Vale's curation tips

- Pay attention to **CpG sites**, especially if you know that methylation is "used" by the genome to silence TEs.
- Apply a majority rule to decide how **indels/gaps** in the alignments should be represented in the new consensus sequence. Example: in a position we have 5 gaps ("-", 5 Cs and 5Gs. Break the tie first between insertion and gaps then apply the majority rule on the winner. In this case insertions win over gaps and you write "S" in the consensus (C/G = Strong).
- **Do not leave "?" in the new consensus sequences** because it is not recognised by blast and RepeatMasker as a valid character. If you have a tie between nucleotide frequencies, please use the ambiguity IUPAC characters (e.g., S, W, R, Y, M, K, ...). Also you can look at the nucleotide at that position in the old consensus to decide. In the worst case scenario "N" is a perfectly acceptable character.
- Double check that the new consensus sequences do not contain flanking regions and TSDs (be particularly careful with Mariners).
- **Double check that the features of the curated TEs are recorded** in the Excel file.
- **Adopt a consistent nomenclature within your library and try to be as consistent as possible with repeat libraries from closely related species** (if any) or, in general, to other manually curated repeat libraries. Check the names in the HD library (after the "#").
- Adopt a **nomenclature for non-autonomous elements that is consistent** with other repeat libraries. Always record it in the Excel file.

Vale's curation tips

- When you classify an element **check the naming convention for the class/superfamily in the HD library and on Repbase** if possible (it could be inconsistent with RepeatMasker though)
- In case you find **LTR retrotransposons**, please split the consensus in the LTR portion ("_LTR") and internal portion ("_I"); you can check on Repbase examples of LTR naming.
- In case you find **LINE retrotransposons**, don't worry if you cannot find both ends of the element. Because of the 5' truncation phenomenon, it is very difficult to obtain complete consensus sequences of LINEs. Try to extend the consensus as much as you can (do not build a consensus of less than 3 sequences). Try to have at least one end included in the consensus sequence
- In DNA transposons, TIRs must be symmetrical but one or two mismatches are ok
- The #curation_examples channel has a lot of useful examples. Also be careful with Zators which are a bit strange at first.
- Don't be afraid to use the "Comment" column to annotate anything you think interesting, you may find unexpected patterns as you keep curating the library.
- Move the curated alignments into the right subfolder → very important for Alex and I during the break at 18:30

SCHEDULE FOR TODAY

16:00 - 18:00 Supervised analysis of own data or participation in a collaborative project (III)
18:00 - 18:30 Break
18:30 - 19:30 Supervised analysis of own data or participation in a collaborative project (IV)
19:30 - 20:00 Overview 3: Course wrap-up *[AS, VP; with brief updates from everyone regarding results of own data analysis or collaborative project]*
20:00 - 21:00 Optional: Course pub and party!

Be sure that your curated alignments (alignment files + Excel file) for each group are correctly uploaded on Dropbox by 18:30!

Alex and I will collect the new consensus sequences to make the library version rm2.0 and analyse it: the results will be shown during the course wrap-up ☺

Anything curated after today 18:30 will be included in the library version 2.1

Today's group + reviewer pairings:
Group 1 + Group 12 in room of Group 1
Group 3 + Group 2 in room of Group 3
Group 5 + Group 4 in room of Group 5
Group 7 + Group 6 in room of Group 7
Group 9 + Group 8 in room of Group 9
Group 11 + Group 10 in room of Group 11