

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/>

Resources ▾ How To ▾mschatr@cshl.edu My NCBI Sign Out

All Databases ▾Search

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.



1 2 3 4 5 6 7 8

Popular Resources

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

NCBI Announcements

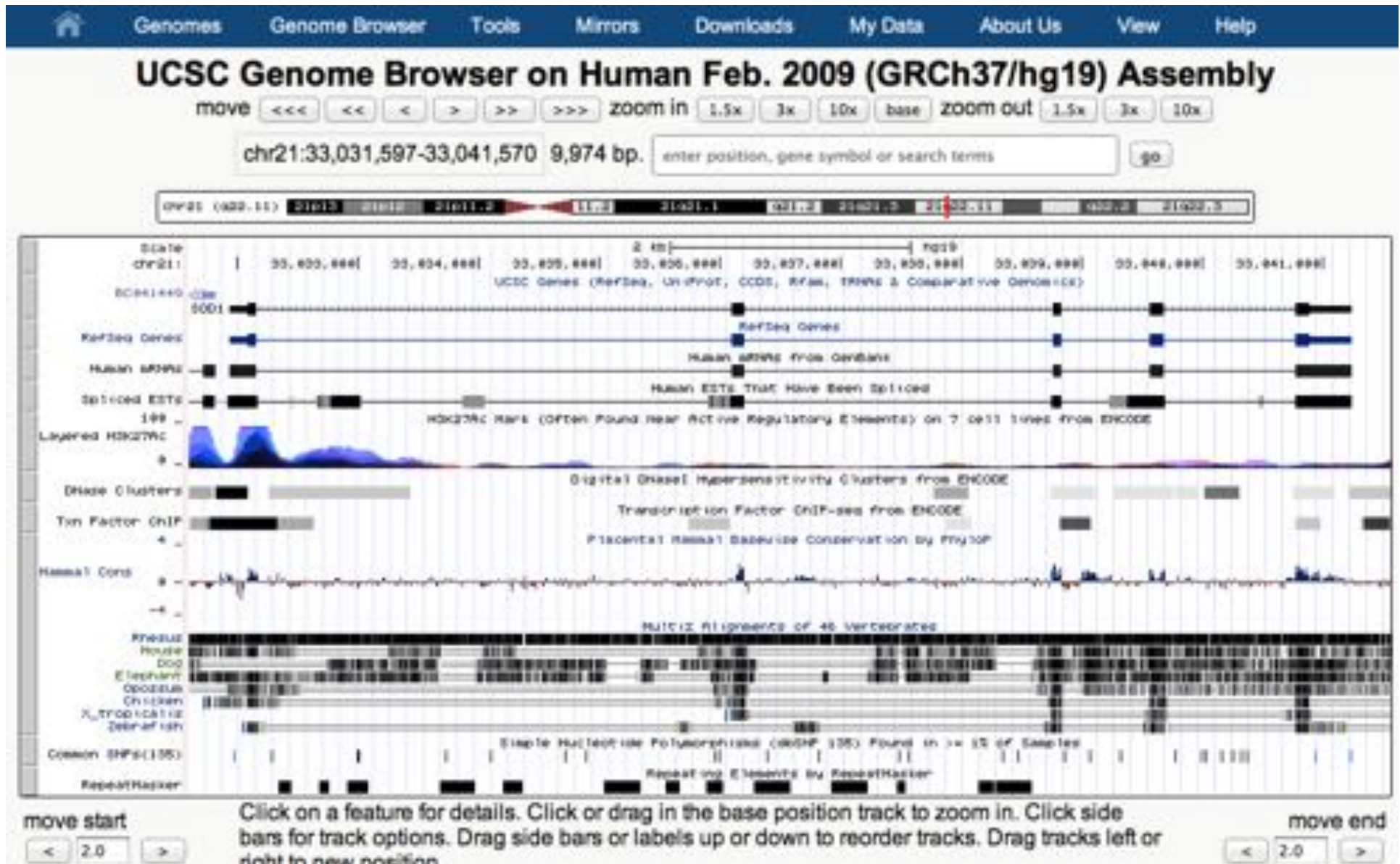
NCBI's July Newsletter is on the Bookshelf

13 Aug 2012

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

UCSC Genome Browser

<http://genome.ucsc.edu/>



Intranet

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

The screenshot shows a web browser window with the address `intranet.cshl.edu/IT-HPCC/blacknblue.html`. The page features a header with the Cold Spring Harbor Laboratory logo and a navigation menu. The main content area is titled "BlackNBlue" and contains detailed information about the cluster's architecture and capabilities.

BlackNBlue

BlackNBlue is an institutional shared compute cluster introduced in 2012. The cluster is intended to support the full spectrum of CCNY research computing efforts and accommodate both standard batch processing and workflows implemented in the Hadoop framework.

BlackNBlue is a 1,696-core IBM System x solution based on the M4 server line with Intel Xeon E3 (Sandy Bridge-EPI) processors. The cluster was designed from 100 servers, configured as development, compute, and management nodes, using 16 Gbps per second Ethernet networking.

Two development nodes provide the sole point of user access to the cluster and serve 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 "for 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 E3-2660 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 E3-4400 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, iStor, and IBM SONAS storage systems are connected to all nodes via NFS.

Last Updated (Wednesday, 09 December 2012 14:27)

Copyright © 2012 Cold Spring Harbor Laboratory. All Rights Reserved.

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>

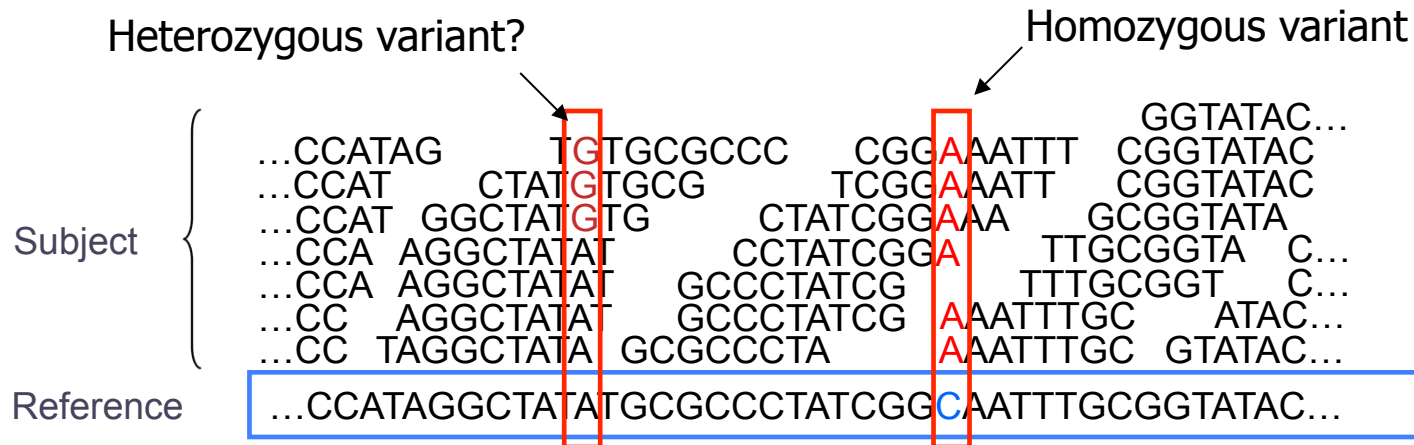
Journals (RSS feeds and eTOC available)

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

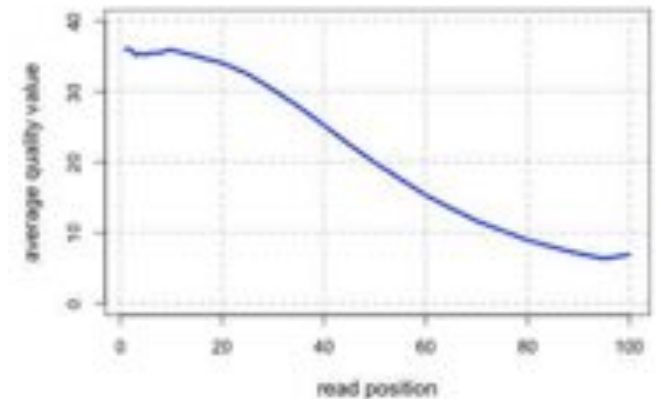
Galaxy
<http://usegalaxy.org> <http://genomics.cshl.edu>

<http://genomics.cshl.edu>

Genotyping



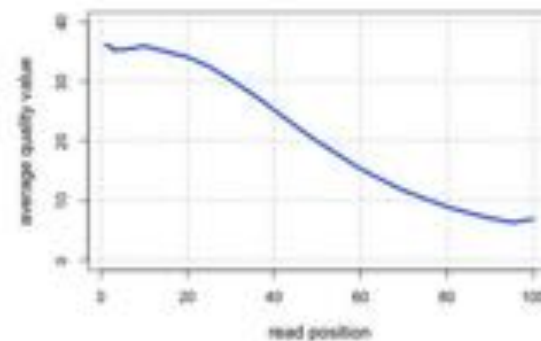
- Sequencing instruments make mistakes
 - Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
 - Often framed as a Bayesian problem of more likely to be a real variant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

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



```

#####
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMN
OPQRSTUVWXYZ[\]^_`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)

```

Paired-end and Mate-pairs

Paired-end sequencing

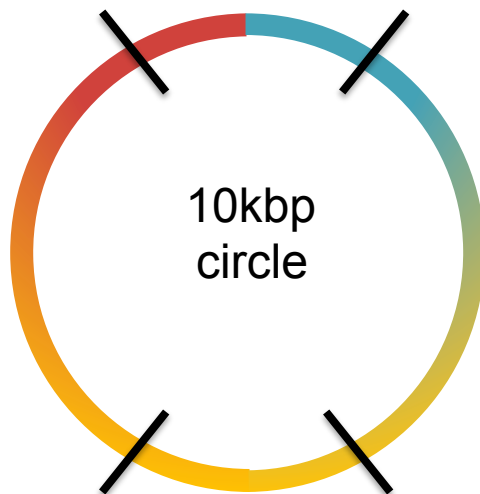
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



Mate-pair sequencing

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

10kbp



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>