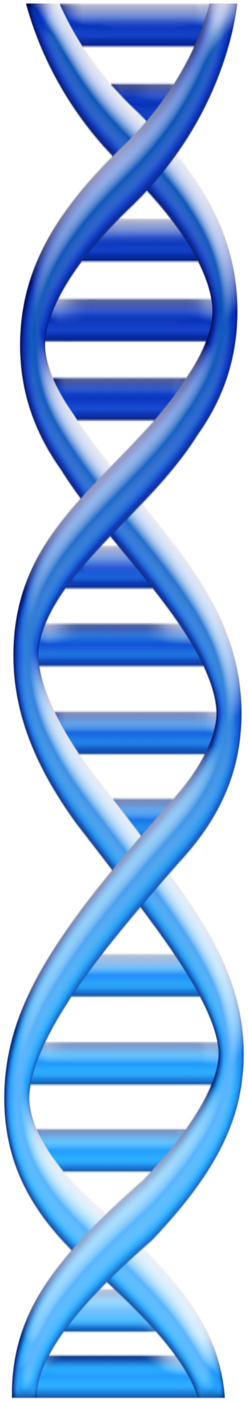
Genomic Resources

Michael Schatz

Sept 3, 2013

QB Bootcamp Lecture 3



Outline

Part 1: Overview & Fundamentals

Part 2: Sequence Analysis Theory

Part 3: Genome Resources

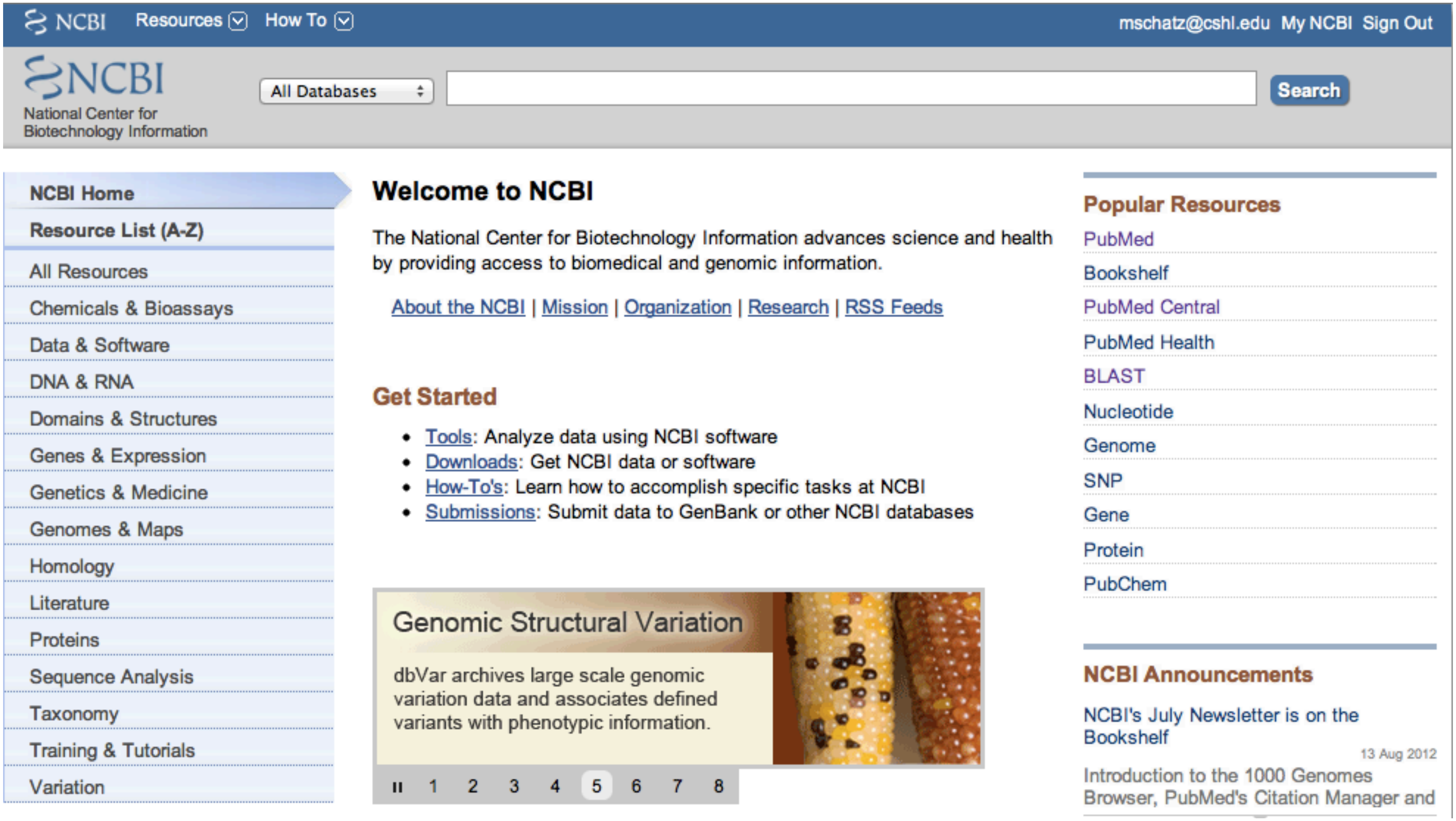
- Public: NCBI, UCSC
- CSHL: Intranet, Meetings, Galaxy

Part 4: Unix Scripting

Part 5: Example Analysis

NCBI

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



- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information.

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

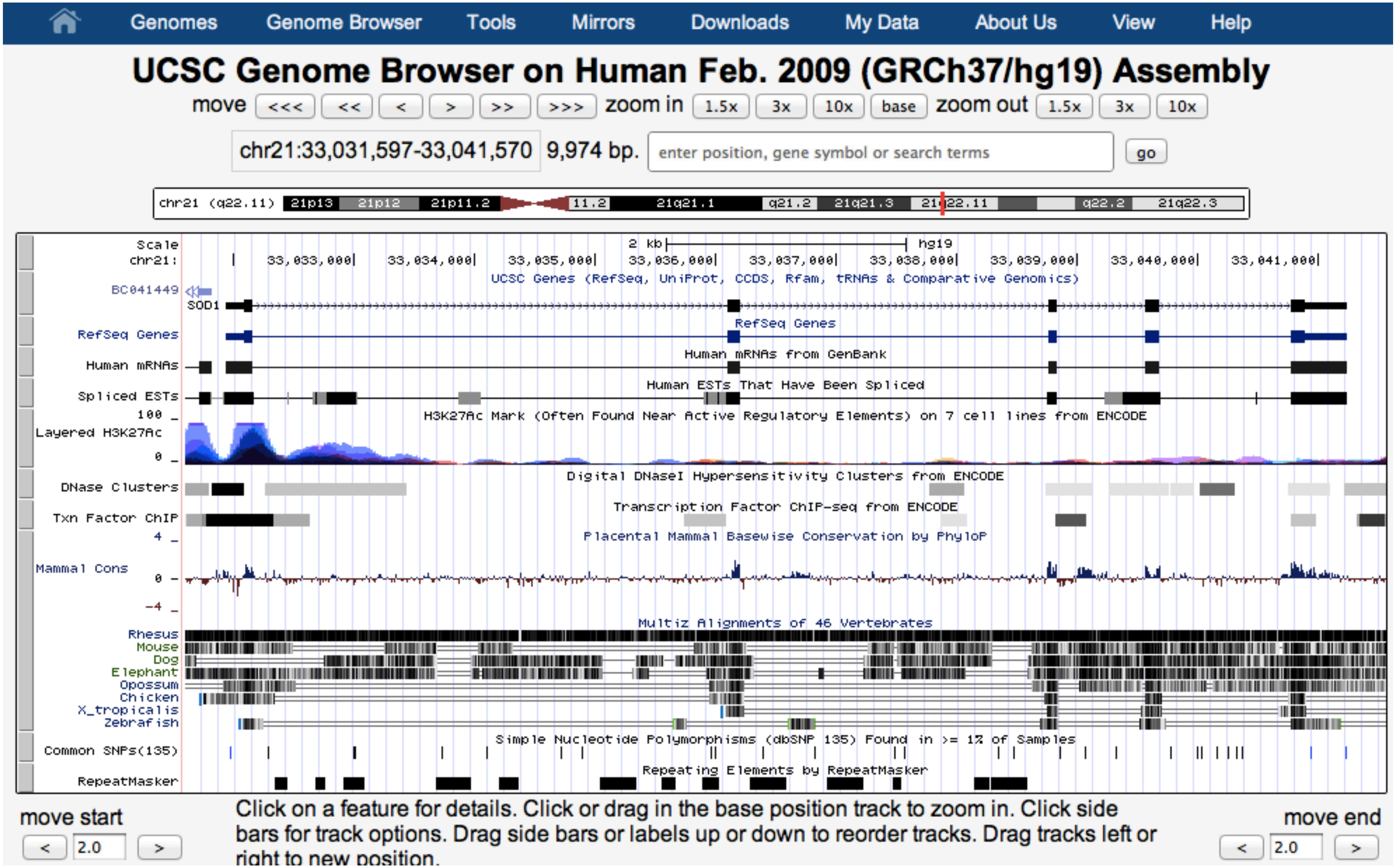
NCBI's July Newsletter is on the Bookshelf

13 Aug 2012

Introduction to the 1000 Genomes Browser, PubMed's Citation Manager and

UCSC Genome Browser

http://genome.ucsc.edu/



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

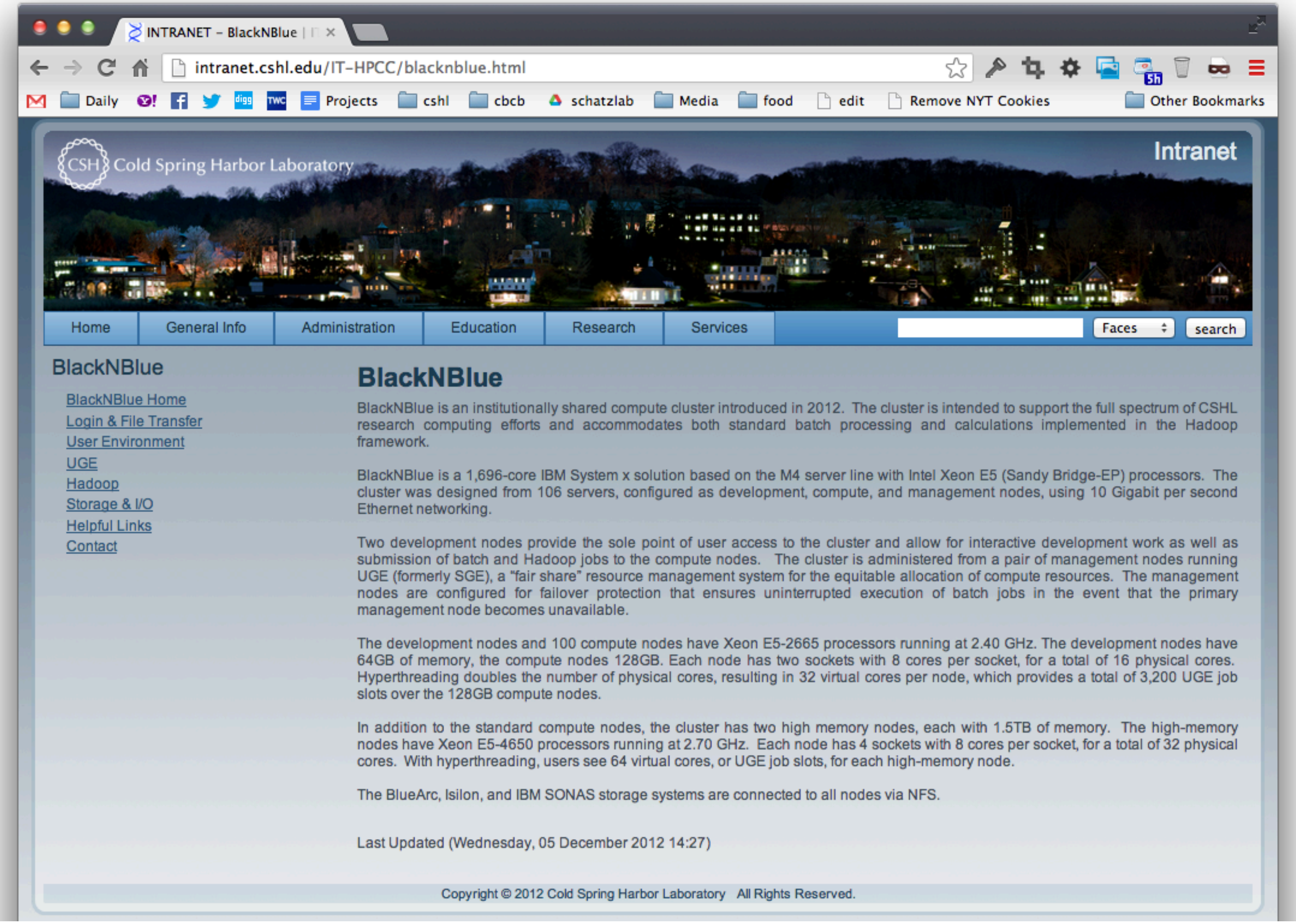
chr21:33,031,597-33,041,570 9,974 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
View
Help
move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x
enter position, gene symbol or search terms
go
chr21 (q22.11) 21p13 21p12 21p11.2 11.2 21q21.1 q21.2 21q21.3 21q22.11 q22.2 21q22.3
Scale
chr21:
2 kb
hg19
33,033,000 33,034,000 33,035,000 33,036,000 33,037,000 33,038,000 33,039,000 33,040,000 33,041,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
BC041449
SOD1
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Mammal Cons
0
-4
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(135)
Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker
move start
< 2.0 >
move end
< 2.0 >

Intranet

http://intranet.cshl.edu/IT-HPCC/blacknblue.html



BlackNBlue

- BlackNBlue Home
- Login & File Transfer
- User Environment
- UGE
- Hadoop
- Storage & I/O
- Helpful Links
- Contact

BlackNBlue

BlackNBlue is an institutionally shared compute cluster introduced in 2012. The cluster is intended to support the full spectrum of CSHL research computing efforts and accommodates both standard batch processing and calculations implemented in the Hadoop framework.

BlackNBlue is a 1,696-core IBM System x solution based on the M4 server line with Intel Xeon E5 (Sandy Bridge-EP) processors. The cluster was designed from 106 servers, configured as development, compute, and management nodes, using 10 Gigabit per second Ethernet networking.

Two development nodes provide the sole point of user access to the cluster and allow for interactive development work as well as submission of batch and Hadoop jobs to the compute nodes. The cluster is administered from a pair of management nodes running UGE (formerly SGE), a "fair share" resource management system for the equitable allocation of compute resources. The management nodes are configured for failover protection that ensures uninterrupted execution of batch jobs in the event that the primary management node becomes unavailable.

The development nodes and 100 compute nodes have Xeon E5-2665 processors running at 2.40 GHz. The development nodes have 64GB of memory, the compute nodes 128GB. Each node has two sockets with 8 cores per socket, for a total of 16 physical cores. Hyperthreading doubles the number of physical cores, resulting in 32 virtual cores per node, which provides a total of 3,200 UGE job slots over the 128GB compute nodes.

In addition to the standard compute nodes, the cluster has two high memory nodes, each with 1.5TB of memory. The high-memory nodes have Xeon E5-4650 processors running at 2.70 GHz. Each node has 4 sockets with 8 cores per socket, for a total of 32 physical cores. With hyperthreading, users see 64 virtual cores, or UGE job slots, for each high-memory node.

The BlueArc, Isilon, and IBM SONAS storage systems are connected to all nodes via NFS.

Last Updated (Wednesday, 05 December 2012 14:27)

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

Nature

PNAS

Genome Biology

Nature Biotechnology

PLoS Biology

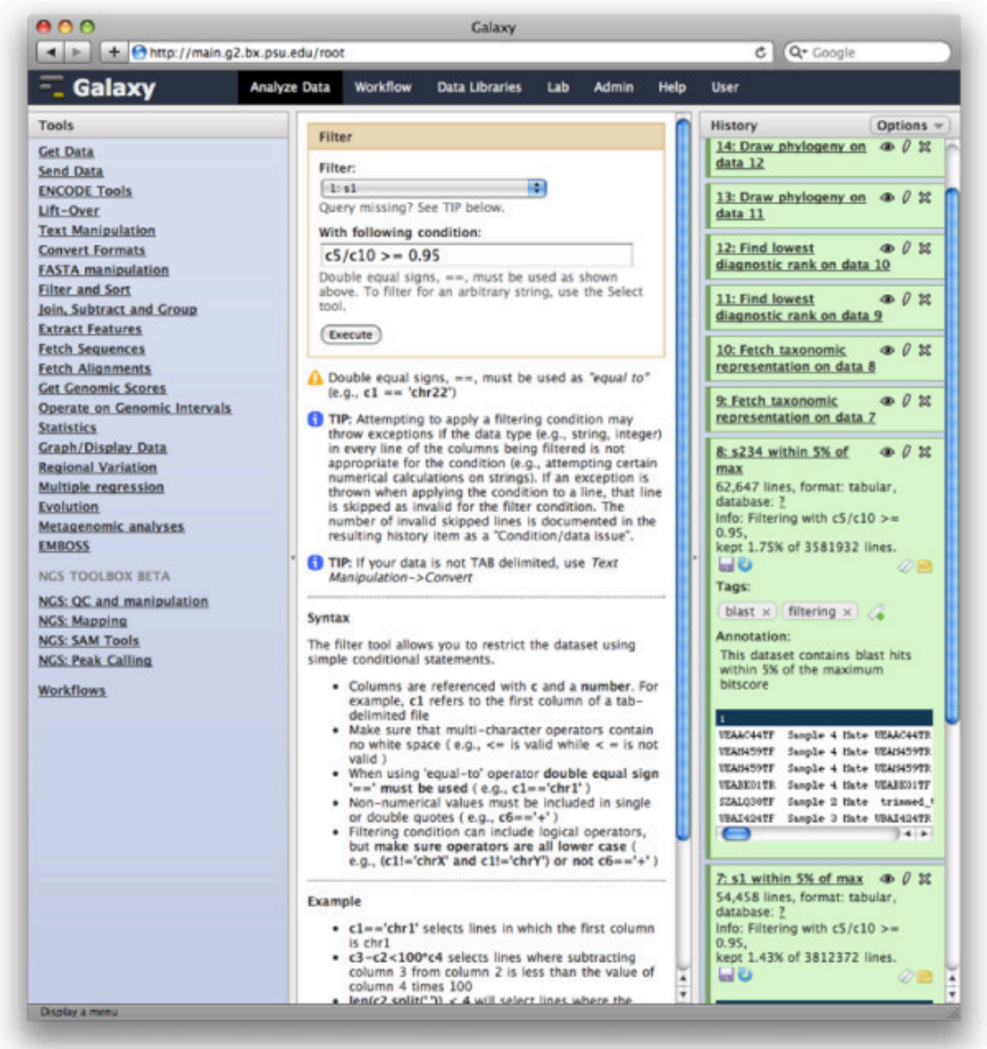
Genome Research

Nature Methods

Science

Galaxy

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



Filter

Filter: 1: s1

Query missing? See TIP below.

With following condition: c5/c10 >= 0.95

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Execute

Double equal signs, ==, must be used as *"equal to"* (e.g., **c1 == 'chr22'**)

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

- Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file
- Make sure that multi-character operators contain no white space (e.g., **<=** is valid while **< =** is not valid)
- When using 'equal-to' operator **double equal sign '==' must be used** (e.g., **c1=='chr1'**)
- Non-numerical values must be included in single or double quotes (e.g., **c6=='+'**)
- Filtering condition can include logical operators, but **make sure operators are all lower case** (e.g., **(c1!='chrX' and c1!='chrY') or not c6=='+'**)

Example

- **c1=='chr1'** selects lines in which the first column is chr1
- **c3-c2<100*c4** selects lines where subtracting column 3 from column 2 is less than the value of column 4 times 100
- **len(c2.split(',')) < 4** will select lines where the

History Options

14: Draw phylogeny on data 12

13: Draw phylogeny on data 11

12: Find lowest diagnostic rank on data 10

11: Find lowest diagnostic rank on data 9

10: Fetch taxonomic representation on data 8

9: Fetch taxonomic representation on data 7

8: s234 within 5% of max

62,647 lines, format: tabular, database: ?

Info: Filtering with c5/c10 >= 0.95,

kept 1.75% of 3581932 lines.

Tags: blast filtering

Annotation:

This dataset contains blast hits within 5% of the maximum bitscore

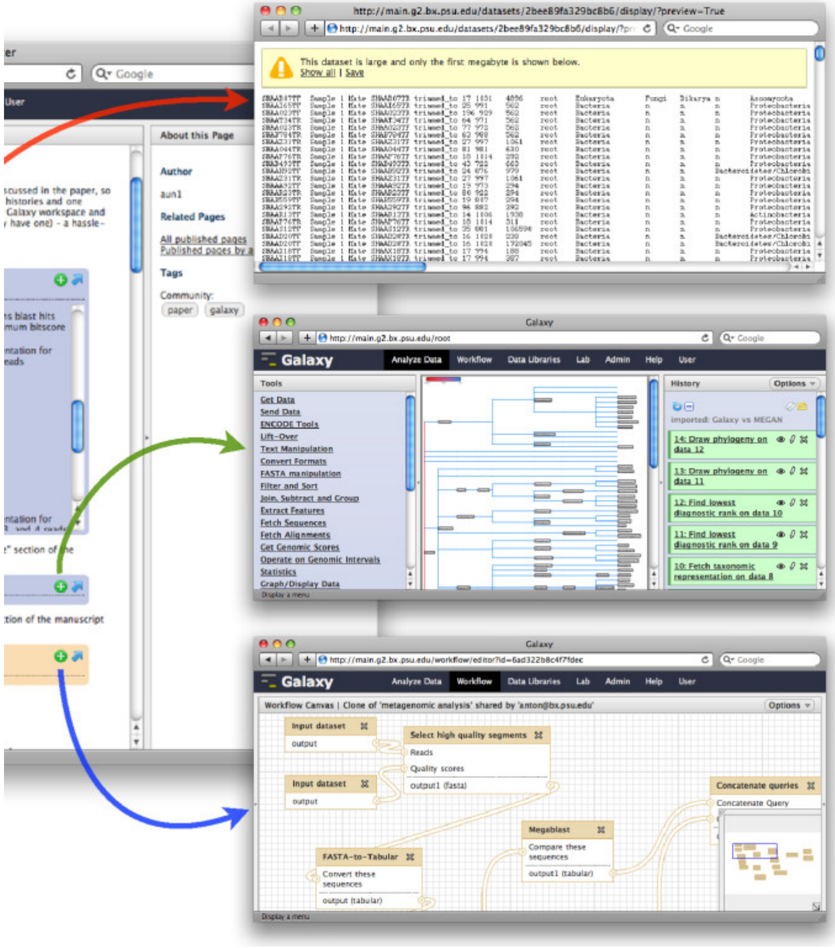
```
UEAAC44TF  Sample 4 Hate UEAAC44TR
UEAH459TF  Sample 4 Hate UEAH459TR
UEAH459TF  Sample 4 Hate UEAH459TR
UEAEK01TR  Sample 4 Hate UEAEK01TF
SZALQ30TF  Sample 2 Hate trimmed_
UBAI424TF  Sample 3 Hate UBAI424TR
```

7: s1 within 5% of max

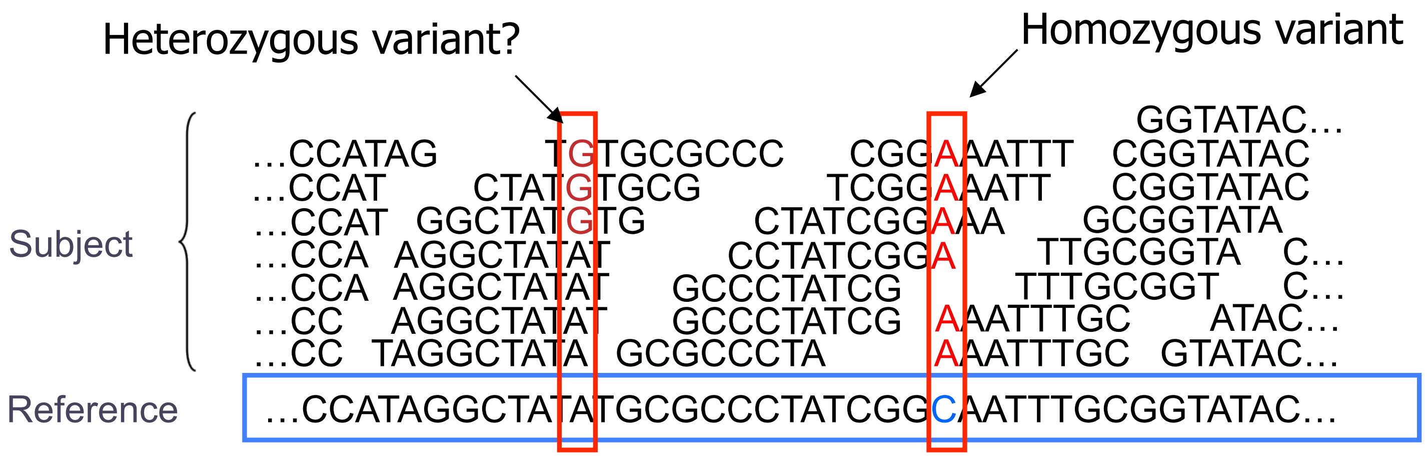
54,458 lines, format: tabular, database: ?

Info: Filtering with c5/c10 >= 0.95,

kept 1.43% of 3812372 lines.



Genotyping



Heterozygous variant?

Homozygous variant

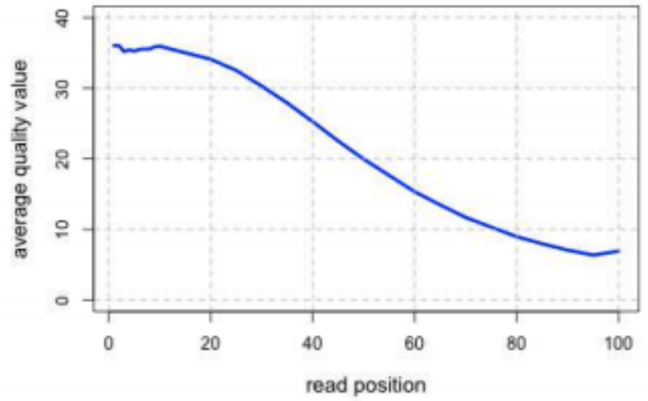
Subject

```
                                                        GGTATAC…
…CCATAG      TGTGCGCCC    CGGAAATTT   CGGTATAC
…CCAT     CTATGTGCG      TCGGAAATT    CGGTATAC
…CCAT  GGCTATGTG      CTATCGGAAA    GCGGTATA
…CCA  AGGCTATAT     CCTATCGGA      TTGCGGTA   C…
…CCA  AGGCTATAT   GCCCTATCG       TTTGCGGT    C…
…CC   AGGCTATAT   GCCCTATCG  AAATTTGC     ATAC…
…CC  TAGGCTATA  GCGCCCTA     AAATTTGC  GTATAC…
```

Reference

```
…CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC…
```

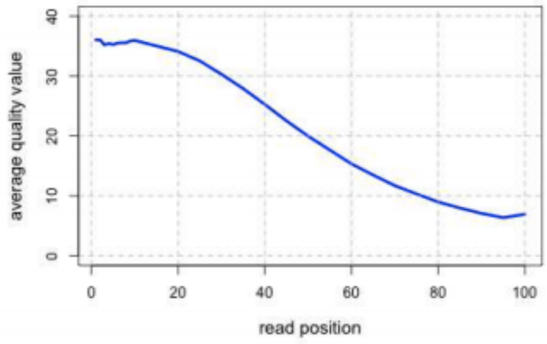
- Sequencing instruments make mistakes
  - Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
  - Often framed as a Bayesian problem of more likely to be a real variant or chance occurrence of N errors
  - Accuracy improves with deeper coverage

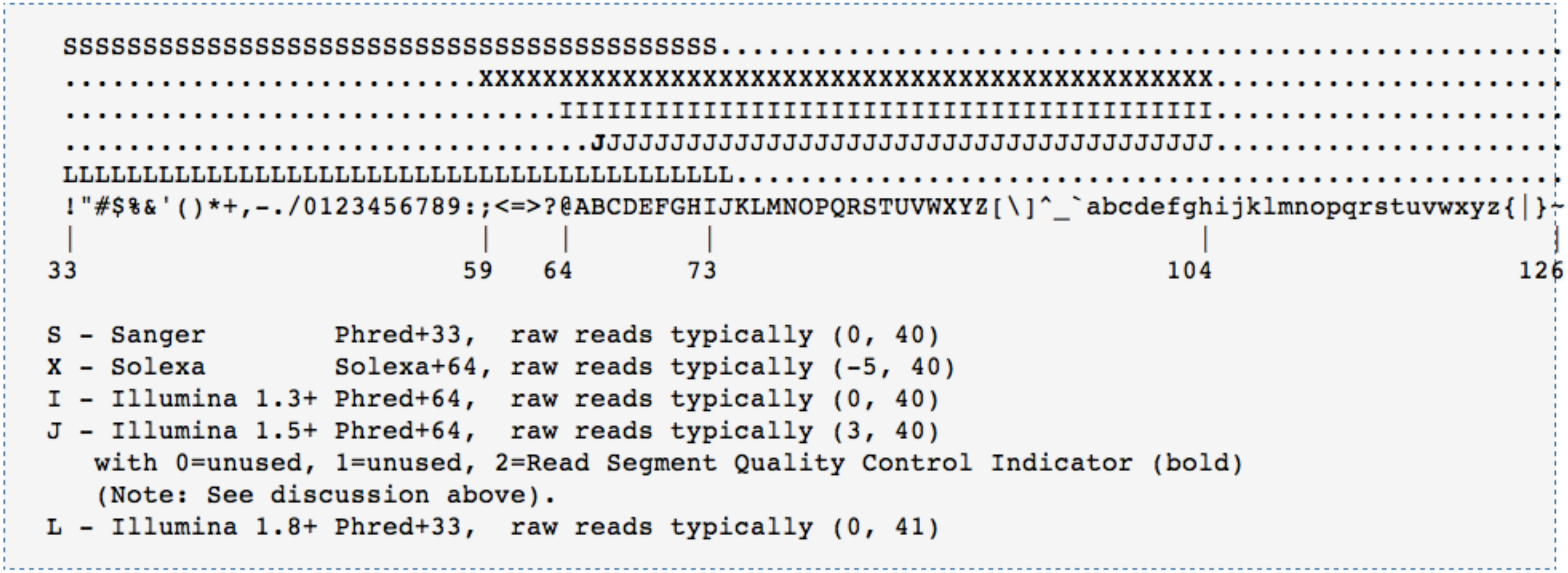


Illumina Quality

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

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





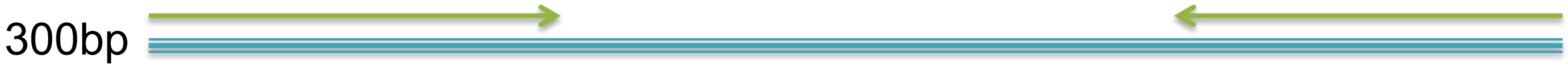
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./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

Paired-end and Mate-pairs

Paired-end sequencing

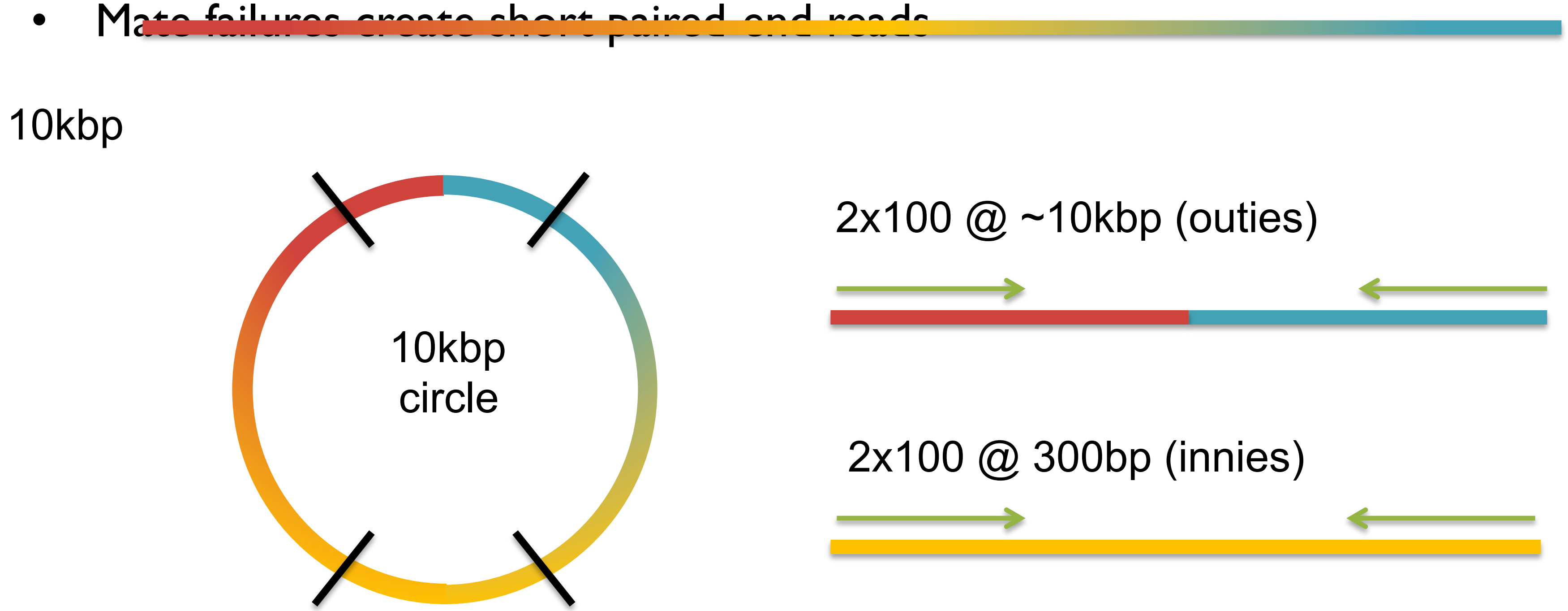
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



300bp

Mate-pair sequencing

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



10kbp

10kbp circle

2x100 @ ~10kbp (outies)

2x100 @ 300bp (innies)

Galaxy Exercise

1. Download data:
   - 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