Genomic Resources

Michael Schatz

Aug 29, 2012

QB Bootcamp Lecture 3

Outline

Part 1: Overview & Fundamentals

Part 2: Sequence Analysis Theory

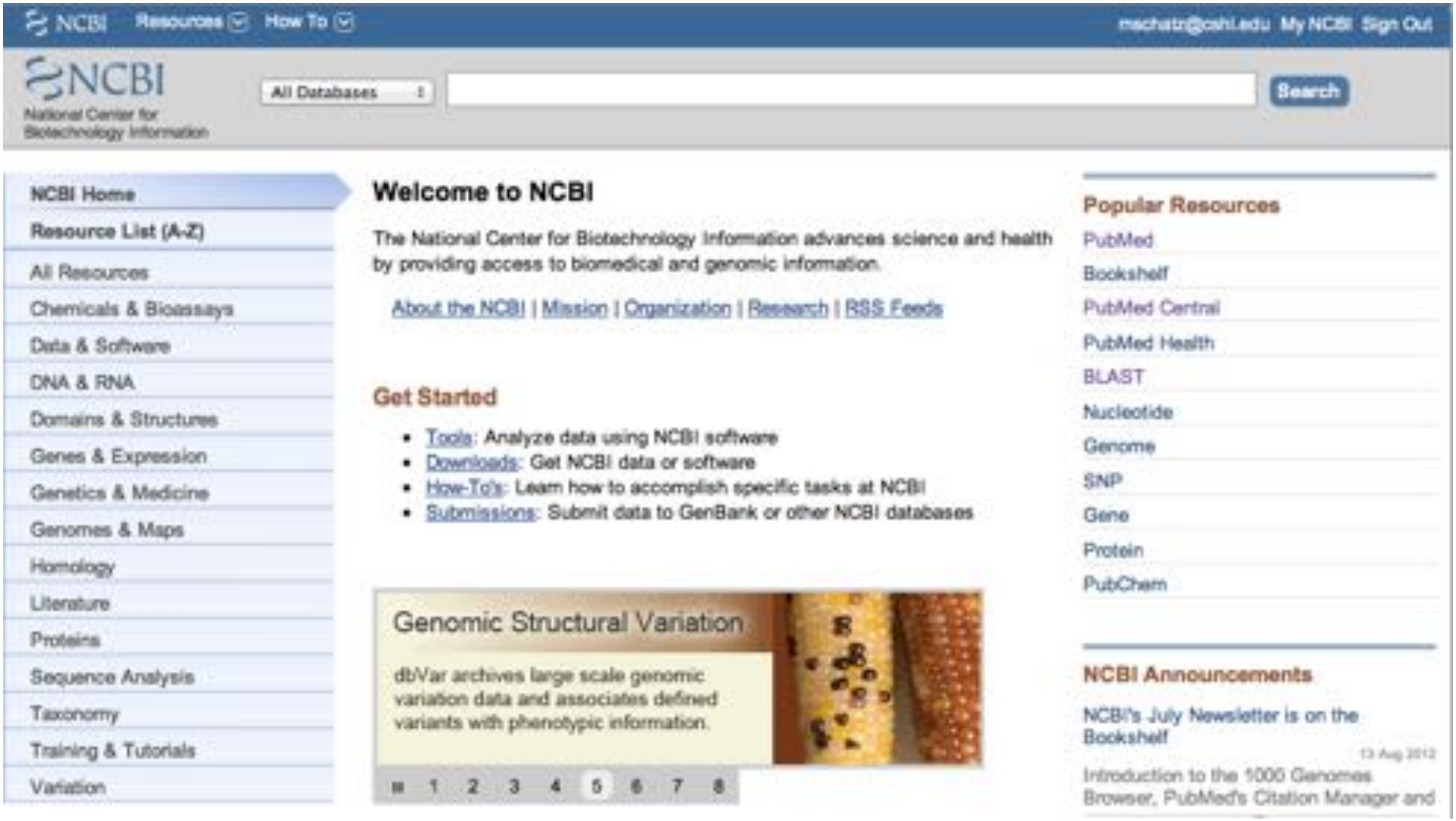
Part 3: Genome Resources

- Public: NCBI, UCSC
- CSHL: Intranet, Meetings, Galaxy

Part 4: Example Analysis

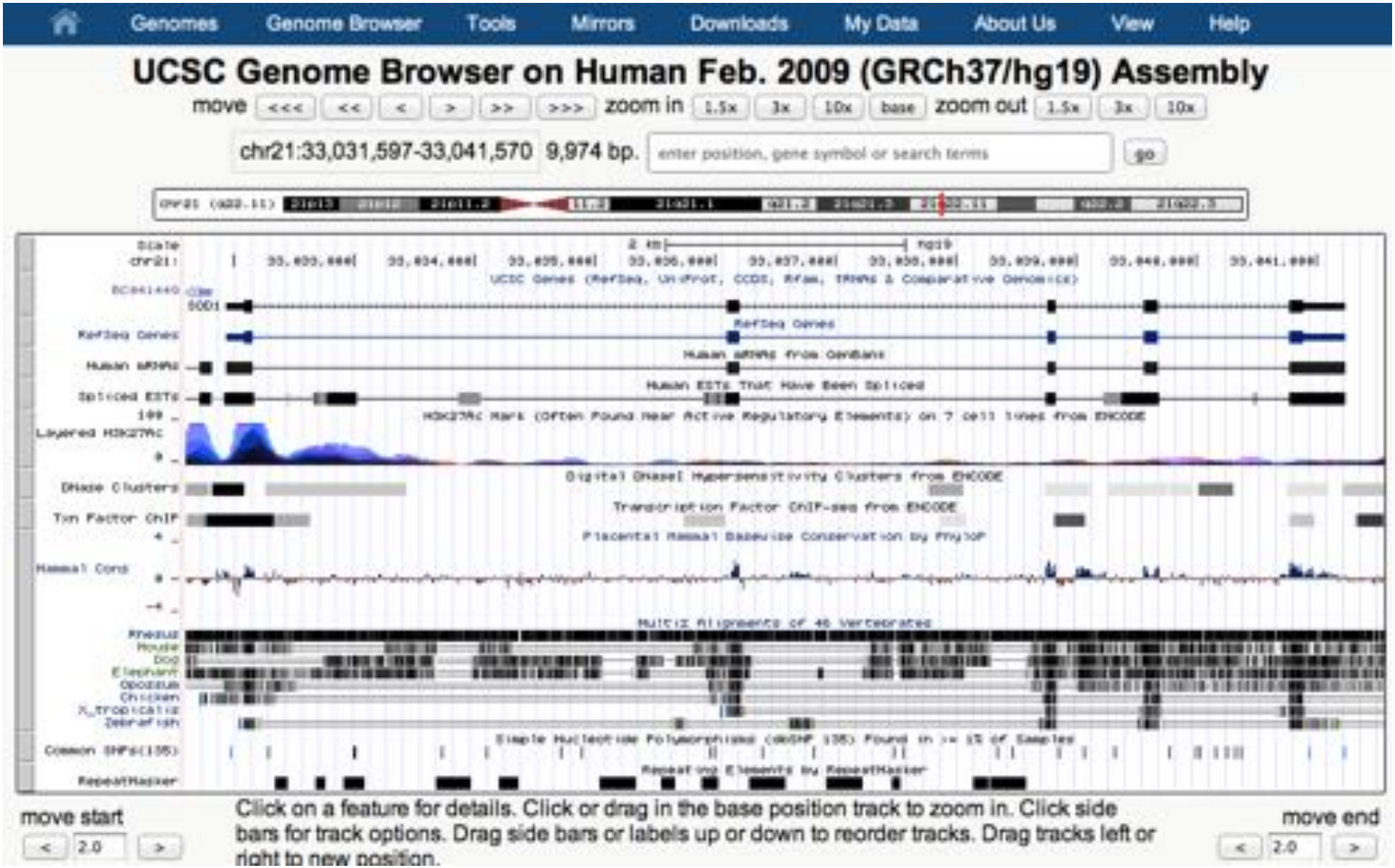
NCBI

http://www.ncbi.nlm.nih.gov/



UCSC Genome Browser

http://genome.ucsc.edu/



Intranet

http://intranet.cshl.edu/it/bluehelix/

COLD SPRING HARBOR LABORATORY

Intranet

Tuesday, August 28, 2012

Search FACES

Home | Most Popular | Administration | Education | General Info | Lab Websites | Science | Services

CSHL Information Technology

Home | About IT | Applications | BlueHelix | FAQ | Helpdesk | Stats

BlueHelix (HPCC)

- BlueHelix Home
- HPCC Talks
- Architecture
- Applications
- Databases
- Request an Account
- Login & File Transfer
- Using BlueHelix
- Dev Node Status
- SGE Job Status
- Maintenance Downtime
- Comments or Questions

BlueHelix, the High Performance Computing Cluster (HPCC) is installed at the main IT DataCenter in the Hilltop complex. The system is an IBM E1350 blade based cluster running Linux. It is comprised of 10 racks with a total of 504 compute nodes, 2 management servers, and a number of storage servers. The system has a Gigabit Ethernet communication network.

- Each of the 504 compute nodes contains 2 dual core Opteron 64-bit processors running at 2.0GHz (2016 cores total), and a local 73GB SFF SCSI disk. There are 56 nodes with 2GB of memory, 364 nodes with 4GB of memory, and 84 nodes with 8GB of memory.

- Six of the nodes are development nodes, used for editing, compiling, and debugging user codes, jobs, and so on.

Conferences and Journals

CSHL Yearly Conferences

Biology of Genomes	May	Latest advances in biology, genomics, and medicine
Symposium	May/June	Latest advances with yearly themes
Genome Informatics	Sept/Nov	Computational Biology
Personal Genomes	Sept/Nov	Computational Biology
In-house Symposium	Nov	Updates from the faculty (Just before Thanksgiving)

You are welcome to attend all meetings at CSHL free of charge:

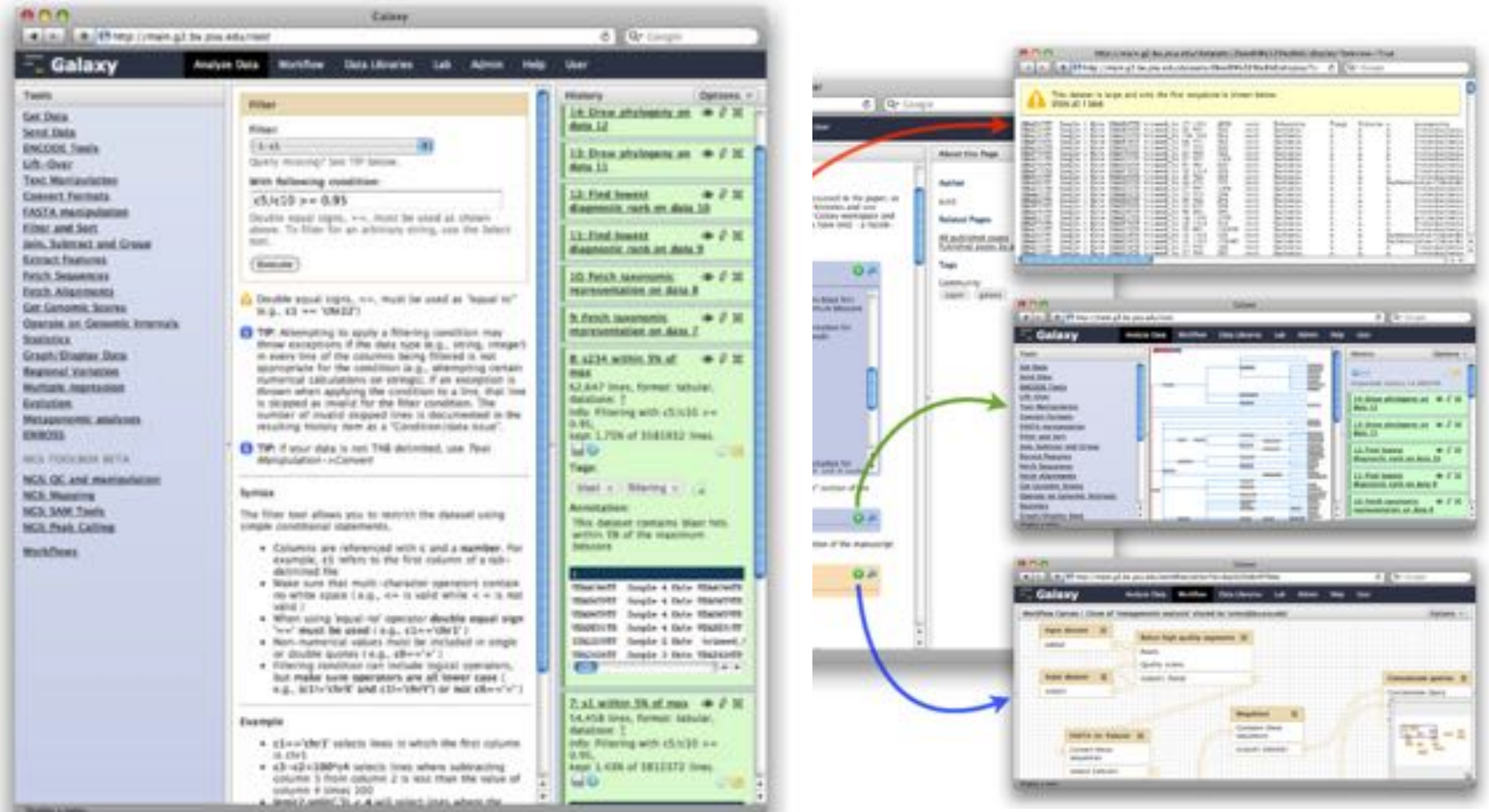
http://meetings.cshl.edu/meetings.html

Journals (RSS feeds and eTOC available)

Bioinformatics	Genome Biology	Genome Research
Nature	Nature Biotechnology	Nature Methods
PNAS	PLoS Biology	Science

Galaxy

http://usegalaxy.org http://genomics.cshl.edu



Short Read Mapping

- Given a reference and many subject reads, report one or more "good" end-to-end alignments per alignable read
  - Fundamental computation to genotyping and many assays
    - RNA-seq
    - ChIP-seq
    - Methyl-seq
    - Dnase-seq
    - FAIRE-seq
    - Hi-C-seq

- Desperate need for scalable solutions
  - Single human requires >1,000 CPU hours / genome
  - **1000 hours * 1000 genomes = 1M CPU hours / project**

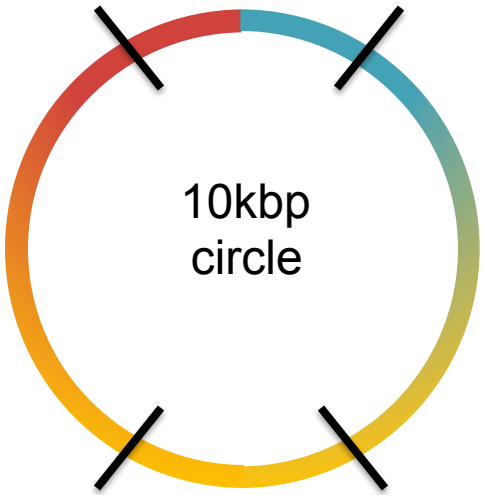
Paired-end and Mate-pairs

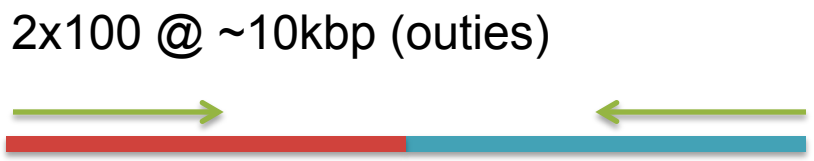
Paired-end sequencing

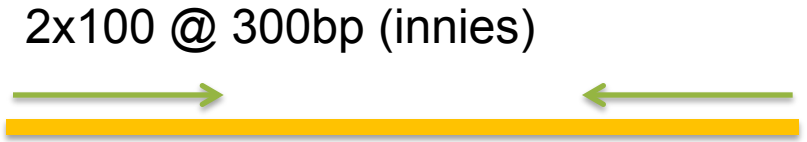
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



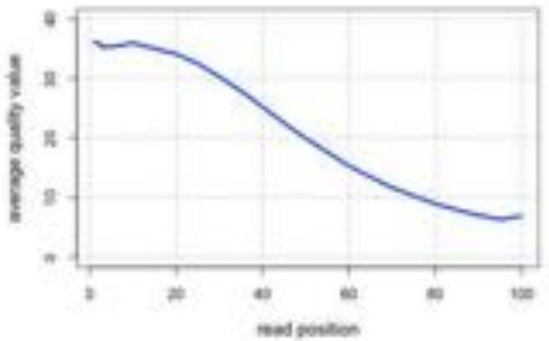




Illumina Quality

QV	p_{error}
40	1/10000
30	1/1000
20	1/100
10	1/10

$$Q_{\text{sanger}} = -10 \log_{10} p$$




```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS..................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..............
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL..............................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 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)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

http://en.wikipedia.org/wiki/FASTQ_format

Galaxy Exercise

1. Download data:
   - curl -O http://schatzlab.cshl.edu/teaching/exercises/mapping/mapping.tgz

2. Unpack and upload to Galaxy
   - Set fastq type to fastqillumina of reads

3. Map with Bowtie for Illumina
   - Aligns the reads to the reference genome

4. SAM-to-BAM
   - Converts from ASCII text file to interval representation

5. Coverage Plot of BAM
   - Mapping Statistics

6. Call variants with FreeBayes
   - Print Stats (search vcf)

Other Resources

Resource	URL	Description
Google	http://www.google.com	Internet Search
Google Scholar	http://scholar.google.com/	Literature Searches
SeqAnswers	http://seqanswers.com/	Bioinformatics Forum
Wikipedia	http://www.wikipedia.org/	Overview on anything
Circos	http://circos.ca/	Circular Genome Plots
GraphViz	http://www.graphviz.org/	Graph Visualization
EndNote	http://endnote.com/	Citation Manager
R	http://www.r-project.org/	Stats & Visualizations
Weka	http://www.cs.waikato.ac.nz/ml/weka/	Data Mining
IGV	http://www.broadinstitute.org/igv/	Read Mapping Viz
Schatz Lab	http://schatzlab.cshl.edu/teaching/	Exercises and Lectures

Questions?

http://schatzlab.cshl.edu