Classifying HEV sequences with “HEVsubtypingMEGAut .xlsx”. and submitting it to GenBank/EMBL.

A Quick Guide of use

This guide helps you to classify HEV sequences, and optionally submit it to GenBank/EMBL. Please, always check the result of each step, and **especially each detail of the final submission file** **– the responsibility for its quality will be always yours**. Any suggestion to make this procedure more accurate will be highly welcome. Please, send it to: [ArielVina.Rodriguez@fli.bund.de](mailto:ArielVina.Rodriguez@fli.bund.de). We assume you have more than a basic knowledge of MEGA and Excel (and Sequin, if you wish to submit the sequences).

You will need to insert your sequences in the HEV alignment, and analyze it with MEGA. In MEGA you will need to select the proper region and the “reference” sequences that will help you classified your sequences. In MEGA5 you can use group files to group sequences and to select some and unselect others, prior an analysis that use this groups. You can also use group files as a way to quickly group or etiquette sequences in a tree windows. You can create these group files using “HEVsubtypingMEGAut.xlsx”. From this Excel file you can filter the sequences you wish to use in each analysis based on many criteria, such as the genomic region, classification, host, geographical region, etc. This file can help you too propagate the features of the HEV genome into yours sequences to build a submission to GenBank or EMBL.

# Add yours sequences to the alignment.

1. Select in MEGA5 in the HEV alignment a reference sequence you expect to be similar to yours, being longer. Make a Copy of it (duplicate it). We routinely use FJ705359 (wbGER27).
2. Import your sequences next (below) to the duplicate sequence.
3. Automatically align with the duplicate reference sequence, selecting the option “Keep Predefined Gaps” (you may need to do pair wise only alignment).
4. Manually correct the duplicate with the “original”. Carefully insert gaps into your small alignment conserving it but simultaneously correcting the positions in the big alignment. Check the correctness of the aa alignment too. Strongly avoid insertions of columns of gaps into the big alignment.
5. Delete the duplicate copy.
6. Save the alignment with a new name (FASTA format).
7. Open it in analysis mode.

# Classify your sequences (genotype, subtype)

1. Select in the sheet “*Regions*” of “HEVsubtypingMEGAut.xlsx” the name of the genomic region you want to analyze. If you want to analyze some other region you can “manually” enter into the corresponding cells the position of beginning and end. This usually are the beginning and end of your sequences in the alignment, but possible an interior region of yours choice (always without primers sequences). This will apply a filter in the sheet “*Seq-class*”. Optionally apply any other filter to the sequences.
2. Select and copy the entire column headed “*Selected*” and paste it into a new text file (recommended end of the file name “\*.select.group file.grp.txt”)
3. Call the “Data/Select & Edit Taxa-Groups” dialog of the MEGA5 Data Explorer, unselect “*All*” sequences, Import Groups from the group file, add yours sequences into the group “*Selected*” and check this entire group.
4. Import any another group file of your preference (with subtypes or with etiquettes). Close dialog.
5. Call the “Data/Select & Edit Genes-Domains” dialog of the MEGA5 Data Explorer. Create and select a new domain with the same beginning and end of the previously selected region
6. Check all the sequences of interest are checked, and that no sequences without nucleotide in this region are checked.
7. Perform the Phylogenetic analysis of your preference and create a tree.
8. Edit the tree, importing alternatively all your preferred group files and grouping clade of sequences into subtrees, and name it with the genotypes, groups of subtypes, subtypes or other forms of classification.
9. Export the group file of the resulting edited tree.

# Update the “HEVsubtypingMEGAut.xlsx” file

1. Copy the sequences names from the MEGA align into the MEGA name column of the “*Seq-class*” sheet. Alternatively you can copy and paste this information from the group file exported from the edited tree.
2. Fill in as much information as you can in white-headers columns, mainly from A to Q (the head-colored columns will be automatically filled).
3. Additionally fill in columns CE (organism) to CM (note) with information to be use during sequence submission

# Prepare subalignment with new sequences to submit to GenBank or EMBL.

1. In MEGA select FJ705359 and the sequences that you with to submit.
2. Export it in NEXUS format
3. Organise and/or filter the items to made possible to copy the information about the sequences to submit from the columns CD (“Seq\_ID”) to CT (“subtype”) of the “Seq-class” sheet of the *“Seq clas - MEGA aut. 25.xlsx”*
4. Paste that in a text document. Include at beginning the headers of the copied columns (copy/paste too) and the corresponding information for FJ705359. Save this file to be use to import Source Modifiers for an alignment in Sequin.
5. Scheck you have the file *Feature accFJ705359.txt*. It has the content of the “Feature\_accFJ705359” region of the “Feature” sheet of the *“Seq clas - MEGA aut. 25.xlsx*”(saved as text file) ([http://www.ncbi.nlm.nih.gov/Sequin/table.html#Table](http://www.ncbi.nlm.nih.gov/Sequin/table.html%23Table))

# Use of Sequin

Scheck you has downloaded and installed [Sequin. (from NCBI)](http://www.ncbi.nlm.nih.gov/Sequin/download/seq_download.html). Run it.

1. Select DB for submission: GenBank (you can by a new load change to EMBL)
2. Start New Submission
3. Fill in submission information OR *File->Import submitter Info* from an select a previously exported file (txt file but with ASN.1 content) and edit to adapt to a new submission.
4. Select Normal submission
5. Select Phylogenetic Study. Now you can select format NEXUS (Using filters, MEGA5 export in NEXUS format, not in FASTA. Select Original submission
6. Import from your previously from MEG5 exported subalignment. Ignore error about missing organism information.
7. Select sequencing method (for direct PCR sequencing: usually Sanger dideoxy sequencing and raw sequence reads)
8. Import your Source Modifiers text file. Ignore message about some field duplication
9. For Annotation you can select “none”, no title and Prefix title with organism name
10. Continue to record view
11. Cancel “Far pointer…”
12. *File -> Open…* and select file “*Feature accFJ705359.txt”*.
13. Select *accFJ705359. Edit -> Feature Propagate…* From *accFJ705359,* All features to yours sequences.
14. Edit -> Alignment Assistant… Expand the features for all sequences. Check and correct any error during feature propagation in nucleotide and amino acid sequences.
15. You can change the title of each sequence: Annotate -> Descriptors -> Title to add genomic region or other information and make it unique and uniform. The database staff will probably change it.
16. File -> Save as… with a temporary name.
17. *Search -> Validate* will help find errors. It can exist a rigorous way of getting rid of all errors, correcting it. After that you can *File -> Prepare submission*, save a file, and send it to [gb-sub@ncbi.nlm.nih.gov](mailto:gb-sub@ncbi.nlm.nih.gov) with some comment. But I have not a better suggestion than the following:
18. Edit -> Sequence Deletion Tool. Delete *accFJ70535* and all the corresponding proteins.
19. *File -> Prepare submission*, ignore errors and save a file, eliminating the alignment.
20. Reopen this file. Accurately eliminate all the errors you see after *Search -> Validate*
21. *File -> Prepare submission*, save a file, and send it to [gb-sub@ncbi.nlm.nih.gov](mailto:gb-sub@ncbi.nlm.nih.gov) with some comment.
22. Alternatively you can reopen this file for submission to EMBL and *File -> Export EMBL*. This new file can be use during [submission to EMBL](http://www.ebi.ac.uk/ena/about/embl_bank_submissions).

# Description of the columns:

Sheet **Seq-class**:

***MEGA name****:*

Have to **exactly** coincide with the sequence name in MEGA.

Very helpfully if just the GenBank Accession Number.

Repeated names are automatically colored (usually an error).

***Tab-Pub***: Is a filter:

**.**  --> OK, this row is in use.

**e** --> some error.

**n** --> empty row, new, with the formats and formulas ready to use.

(Provided as a place where you can add your new data)

***Strain name****:*

Contain the strain or isolate name.

Take care to consistently enter the same name for all the sequences belonging to the same strain/isolate. In some few case there is know the strain and the isolate name: in these case the isolate name is written in the Isolate column.

Repeated names are automatically colored (NOT an error).

***Isolate name****:*

It will be combined with the strain name in the Table column “***Strain name / isolate***”

***Country cod****:*

The 3-letter country code after ISO 3166-1 alpha-3.

***Lu & Li:***

Subtype as appear in ([Lu et al., 2006](#_ENREF_18)). Only for sequences cited there. Any value here will cause a mark (\*) in the Table and in etiquettes.

Next 6 columns: ***CG***,***ORF1.250nt***, ***HVR.247nt***, ***RdRp.280nt***, ***ORF3.225nt*** and ***ORF2.187nt:***

Result of the genotyping using these regions.

Next 6 columns: ***CG***,***ORF1.250nt***, ***HVR.247nt***, ***RdRp.280nt***, ***ORF3.225nt*** and ***ORF2.187nt:***

Regions present in the sequence.

***Beg****,* ***End:***

Beginning and end of the sequence in “***alignment”*** *coordinates*.

***Length:*** Formula

***Contain region:***

Formula. For use as filter. Will be **True** if the sequence fully span (contain) the region selected in the sheet “*Regions*”.

Next 3 columns: ***C1***,***C2*** and ***C3:***

Free columns, to be use for comments or filters.

Next 3 columns: ***F.ORF1***, ***F.HVR*** , ***F.RdRp*** , ***F.ORF3*** and ***F.ORF2:***

Filters used to decide the sequences to appear in the corresponding Fig. or Tree. Possible usage:

**n** --> will not appear.

**s** --> will appear.

They also modify the corresponding column in the Table: add there a **+** when these filter begging with **s** (sequence proposed to be a “*standard*” for classification of other sequences using these regions)

Next 3 columns: ***C4*** and ***C5:***

Free columns, to be use for comments or filters.

***NCBI:*** Formula

Is a WWW link to the NCBI site and will show the full sequence Item. Very handy while actualizing the data of the sequence, and to see other information not in these file. Especially useful to *quickly find the original publication*.

Next 3 columns: ***Etiq-Automat***, ***Acc.Strain*** and ***Acc:*** Formulas.

Possible different forms of sequence etiquette.

***Selected:*** Formula

After all the filters are applied, and only the sequences you want to analyze are displayed, this is the column you need to select, copy and paste in a new, empty text document. Save the document to be use as “*group file*” in MEGA.

Next 4 columns: ***Etiq-Strain***, ***Etiq-Strain+Automat+CG***, ***Etiq-Strain+Automat*** and ***Etiq-Automat***: Formulas.

Different variants of etiquettes **to be used in MEGA** as “*group file*”.

Next 3 columns: ***g.subtype***, ***g.grupe*** and ***g.genotype***: Formulas.

Different variants of grouping the sequences, **to be used in MEGA** as “*group file*”, especially to import into tree windows.

***f.subt***: Formula

Is the best column to order and filter the sequences by subtype.

Next 9 columns: ***Accession no.***, ***Strain name / isolate***, ***Classification & country***, ***CG***, ***ORF1.250nt***, ***HVR.247nt***, ***RdRp.280nt***, ***ORF3.225nt***, and ***ORF2.187nt***: Formulas, **Table**.

Together generate a **Table** as appear in Vina-Rodriguez, 2013.

***Orig Gr Select f:*** (Original-Group Selected filter)

When you finish in MEGA of classifying a new bath of sequences, you can export a group file containing all these information, open it in a text editor, and copy all to paste it into this column.

The next 5 columns will help you parse this information back to this Excel file.

Next 5 columns: ***MEGA***, ***MEGA name***, ***MEGAselect***, ***AccN*** and ***StrainName:***  Formulas.

You can copy from here and *paste by value or content only* into the corresponding initial columns of this Excel sheet.

The rest of the columns are used to generate the file you need to submit the sequences to GenBank, or just to keep track of the primers used to generate each fragment.