Introduction to NGS Data

Matt Gitzendanner magitz@ufl.edu

Oleksandr Moskalenko om@hpc.ufl.edu

Generalized data path

- Illumina data from ICBR
  - Copy data off ext3 USB Drive
  - Concatenate qseq files and convert to fastq
  - Quality assessment
  - Quality filter
  - Ready for analysis

Managing ext3 USB drives

- Easiest to use Linux, either real or virtual
- On MacOS: MacFuse
  - Though not an easy fix
- Windows
  - ???

- We are working with ICBR to offer direct data deposit into your space at HPC
  - Ask ICBR for this if you would like it

For Illumina

- The sequence data are in:
- 110107_HWUSI-EAS163FR_00008_FC_D_B_S_M/
  Data/
  Intensities/
  BaseCalls/
  s_*.qseq.txt

qseq to fastq

- There are many qseq files per lane
- Need to concatenate and convert to fastq
- Many scripts online to do this
  - Be careful, some will convert quality encoding too

With CASAVA 1.8, this step is a thing of the past

  - De-multiplexing also needed if indexed libraries were used
    - Also handled by Casava 1.8

qseq to fastq

```
cat s_3_1_0???_qseq.txt |
 perl ~/scripts/qseq2fastq.1.3quals.pl
 > all.1.3quals.fq
```

- **cat** is a convenient command to combine a bunch of files
- **?** allows wildcard in names
- **|** sends the output of one command to the input of another
- **>** writes the output of one command to a file

Quality Assessment

- Poor quality reads lead to problems in assembly and even mapping
  - Need to remove adapters
  - Trim low quality sequence
  - Remove identical sequences
    - PCR duplicates
- Quality assessment
  - FastQC
    - Galaxy and CLI

➢ Take care with:
- Quality encoding
- Paired-end data

Quality Filter

- Sickle
  - Especially for filtering paired-reads
    - Creates a file of single reads for pairs where one read is discarded and the other is good

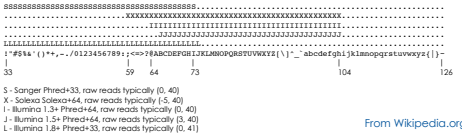
```
FastQ paired records kept: 3584484 (1792242 pairs)
FastQ single records kept: 216860 (from PE1: 25453,
from PE2: 191407)
FastQ paired records discarded: 44980 (22490 pairs)
FastQ single records discarded: 216860 (from PE1:
191407, from PE2: 25453)
```

- Fastxtoolkit
  - Has many great utilities
  - Does not handle filtering of paired-reads

Compressed data formats

- Gzip
  - A standard file compression tool
  - Files end in .gz
  - Many NGS applications can natively process gzipped data files
    - Use when possible: reduces storage needs, but also disk and network I/O in analyzing your data

Things to watch out for



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

From Wikipedia.org

- Quality score of 64, a valid score in Sanger and Illumina, is @
- Fastq record delimiter is @

```
@Sequence name          @Sequence name
ACTGAGCTAGCATCGATCAG    ACTGAGCTAGCATCGATCAG
+                       +
EEEEEEEEEEEEEEEEEEEE    @ABCDEEEEEEEEEEEEEE
```

- BioPerl has problems—don't use with fastq

Things to watch out for

- Converting quality formats unintentionally
- Keep track of input/output formats of data manipulation tools
  - Some will convert for you
  - Can be helpful, but need to know what tool is doing
- Don't need to transfer/keep all data provided by ICBR to HPC
  - Definitely keep the data
  - Can reprocess raw data or view some run stats
  - Rarely used though
  - Full folder from 1 lane: 54G
  - BaseCalls folder from 1 lane: 17GB

Training Schedule

- ✓ Aug 28: Intro to UFHPC, getting started
- ✓ Sept 10: Modules, RHEL6 Transition, User Q&A
- ✓ Sept 17: The Linux/Unix Shell - An Introduction
- ✓ Sept 24: Running Jobs, Submission Scripts, Modules
- ✓ Oct 1: Galaxy Overview, The Basics
- ✓ Oct 8: NGS Data Techniques: General Methods and Tools
- Oct 15: NGS Data Techniques: Reference Based Mapping and de Novo Assembly
- Oct 22: Phylogenetic Analyses
- ◆ Oct 29: Research Computing Day: Moving Big Data
- Nov 5: Multiprocessing at the HPC Center
- Nov 12: Using Git and CMake to Organize and Drive Data Analysis Pipelines
- Nov 19: Introduction to GPU Nodes
- Nov 29: NGS Data Techniques: RNA-Seq
- Dec 3: NGS Data Techniques: Alternative Splicing Analysis

UF Research Computing

- Help and Support (Continued)
  - http://wiki.hpc.ufl.edu
    - Documents on hardware and software resources
    - Various user guides
    - Many sample submission scripts
  - http://hpc.ufl.edu/support
    - Frequently Asked Questions
    - Account set up and maintenance