iPlant Discovery Environment

THE MOST IMPORTANT THING OF ALL!

Read Cleanup

iPlant Collaborative™

Not Sexy, but….

- Cleanup of sequencing reads is critical to the quality of your work
- Genome Assembly
- Transcriptome Assembly
- RNASeq
- Everything: junk in, junk out

iPlant Collaborative™

The Easiest, Most Efficient Way to Do Cleanup

- Use the HTProcess Pipeline
- What is the HTProcess Pipeline?
  - A set of Apps in the DE designed to run multiple files simultaneously as a *read library*
  - Can be used with RNA-Seq data or with genome sequencing data
  - Ideally a library is created for each set of reads that share all main sequencing characters, e.g. read length, fastq quality format, read pair spacing, even source tissue or condition for creation

iPlant Collaborative™

HTProcess Apps in the DE

- HTProcess Tuxedo pipeline
  - HTProcess-prepare_directories_and_run_fastqc
  - HTProcess_trimmomatic
  - HTProcess_Tophat-2
  - HTProcess_Cufflinks
  - HTProcess_Cuffmerge
  - HTProcess_CuffDiff
- Other HTProcess apps
  - HTProcess_Kmergenie
  - HTProcess_BAMstats-1.0

iPlant Collaborative™

Are HTProcess Apps Only for RNA-Seq?

- No
- HTProcess Apps for Genome Assembly, too.
  - HTProcess-prepare_directories_and_run_fastqc
  - HTProcess_trimmomatic
  - HTProcess_Tophat-2 (reference-based assembly)
  - HTProcess_Kmergenie (small genomes)

iPlant Collaborative™

How Do HTProcess Apps Work?

- It's a pipeline
- You must start with the first app, and then use the output folder as the input for the next app.
- Much of the important data about your files is kept for you in the manifest file.
- A log is created to help keep track of what has been done as part of the analysis.

iPlant Collaborative™

Back to Read Cleanup…

- HTProcess-prepare_directories_and_run_fastqc and HTProcess_trimmomatic make up a good general read preparation pipeline
- The steps HTPPDRF runs:
  - Produces a manifest file to contain the important parameters for downstream apps and help keep records
  - Runs FastQC on each read file
  - Produces a single report that you can click on and open directly in the DE.
  - Gets your data ready for HTProcess_trimmomatic

iPlant Collaborative™

HTProcess-prepare_directories_run_fastqc

1. Starts by importing your left reads and your right reads and single-ended reads separately – so the pairing can be understood by all the downstream apps.
2. Gives you a list of boxes to fill in to provide important information on your reads.

iPlant Collaborative™

HTProcess-prepare_directories_run_fastqc

- Do I really need to fill all those boxes?
- Fill in all of them if you know what's good for you (or your analysis). Most of the information you will need at some point, so best practice is to get it in the beginning when you just received your read files.
- You can make up anything you want for a library name and number, but it is a way to identify your library.
- "Condition" is what will define your samples for RNA-Seq analysis. If you don't keep them separate, how are the apps to know which is which? Actually, for the Tuxedo workflow you can put all your reads in one, big library directory, but you will need to define which reads belong with what condition later for the CuffDiff step.

iPlant Collaborative™

Unix/Linux… the command line

Minicomputer?

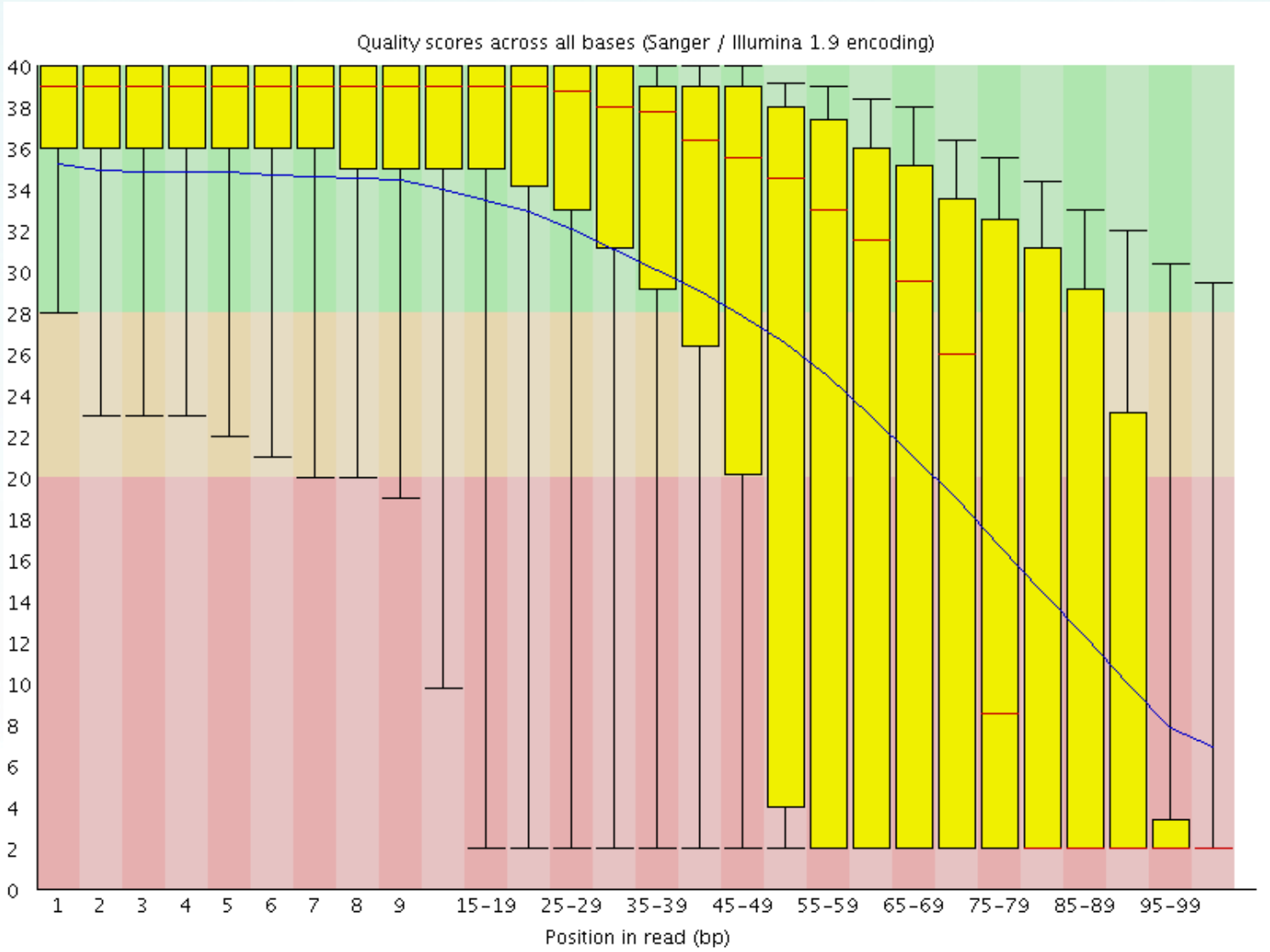
A space=
end of entry

FastQC

- Does a quick analysis of a read file
- Provides graphs of your reads in terms of quality, repeats, potential contamination
- Gives Pass/Fail evaluations of your reads for different criteria

iPlant Collaborative™

FastQC



iPlant Collaborative™

FastQC
Summary

- PASS Basic Statistics frag_1.fastq
- FAIL Per base sequence quality frag_1.fastq
- PASS Per sequence quality scores frag_1.fastq
- FAIL Per base sequence content frag_1.fastq
- WARN Per base GC content frag_1.fastq
- PASS Per sequence GC content frag_1.fastq
- PASS Per base N content frag_1.fastq
- PASS Sequence Length Distribution frag_1.fastq
- PASS Sequence Duplication Levels frag_1.fastq
- WARN Overrepresented sequencesfrag_1.fastq
- WARN Kmer Content frag_1.fastq

iPlant Collaborative™

HTProcess_trimmomatic

- Drag the HTProcess_Reads directory created by HTPPDRF into the input
- Go through and set all the trimming functions needed, and run!
- Many different trimmers

iPlant Collaborative™

HTProcess_trimmomatic

- Headcrop – trim n basepairs from 5'prime end
- Illuminaclip – find matches to a list of adapter and primer sequences, and excise.
- Leading – clip at the 5'prime end where the quality drops below a threshold
- Trailing – clip at the 3'prime end where the quality drops below a threshold
- Sliding Window – clips where the average quality falls below a threshold
- Max Info – set a preferred average length for clipped reads and a number to indicate the importance of sequence quality (strictness)
- Crop – trim the read to leave a specific length read
- Set a minimum read length (reject ones that are trimmed shorter)

iPlant Collaborative™

General Conclusions

- Don’t skip boxes
- Don’t use spaces
- Read the trimmomatic documentation for a deeper understanding
- If you don’t like the results, do it again with new settings
- Yes, there are other tools to use for trimming a file at a time (e.g. Scythe, Sickle)
- Don’t use a lot of quality trimming if you are going to an error correction program (e.g. AllpathsLG)

iPlant Collaborative™