Summer School 2019

What kind of *Daphnia* are swimming around in northern german lakes?

We extracted DNA from 30 individuals per lake, and amplified 4 different DNA fragments through PCR: 3 nuclear fragments, and one mitochondrial

**First steps**

* Open your browser and copy the following URL: <http://galaxy.biologie.uni-hamburg.de:8080/> --- Only accessible within University network
* use the provided user name and password to log into Galaxy Hamburg

**How does Galaxy work? A few essential steps…**

The homepage of Galaxy is divided into 3 Panels, with a top menu

Left: tools

Middle: workspace

Right: the history, where you can see all datasets you are working with

**Data**

We have many different data formats to store different kinds of information

Fasta and fastq store sequences

BAM/SAM store alignments

Bcf/vcf store variant information

**Analysis**

We have two groups of reads: the reference samples (single letter name) and the biomonitoring samples (abbreviations with 2 or 3 letters). I ran the full protocol for all samples but one, and we will analyse one of the biomonitoring samples together to visualize the steps.

1. Import data and Workflow

- click on “shared data” and choose the History “summer school 2019” to import it into your workspace

- follow the same procedure to import the workflow

**2. Quality check**

What does the data for WRD look like? Use FastQC on one of the WRD files

Look at the output (webpage version)

- How many reads do you have in each WRD sample?

- How long are the reads?

**3. Data filtering**

Some reads are quite short and not interesting, and the quality values can be used to sort the bad ones out

Use Trimmomatic to automatically sort your data

Example:

|  |  |  |
| --- | --- | --- |
| Single-end or paired-end reads? | pair\_of\_files |  |
| Input FASTQ file (R1/first of pair) | [102: PA\_S16\_L001\_R1\_001.fastq.gz](http://galaxy.biologie.uni-hamburg.de:8080/datasets/8e41af07def5ed02/show_params) |  |
| Input FASTQ file (R2/second of pair) | [103: PA\_S16\_L001\_R2\_001.fastq.gz](http://galaxy.biologie.uni-hamburg.de:8080/datasets/2beeb3aa33d99f42/show_params) |  |
| Perform initial ILLUMINACLIP step? | yes |  |
| Select standard adapter sequences or provide custom? | standard |  |
| Adapter sequences to use | Nextera (paired-ended) |  |
| Maximum mismatch count which will still allow a full match to be performed | 2 |  |
| How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment | 30 |  |
| How accurate the match between any adapter etc. sequence must be against a read | 10 |  |
| Minimum length of adapter that needs to be detected (PE specific/palindrome mode) | 8 |  |
| Always keep both reads (PE specific/palindrome mode)? | True |  |
| Select Trimmomatic operation to perform | SLIDINGWINDOW |  |
| Number of bases to average across | 4 |  |
| Average quality required | 15 |  |
| Select Trimmomatic operation to perform | TRAILING |  |
| Minimum quality required to keep a base | 10 |  |
| Select Trimmomatic operation to perform | MINLEN |  |
| Minimum length of reads to be kept | 70 |  |
| Output trimlog file? | False |  |
| Output trimmomatic log messages? | False |  |

**4. Map the reads to reference sequences using BWA-MEM**

The clean data named “Trimmomatic on data x & y” will be used to align our sample of interest, here WRD, to a reference. The tool is named BWA-MEM. It quickly compares our data to know sequences and aligns them: in the HTS world, this is called mapping. This procedure has been conducted on all other samples already, and I provided you the mapping files.

|  |  |  |
| --- | --- | --- |
| Will you select a reference genome from your history or use a built-in index? | history |  |
| Use the following dataset as the reference sequence | [RefSeq.fasta](http://galaxy.biologie.uni-hamburg.de:8080/datasets/e636a3db1804c19b/show_params) |  |
| Algorithm for constructing the BWT index | Auto. Let BWA decide the best algorithm to use |  |
| Single or Paired-end reads | paired |  |
| Select first set of reads | [Trimmomatic on WRD\_ R1\_001.fastq.gz (R1 paired)](http://galaxy.biologie.uni-hamburg.de:8080/datasets/47d4d9ce7078d8c6/show_params) |  |
| Select second set of reads | [Trimmomatic on WRD\_R2\_001.fastq.gz (R2 paired)](http://galaxy.biologie.uni-hamburg.de:8080/datasets/89f6aa8a98ef1b7f/show_params) |  |
| Enter mean, standard deviation, max, and min for insert lengths. | Empty. |  |
| Set read groups information? | set\_picard |  |
| Auto-assign | false |  |
| Read group identifier (ID) | BioWRD |  |
| Auto-assign | false |  |
| Read group sample name (SM) | WRD |  |
| Auto-assign | true |  |
| Platform/technology used to produce the reads (PL) | ILLUMINA |  |
| Platform unit (PU) | 000000000-D37YK |  |
| Sequencing center that produced the read (CN) | FoodSec |  |
| Description (DS) | Empty. |  |
| Predicted median insert size (PI) | Not available. |  |
| Date that run was produced (DT) | 2017-11-16 |  |
| Select analysis mode | illumina |  |

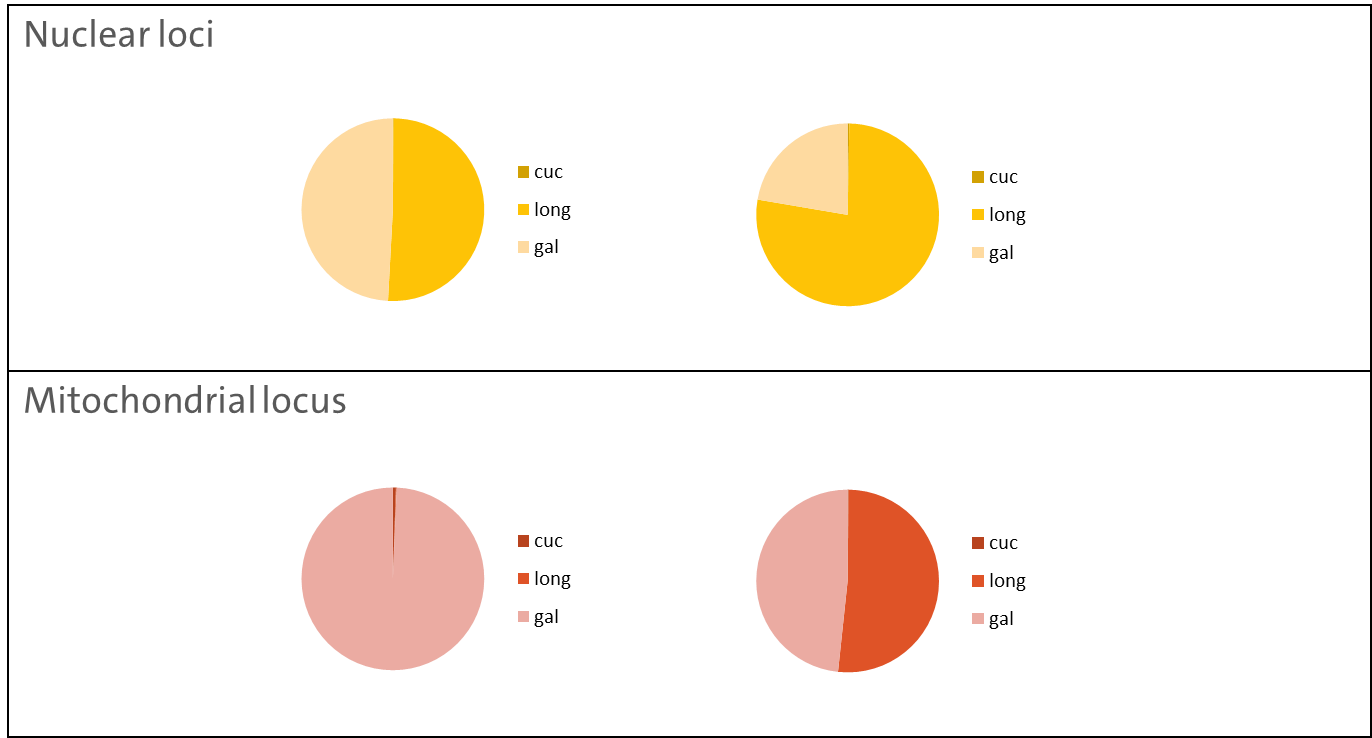
5. Call variants

For this step you will use the provided workflow. Briefly, it concatenates all single mapping files into a single one, first in a pileup format, and then in a vcf file. Because we are only interested in a subset of the vcf file, there are some additional filtering steps in the workflow.

You should include all the mapping files (total 15: 9 biomonitoring and 6 reference samples) for the mpileup step, and provide a list of the samples in the correct order for the VarScan step. The cut column step also need you to name the columns of interest: this information is coded in the workflow.

6. Export the obtained files to your computer and open them in a text editor. Extract the relevant column in this file (the one with percentages).

6. You will be given a list of species specific loci (alleles.xls) to analyse the allelic frequencies and answer the two questions:

Which reference samples do these pie charts represent? 

Does WRD contains hybrids?