates, when autosomal predictions are

960 included into calculations the association between numbers of predictions calculated on two

961 genomes are considered extremely statistically significant (*** two tailed P<0.0001, Chi-square

962 test). The reason for *C. elegans* X chromosomes scoring the lowest values remains to be

963 determined, but likely reflects the proportionally lower number of the haplotypic duplications

964 maintained on sex chromosomes when compared to autosomes.

965

966

Chr	N2 predictions	Hawaii predictions	N2 chromosome Length (MB)	Hawaii Length (MB)	#Per 1MB N2	#Per 1MB Hawaii
I	65	54	15,072434	14,890789	4,31	3,63
II	85	123	15,279421	14,885952	5,56	8,26
III	111	113	13,783801	13,596826	8,05	8,31
IV	97	95	17,493829	17,183857	5,54	5,53
V	168	172	20,924180	20,182852	8,03	8,52
X	41	31	17,718942	17,537347	2,31***	1,77***
Total	567	588	100,272607	98,277623	5,65***	5,98***

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985