



Cold  
Spring  
Harbor  
Laboratory

# Advanced Sequencing Technologies & Applications

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

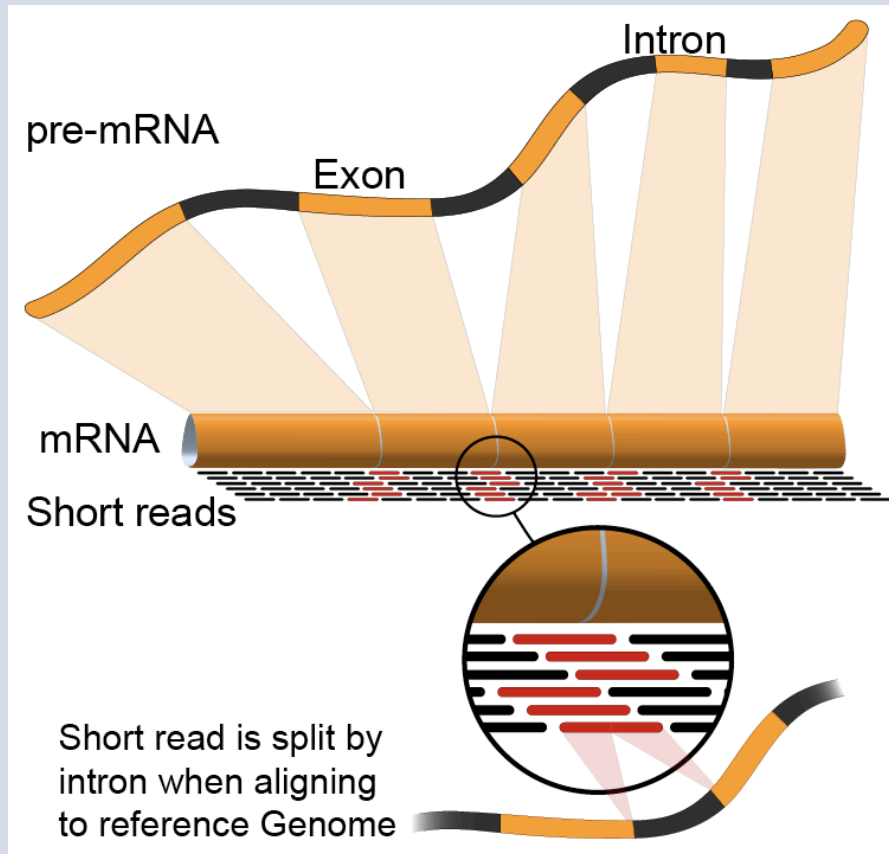


Cold  
Spring  
Harbor  
Laboratory

# Module 1

## Introduction to RNA sequencing (tutorial)

Malachi Griffith, Obi Griffith, Jason Walker, Ben Ainscough  
Advanced Sequencing Technologies & Applications  
November 11-23, 2014



# Learning Objectives of Tutorial 1

- Install commonly used RNA-seq tools (Samtools, Bowtie, Tophat, STAR, Cufflinks, R, CummeRbund, FastQC, picard-tools, SamStat)
- 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 your `.bashrc` file so that you don't have to worry about it again

# Tutorial files

- One part
  - **Tutorial\_Module1\_Linux.txt**
    - Introduction to Linux and basic commands
    - Install necessary tools
    - Obtain the human reference genome and create an index
    - Obtain and explore gene annotations
    - Obtain and explore RNA-seq data

# Introduction

- This presentation provides a brief description of tutorial steps
- The tutorial file contains more complete instructions
- Lines in the tutorial file 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](http://www.nettech.in/course/Basic%20Commands.pdf)

# 1. Installation

- Installation instructions are provided for :
  - Samtools
    - <http://samtools.sourceforge.net/>
  - Bam-readcount
    - <https://github.com/genome/bam-readcount>
  - Bowtie
    - <http://bowtie-bio.sourceforge.net/>
  - Tophat
    - <http://tophat.cbcb.umd.edu/>
  - STAR
    - <http://code.google.com/p/rna-star/>
  - Cufflinks
    - <http://tophat.cbcb.umd.edu/>
  - Htseq-count
    - <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>
  - R/Bioconductor/CummeRbund/edgeR
    - <http://cran.r-project.org/>
    - <http://www.bioconductor.org/>
    - <http://compbio.mit.edu/cummeRbund/>
    - <http://www.bioconductor.org/packages/release/bioc/html/edgeR.html>
  - Samstat
    - <http://samstat.sourceforge.net/>
  - FastQC
    - <https://sites.google.com/a/brown.edu/bioinformatics-in-biomed/fastqc>
  - PicardTools
    - <http://picard.sourceforge.net/>

## 2. Obtain reference genome

- All reference files are obtained from the Illumina iGenomes project
  - <http://cufflinks.cbcb.umd.edu/igenomes.html>
- This step downloads reference human genome files from iGenomes
  - The GRCh37 (hg19) build of the human genome is used
- For the tutorial, a single chromosome is used
  - The reason for this is to reduce run time for the tutorial
  - Instructions for downloading all chromosomes are provided



### 3. Obtain known transcript annotations

- All annotation files are obtained from the Illumina iGenomes project
  - <http://cufflinks.cbc.umd.edu/igenomes.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>

## 4. 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
  - Bowtie is used to index the genome for Tophat alignments
  - We will also optionally try the STAR aligner which requires its own indexed version of the genome

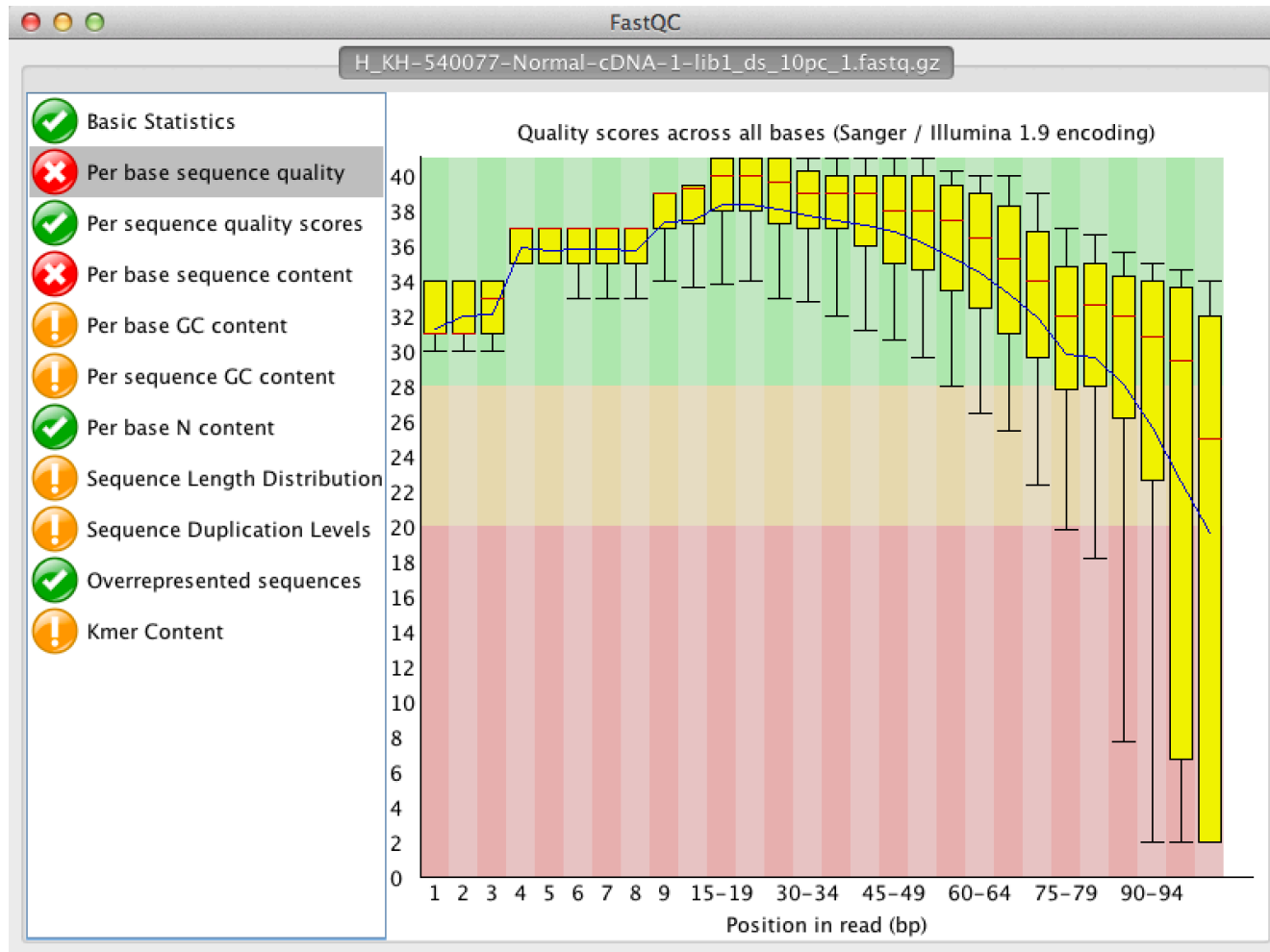
# 5. Obtain RNA-seq data

- 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 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](http://en.wikipedia.org/wiki/FASTQ_format)

# 5. 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)

# Pre-Alignment QC with FastQC



Break