# IESHTSTE

### ***Intructions for extracting sequences from HTS target enrichment reads***

## Data

This batch was generated by [RAPiD Genomics](www.rapid-genomics.com).

Within the data directory there is a SampleSheet **csv** file with the barcodes, filenames, and sample codes.

***Note that Plates1-4 were sequenced on two lanes (L001 and L002), so there are two sets of fastq files per sample.***

**Raw Data:**

* This data has been demultiplexed using Illuminas BCLtofastq. No quality trimming or processing has been done beyond demutiplexing.
* The adapters used are below, *“BCBCBCBC”* stands for the barcodes.
  + i7: GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-BCBCBCBC-ATCTCGTATGCCGTCTTCTGCTTG
  + i5: AATGATACGGCGACCACCGAGATCTACAC-BCBCBCBC-ACACTCTTTCCCTACACGACGCTCTTCCGATCT

## Assembly methods

Currently, Three ways you can analyze high-throughput sequencing reads using target enrichment:

1. HybPiper

[Publication](https://bsapubs.onlinelibrary.wiley.com/doi/full/10.3732/apps.1600016)

[Code in github](https://github.com/mossmatters/HybPiper)

1. aTRAM

[Publication](https://journals.sagepub.com/doi/10.1177/1176934318774546)

[Code in github](https://github.com/moskalenko/aTRAM)

1. SECAPR

[Publication](https://peerj.com/articles/5175/)

[Code in github](https://github.com/AntonelliLab/seqcap_processor)

### HybPiper

*Credit to Andre A. Naranjo in Soltis Lab*  
*Miao modified and add in Shell cmd and bash scripts*

**Steps for Hybpiper:**

1. Concatenate all lanes (L001 and L002; only if you have them on separate plates!)  
   e.g.  
   cat RAPiD-Genomics\_F076\_UFL\_###\_P003\_WD02\_i5-503\_i7-72\_S22\_L001\_R1\_001.fastq.gz RAPiD-Genomics\_F076\_UFL\_###\_P003\_WD02\_i5-503\_i7-72\_S60\_L002\_R1\_001.fastq.gz > P003\_WD02\_72\_R1.fastq.gz

or run in a batch manner:   
  
`bash fastq\_lane\_cat.sh sample\_ID\_file Seq\_ID\_table`   
  
e.g.,   
  
`bash fastq\_lane\_cat.sh Evgeny\_13.txt UFL\_394803\_SampleSheet.csv`

This bash script will take two input files: one is sample ID file, and the other is sequence ID table.

e.g.,   
 `[cactus]$ head XXX\_88.txt  
 CPG00213  
 CPG00216  
 CPG05361  
 CPG05783  
 CPG07101  
 CPG10128  
 CPG11009  
 CPG11189  
 CPG11230   
   
 [cactus]$ head -6 UFL\_XXX\_SampleSheet\_XXX86.csv  
 RG\_Sample\_Code,Customer\_Code,i5\_Barcode\_Seq,i7\_Barcode\_Seq,Sequence\_Name,Sequencing\_Cycle  
 UFL\_394803\_P002\_WG08,D\_4566,TAAGATTA,TTCACGCA,RAPiD-Genomics\_F076\_UFL\_394803\_P002\_WG08\_i5-506\_i7-68\_S171\_L001\_R1\_001.fastq.gz,2x150

UFL\_394803\_P002\_WG09,D\_4567,TAAGATTA,CGACTGGA,RAPiD-Genomics\_F076\_UFL\_394803\_P002\_WG09\_i5-506\_i7-41\_S172\_L001\_R1\_001.fastq.gz,2x150 UFL\_394803\_P002\_WG10,D\_4568,TAAGATTA,CCGAAGTA,RAPiD-Genomics\_F076\_UFL\_394803\_P002\_WG10\_i5-506\_i7-37\_S173\_L001\_R1\_001.fastq.gz,2x150 UFL\_394803\_P002\_WG11,D\_4570,TAAGATTA,GCCAAGAC,RAPiD-Genomics\_F076\_UFL\_394803\_P002\_WG11\_i5-506\_i7-96\_S174\_L001\_R1\_001.fastq.gz,2x150 UFL\_394803\_P002\_WG12,D\_4571,TAAGATTA,CGCATACA,RAPiD-Genomics\_F076\_UFL\_394803\_P002\_WG12\_i5-506\_i7-42\_S175\_L001\_R1\_001.fastq.gz,2x150

`

1. \***fastqc** to quick check the quality; and later on can be used for comparison after trim and clean.

* scripts needed:  
  fastqc.sh check\_result.sh mean.R

e.g.,  
module load ufrc fastqc  
srundev -t time  
fastqc \*.gz -o FastQC\_result

For slurm job scripts see:  
**fastqc.sbatch**  
+ after runing *fastqc.sh*, it will put fastqc results in to folder called *FastQC\_result* ; + Move scripts *check\_result.sh*, and *mean.R*, into *FastQC\_result*, then excute the bash script, it will generate a summary table *Illumina\_FastQC\_report.csv* for reads quality. Details see folder *unzip\_file*.

1. Trim and clean reads using Trimmomatic, and preapre for next step — Hybpiper.

* scripts needed:  
  Trimmomatic.sbatch (if you have a few sample you can just go with bash Trimmomatic.sh) For large number of samples, submission to SLURM in HPC is required.
* run: bash Trimmomatic.sh

1. run hybpiper

6)if want introns run intron script on accession folders out putted from previous step 7) to retrieve the supercontig sequences from the above run put them all in one place (so mv P*W* seq\_dir): > module load python > python HybPiper/retrieve\_sequences.py baits1.fasta seq\_dir dna (just exons use DNA, if you run intronerate use supercontig)

1. run mafft script on individual gene. mv aligned to desktop. 8a) rename using rscript to make sed command and sed.sh. Move to geneioius 8b) strip 90% gaps BY GENE and then concatenate. Made a workflow in geneious to do this.
2. concat alignments (load into geneious to look at). easy enough to do in geneious. export as phy.
3. run raxml
4. win superhard