Integrated virus explorer (IVEX) pipeline guide

read hashtagged text in scripts
the start section of scripts detail how to run the script in command line, what arguments and prerequisites are needed, and specific setup requirements, if any. Scripts also include comments that explain most lines of code

KEY
= script is an example of workflow, not to be run as is
! = script must be modified for each user OR specific project OR change to an existing project (modify a copy of script)
$ = script is optional
£ = script must be run again in its entirety if genomes are added or removed from the project
% = script uses software that was not available at the time on the ABiMS cluster: software must be installed locally and script must be edited to match the user-specific software installation. The user must also move the output files to the cluster via rsync/uploading/scp.

script run per:
A project (i.e. it runs on all genomes together, or it generates an output used by all genomes)
B genome subset (i.e. it runs on smaller groups of genomes, called rungroups)
C genome (must be run separately per genome)

SECTION 1 - scripts mostly run as provided automatically, with only minor script changes possibly required
generates the main files that will be used to generate all following figures or information
scripts are generic and will run on any genomes as long as format requirements are met and no genome specific information is required

START

IVEX000_NCVOG.sh (A) SETUP
- setup project directories (manual step: see in-script comments)
- script creates Nucleocytoviricota Orthologous Groups reference file

IVEX000_genomes.sh (#) PREPARE THE GENOMES
- at this stage, copy ALL genome assemblies (nucleotide fasta),
proteomes (protein fasta), and annotations (gff files) to IVEX000_genomes/
- read script for examples of required formats for these files and examples
used to reformat specific genomes so that they are compatible with the rest of the pipeline.
- I recommend you copy this script and record in it any commands used to reformat your genomes.
- phaeoexplorer genomes will not need any reformatting,
but ensure their files are still copied to IVEX000_genomes/ (or any other genomes).
- Once all genomes are ready and in IVEX000_genomes/, gzip all the genome assemblies (nucleotide fastas)

IVEX000.sh (% ! B) RE-ANNOTATE THE GENOMES
- using a virus-optimisied annotation program (GeneMarkS-Viruses), this
script will find additional viral genes that tend to be missed by eukaryote-optimised annotators.
- IMPORTANT - GeneMarkS is not on the ABiMS cluster (it has a free, but temporary licence) - you can easily
install it locally in your personal folders on the cluster, see : http://exon.gatech.edu/GeneMark/license_download.cgi
You will then need to change the GeneMarkS variable in the script to your GeneMarkS installation location (the .pl file).
This is the only change you will need to make to this script.
- as explained in the script, you must assign your genomes to smaller subsets called rungroups - this is to reduce the
number of jobs running at once on the cluster - e.g. instead of waiting for 60 queued jobs to run (with most queued until the previous jobs finish),
I ran them all simultaneously in rungroups of 5 genomes. Even though a single genome is processed at a time per rungroup,
it is still much faster than running 60 individual jobs. If you can run all your genomes individually simultaneously,
then do so - simply assign 1 genome per rungroup.
- Once complete, each genome will have a new protein fasta and gff file (in IVEX000/) comprised of the new GeneMarkS annotations
merged with the original ones (including only GeneMarkS genes that did not overlap the original genes).
The rest of the pipeline will analyse these files.

IVEX001.5_virsorter.sh ($ C) RUN VIRSORTER2 PIPELINE
- Optional because this pipeline is not optimised for eukaryotic datasets, which means it is
unlikely to be useful. At best, it will flag the same contigs as viral as
the blast results. But this will only happen in a genome with large contigs and that has large viral inserts that are
similar to known phycodnaviruses like EsV-1. Skip this script, unless the Virsorter pipeline
has been updated to work better on eukaryotes.
- If you do run this script, open the output gffs with the fastas in a genome viewer like Geneious
or extract a table of contigs flagged as viral by virsorter and add it to the IVEX002 summary table

IVEX001.sh (B) BLAST ANALYSES
- IMPORTANT: check the setup details as detailed in the script comments
- this is the slowest stage; may take up to a week for a set ~5 genomes
(but this ok if all rungroups are running simultaneously)

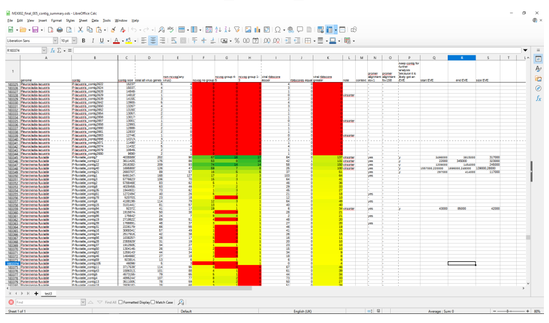
IVEX002.sh (£ A) PREPARE INPUTS TO GENERATE MAIN FIGURES IN R
- run once IVEX001.sh is complete for all genomes
- also generates tabular summary files and modified gffs that are more
suitable for viewing in excel/genome viewers (see script comments)
- details of figures given in script comments
- may take 24 hours to run

IVEX002.r (! £ A) GENERATE MAIN FIGURES IN R
- script will need to modified quite a lot to match the number and names of genomes, as well as the desired composition(s) of figures
(especially the scatterplots), and figure legends which are set within this script
- changes required are explained in scripts comments, but may require some understanding of R to follow
- note that the libraries specified in the script will need to be installed in R

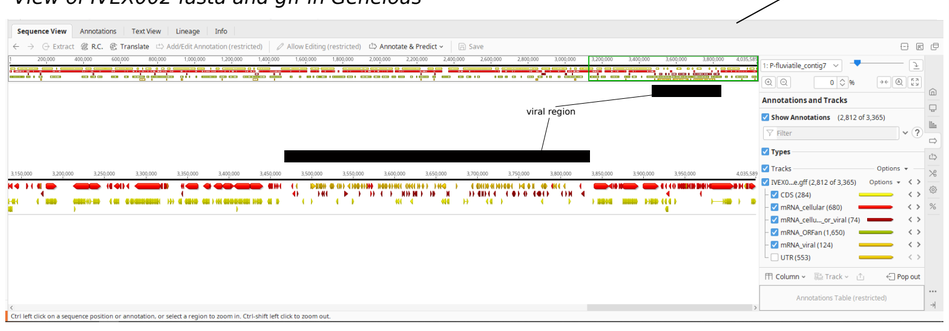
IVEX_promer.sh ($ %) (STEP_001: ! C) (STEP_002: ! # £ A)
- Optional, but very useful for identifying regions with good similarity to a reference viral genome - STEP_001 will run as is
- Running this script is a prerequisite for the IVEX_genoplotr scripts
- With modification, STEP_002 details how you could summarise obtain a checklist of all contigs which aligned to reference genome(s), which could be appended to the IVEX002_summary file

SECTION 2 - scripts are mix of manual and automatic processes - some script editing needed for the R script. The outputs include the main figures and in general these outputs are generic and will be useful for any project. The manual inspection step can be highly time consuming but it will record a useful overall view of the viral inserts present in your genome(s).

View of edited IVEX002_contig_summary in spreadsheet



View of IVEX002 fasta and gff in Geneious



MANUAL INSPECTION
- Before continuing with further scripts, you must manually inspect the
contigs to identify contigs/regions which are likely viral and related to your virus of interest
- Open IVEX002_final_005_contig_summary and save a copy as a spreadsheet
- I recommend at this stage adding to IVEX002_final_005_contig_summary the
optional checklists indicating which contigs aligned in promer
(per reference virus of interest) and virsorter.
- Next, open the gff and fastas generated by IVEX002.sh in a genome viewer like Geneious
- For each genome, go through the list (they are in descending order of number of good viral hits),
quickly inspect the contig in Geneious. Look for clusters of monoexonic genes in the mRNA_viral
or mRNA_viral_or_cellular tracks, check some blast hit names or for core viral genes. If an interesting
viral region is present, record its approximate loci in the spreadsheet (contig size, i.e. the end is already
included in the file)
- Repeat until the contigs start to have only false viral hits, short viral regions, or if the contigs themselves
are very short. The point is to find the bulk of the good viral regions/contigs, not to find every single one.
Repeat for each genome.
- Once complete for all genomes, extract a list of all contigs with interesting viral regions for each genome

IVEX004.sh (C) REDUCE GFFS AND FASTAS DOWN TO ONLY CONTIGS WITH VIRAL REGIONS OF INTEREST
- This script also modifies the gff by adding blast hits for non-viral genes, which may add useful
information to cellular genes accompanying viral regions.
- Note that this script requries a list of contigs of interest identified by MANUAL INSPECTION
- Do not run this script on full genomes - it will only run in a reasonable amount of time on small contig
sets per genome.
- Once run, use the final output IVX004 fastas and gffs for viewing in Geneious from here onwards

SECTION 3 - all of these scripts must be edited to match a specific project/user (these can be complex changes, especially for the R scripts). These scripts are intended to generate specific figures that have been pre-planned by the user. All are optional - skip scripts that generate figures/outputs that are not relevant to your project.
Manual identification of small sets of contigs and viral regions is a prerequisite for many of these scripts.
For example, I used these scripts to generate figures that specifically depict the alignments and structures of 6 viral insertions in the genome of Porterinema fluviatile, in comparison to the phaeoviruses EsV-1 and FsV-158.

IVEX_genoplotr.sh & IVEX_genoplotr.r & IVEX_genoplotr_similarity_summary.sh ($! A/B/C) GENOPLOTR VISUALISATIONS
- IVEX_genoplotr.sh generates the input files for IVEX_genoplotr.r
- These scripts require quite a lot of editing to work on specific genome sets -
this is explained in the script comments
- Ideally with these scripts you will have a specific figures for a paper planned,
and you will be editing a copy of this script to work on a specific set of viral
regions and reference genome(s)
- IVEX_genoplotr_similarity_summary.sh is a simple script used to summarise numerically the alignments
between a specific set of viral inserts and reference genomes

IVEX003.sh ($ # ! £ A/B) PHYLOGENY
- general example of workflow to prepare proteins for phylogeny
- commands for alignment and mafft and RaxML should run as they are on the cluster

IVEX_nucmerRep.sh (% $! C) IDENTIFY REPEAT REGIONS
- This script attempts to identify repetitive regions in contigs - it was specifically intended to identify the
inverted terminal repeats of phaeoviruses.
- Though this script is simple, it will likely be of little use, as we currently lack exogenous viral genomes
which could allow identification of the termninal repeat sequences.
- I recommend avoiding this script, unless you are working on repeat sequences - and even then there
must be software packages that can do this better.

RNA_plotting.sh & RNA_plotting.r ($! C) VISUALISE RNA EXPRESSION
- These scripts both require complex editing to work on specific projects/genomes.
- They plot RNA expression from TPM data as a heatmap.
- Their setup is explained in the script comments, but it may be a lot of effort.
- Maybe try this script if needing RNA info, but there are certainly software packages
that can do this with more ease

IVEX_cgview.sh (% $! A/B/C) CIRCULAR VISUALISATIONS OF CONTIGS
- Command line arguments control various features of figure - see CGview guides online
and script comments for details.

IVEX_stramenopiles.sh ($! A/B/C) BROADER STUDY OF PHAEOVIRUS GENES IN STRAMENOPILES
- Highly specific script to identify a set of phaeovirus genes from across a wide range of stramenopile
genomes. Ignore this script, unless it is a useful example for a related project.

These steps are all to be done in order

These steps can be done in any order