Supplementary Information

to

Complex Genomic Rearrangements Lead to Novel Primate Gene Function.

Ciccarelli FD, von Mering C, Suyama M, Harrington E, Izaurralde E, and Bork P

Contents:

Figure S1. DNA composition of the regions flanking the junction sites of the duplicated segments.

The junction sites for each of the 8 duplicated segments containing the RGP genes were established as reported in Methods. The borders of the regions containing RGP5, RGP6 and the 5' border of RGP7 could not be assessed for the presence of gaps in the assembly and were not considered in the analysis. A sequence interval of 500 bp was taken at each side of the junction points, obtaining 11 regions of 1kb each. Each region was divided in windows of 10 bp and the content in DNA repeats was assessed for the 100 resulting windows using RepeatMasker (Smit, AFA & Green, P RepeatMasker at http://repeatmasker.org). As a control, 20 intergenic regions of 1 kbp each were randomly picked in the Chromosome 2 sequence and their content in DNA repeats measured using RepeatMasker.

The % of DNA repeats for each window is reported. The window number 50 corresponds to the junction site ± 5 bp (red). The bar below the X axis reports the bp position of the entire 1kbp regions centred at the junction site. The entire region adjacent to the junction results enriched in DNA repeats when compared to the control, with the peak occurring at the junction site.

A. The content for the main DNA repeat families is plotted in different colours. The average content in the control (39.0%) is reported as a continue blue line. The average content of DNA repeats in the 1 kbp region spanning each of the 11 junction sites (55.0%) is reported as a dotted blue line.

B. The content in *Alu* elements in the region spanning the junction site is reported. The average content of *Alu* in the control (10.3%) is showed as a continue green line, while the average of *Alu* in the junction regions (25.6%) is reported as a dotted green line.

Table S1. Content in DNA repeats of the region flanking the junction sites compared to a random control.

Repeat class	Junction Site (%)	Control (%)
SINE	**27.4**	**14.3**
Alu	25.6	10.3
MIR	1.8	4.0
LINE	**12.6**	**11.96**
LTR	**11.0**	**10.46**
Others	**4.0**	**2.28**
Total	**55.0**	**39.0**

The LINE class contains L1 and L2 repeats, LTR contains ERVL, ERVK, ERV1 and MaLR elements. Other repeats include low complexity regions, satellites, simple repeats, MER1 and MER2.

Figure S2. Phylogenetic trees used for detecting positive selection.

The branch-site model of the ML method was used{Yang, 2002 #62}. This model allows the Ka/Ks ratio to vary both among the amino acidic sites and among lineages, resulting particularly suitable for the study of paralogs divergence after gene duplication{Yang, 2002 #62}.

As the RGP genes have an incomplete exon 20 when compared to RanBP2 and acquired the last 3 exons from GCC2, we divided the RGP gene sequence into three regions (A). The first region goes from exon 1 to exon 19 (B); the second corresponds to exon 20 (C) and the third comprises exons *p* and *q* (D). We then derived the orthologs of RanBP2 and GCC2 across the species, keeping only those for which a complete sequence was retrievable. For example, although RanBP2 is present in chimp, the sequence of the first 19 exons is not complete, therefore we kept it out from the tree of exons 1-19 (B) but considered it for exon 20 (C). The sequence of RGP7 could not be used in trees (C) and (D), as it falls in an unresolved region of the genome assembly.

The branch-site model requires that the phylogeny and the branches putatively under positive selection are known *a priori*. For each multiple alignment, we then reconstructed the phylogeny as described in Methods and assumed as foreground branch the RGP branch (depicted in red in each tree).

Abbreviations: gg, *G.gallus*; hs, *H.sapiens*; mm, *M.musculus*; pt, *P.troglodytes*; rn, *R.norvegicus*.

Figure S3. Residues under positive selection in the RanBD domains of the RGP proteins.

A. The residues predicted to be under positive selection are mapped onto the domain architecture representation of the RGP gene. Five out of the 7 residues predicted for the exon 20 map into the two RBD domains and are depicted in cyan. It is worth noting that the three mutations D2031G, R2039G and K2338E all localize in corresponding Ran binding regions of the two RBD domains. The numbering refers to the human RanBP2 (NP_006258).

B. Multiple alignment of the two RBD domains present in the RGP genes and the corresponding RBD domains in the RanBP2 orthologs. The residues conserved in more than 90% of the sequences are depicted in pink. The residues under positive selection are depicted in cyan. The amino acid positions are reported only for the orthologs with a complete sequence. Abbreviations: dr, *D.rerio*; fr, *F.rubripes*; gg, *G.gallus*; hs, *H.sapiens*; mm, *M.musculus*; pt, *P.troglodytes*; rn, *R.norvegicus*.

C. Mutated residues mapped onto the 3D structure of the RanBP2 RDB domain. The 3D structure of the Ran-RanBD1 complex {Vetter, 1999 #46} was used (PDB: 1rrp). Ran is depicted in grey; the first RanBD domain of RanBP2 is depicted in magenta. The residues under positive selection in the RanBD of the RGP proteins are represented as cyan sticks and labelled including the corresponding mutations. As 1rrp represents the first RanBD of RanBP2, the corresponding residue numbers are: D2031 = D1190; R2039 = R1198; Q2121 = Q1280; K2338 = K1200. The last residue under positive selection (T2391S) is not represented as the modification does not affect the binding to Ran and it corresponds to S1249 in the first domain of RanBP2. Two orthogonal views of the complex are represented, in order to highlight all the residues mutated in the RGP genes. The figures were built using PyMol (http://pymol.sourceforge.net).

Figure S1

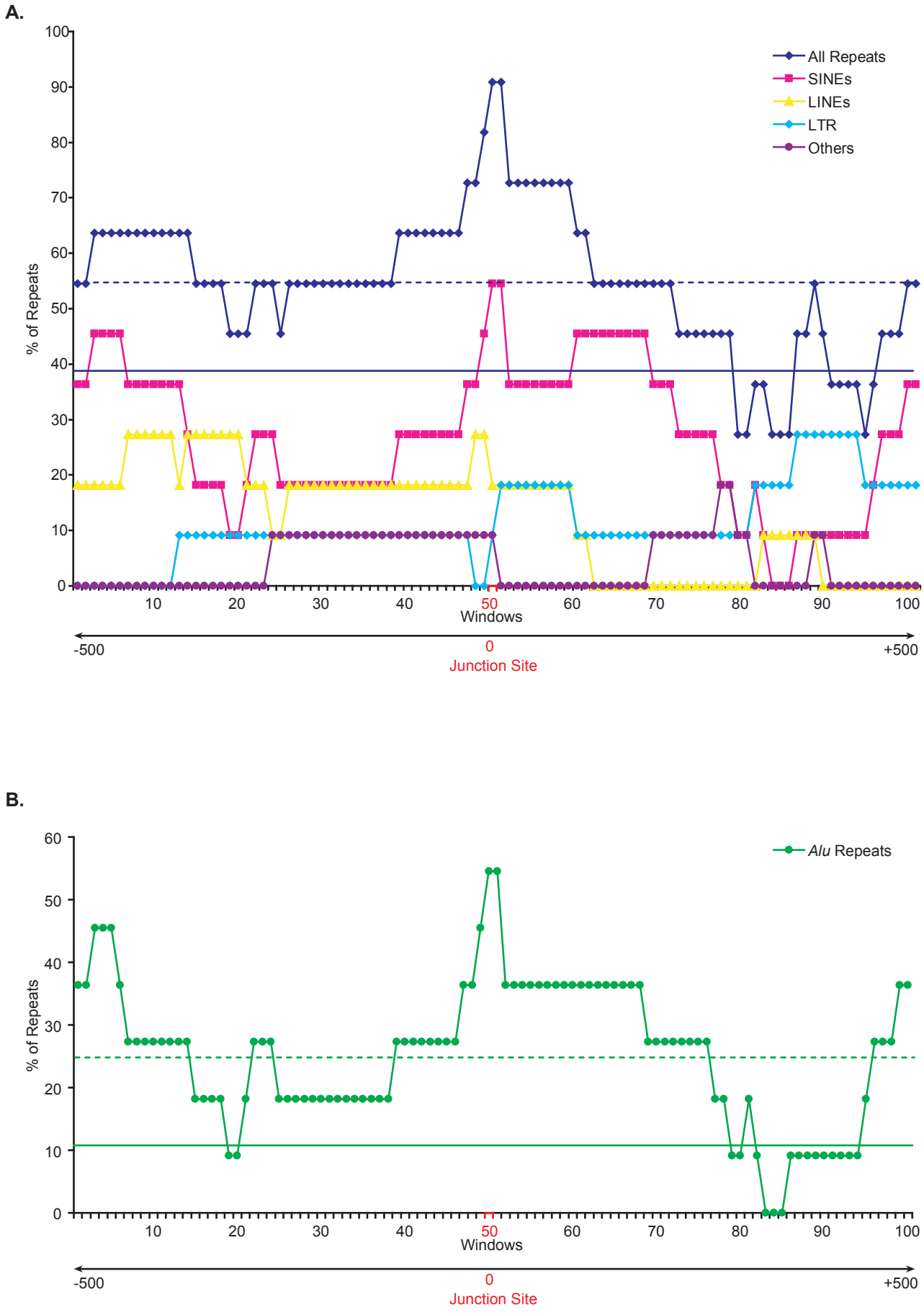

A.
All Repeats
SINEs
LINEs
LTR
Others
% of Repeats
Windows
-500
0
+500
Junction Site
B.
Alu Repeats
% of Repeats
Windows
-500
0
+500
Junction Site

Figure S2

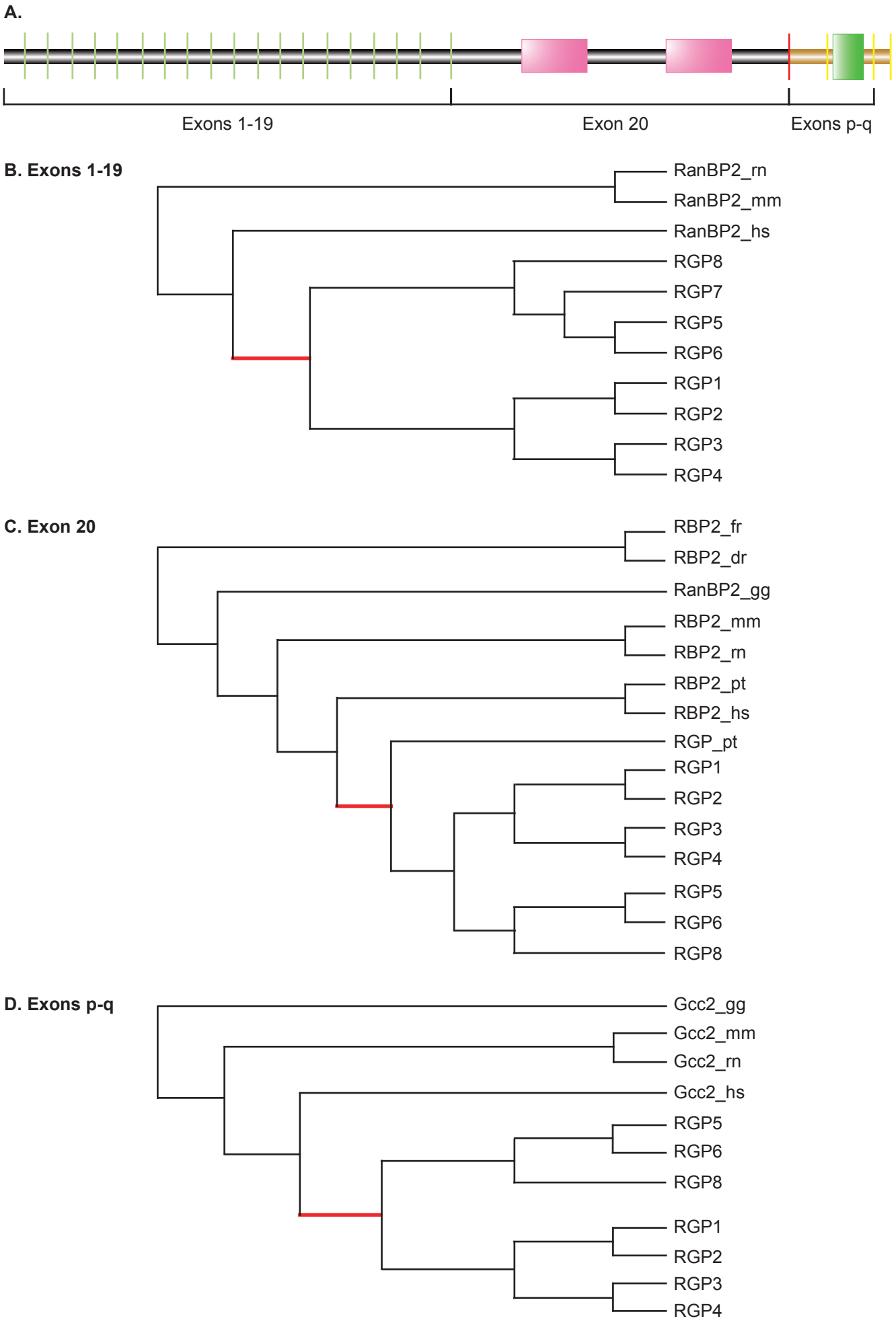
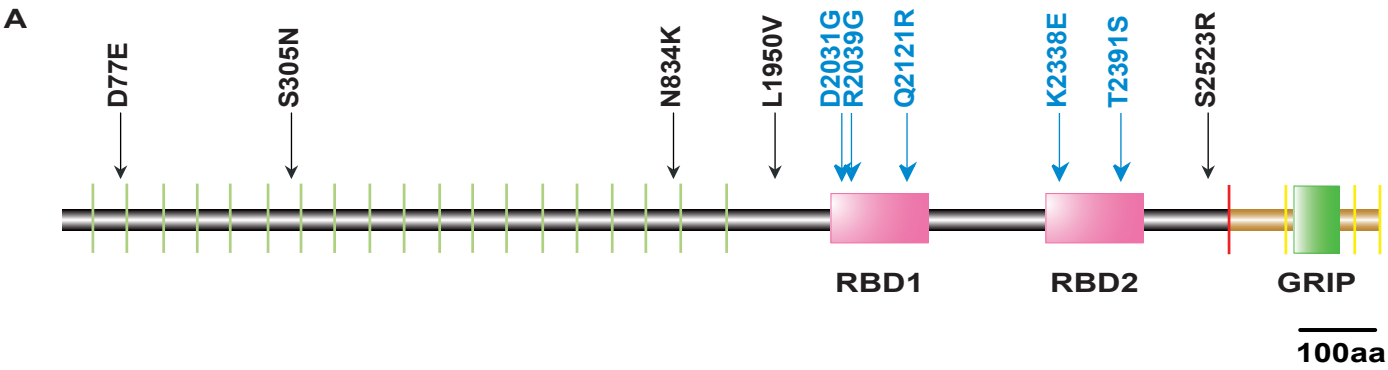

A.
Exons 1-19
Exon 20
Exons p-q
B. Exons 1-19
RanBP2_rn
RanBP2_mm
RanBP2_hs
RGP8
RGP7
RGP5
RGP6
RGP1
RGP2
RGP3
RGP4
C. Exon 20
RBP2_fr
RBP2_dr
RanBP2_gg
RBP2_mm
RBP2_rn
RBP2_pt
RBP2_hs
RGP_pt
RGP1
RGP2
RGP3
RGP4
RGP5
RGP6
RGP8
D. Exons p-q
Gcc2_gg
Gcc2_mm
Gcc2_rn
Gcc2_hs
RGP5
RGP6
RGP8
RGP1
RGP2
RGP3
RGP4

Figure S3



B

RBD1

```
RBP2_fr          FEPVVQMPDKIDLVTGEEDEEVLYSQRVKLFRFDSTVSQWKER-RGILKILKNPTNGRLRVLMRREQVLKVCANHWITTTMNLKPLAGSDRAWMWMANDFSDGDAKPEQLAAEFKSPELAEEFKLKFEECQ
RBP2_dr          FEPVVQMPDKVDLVTGEEDEKILYSQRVKLFRFDPETSQWKERGVGNLKLLKNNQNGKLRVLMRREQVLKVCANHWITTTMNLKPLAGSDRAWMWLASDFSDGDARLEQLAAKFKTPELAEEFKQRFEECQ
RBP2_gg 1727     FEPIVQMPEKVEPFTGEEDEKVLYSQRVKLFRFDPETSQWKERGVGNLKILKNEVNGKVRILMRREQVLKVCANHWITTTMNLKQLSGSDKAWMWMASDFSDGDAKLEQLAAKFKTPEQAEEFKQKFEECQ 1856
RBP2_mm 1850     FEPVVQMPEKVELVTGEEDEKVLYSQRVKLFRFDAEISQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWLASDFSDGDAKLEQLAAKFKTPELAEEFKQKFEECQ 1979
RBp2_rn 1870     FEPVVQMPEKVELVTGEEDEKVLYSQRVKLFRFDAEISQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWLASDFSDGDAKLEQLAAKFKTPELAEEFKQKFEECQ 1999
RBP2_pt          FEPVVQMPEKVELVTGEEDEKVLYSQRVKLFRFDAEVSQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWLASDFSDGDAKLEQLAAKFKTPELAEEFKQKFEECQ
RBP2_hs 2013     FEPVVQMPEKVELVTGEEDEKVLYSQRVKLFRFDAEVSQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWLASDFSDGDAKLEQLAAKFKTPELAEEFKQKFEECQ 2142
RGP_pt           FEPVVQMPEKVELVTGEECEKVLYSQRVKLFRFDAEISQWKERGLGNLKILINELNGKPRMLMRREQVLKVCANHWITTTMNLKPLSGSDRVWMWLASDFSDGDAKLEQLAAKFKTPELAEEFKQKFEECQ
RGP1_hs 1029     FEPVVQMPEKVELVTGEECEKVLYSQCVKLFRFDAEISQWKERGLGNLKILKNEVNGKPRMLMRRDQVLKVCANHWITTTMNLKPLSGSDRAWMWLASDFSDGDAKLERLAAQFKTPELAEEFKQKFEECQ 1158
RGP2_hs 1030     FEPVVQMPEKVELVTGEECEKVLYSQCVKLFRFDAEISQWKERGLGNLKILKNEVNGKPRMLMRRDQVLKVCANHWITTTMNLKPLSGSDRAWMWLASDFSDGDAKLERLAAQFKTPELAEEFKQKFEECQ 1159
RGP3_hs 1038     FEPVVQMPEKVELVTGEECEKVLYSQCVKLFRFDAEVSQWKERGLGNLKILKNEVNGKVRMLMQREQVLKVCANHWITTTMNLKPLSGSDRAWMWSASDFSDGDAKLERLAAKFKTPELAEEFKQKFEECQ 1167
RGP4_hs 1038     FEPVVQMPEKVELVIGEECEKVLYSQCVKLFRFDAEVRQWKERGLGNLKILKNEVNGKPRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWSASDFSDGDAKLERLAAKFKTPELAEEFKQKFEECQ 1167
RGP5_hs 1037     FEPVVQMPEKVELVTGEECEKVLYSQCVKLFRFDAEVRQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWSASDFSDGDAKLERLAAKFKTPELAEEFKQKFEECQ 1166
RGP6_hs 1037     FEPVVQMPEKVELVTGEECEKVLYSQCVKLFRFDAEVRQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWSASDFSDGDAKLERLAAKFKTPELAEEFKQKFEECQ 1166
RGP8_hs 1037     FEPVVQMPEKVELVTGEECEKVLYSQCVKLFRFDAEVRQWKERGLGNLKILKNEVNGKLRMLMRREQVLKVCANHWITTTMNLKPLSGSDRAWMWSASDFSDGDAKLERLAAKFKTPELAEEFKQKFEECQ 1166
```

RBD2

```
RBP2_fr          FEPVVPLPDLVEISTGEENEQVVFSHRAKLYRYDKEAAQWKERGIGDLKILQNYETKCVRLLMRRDQVLKICANHWVTSAMKLEPMKGAEKAWVWSAMDFAGVEEGKIEQLAVRFKLQETANTFKQIFEES
RBP2_dr          FEPVVPMPDLVEVSTGEEDEQVLFSHRAKLYRYDKTLSQWKERGIGDLKILQHYETKRVRLVMRRDQVLKLCANHWIDSSMKLEPMKGAEKAWIWSAFDFAEGQ-GKVEQLAVRFKLQDTASSFRDVFEES
RBP2_gg 2022     FEPVVPLPDLVEVTSGEENEQVVFSHRAKLYRYDKDTNQWKERGIGDIKILQNYDSKQARIVMRRDQVLKLCANHRITPDMNMQQMKGSDRAWVWTACDFADGER-KVELLAVRFKLQDVADSFKQTFDEA 2125
RBP2_mm 2147     FEPVVPLPDLVEVSSGEENEQVVFSHRAKLYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMTLQTMKGTERVWVWTACDFADGER-KIEHLAVRFKLQDVADSFKKIFDEA 2276
RBp2_rn 2167     FEPVVPLPDLIEVSSGEENEQVVFSHRAKLYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMTLQTMKGTERVWVWTACDFADGER-KIEHLAVRFKLQDVADSFKKIFDEA 2296
RBP2_pt          FEPVVPLPDLVEVSSGEENEQVVFSHRAKLYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMTLQNMKGTERVWLWTAYDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA
RBP2_hs 2310     FEPVVPLPDLVEVSSGEENEQVVFSHRAKLYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMTLQNMKGTERVWLWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 2439
RGP_pt           FEPVVPLPDLVEVSSGEENEQVVFSHRAELYRYDKDVGQWKERGIGDIKILQNYDNKQVHIVMRRDQVLKLCANHRITPDMSLHNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA
RGP1_hs 1326     FEPVVPLPDLVEVSSGEENEQVVFSHRAELYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1455
RGP2_hs 1327     FEPVVPLPDLVEVSSGEENEQVVFSHMAELYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1456
RGP3_hs 1335     FEPVVPLPDLVEVSSGEENEQVVFSHRAEFYRYDKDVGQWKERGIGDIKILQNYDNKHVRILMRRDQVLKLCANHRITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1464
RGP4_hs 1335     FEPVVPLPDLVEVSSGEENEKVVFSHRAELYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHTITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1464
RGP5_hs 1334     FEPVVPLPDLVEVSSGEENEQVVFSHRAEIYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1463
RGP6_hs 1334     FEPVVPLPDLVEVSSGEENEQVVFSHRAEIYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1463
RGP8_hs 1334     FEPVVPLPDLVEVSSGEENEQVVFSHRAEIYRYDKDVGQWKERGIGDIKILQNYDNKQVRIVMRRDQVLKLCANHRITPDMSLQNMKGTERVWVWTACDFADGER-KVEHLAVRFKLQDVADSFKKIFDEA 1463
```

C

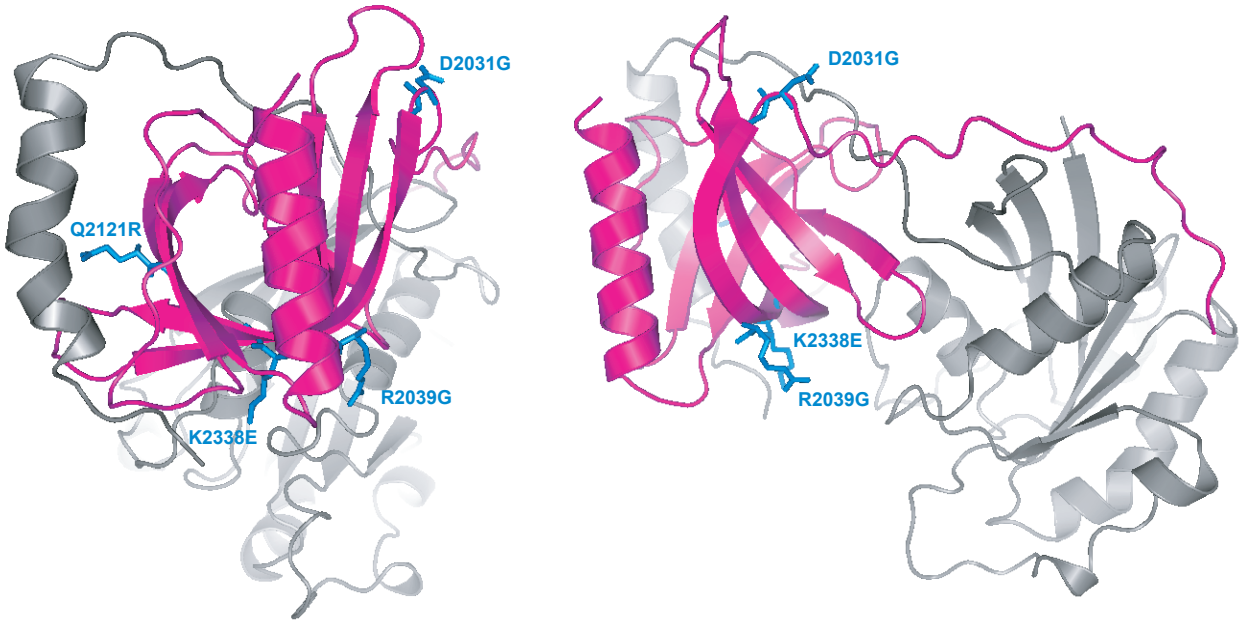


Figure S4.


CD8B_HUMAN
PRGB_HUMAN
Q969Y7
Q9PNI7
Q8N8Z4
Q96HE4
Q9GZV3
S1C2_HUMAN
GCC2_HUMAN
PINC_HUMAN
RBP2_HUMAN
Q86VL7
EDAR_HUMAN
Q9NXW5
Q8NDU1/Q8TEJ3
Q9P0V9
Q8NE15
Q8N952
BENE_HUMAN
NPH1_HUMAN
Q8NCU8
BUB1_HUMAN
Q8TCE7
BIM_HUMAN
ANC_HUMAN
MERK_HUMAN
Q96K49
Q8N9G0
Q8N5P1
Q9N426

Figure S4. Enlarged version of Figure 2.