Seq2Geno

An integrated tool for microbial sequence analyses, aiming to allow biologists to focus on interpreting the results and assist bioinformaticians to ensure method reproducibility

Tzu-Hao Kuo
Department of Computational Biology of Infection Research
Helmholtz Centre for Infection Research
Inhoffenstraße 7
38124 Braunschweig

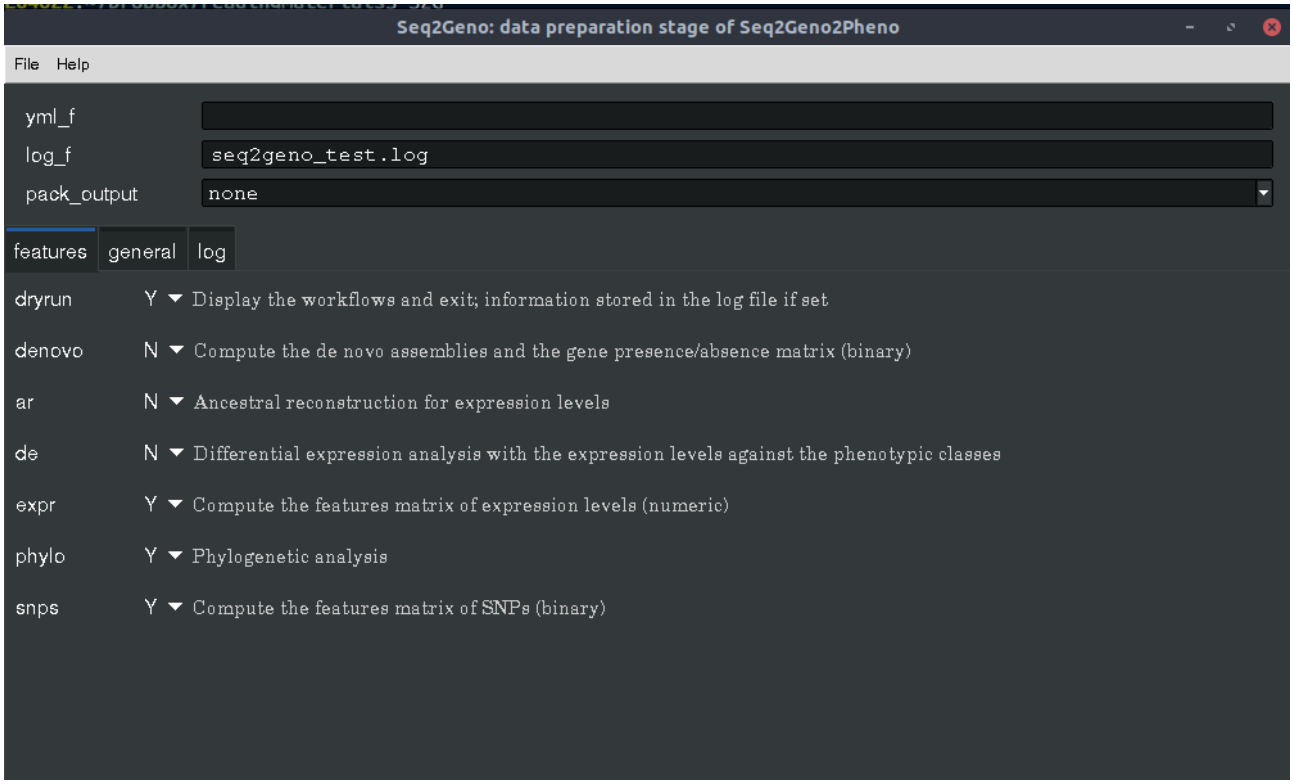
Options

panels/fields		description
yml_f		The yaml file as the input of Seq2Geno, where the arguments described in the main panel will be written
log_f		If specified, the log file will store messages that are redirected from stdout and stderr from Seq2Geno. For dryrun mode, the procedures will be printed.
pack_output		Pack the results into a zip file. Opt 'none' for keeping the folder; 'all' for packing everything, including the intermediate data, and delete the working directory; 'main' for packing the main results and deleting the working directory; 'g2p' for packing only those needed by the predictive package Geno2Pheno and deleting the workign directory (automatically opted in the remote mode
features		The workflows for computing features
	denovo	Compute the de novo assemblies and the gene presence/absence matrix (binary)
	snps	Compute the features matrix of SNPs (binary)
	expr	Compute the features matrix of expression levels (numeric)
	phylo	Phylogenetic analysis
	ar	Ancestral reconstruction for expression levels
	de	Differential expression analysis with the expression levels against the phenotypic classes
	dryrun	Display the workflows and exit; information stored in the log file if set
general		Input data and other files
	cores	number of cpus
	mem_mb	memory size (mb)
	old_config	re-use the procedure-specific config files previously generated in the project folder (wd)
	dna_reads	The list of DNA-seq data (paired-end reads)
	wd	working directory
	phe_table	The list of phenotypes
	ref_fa	The fasta file of reference genome; only ONE sequence should be contained
	ref_gbk	The genbank file of reference genome; only ONE chromosome should be contained
	ref_gff	The gff file of reference genome; only ONE chromosome should be contained
	rna_reads	The list of RNA-seq data (short reads)
	adaptor	The fasta file of adaptor sequences for trimming reads

Graphical user interface

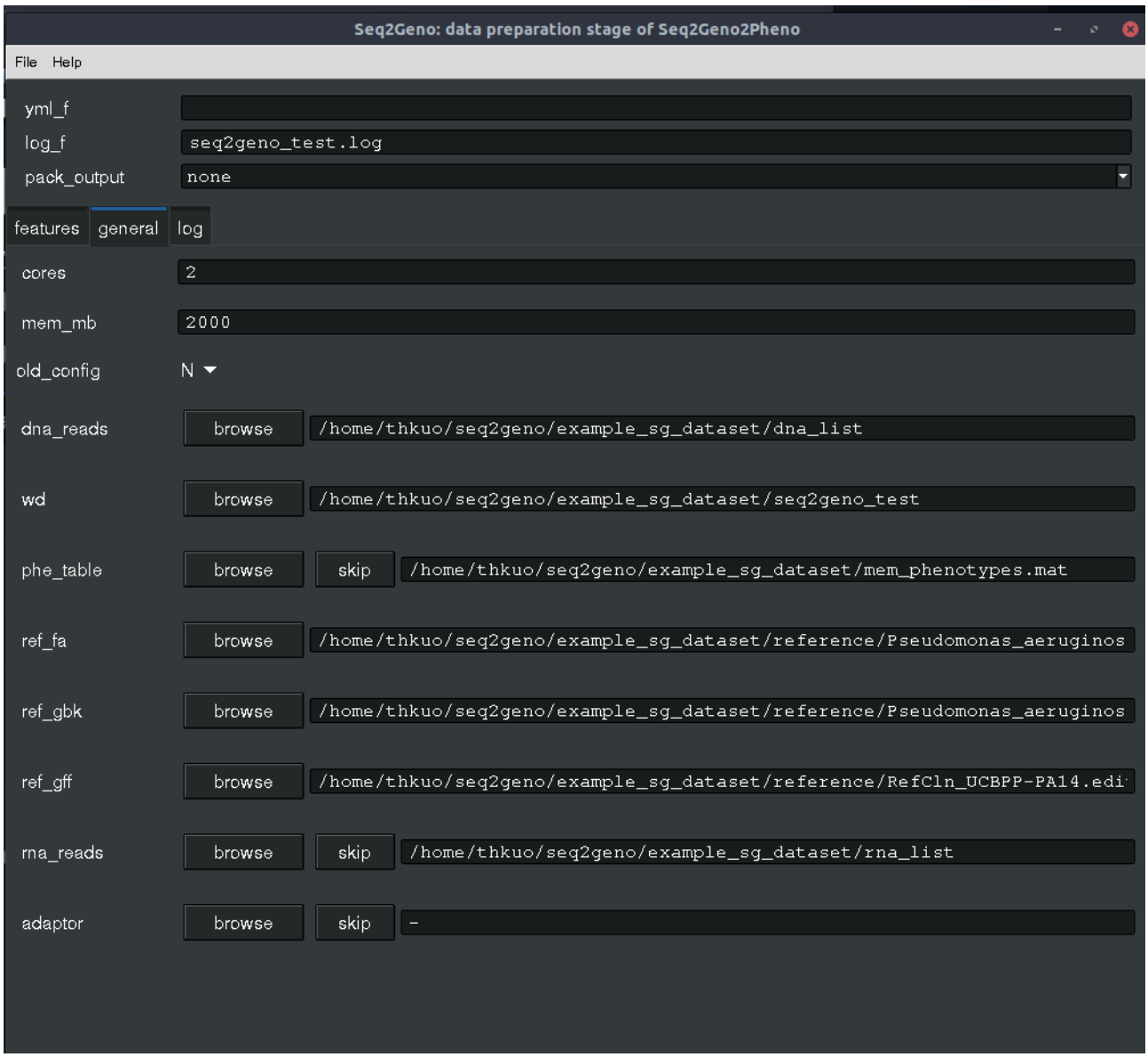
A new project without a yaml file

In the menu bar, *File* includes the options to read or save files. As we move down, there are long horizontal fields. We haven't yet had a yaml file so just skipping the first. We can set the log file name **seq2geno_test.log** in the second field. The third field for output format is set as **none**, with which the output folder won't be packed into a zip after the analyses are done.

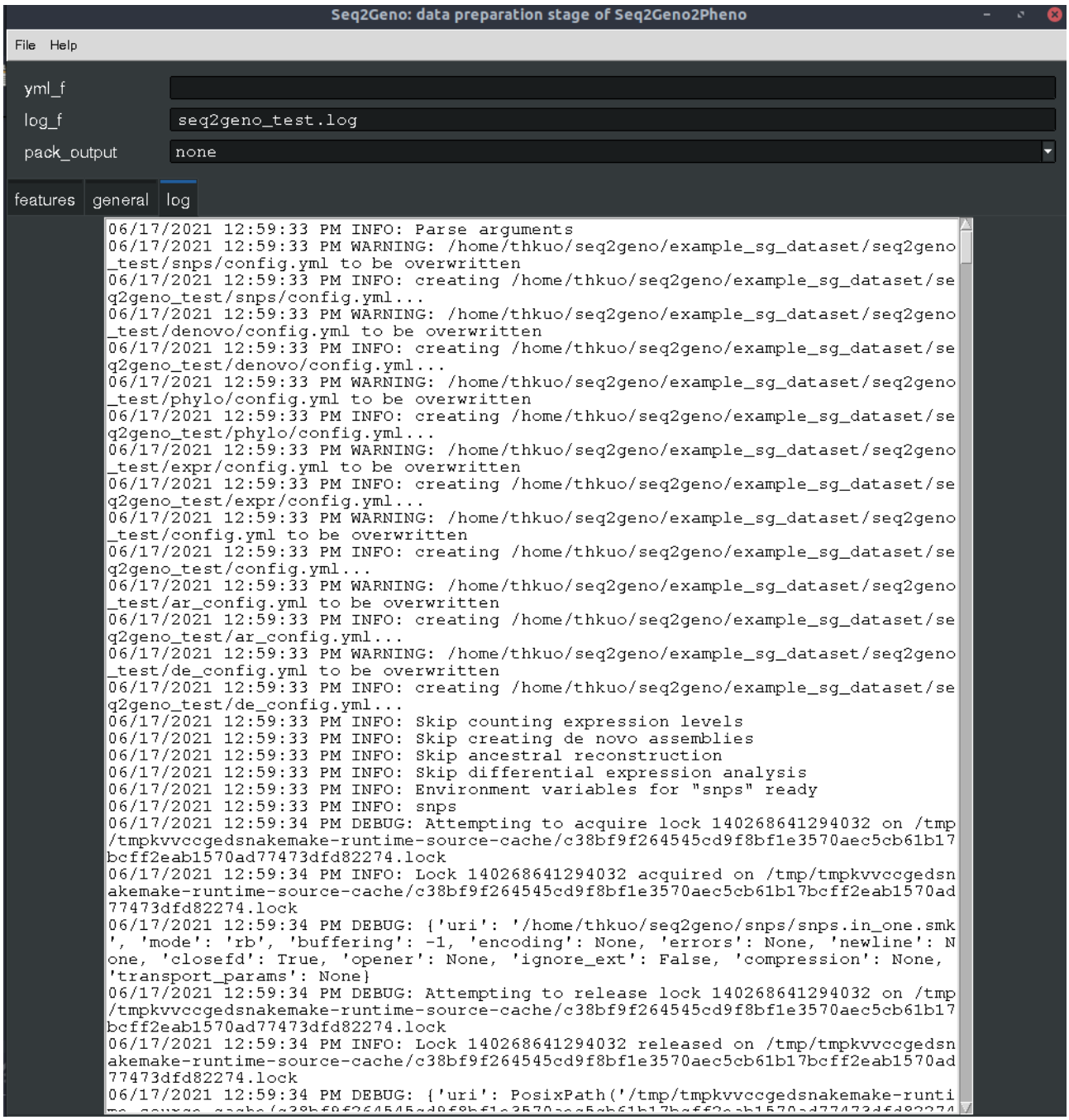


In the *features* page, opt **Y** to include the analyses (details in the previous chapter). In this tutorial, we include *snps*, *expr*, *phylo*, as well as *dryrun* to have a preview.

Then, we will be determining the input data in the *general* page. If the working directory (set in *wd*) is new, make sure that *old_config* is **N**; otherwise, the package won't find the required configuration files.



After everything is set, move to the menu bar and click *Run*. A window will pop out to allow you to save the settings are start running. Because we have opted *dryrun*, the processes won't start and Seq2Geno will only write the working plan in the log file **seq2geno_test.log**. Click *Load log* and select the log file, and then click *log* to view the information.



If the settings and analyses plans look fine, go changing the other log filename such as **seq2geno_real.log**. Then, go to *features* and set **N** for *dryrun*. Finally, we are about to start the analyses: go clicking *File* and then *Run* to launch the analyses. You can view the status by loading the log file like mentioned above or using any your preferred text viewer/editor.

A new project with a yaml file

If you want to start by editting a pre-determined or shared settings, go to the menu bar and click *File* and *Load yaml* to import the old settings. You can review or edit the settings like mentioned in the previous section. Otherwise, just run it by clicking *File* > *Run*.

Add analyses to an old project

It is possible to add some more analyses after one is done without rerunning the others. For instance, we want to compute de novo assemblies and gene presence/absence table after the above analyses were finished. First, load the old yaml using *File > Load yaml*. Then, repeat the above: setting another log file in *log_f,* opt **Y** for *denovo* and then go to *File > Run*.

Add more samples to an old project

Adding samples are also possible without recomputing everything. Given an old project, there should already be a list of DNA-seq reads set in *dna_list*. Add the new samples into the list, set another log file in log_f, and then *Run* the analyses again. You could also use *dryrun* to ensure what and how many new analyses will be launched.