GÖTEBORGS UNIVERSITET

An introduction to bioinformatic tools for population genomic data analysis, 2.5 higher education credits Third Cycle

Faculty of Science; Department of Marine Sciences
The Swedish Royal Academy of Sciences

---

1. Confirmation

The syllabus was confirmed by the Steering Committee of the Department of Marine Sciences.
Discipline: Natural Science
Responsible department: Department of Marine Sciences
Main fields of study: Bioinformatics

2. Position in the educational system

Elective course; third-cycle education.

3. Entry requirements

Admitted to third cycle education.

4. Course content

This course aims at detailed understanding and hands-on experience of using state of the art bioinformatics pipelines for one's own biological research questions. An important aspect of the course is to show how genomic data can be applied to address and answer research questions in the fields of genetics, ecology, population biology, biodiversity monitoring and conservation. The students will be trained in the latest bioinformatic methods to analyze high throughput sequencing data, which is present in many research projects. The course will cover basic computing tools required to run command line applications, processing high throughput sequencing data of whole genome / exome / restriction site digested (RAD) DNA for population genomic studies.

The first part of the course introduces general computing tools for beginners such as the UNIX command line environment, bash commands, data formatting using regular expressions and basic scripting in the unix shell with a series of examples and exercises using a remote server.

The course introduces bioinformatics software for analysis of RAD-data, and downstream population genetic analysis of genotype data.
The course also introduces basic and advanced concepts of population genomics data analysis such as genome/transcriptome assembly, alignment/mapping, differential gene expression, functional enrichment tests, SNP genotyping, PCA, population structure analysis, outlier tests, and demographic analysis based on allele frequency spectra (AFS).
The course corresponds to 1 week of full time studies and and is composed of lectures, demonstrations and computer labs.

5. Outcomes

1. Knowledge and understanding
1a. Demonstrate advanced knowledge of experimental strategies, applications and bioinformatic tools for population genomics.
1b. Demonstrate advanced knowledge of the potential of genomics approaches to answer ecosystem-wide questions, in particular for biodiversity monitoring.

2. Skills and abilities
2a. Ability to use basic commands in the Unix command line environment (reformatting data with regular expressions, basic scripting, running python scripts from the unix shell)

2b. Ability to use different software tools to analyse sequence data from restriction-site digested DNA (data cleaning steps, clustering of reads, mapping to reference genomes, extracting and filtering genotype data.

2c. Ability to use population genomics software tools to assemble and a genome/transcriptome, and perform gene alignment/mapping, differential gene expression, functional enrichment tests, SNP genotyping, PCA, outlier tests, population structure, and demographic analysis.

3. Judgement and approach
3a. Formulate one's own research questions, identify data and tools needed to answer these questions and critically evaluate and analyse the results.

6. Required reading

Part 1: General computing tools.
This will be the main textbook for the introduction to general computing tools:
- Haddock and Dunn (2010). Practical computing for Biologists. Sinauer Associates.

Part 2: RAD data analysis.
- Wang et al. (2012). 2b-RAD: a simple and flexible method for

genome-wide genotyping. Nature Methods 9, 808-810.
- Davey et al. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12, 499-510.

Part 3: Population transcriptomics
- De Wit et al. (2012). The simple fool's guide to population genomics via RNA-seq: an introduction to high-throughput sequencing data analysis. Molecular Ecology Resources 12, 1058-1067.

Online course material
- The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Details of the pipeline can be found at (http://sfg.stanford.edu) Practical computing for Biologists (http://practicalcomputing.org)
- Github repositories:
https://github.com/z0on/2bRAD_GATK
https://github.com/z0on/2bRAD_denovo
https://github.com/DeWitP/Bioinformatic_Pipelines
https://github.com/The-Bioinformatics-Group/Learning_Unix

7. Assessment

Attendance is mandatory for a pass grade.

8. Grading scale

The grading scale comprises Fail (U), and Pass (G).

9. Course evaluation

The course evaluation will be carried out through an online questionnaire.

Additional information
Language of instruction will be English, as international guest lecturers will participate.

11. Preliminary course schedule

Part I: Introduction to Unix and high-throughput sequencing methods (prerequisite for parts 2 and 3)

Length: 1 day

Date: 6th November 2017

Teachers: Pierre De Wit, Hernan Morales, Mats Töpel

Part I of the course will be an introduction to general computing tools, such as the unix command line environment. We will go through bash commands (less, nano, ls, ll, wc, |, tail, head, mkdir, cat, grep), regular expressions, basic scripting, and running python scripts from the unix shell with a series of examples and exercises. There will also be an introduction on different types of high-throughput sequencing methods, and experimental design for these types of studies.

Part II: Genomic/Transcriptomic data analysis pipelines

Length: 2 days

Date: Between 7th-8th November 2017

Teachers: Pierre De Wit, Mats Töpel, Hernan Morales, Tomas Larsson.

Topics covered:

1) Quality control of high-throughput sequencing data

2) Genome/Transcriptome assembly

3) Alignment/Mapping

4) SNP genotyping, PCA, outlier scans

5) Differential Gene expression, functional enrichment tests

Part III: Restriction-site digested DNA analysis pipelines

Length: 2 days

Date: Between 9th-10th November 2017

Teachers: Pierre De Wit, Mikhail Matz, Mats Töpel.

Topics covered:

1) Analysis of 2b-RAD data

2) AFS-based methods for genotyping, studying demographic history.