Figure 1: Diagram showing the structure of the Ec2_chr5_27325585_27334012 locus. Grey shaded areas indicate locations of the genes, location of gene boundaries in kb is indicated. Cross hatched areas represent regions identified using Exonerate. Vertical lines represent stop codons when the coding region is translated in the reading frame of the longest ORF for that gene.

Figure 2: PhyML phylogenetic tree showing the relationship between the *gag* genes of endogenous gammaretroviruses from the horse genome and known gammaretroviruses. Horse sequences are marked in bold. For clarity well supported clades of horse sequences are represented by consensus sequences or individual representative sequences. Branch support of greater than 50% is marked on the appropriate branches. Groups containing a sequence with an RPMK >1 are underlined. The details of the horse sequences are listed in Supplementary Table 1, Genbank accessions and sources of known retroviruses in Supplementary Table 2.

Figure 3: PhyML phylogenetic tree showing the relationship between the *pol* genes of endogenous gammaretroviruses from the horse genome and known gammaretroviruses. Horse sequences are marked in bold. For clarity well supported clades of horse sequences are represented by consensus sequences or individual representative sequences. Branch support of greater than 50% is marked on the appropriate branches. Groups containing a sequence with an RPMK >1 are underlined. The details of the horse sequences are listed in Supplementary Table 1, Genbank accessions and sources of known retroviruses in Supplementary Table 2.

Figure 4: PhyML phylogenetic tree showing the relationship between the *env* genes of endogenous gammaretroviruses from the horse genome and known gammaretroviruses. Horse sequences are marked in bold. For clarity well supported clades of horse sequences are represented by consensus sequences or individual representative sequences. Branch support of greater than 50% is marked on the appropriate branches. Groups containing a sequence with an RPMK >1 are underlined. The details of the horse sequences are listed in Supplementary Table 1, Genbank accessions and sources of known retroviruses in Supplementary Table 2.

Figure 5: PhyML phylogenetic tree showing the relationship between the *pol* genes of endogenous betaretroviruses from the horse genome and known betaretroviruses. Horse sequences are marked in bold. For clarity well supported clades of horse sequences are represented by consensus sequences or individual representative sequences. Branch support of greater than 50% is marked on the appropriate branches. Groups containing a sequence with an RPMK >1 are underlined. The details of the horse sequences are listed in Supplementary Table 1, Genbank accessions and sources of known retroviruses in Supplementary Table 2.

Figure 6: PhyML phylogenetic tree showing the relationship between the *pol* genes of endogenous epsilonretroviruses from the horse genome and known epsilonretroviruses. Horse sequences are marked in bold. For clarity well supported clades of horse sequences are represented by consensus sequences or individual representative sequences. Branch support of greater than 50% is marked on the appropriate branches. Groups containing a sequence with an RPMK >1 are underlined. The details of the horse sequences are listed in Supplementary Table 1, Genbank accessions and sources of known retroviruses in Supplementary Table 2.