2015 Český Krumlov Genomics Workshop - UNIX Homework 1

1. Copy this file into a working folder in your home directory:

   ~/workshop_data/`batch_2.fst_2-3.tsv.gz`

   This file contains F_{ST} values -- a measure of genetic differentiation between two populations of Stickleback fish (populations 2 and 3). Later in the course you will generate this type of data yourself.

   Column 2 of this file contains each locus examined in the fish. A locus is a 100bp long RAD-tag, so there can be multiple SNPs in each locus. We calculated F_{ST} for each SNP between the two populations. **Columns 5 and 6** contain the chromosome (a.k.a. linkage group) and base-pair position of each SNP. **Column 9** contains the F_{ST} measurement.

2. Decompress the file.

3. Answer the following questions using the UNIX commands learned in class. Each question can be answered with a single command (usually containing multiple parts), report the answer and the command you used to achieve it.

   1. How many SNPs were examined in this file?

   2. How many unique loci are in this file?

   3. How many linkage groups are in this file, what are their names, how many SNP measurements were made on each one?

   4. Which SNP has the highest base-pair position on linkage groupII? (Hint: use the `-k` option to `sort` in part of the command.)

   5. What is the lowest value of F_{ST} in the dataset? How many loci have this value?

   6. What is the most frequent value of F_{ST} in the dataset? How many loci have this value?