Supplementary information

Expression pattern analysis of transcribed HERV sequences is complicated by *ex vivo* recombination

Flockerzi et al.

Recombination analysis of HERV-KX sequences with RECCO

RECCO [1] is a novel method for analyzing sequences subject to recombination. Given costs for recombination and for mutation events, RECCO finds all cost-optimal explanations of the putative recombinant sequence in terms of recombination and mutation of a given set of aligned sequences. A high recombination cost results in an explanation that only uses the single, most related reference sequence without any recombination events. As the recombination cost decreases, more and more recombination events are introduced in the explanation in favor of mutation events. The first recombination event reduces the number of mutations needed for an explanation by the largest factor. RECCO builds a list of recombination events and displays the amount of mutation cost saved by each recombination – the so called “savings” of a recombination. RECCO also computes the total mutation cost of the explanation that includes this recombination event (see Table 2). True recombinant sequences usually display a strong reduction in mutation cost (i.e. a high savings value) for the first few recombination events introduced.

To quantify the statistical significance of each recombination event, RECCO generates sets of alignments by permuting the columns of the alignment. As a result, the permuted alignments do not contain any recombination signal, but have the same diversity as the original alignment. P-values are then estimated by computing the probability of obtaining higher savings than observed in the given alignment purely by chance, based on the analysis of the set of permuted alignments. We report the p-values for the query sequence here, as our goal was to analyze the recombination signal for the query sequence only.
Treatment of gaps in RECCO analysis

Treating gaps correctly was critical for the analysis of the HERV-KX sequences, as the multiple alignment contained two long gaps and several small gaps. Recently published recombination detection methods usually implement one of the three following options: (i) discard sites that contain a gap character, (ii) treat each gap character as a fifth nucleotide state or (iii) treat each consecutive run of columns containing gaps as a large polymorphism (Geneconv [2]). The first option results in an unacceptable loss of information, in our case. For example, the 96 bp indel differentiates between evolutionarily young and old HERV-KX sequences [3]. The second option may lead to an artificially high similarity or dissimilarity between sequences in gap regions and eventually produces spurious recombination events. The third option prohibits recombinations in any run of columns containing a gap, such that a sequence containing a long gap may confound recombinations that involve other sequences. It is also difficult to choose an adequate scoring term for the resulting large polymorphisms. In conclusion, all existing approaches for treating gaps either discard a lot of information and thus miss recombination events or may infer spurious recombination events solely based on gap information.

We decided to implement an approach that discriminates between possibly spurious recombination events based on gap information and recombination events based on polymorphisms. First, it is important to realize that gaps in the query sequence have a different quality than gaps in the sequences used for an explanation. If gaps in the query sequence are matched with nucleotides in the explanation, the involved sequence in the explanation represents irrelevant information. Hence, gaps in the query sequence are assigned zero cost, such that all columns with a gap in the putative recombinant are effectively removed from the alignment. The situation is totally different if nucleotides in the query sequence are matched with a gap in the explanation. In this case, there is missing information as the query sequence is not fully explained by the other sequences. Consequently, we have chosen to penalize gaps in the explanation.

To capture the effect of the gap penalty on inferred recombination events, we executed RECCO [1] for all gap penalties in the set {0.016, 0.02, 0.04, 0.07, 0.1, 0.2, 0.4, 0.6, 0.8}. As gaps were strongly preferred for gap penalties below 0.016 and resulted in spurious recombination events, we did not consider smaller gap penalties than 0.016. The results were then preprocessed as follows: recombination events with a p-value above 0.05 were discarded as not significant. If a recombination event with the same start and end position was detected as significant for all gap penalties, it was accepted as correct. Other recombination events were
manually examined and consolidated. In the tables displaying recombination events, we only report the sequence p-value for the default gap penalty of 0.2.

References

1. Maydt J, Lengauer T: Recco: recombination analysis using cost optimization. Bioinformatics 2006, 22:1064-1071.
2. Sawyer S: Statistical tests for detecting gene conversion. Mol Biol Evol 1989, 6:526-538.
3. Mayer J, Meese E, Mueller-Lantzsch N: Human endogenous retrovirus K homologous sequences and their coding capacity in Old World primates. J Virol 1998, 72:1870-1875.

Following pages:
Original RECCO output for HERV-KX sequences "B270", "93" and "94"

Multiple alignment of HERV-KX and HML-2 reference sequences, provided in fasta format.

Reproduction of HERV-K(HML-2) gag sequence recombinants
GGCCAGATACAGTCATTTTCGATAAAATGCT-AAAAGATATGAAAGAGGGAGTAA--AACAGTATGGACCCAACTCCC

CTTAATAGGAGGACATTTATAGTTCATCTGTCATGGGACAT-----AGACTCATTCCC-TTA-----TATTGGGAGA-TTTCGGC

AAAATCTCCTC-TCTCCACTC-TCACATTTCACATAATTAAAAGCTTTAGGGG-GTTGATAGG-----GTCAC

AAGAACA

>4_008

A-AGCCACCAAGCTTCTTCCGCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGCCATTCAGGCCAGATAGTCCGACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_032

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_052

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_068

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_074

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_052

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_008

A-AGCCACCAAGCTTCTTCCGCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGCCATTCAGGCCAGATAGTCCGACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_052

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_074

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_008

A-AGCCACCAAGCTTCTTCCGCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGCCATTCAGGCCAGATAGTCCGACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_052

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA

>4_074

A-AGGCCCCACAAATCTTCTTCGCCCGGTTGTCAGTCTATTGTCATACAATCTCAAA-----------------------

---GGGCATACCCGAGGATACATCCGTCATGGGCAAG---AGCTCAAGTTTCAGTAGCCCATAGACTATAGGACATCCGTAGC

GGAGGAGTGGCAGGTCAGGTGCGTCCTGTGCCAAATAGCAA----GGCACTATATCCAGCAAGCCTGGCGAGACTAGAAGTGCACTGCTATGAGACCATTGCTACA----------GCCAGAAAACAGGGAGA--TCTTGAGGCATGGC

AAGAACA
AGTCAGTATGGATATCCAGGAATGCCCCCAGCACCACAGGGCAG----GGCGCCATACCCTCAGCCGCCCACTAGGAGAC------------------------------------------ ... --------TCAAGAAAGGAAGGAGA--TACTGAGGCGTGGCAATTCCCAGTAACATTAGAACCGATGCCAACTGGAGAAGGA-----------GCCCAAGAGGGAGAGCCTCCCCTAGCTGA

ATTCCCAGTAATATTAGAACCAATACCACCTGGAGAAGGG-----------GCCCAAGAGGGAGAGCCTCCCCTAGCTGA

---AGAAAATAAGACCCAACCGCCAGTAGCCTATCAATACTGGCCGCCGGCTGAACTTCAGTATCGGCCACCCCC-AGAA

A-GGCAC---AAGTCATCTTCCAGCAGGTCAGGTGCCCGTAACATTACAACCTCAAAAGCAGGTTAA-------------

>B276

AAGAACA

>AAGAACA

AAGAACA

AAGAACA

AAGAACA

AGTCAGTATGGATATCCAGGAATGCCCCCAGCACCACAGGGCAG----GGCGCCATACCCTCAGCCGCCCACTAGGAGAC

---AGAAAATAAGACCCAACCGCCAGTAGCCTATCAATACTGGCCGCCGGCTGAACTTCAGTATCGGCCACCCCC-AGAA

A-GGCAC---AAGTCATCTTCCAGCAGGTCAGGTGCCCGTAACATTACAACCTCAAAAGCAGGTTAA-------------

>B12

A-GGCAC---AGTCATCTTCCAGCAGGTCAGGTGCCCGTAACATTACAACCTCAAAAGCAGGTTAA-------------

>B362

A-GGCAC---AGTCATCTTCCAGCAGGTCAGGTGCCCGTAACATTACAACCTCAAAAGCAGGTTAA-------------
AGTCAGTATGGATATCCAGGAACGCTCCCAGCACCACAGGGCAG----GGTGCCATACCCTCAGCCACCCACTAGCAGAC
---------------------------------------------------------TTAGTCATACGGCACCACCTAGT
AGACAGGGTAGTAGGATCATGAAATTTATTGATAAAA------CCAGAAAAGAAGGAGA--TACTGAGGCGTGGCA
ATTCCCAGTACAATAGCGGATGCAAAGAGAGAGA-------------GCCCAAGGGGAGGACCTCCACAGTTGA
GGCCAGATACAAATCTTTTTTCGATAAAATGCT-AAAAAGATATGAAAAGGAGAGAA--ACAGATATGGACCAACTCCC
CTTATATGAGGACATTTATTGAGATCCCATATGCTTACATGACAT----AGACTCATCC-TTA--TGATGGGAGA-TTCTGCC
AAAATAGTCTTCT-CTCACCCTC-TCAATTTTTACAAAAATAGCTTGGC----------------GATTGATAGA-----GTAC
AAGAACA
Reproduction of HERV-K(HML-2) gag sequence recombinants

We aimed to reproduce recombination events between HML-2 gag sequences in an *in vitro/ex vivo* approach. We utilized different HML-2 gag sequences that we had cloned in the course of the experiment (see manuscript text). The gag sequences differed from each other by presence of unique restriction enzyme sites at the 5' or the 3' end of the sequence. Specifically, gag sequence B280 uniquely displayed an HindIII site at the 5' end. Sequence B255 uniquely displayed an RsrII site at the 3' end. Sequence B15 uniquely displayed a PflFI site at the 5' end (Suppl. Fig. A).

We produced RNA from those cloned gag sequences using the SP6/T7 Transcription Kit (Roche). We purified the RNA using the RNeasy Mini Kit (Qiagen) columns. We eliminated plasmid template DNA by DNase I treatment using Turbo DNA-free (Ambion Inc.) and following recommendations for rigorous DNase treatment. We generated cDNA from mixed equal amounts of RNA from clones B255 and B15 and from clones B255 and B280. RNA's were diluted at 1:1000 and about 1.5 ng of each diluted RNA were included in mixed cDNA synthesis. We used the Omniscript RT Kit (Qiagen Inc.). An RT negative control was also included at that point. Kits were used according to the manufacturers' recommendations. We subjected cDNA and RT(–) control to a standard PCR employing gag primers (see manuscript text). We proceeded further only when the negative control displayed no product. We cloned PCR products into pGEM T-Easy (Promega) and prepared plasmid DNA from 200 (100 per mixed RNA's ) clones using a standard miniprep procedure. We double-digested plasmid DNAs with appropriate restriction enzymes, followed by agarose gelelectrophoresis of restriction fragments.

Out of 100 analysed plasmid clones for mixture B255/B15, we identified 6 clones producing DNA fragments characteristic of recombination of two restriction enzyme sites in one sequence. We identified 8 clones that could not be digested at all, indicating recombination-mediated elimination of restriction enzyme sites (Suppl. Fig. B). Out of 100 analysed plasmid clones for mixture B255/B280, we identified 3 clones that indicated recombination of restriction enzyme sites, and 4 clones that indicated elimination of sites.

Taken together, 21 out of 200 plasmid clones (=10.5%) showed clear signs of recombination events. Thus, recombination events observed in the HERV-K(HML-2) expression analysis, and resulting in HERV-KX sequences, could be reproduced in an *in vitro/ex vivo* approach.
Suppl. Fig. A: Depiction of different HML-2 gag sequences employed to reproduce/screen recombinations in vitro/ex vivo. B255, B15 and B280 represent different gag sequences cloned into the pGEM T-Easy vector (Promega). B255 stems from locus 7_450 that is an HERV-K(OLD) provirus and that harbors a 96 bp insertion in its 5' region compared to B15 and B280 that stem from modern HML-2 proviruses (see manuscript text). The different gag sequences harbor unique restriction enzyme sites near their 5' or 3' ends. In the course of the in vitro/ex vivo experiment, recombinations result in combination of those sites within a clone, or in removal of those sites, depending on the direction of chimera formation. All three restriction enzymes do not cut the plasmid vector backbone. Indicated lengths of restriction fragments depend on presence/absence of the 96 bp sequence, and on direction and location of the recombination. We assayed RNA mixtures of clones B255/B15 and B255/B280. Gag sequences and location of restriction sites are not drawn to scale.
Suppl. Fig. B: Results of restriction enzyme-mediated screening for *in vitro/ex vivo* recombinants. Exemplarily shown are results for digestions of plasmid clones with PflFI and RsrII. As depicted in Suppl. Fig A, a restriction fragment of 689 bp or 593 bp is characteristic of recombination of both restriction sites in one plasmid clone (87, 88, 90), while undigestable plasmids indicate recombination-mediated elimination of both restriction sites from the plasmid clone (82, 84, 96, 98). Controls for properly digested plasmid B15+, B255+) and for undigested plasmid (15−, 255−) were included.