# Activity 4: GBS data processing with Stacks

In this exercise, you will process GBS data from a strawberry breeding population using core scripts of the Stacks bioinformatic pipeline.

As you work, you’ll find it useful to have the Stacks manual page open:

<http://creskolab.uoregon.edu/stacks/manual/>

**Step 1: Set up a working directory for Stacks. In your home directory, make a “stacks” subdirectory. This will hold output files from subsequent steps.**

**mkdir stacks**

**Step 2. Pstacks. This step takes as input Sequence Alignment Map (.sam) files, and extracts aligned “stacks” of sequence tags. In this case, Bowtie2 was used to align sequence tags to the strawberry reference genome and produce .sam files for two strawberry crossing parents and four progeny plants. Pstacks has to be run once for each of the six alignment files. If more than three stacks cannot be resolved at a locus, the locus is “blacklisted”.  
  
In the pstacks commands you see these options:   
-t indicates that the file type (.sam in this case)   
-f indicates the input file  
-o indicates the output file directory  
-i is a unique sample ID used by Stacks to track samples through the pipeline  
-m is the minimum depth of coverage to support a stack  
-p enables parallel execution on multiple threads/processors (speeds processing).**  
/local/cluster/stacks/stacks-1.29/bin/pstacks -t sam -f /nfs1/Teaching/data/viningk/GBS/FilesForStacks/sample\_Holiday\_all.sam -o ./stacks -i 1 -m 5 -p 2

**/local/cluster/stacks/stacks-1.29/bin/pstacks -t sam -f** /nfs1/Teaching/data/viningk/GBS/FilesForStacks/**sample\_Korona\_all.sam -o ./stacks -i 2 -m 5 -p 2**

**/local/cluster/stacks/stacks-1.29/bin/pstacks -t sam -f** /nfs1/Teaching/data/viningk/GBS/FilesForStacks/**sample\_GAGGA.sam -o ./stacks -i 3 -m 5 -p 2**

**/local/cluster/stacks/stacks-1.29/bin/pstacks -t sam -f** /nfs1/Teaching/data/viningk/GBS/FilesForStacks/**sample\_CGGTAGA.sam -o ./stacks -i 4 -m 5 -p 2**

**/local/cluster/stacks/stacks-1.29/bin/pstacks -t sam -f** /nfs1/Teaching/data/viningk/GBS/FilesForStacks/**sample\_CCTAC.sam -o ./stacks -i 5 -m 5 -p 2**

**/local/cluster/stacks/stacks-1.29/bin/pstacks -t sam -f** /nfs1/Teaching/data/viningk/GBS/FilesForStacks/**sample\_GAACTTC.sam -o ./stacks -i 6 -m 5 -p 2**

**Fill in the table below with your pstacks results.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Genotype** | **TotalReads** | **UniqueStacks** | **Loci** | **ExcludedLowCoverageLoci** | **KeptLoci** | **MeanCovDepth** |
| Holiday |  |  |  |  |  |  |
| Korona |  |  |  |  |  |  |
| GAGGA |  |  |  |  |  |  |
| CGGTAGA |  |  |  |  |  |  |
| CCTAC |  |  |  |  |  |  |
| GAACTTC |  |  |  |  |  |  |

Notice that three output files are generated for each sample: \*tags.tsv, \*.snps.tsv, and \*.alleles.tsv. Refer to today’s Powerpoint presentation, or Stacks manual section 6.1, to see what information is in each column in each of these files.

**Step3. Cstacks. This step creates a catalog of SNPs (and small insertion/deletion polymorphisms). It uses pstacks output files as input. We expect that all polymorphisms in the progeny came from one of the two parents, and therefore we can build the catalog from the two parents.**

**In the cstacks command, you see these options:  
-b an arbitrary batch number for the catalog used by downstream steps  
-p enables parallel execution on multiple threads/processors (speeds processing).  
-n number of mismatches allowed between stacks   
-s designates each sample used to construct the catalog.**

**Notice that in the cstacks command, the file extensions are not included for the samples. Stacks will use all three files (\*tags.tsv, \*.snps.tsv, and \*.alleles.tsv)** **for each sample.**

/local/cluster/stacks/stacks-1.29/bin/cstacks -b 1 -p 2 -n 3 -s stacks/sample\_Holiday\_all -s stacks/sample\_Korona\_all -o ./stacks

**Step4. Sstacks. This step matches individual samples in the population back to the catalog. It outputs a set of matching loci in a \*.matches file. This**

**In the sstacks commands, you see the these options:  
-b flag assigns a batch number to each individual sample   
-c indicates the catalog  
-s indicates the sample  
-o indicates the output directory**

/local/cluster/stacks/stacks-1.29/bin/sstacks -b 1 -c stacks/batch\_1 -s ./stacks/sample\_Holiday\_all -o ./stacks

/local/cluster/stacks/stacks-1.29/bin/sstacks -b 3 -c stacks/batch\_1 -s ./stacks/sample\_GAGGA -o ./stacks

/local/cluster/stacks/stacks-1.29/bin/sstacks -b 4 -c stacks/batch\_1 -s ./stacks/sample\_CGGTAGA -o ./stacks

/local/cluster/stacks/stacks-1.29/bin/sstacks -b 5 -c stacks/batch\_1 -s ./stacks/sample\_CCTAC -o ./stacks

/local/cluster/stacks/stacks-1.29/bin/sstacks -b 6 -c stacks/batch\_1 -s ./stacks/sample\_GAACTTC -o ./stacks

**Fill in the table below with your sstacks results.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Genotype** | **Catalog matches** | **Matched multiple loci** | **SNPs not in catalog, excluded** | **Total haplotypes** |
| Holiday |  |  |  |  |
| Korona |  |  |  |  |
| GAGGA |  |  |  |  |
| CGGTAGA |  |  |  |  |
| CCTAC |  |  |  |  |
| GAACTTC |  |  |  |  |