Cold Spring Harbor Laboratory

Advanced Sequencing Technologies & Applications

http://meetings.cshl.edu/courses.html

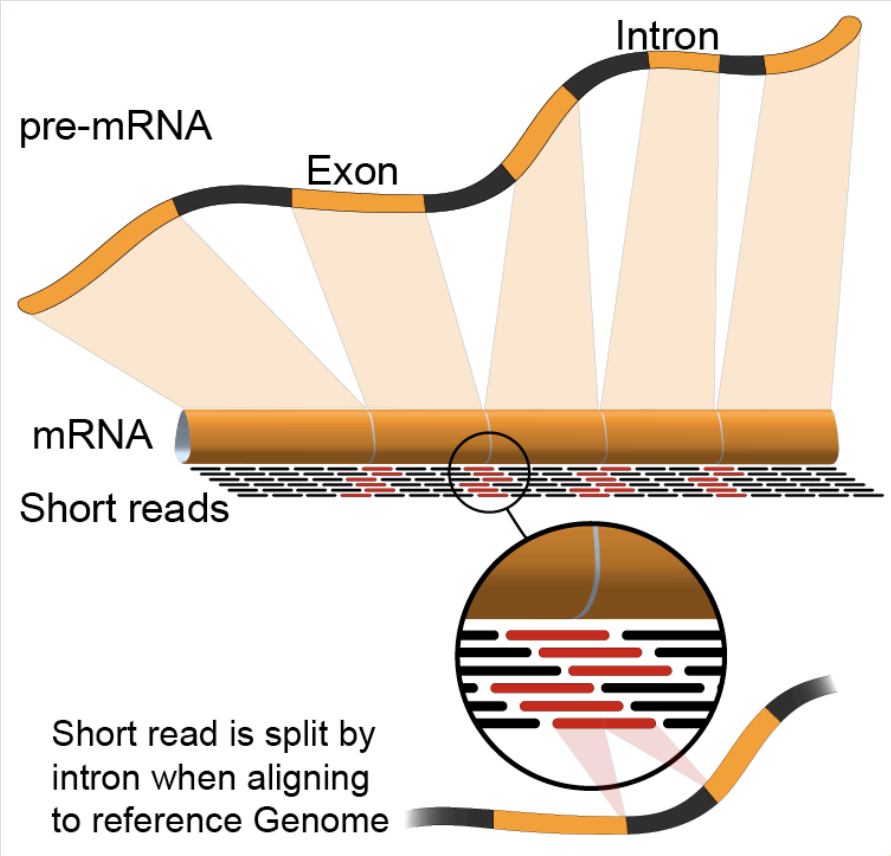
Cold Spring Harbor Laboratory

RNA-Seq Module 1
Introduction to RNA sequencing (tutorial)

Malachi Griffith, Obi Griffith, Jason Walker, Alex Wagner

Advanced Sequencing Technologies & Applications

November 7 - 18, 2017



THE GENOME INSTITUTE at Washington University

Learning Objectives of Tutorial

- Install commonly used RNA-seq tools (Samtools, bam-readcount, HISAT2, StringTie, gffcompare, htseq-count, FastQC, picard-tools, Flexbar, R, Bioconductor, Ballgown, edgeR, …)
- Obtain a reference genome
- Obtain gene/transcript annotations
  - Understand GTF file format
- Index reference genome files for use with aligners
- Obtain and explore raw sequence data
  - Understand fasta/fastq format

The most common problems encountered while working on the tutorials

- Type short commands carefully if you like, but in order to get through all the steps smoothly, it is safer to copy and paste from the tutorial files
- Copy/Paste errors
  - Learn the short cuts for copying/pasting on your system and use them (e.g. <command><c> & <command><v> on Mac)
  - Make sure you copy the entire command. Watch out for commands that span across multiple lines
- Being in the wrong directory at the wrong time
  - The simplest way to avoid this is only change directories as instructed
  - If you do change directories to look around, make sure you go back before continuing with commands
- Not having the $RNA_HOME environment variable set
  - Make sure you check this when logging in:
    - echo $RNA_HOME
  - If it is not defined do this:
    - export RNA_HOME=~/workspace/rnaseq
  - Then add this to you .bashrc file so that you don't have to worry about it again

Introduction

- This presentation provides a brief description of tutorial steps
- The wiki contains more complete instructions
- Lines beginning with "#" are comments
- All other lines are commands that will be pasted and executed from a linux terminal or R tutorial
- Each command is annotated with comments except that basic familiarity with linux is assumed
  - e.g. You should know that 'mkdir' means to 'make a directory, 'cd' means to 'change directory', etc.
- Some reference materials for linux can be found here:
  - *http://files.fosswire.com/2007/08/fwunixref.pdf*
  - *http://vic.gedris.org/Manual-ShellIntro/1.2/ShellIntro.pdf*
  - *www.nettech.in/course/Basic%20Commands.pdf*

1-i. Installation

- Installation instructions are provided for:
  - Samtools
    - http://www.htslib.org/download/
  - bam-readcount
    - https://github.com/genome/bam-readcount
  - HISAT2
    - https://ccb.jhu.edu/software/hisat2/index.shtml
  - StringTie
    - https://ccb.jhu.edu/software/stringtie/
  - Gffcompare
    - http://ccb.jhu.edu/software/stringtie/gff.shtml
  - htseq-count
    - https://pypi.python.org/pypi/HTSeq
  - FastQC
    - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - picard-tools
    - https://github.com/broadinstitute/picard
  - Flexbar
    - https://github.com/seqan/flexbar
  - R
    - https://www.r-project.org/
  - Bioconductor
    - https://www.bioconductor.org/install/
  - Ballgown (Bioconductor package)
    - http://bioconductor.org/packages/release/bioc/html/ballgown.html
  - edgeR (Bioconductor package)
    - https://bioconductor.org/packages/release/bioc/html/edgeR.html

1-ii. Obtain reference genome

- All reference files are obtained from Ensembl
  - ftp://ftp.ensembl.org/pub/release-86/fasta/homo_sapiens/dna/
  - This step downloads reference human genome files from Ensembl
  - The GRCh38 build of the human genome is used
    - This is the latest version of the human reference
- For the tutorial, a single chromosome is used (chr. 22)
  - The reason for this is to reduce run time for the tutorial
  - Instructions for downloading all chromosomes are provided

1-iii. Obtain known transcript annotations

- All annotation files are obtained from Ensembl
  - http://useast.ensembl.org/info/data/ftp/index.html
  - There are many other ways to obtain gene annotation files. For example:
  - UCSC Genome Browser, Ensembl API, BioMart, Entrez, Galaxy, etc. could also be used
- You will download GTF files describing human transcripts (exon coordinates, gene ids, gene symbols, etc.)
- Descriptions of the GTF file format can be found here:
  - http://genome.ucsc.edu/FAQ/FAQformat.html#format4

1-iv. Create Indexed reference genome

- Before sequences can be mapped to the genome, it must be ‘indexed’ in a way that is compatible with the aligner being used
  - Since we are using HISAT2 for alignment, we will need an index built for that purpose
  - Other RNA-seq aligners will have their own indexing utility
    - E.g. TopHat and STAR.
    - Do not use an index created for another aligner

1-v. Obtain RNA-seq data

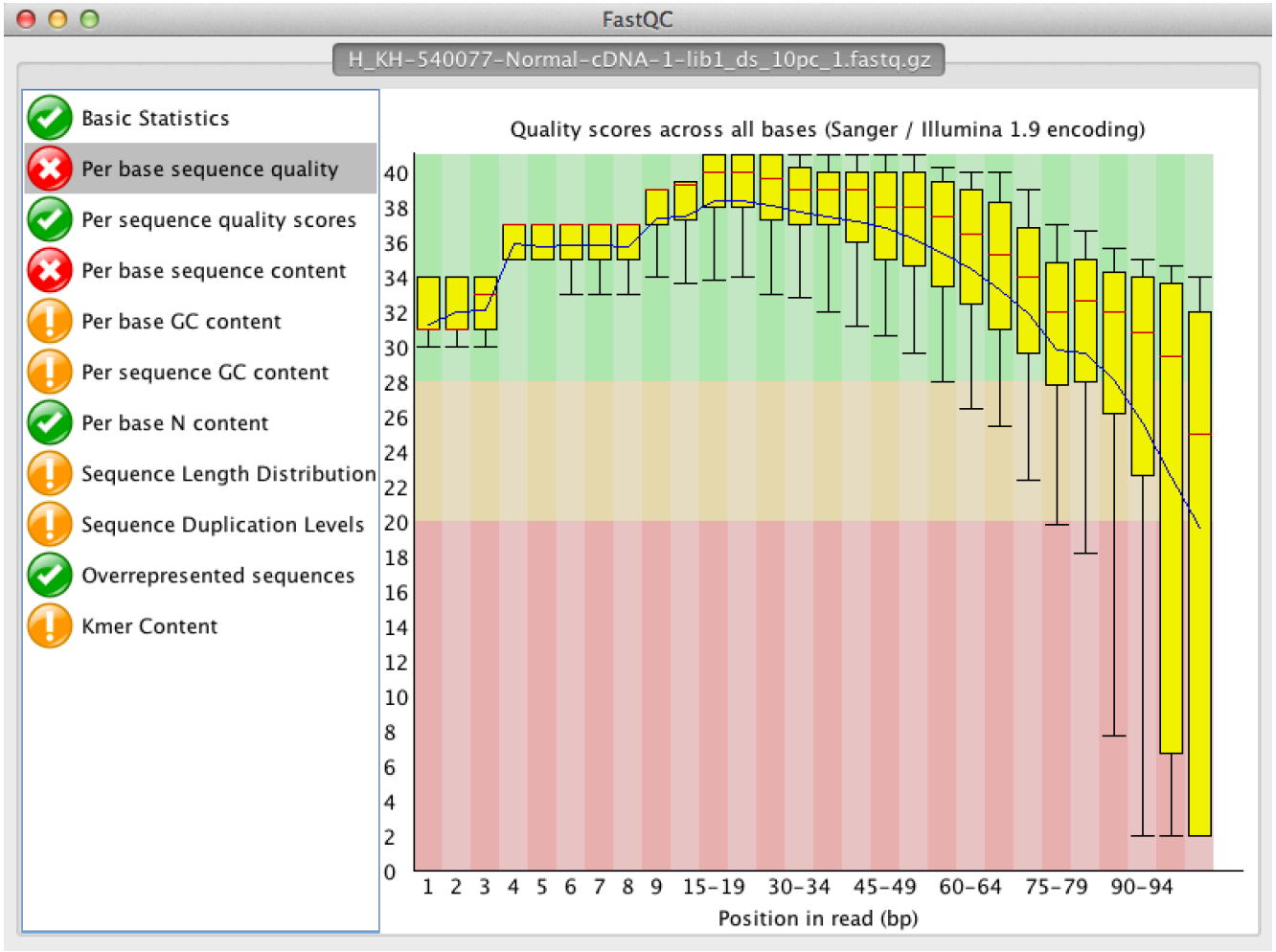
- For purposes of the tutorial, the test data has been pre-filtered
  - Identified reads that appear to match transcripts on a single chromosome
- The test data corresponds to two RNA sources
  - The Universal Human Reference (UHR) and Human Brain Reference (HBR)
  - Each sample also included one of two ERCC RNA "spike-in" mixes (Mix1 or Mix2)
  - Each RNA was source was sequenced in triplicate to create six independent Illumina sequence libraries (**'UHR_Rep1_Mix1', 'UHR_Rep2_Mix1', 'UHR_Rep3_Mix1', 'HBR_Rep1_Mix2', 'HBR_Rep2_Mix2', and 'HBR_Rep3_Mix2'**)
- The input data is provided in 'fastq' format:
  - http://en.wikipedia.org/wiki/FASTQ_format

1-v. Obtain RNA-seq data (cont'd)

- Universal Human Reference (UHR):
  - A pool of 10 human cell lines. This sample was purchased from Strategene (Agilent Technologies)
  - http://www.genomics.agilent.com/en/References-Controls/Universal-Reference-RNAs/?cid=AG-PT-172&tabId=AG-PR-1217
- Human Brain Reference (HBR):
  - A pool of brain tissue from multiple brain regions from multiple human donors. This sample was purchased from Ambion (Life Technologies).
  - http://www.lifetechnologies.com/order/catalog/product/AM6050
- External RNA Reference Consortium (ERCC):
  - ERCC reference RNA spike-ins purchased from Ambion (Life Technologies).
  - http://www.lifetechnologies.com/order/catalog/product/4456739
  - The UHR samples used ERCC Mix1. The HBR samples used ERCC Mix2.

- In this tutorial we will compare the three UHR libraries vs three HBR libraries (6 samples in total)

1-vi. Pre-Alignment QC with FastQC



We are on a Coffee Break & Networking Session