

LS610 - Next Generation Sequencing Data Analysis

Description:

Massively parallel sequencing, also known as next generation sequencing, is a technology enabling high-throughput sequencing of genomes or loci of interest. This hands-on class focuses on a single locus. It examines the quality of the sequence reads; mapping of reads; and the quality of the mapping. It also examines sequence variation.

Objectives:

- Understand quality assessment of reads
- Map reads to a chromosome
- Assess quality of the mapping
- Find regions of significant variation
- Interpret the results in the biological context





An ORS Service

Next Generation Sequencing Data Analysis

Lynn Young, Ph.D.

lynny@mail.nih.gov

NIH Library Bioinformatics Support Program

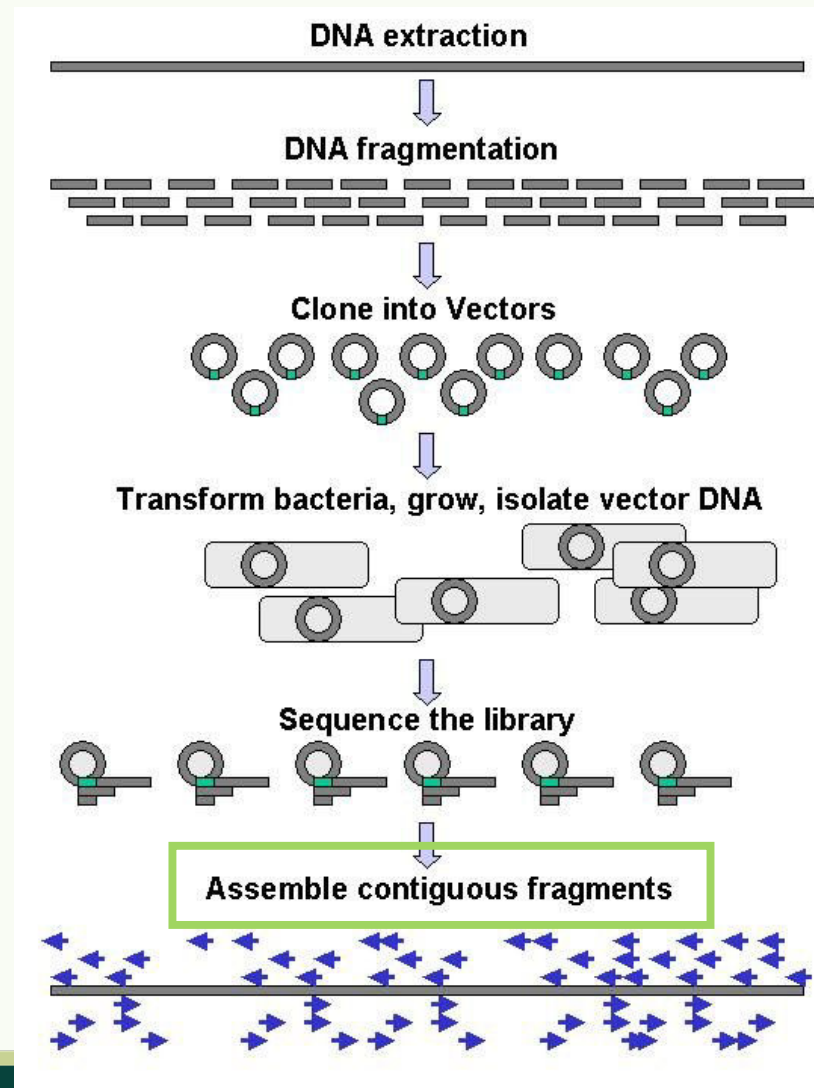
20 September 2012

Acknowledgement

This training uses cloud services provided by an “AWS in Education” grant to the Galaxy Project.



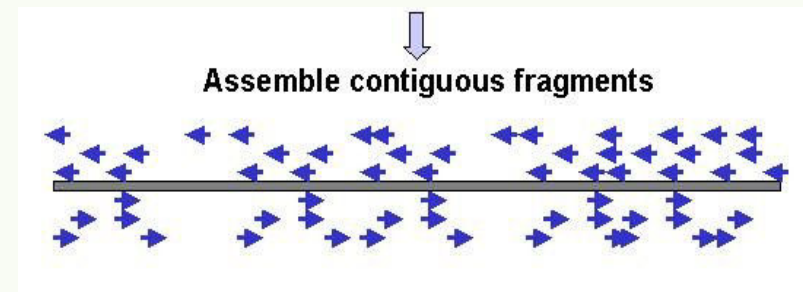
Introduction



http://en.wikipedia.org/wiki/File:DNA_Sequencing_gDNA_libraries.jpg

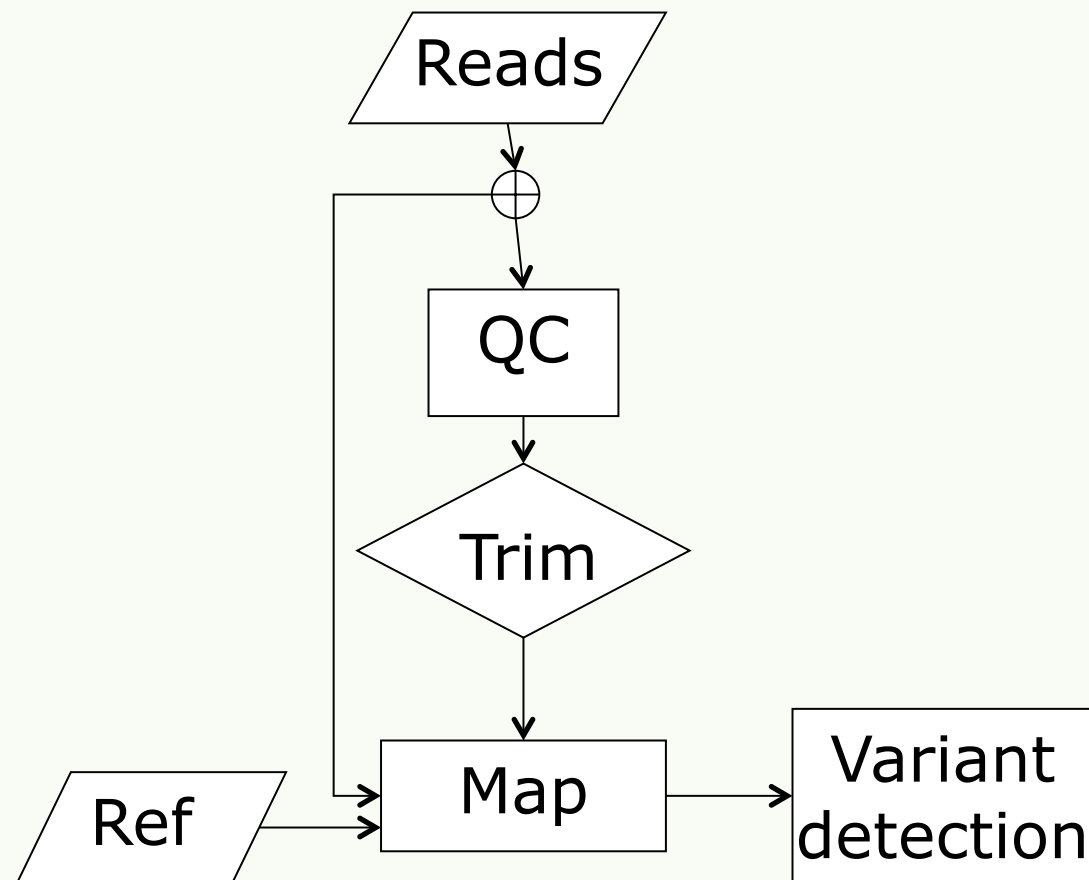
Objectives

- Sequence quality
- Mapping
- Mapping quality
- Variant analysis
- Biological context



http://en.wikipedia.org/wiki/File:DNA_Sequencing_gDNA_libraries.jpg

Data Analysis Workflow



Reference – FASTA format

```
>gi|206583719|gb|CM000511.1| Homo sapiens chromosome 21, whole genome shotgun s  
ATTCATTCCATTCCACTGCACTCCAATCTTCACATAAAATGTAGACAGAAGCTTTCTGAGAACTTTTCT  
CTGATGTGTGCATTCATCTCACAGATGTGAACCATTCTTTTGTTTGAGCAGTTTGGTAACATTCTTTTG  
TAGAATCTGCAAAAGGATATTTGTGAGCACTTTGAAGCCTATGGTGAAAAAGGAAATATCTTCAGAGAAA  
AACTAGAAAGAAGGTTTCTGAGAACTGCTTTGTCATGTGTGAATTAGTCTCACAGATTTGAACCTTTCT  
GTTGATTGAACATATTGGAAACCTTCTTTTTGTAGAATCTGCAAAGGGATATTTGTGAACACTTGGAGGC  
CAATGGTGAAAAAGGAAATATATTCACATGAAAACCTAGACAGAATCTTTCTGAGACACTTCTGTGTTTGG
```



Reads – FASTQ Format

@SRR016862.16884

ATTTTGAGTGGTACATCTAGGTAGCCGTTTTTGGAAACGGG

+

IIIIII,IIIII?III?I&II9\$H+/I>IA%1.\$,\$%\$#\$F

@SRR016862.58801

ATTTTGAGTGGTACATCTAGGTAGCCGTTTTTGAAACCAGG

+

IIIIIIIIIIIIIIIIIIII9III0II4.II@&?6&\$&#%'@.



Alignments – SAM Format

1	@SQ	SN:chr21 LN:48129895									
2	@PG	ID:bwa	PN:bwa	VN:0.5.9-r16							
3	SRR016862.16884	0	chr21	27002753	25	41M	*	0	0	ATTTTGAGTGGTACATCTAGGTAGCCGTTT	
4	SRR016862.58801	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGTAGCCGTTT	
5	SRR016862.198085	0	chr21	27002753	37	11M1D30M	*	0	0	ATTTTGAGTGGACATCTAGGTAGCCGTTT	
6	SRR016862.219598	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGTAGCCGTTT	

CGTTTTTGGAAACGGG	IIIIII,IIII?III?I&II9\$H+I/>IA%1.\$,%\$#\$F	XT:A:U	NM:i:3	X0:i:1	X1:i:0	XM:i:3	XO:i:0	XG:i:0	MD:Z:30C2A4A2	
CGTTTTTGAACCCAGG	IIIIIIIIIIII9IIIOII4.II@&?6&\$&#%#@.	XT:A:U	NM:i:2	X0:i:1	X1:i:0	XM:i:2	XO:i:0	XG:i:0	MD:Z:30C5A4	
CGTTTCTTAAAAAGGT	IIIIIIAI@II;I%ICI&I81(3\$""8>\$(&\$*##\$.(\$	XT:A:U	NM:i:3	X0:i:1	X1:i:0	XM:i:2	XO:i:1	XG:i:1	MD:Z:11^T20G4C	
CGTTTCTGAACACAGG	IIIIIIIIIAE@4I-H6CC8*F&I3+,:\$I1#(\$/'\$&0	XT:A:U	NM:i:1	X0:i:1	X1:i:0	XM:i:1	XO:i:0	XG:i:0	MD:Z:35A5	

<http://samtools.sourceforge.net/SAM1.pdf>

<http://bio-bwa.sourceforge.net/bwa.shtml#4>



Variant Calls – VCF Format

<http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41>

1	##fileformat=VCFv4.1				
2	##fileDate=20120917				
3	##source=freeBayes version 0.9.4				
4	##reference=localref.fa				
5	##phasing=none				
6	##commandline="freebayes --bam localbam_0.bam --bam localbam_1.bam --bam localbam_2.bam --fasta-reference localref.fa --vcf /galaxy/main_pool/pool3/t				
7	##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">				
8	##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">				
9	##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">				
10	##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">				
11	##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">				
12	##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observations">				
13	##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations">				
14	##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability				
15	##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability				
16	##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the				
17	##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability o				
18	##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">				
19	##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the devi				
20	##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the prob				
21	##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the devia				
22	##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the proba				
23	##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio. Ratio between depth in samples with each called alternate allele and those				
24	##INFO=<ID=XRM,Number=1,Type=Float,Description="Reference allele read mismatch rate: The rate of SNPs + MNPs + INDELs in reads supporting the reference				
25	##INFO=<ID=XRS,Number=1,Type=Float,Description="Reference allele read SNP rate: The rate of per-base mismatches (SNPs + MNPs) in reads supporting the ref				
26	##INFO=<ID=XRI,Number=1,Type=Float,Description="Reference allele read INDEL rate: The rate of INDELs (gaps) in reads supporting the reference allele.">				
27	##INFO=<ID=XAM,Number=A,Type=Float,Description="Alternate allele read mismatch rate: The rate of SNPs + MNPs + INDELs in reads supporting the alternate a				

VCF Format - Data

50	##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">							
51	##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">							
52	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
53	chr21	27006332	.	A	G	3.36835	.	AB=0.0833333;ABP=21.105
54	chr21	27006817	.	GT	AG	184.005	.	AB=0;ABP=0;AC=2;AF=1;AM
55	chr21	27010656	.	TT	AC	256.001	.	AB=0;ABP=0;AC=2;AF=1;AM
56	chr21	27010825	.	CA	AC	5571.43	.	AB=0;ABP=0;AC=2;AF=1;AM
57	chr21	27011565	.	AA	TG	76.0432	.	AB=0;ABP=0;AC=2;AF=1;AM
58	chr21	27011976	.	G	A	7662.29	.	AB=0;ABP=0;AC=2;AF=1;AM

	FORMAT	unknown	
:DPRA=0;EPP=5.18177;EPPR=26.8965;HWE=-0;I	GT:GQ:DP:RO:QR:AO:QA:GL	0/1:3.36835:12:11:307	
677;EPPR=0;HWE=-0;LEN=2;MEANALT=1;MQM	GT:GQ:DP:RO:QR:AO:QA:GL	1/1:29.0363:5:0:0:5:20	
106;EPPR=0;HWE=-0;LEN=2;MEANALT=1;MQM	GT:GQ:DP:RO:QR:AO:QA:GL	1/1:35.0529:7:0:0:7:28	
:367.818;EPPR=0;HWE=-0;LEN=2;MEANALT=2;M	GT:GQ:DP:RO:QR:AO:QA:GL	1/1:147.608:171:0:0:16	
324;EPPR=0;HWE=-0;LEN=2;MEANALT=1;MQM	GT:GQ:DP:RO:QR:AO:QA:GL	1/1:20.0432:2:0:0:2:80	
:355.276;EPPR=5.18177;HWE=-0;LEN=1;MEANA	GT:GQ:DP:RO:QR:AO:QA:GL	1/1:50000:234:1:40:23	

Data – Sequence Read Archive

<http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP000535>

The screenshot shows the NCBI Sequence Read Archive (SRA) website. The browser is Firefox, and the URL is <http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP000535>. The page title is "Sequence Read Archive". The navigation bar includes links for Main, Browse, Search, Download, Submit, Documentation, Software, Trace Archive, Trace Assembly, Trace Home, and Trace BLAST. The "Studies" section is active, showing a list of studies. The study SRP000535 is highlighted, with the title "Significantly improved multiplex padlock capturing and large scale sequencing reveal hypermutable CpG variations". The study type is "Resequencing", submitted by Virginia Commonwealth on 2009-01-23 15:50:42. The abstract describes the use of next-generation sequencing to identify genetic variations in hypermutable CpG rich regions. The description mentions that 53,777 padlock probes were designed to cover all CpGs in non-repetitive regions of chromosome 21. A table of experiments is shown, listing accession numbers, spots, and bases.

SRP000535 Significantly improved multiplex padlock capturing and large scale sequencing reveal hypermutable CpG variations

Study Type: Resequencing
Submission: SRA007914 by Virginia Commonwealth on 2009-01-23 15:50:42
Abstract: To take full advantage of the power of next generation sequencing requires large scale multiplex capturing and amplification of genomic regions of interest in multiple samples. We improved a previously developed padlock probe-based approach by 10,000-fold and evaluated it by identifying genetic variations in hypermutable CpG rich regions, tiling CpGs across human chromosome 21 with 53,777 probes in an unbiased manner. Two to three million mapped sequences derived from a single Illumina Genome Analyzer lane enable observation of 90.8%-94.0% of target sites with at least one read. Although the sites were not uniformly captured, ~85% and ~55% of all targets fell within a 100- and 10-fold range of each other, respectively. A total of 442937 genotypes were computed with confidence across six subjects and examined for single nucleotide polymorphisms and were found to be in between 98.4% and 100% agreement with independently obtained genotype data. We detected as many as 502 sites of variation not present in dbSNP, including 362 in targeted CpG locations not among the 2640 in dbSNP. CpG->CpA and CpG->TpG variations were found to be ~12.5x more abundant than non-CpG variations. Variation rates differed among subjects in a possibly ancestry-related manner. Our success suggests that genotyping hypermutable CpGs may be an efficient way of identifying common and rare mutations in human diseases and that padlock-probe based sequencing can reveal important aspects of human variation generally. We identified areas for further improvement and study.

Description: 53,777 padlock probes were designed to cover all CpGs in non-repetitive regions of chromosome 21. Sequencing libraries were constructed for six subjects (see

Show [Entrez documents](#) for all experiments
Download reads for entire study in [sra format](#)

Experiments
[Show RUNs for each experiment](#)

Accession	Spots	Bases
Total: 11	33.8M	1.3G
SRX005032	3.5M	143.0M
SRX005036	2.7M	109.1M
SRX005037	2.5M	90.5M
SRX005038	1.9M	79.4M
SRX005039	2.9M	119.5M
SRX005040	2.5M	101.5M
SRX005041	3.0M	122.1M
SRX005042	2.5M	89.8M
SRX005043	4.5M	163.4M
SRX005044	3.7M	133.1M
SRX005045	4.1M	148.6M



An ORS Service

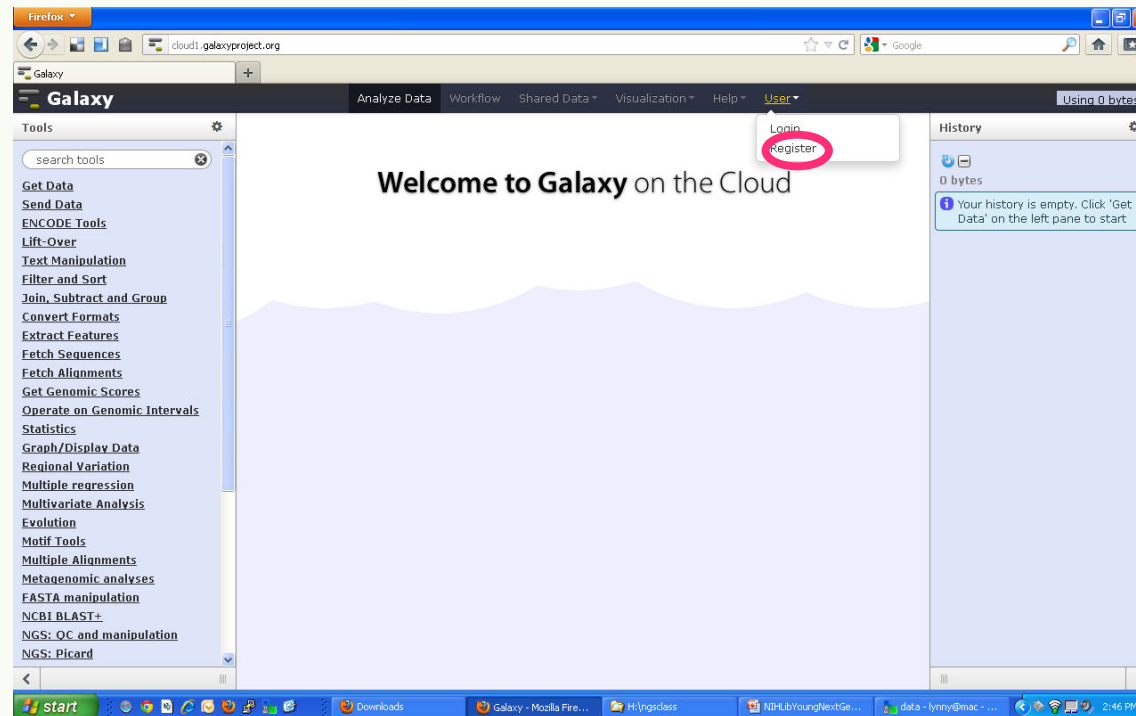
Next Generation Sequencing Data Analysis

Galaxy

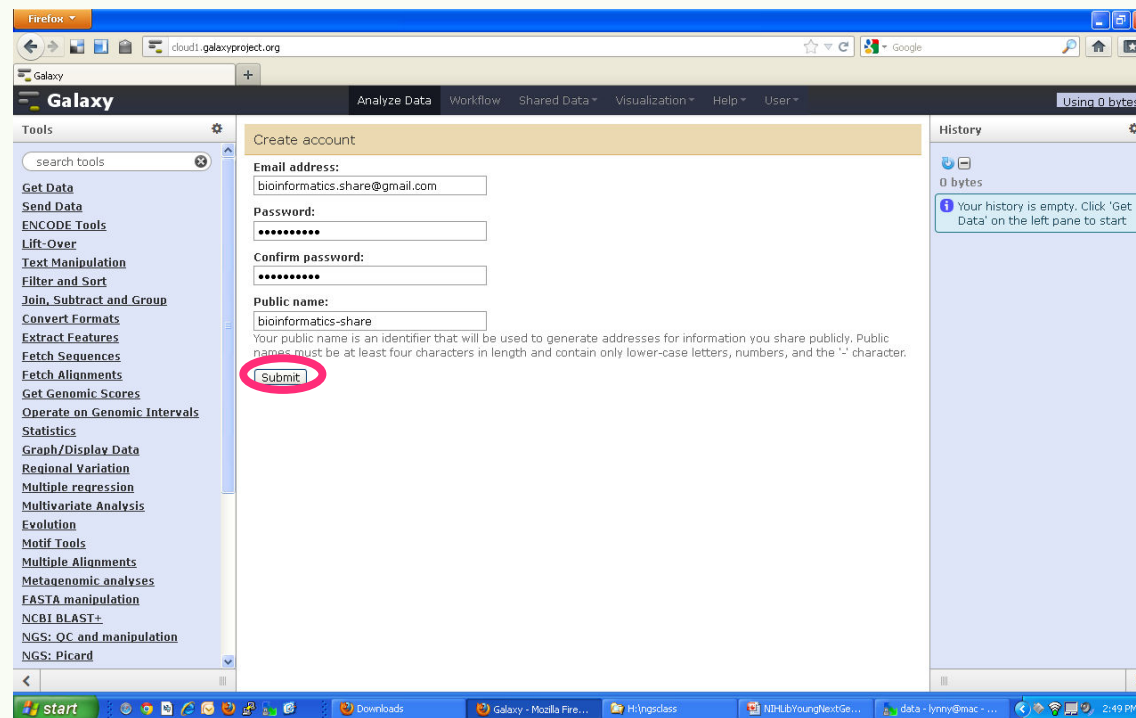
- Public
 - *usegalaxy.org*
- 20 September 2012 class
 - *cloud1.galaxyproject.org*
 - *cloud2.galaxyproject.org*



Galaxy Account Registration



Galaxy Account Registration



The screenshot shows the Galaxy Project website in a Firefox browser window. The address bar displays 'cloud1.galaxyproject.org'. The page features a dark navigation bar with 'Galaxy' and several menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0 bytes' indicator is visible on the right. On the left, a 'Tools' panel lists various bioinformatics tools under a search bar. The main content area is titled 'Create account' and contains a registration form with the following fields: 'Email address' (filled with 'bioinformatics.share@gmail.com'), 'Password' (masked with dots), 'Confirm password' (masked with dots), and 'Public name' (filled with 'bioinformatics-share'). Below the 'Public name' field is a note: 'Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.' A red circle highlights the 'Submit' button at the bottom of the form. On the right, a 'History' panel shows '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start.'

Firefox

cloud1.galaxyproject.org

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 0 bytes

Tools

search tools

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

NCBI BLAST+

NGS: QC and manipulation

NGS: Picard

Create account

Email address:
bioinformatics.share@gmail.com

Password:

Confirm password:

Public name:
bioinformatics-share

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.

Submit

History

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start.

start

Downloads

Galaxy - Mozilla Fire...

HTIngclass

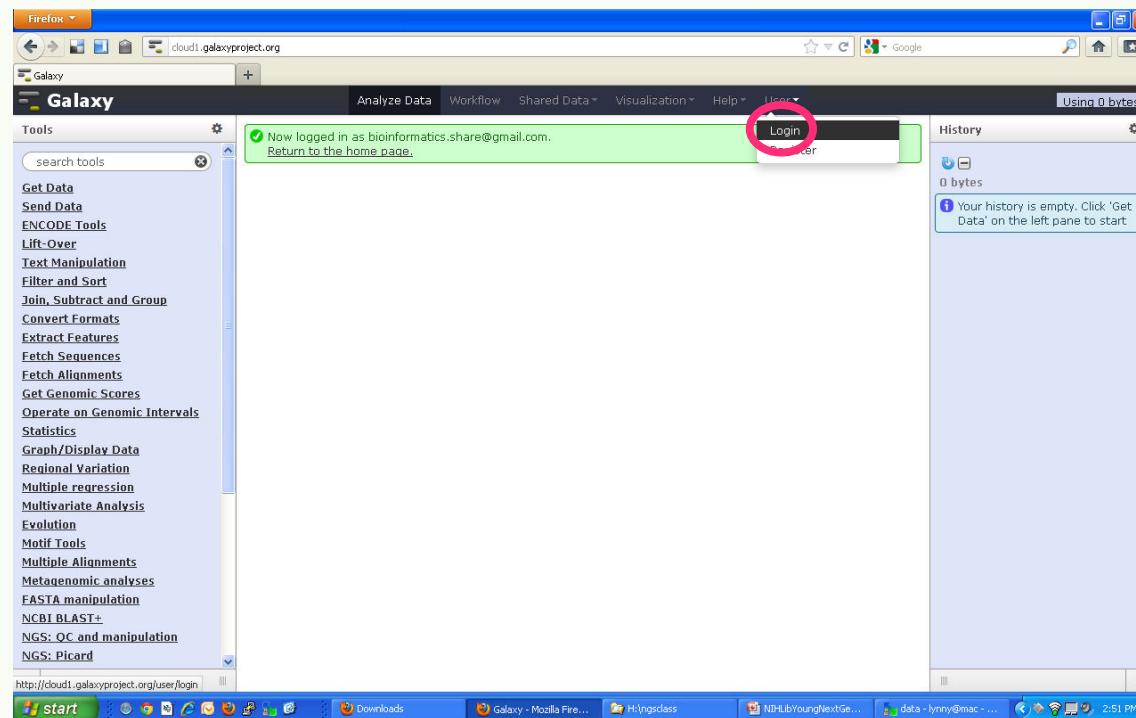
NIHLibYoungNextGe...

data - lynny@mac - ...

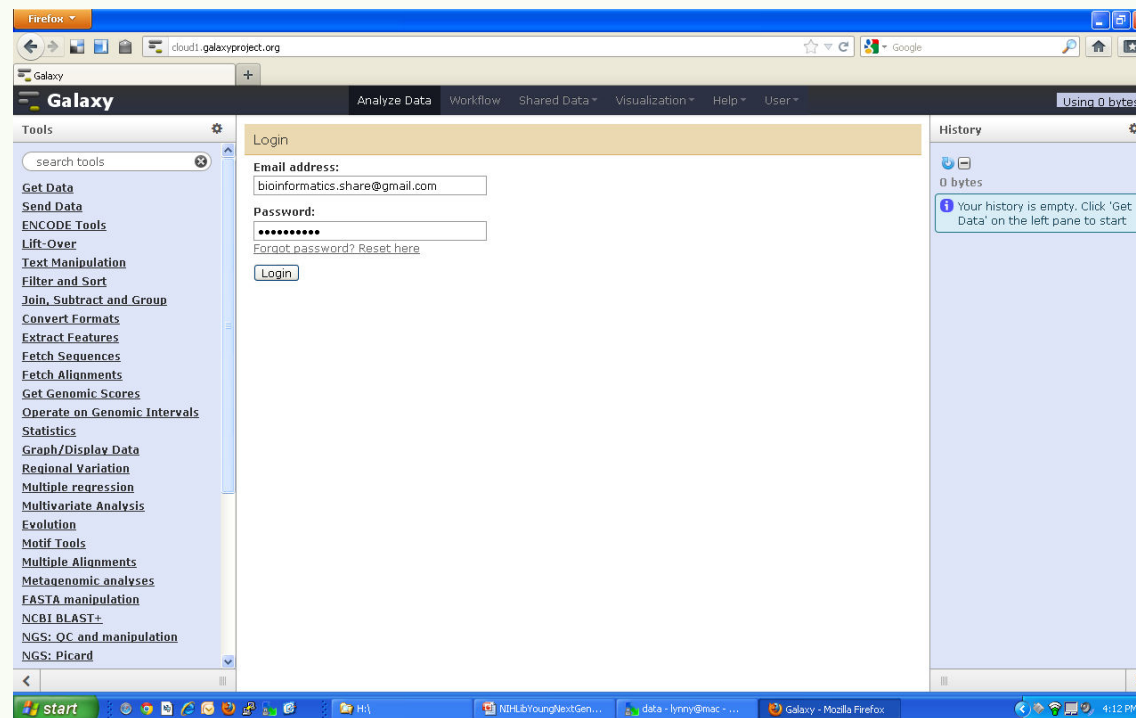
2:49 PM



Galaxy Login if Already Have Account



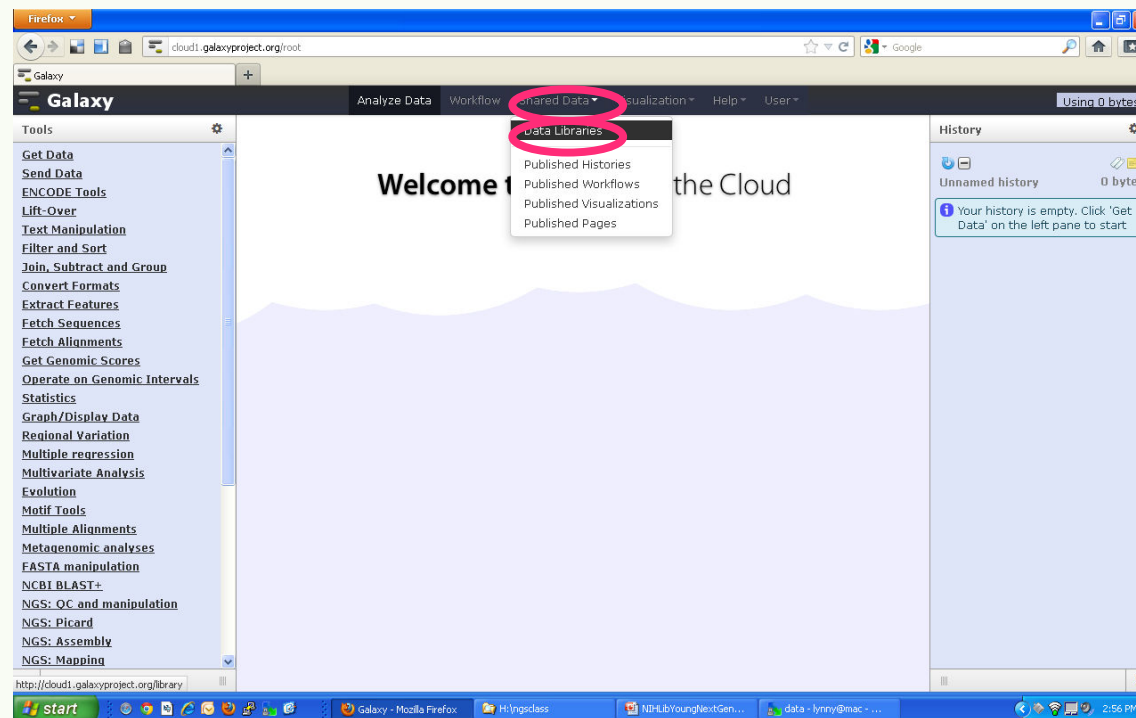
Galaxy Login



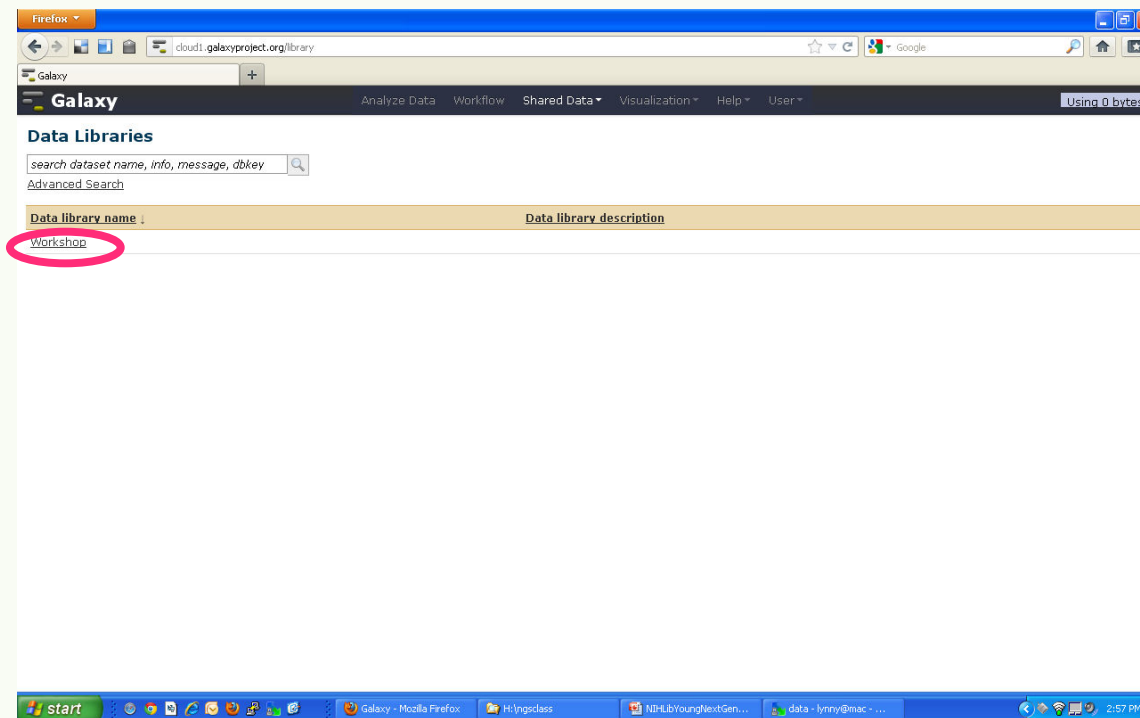
An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Shared Data



Galaxy – Obtain Shared Data



Galaxy – Obtain Share Data for Input Datasets

Step 1

name	Message	Data type	Date uploaded	File size
☒ SRR016861srtid27to28M.fastq		fastq	2012-09-18	2.3 MB
☒ SRR016865srtid27to28M.fastq		fastq	2012-09-18	3.9 MB
☒ chr21.fa		fasta	2012-09-18	46.6 MB
☒ SRR016862srtid27to28M.fastq		fastq	2012-09-18	3.2 MB

For selected datasets:

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

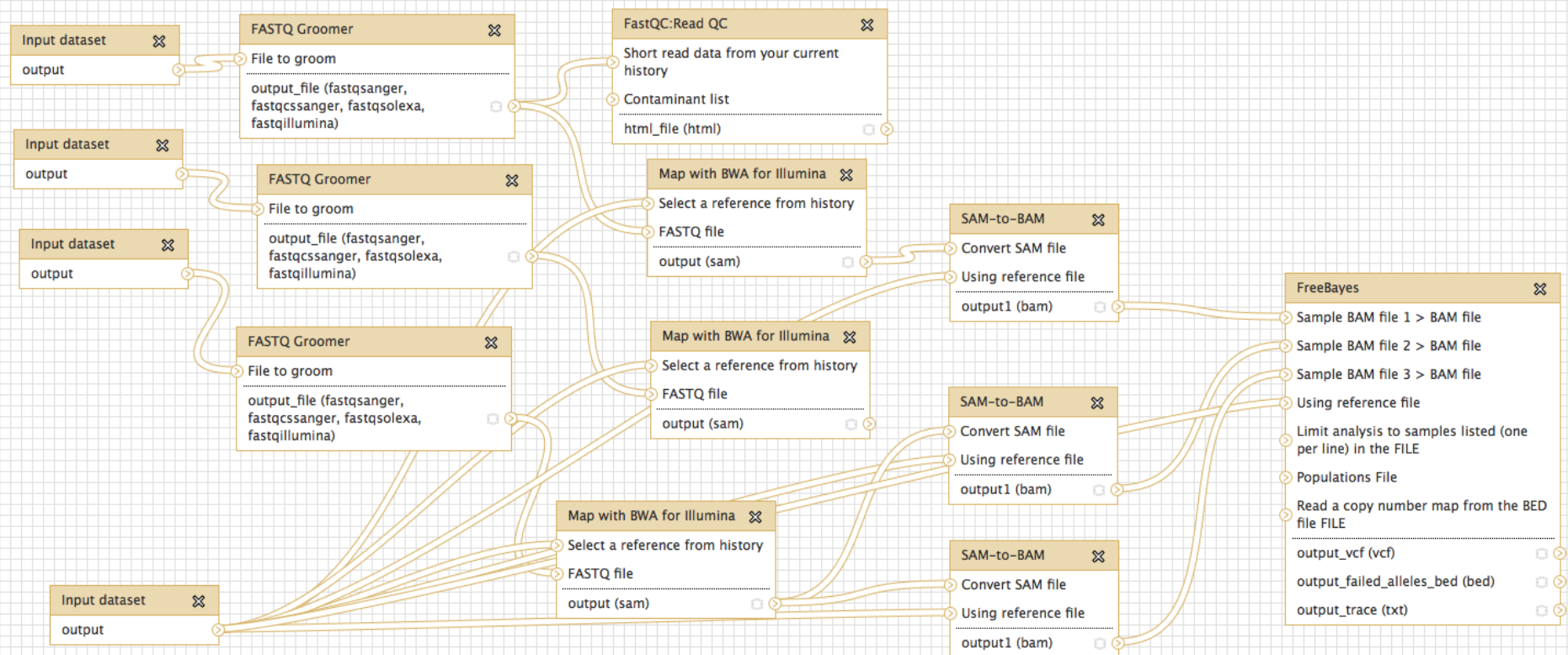
Step 2



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy Data Analysis Workflow - Details



Galaxy – Analyze Data

Firefox

cloud1.galaxyproject.org/library

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 0 bytes

Data Library "Workshop"

4 datasets imported into 1 history: Unnamed history

Name	Message	Data type	Date uploaded	File size
☐ SRR016861srtdd27to28M.fastq		fastq	2012-09-18	2.3 MB
☐ SRR016865srtdd27to28M.fastq		fastq	2012-09-18	3.9 MB
☐ chr21.fa		fasta	2012-09-18	46.6 MB
☐ SRR016862srtdd27to28M.fastq		fastq	2012-09-18	3.2 MB

For selected datasets:

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats



Galaxy Input Dataset – View Sequence Reads

The screenshot displays the Galaxy web interface in a Firefox browser window. The address bar shows the URL `cloud1.galaxyproject.org/root`. The interface includes a top navigation bar with tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A left sidebar lists various tools under categories like 'Get Data', 'Text Manipulation', 'Filter and Sort', etc. The main panel shows a list of sequence reads with a warning message: 'This dataset is large and only the first megabyte is shown below. Show all | Save'. The reads are displayed in a table format with columns for ID, sequence, and quality. The right sidebar shows a 'History' panel with a list of datasets, including 'SRR016861.18744', 'SRR016861.62965', 'SRR016861.115405', 'SRR016861.210037', 'SRR016861.393052', 'SRR016861.516452', 'SRR016861.526448', 'SRR016861.662818', 'SRR016861.666840', and 'SRR016861.666840'. A red circle highlights the 'SRR016861.666840' dataset in the history panel.



Galaxy Input Dataset – View Reference Sequence

The screenshot shows the Galaxy web interface in a Firefox browser. The address bar displays 'cloud1.galaxyproject.org/root'. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A yellow warning banner at the top states: 'This dataset is large and only the first megabyte is shown below. Show all | Save'. The left sidebar contains a 'Tools' menu with categories like 'Get Data', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', 'FASTA manipulation', 'NGS: BLAST', 'NGS: QC and manipulation' (circled in red), 'NGS: Read', 'NGS: Assembly', and 'NGS: Mapping'. The main content area displays a reference sequence for 'chr21' as a long string of 'N' characters. The right sidebar shows a 'History' panel with a list of datasets: '4: SRR016862srt27to28M.fastq', '3: chr21.fa' (circled in red), '2: SRR016865srt27to28M.fastq', and '1: SRR016861srt27to28M.fastq'. The bottom status bar shows the system clock as 3:17 PM.

Reference
Sequence

For next
slide



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy - FASTQ Groomer

Step 1

Step 2

Step 3

Repeat steps for the other two FASTQ files

The screenshot shows the Galaxy web interface for the FASTQ Groomer tool (version 1.0.4). The left sidebar lists various tools under 'Metagenomic analyses', with 'FASTQ Groomer' selected. The main panel displays the tool's configuration: 'File to groom' is set to '1: SRR016061srtdd27to28M.fastq', 'Input FASTQ quality scores type' is 'Sanger', and 'Advanced Options' is 'Hide Advanced Options'. The 'Execute' button is highlighted. The right sidebar shows a 'History' panel with three entries: '4: SRR016062srtdd27to28M.fastq', '3: chr21.fa', and '2: SRR016065srtdd27to28M.fastq'. A red box highlights the 'Execute' button and the 'File to groom' dropdown. A red arrow points from the 'Repeat steps for the other two FASTQ files' text to the 'History' panel.



Galaxy - FastQC

Step 4

Step 1

Step 2

Step 3

The screenshot shows the Galaxy web interface in a Firefox browser. The left sidebar lists various tools under 'Metagenomic analyses'. The central panel shows the 'FastQC:Read QC (version 0.51)' tool configuration. The right sidebar shows a history list with several entries. Red circles and arrows highlight the following steps:

- Step 1:** Points to 'FastQC:Read QC reports using FastQC' in the left sidebar.
- Step 2:** Points to the 'Short read data from your recent history' dropdown menu.
- Step 3:** Points to the 'Execute' button.
- Step 4:** Points to 'FastQC:Read QC reports using FastQC' in the left sidebar.



Galaxy – FastQC Results

For next
slide

The screenshot displays the Galaxy web interface. The left sidebar lists various tool categories, with 'NGS: Mapping' circled in red. The central panel shows the 'FASTQ_Groomer_on_data_1' FastQC Report, dated 'Tue 18 Sep 2012'. The report includes a 'Summary' section with a list of metrics, each preceded by a green checkmark or a red X. The right sidebar shows a 'History' list with several items, including '8: FastQC_FASTQ_Groomer_on_data_1.html', which is circled in red. A red arrow points from the 'Step 2' label to the history list, and another red arrow points from the 'Step 1' label to the '8: FastQC_FASTQ_Groomer_on_data_1.html' item.

Tools

- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard
- NGS: Assembly
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools
- NGS: Peak Calling
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- VCF Tools
- Workflows
- All workflows

FASTQ_Groomer_on_data_1 FastQC Report

FASTQC Report

Tue 18 Sep 2012

FASTQ_Groomer_on_data_1

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

History

- Unnamed history 10.3 MB
- 8: FastQC_FASTQ_Groomer_on_data_1.html
- 7: FASTQ_Groomer_on_data_4
- 6: FASTQ_Groomer_on_data_2
- 5: FASTQ_Groomer_on_data_1
- 4: SRR016862srt27to28M.fastq
- 3: chr21.fa
- 2: SRR016865srt27to28M.fastq
- 1: SRR016861srt27to28M.fastq

Step 2

Step 1



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy Mapping – Burris Wheeler Aligner (BWA)

Step 1 (circled in red) points to the tool selection in the left sidebar.

Step 2 (circled in red) points to the dropdown menu for selecting a reference genome.

Step 3 (circled in red) points to the FASTQ file selection dropdown.

Step 4 (circled in red) points to the Execute button.

Repeat steps for the other two Groomed files (indicated by arrows pointing to the history list on the right).



Galaxy – View BWA Results

Tools

- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard
- NGS: Assembly
- NGS: Mapping
 - Lastz map short reads against reference sequence
 - Lastz paired reads map short paired reads against reference sequence
 - Map with Bowtie for Illumina
 - Map with Bowtie for SOLiD
 - Map with BWA for Illumina
 - Meqablast compare short reads against htgs, nt, and wgs databases
 - Parse blast XML output
 - Map with Perl for SOLiD and Illumina
 - Map with Mosaik
- NGS: Indel Analysis
- NGS: SAM Tools
- NGS: GATK Tools
- NGS: Peak Calling

QNAME **FLAG** **RNAME** **POS** **MAPQ** **CIGAR** **MRNM** **POS** **SIZE** **SEQ**

Q@SQ SN:chr21 LN:48129895

SRR016861.18744	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.62965	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.115405	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.210037	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.393052	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGATACATCTAGGT
SRR016861.516452	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.526448	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.662818	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.666840	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.691394	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.709457	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.731252	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.939077	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.961899	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1028652	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1048271	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1096090	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1124174	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1134331	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1151818	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1276814	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.1354301	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGT
SRR016861.110466	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG
SRR016861.275426	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG
SRR016861.341934	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG
SRR016861.411273	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG
SRR016861.448301	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG
SRR016861.453926	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG
SRR016861.495820	0	chr21	27006294	37	41M	*	0	0	AAAAGAAGTTAGGTAATTGGG

History

- 11: Map with BWA for Illumina on data 7 and data 3: mapped reads
- 10: Map with BWA for Illumina on data 6 and data 3: mapped reads
- 9: Map with BWA for Illumina on data 5 and data 3: mapped reads
- 8: FastQC FASTQ Groomer on data 1.html
- 7: FASTQ Groomer on data 4
- 6: FASTQ Groomer on data 2
- 5: FASTQ Groomer on data 1
- 4: SRR016862.rtd27to28M.fasta
- 3: chr21.fa
- 2: SRR016865.rtd27to28M.fasta

For next slide

Step 2

Step 1



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – SAM to BAM

The screenshot shows the Galaxy web interface in a Firefox browser window. The main panel displays the 'SAM-to-BAM (version 1.1.2)' tool. Annotations with red circles and text labels indicate the following steps:

- Step 1:** Points to the 'Tools' sidebar on the left, specifically to the 'SAM Tools' section.
- Step 2:** Points to the 'Choose the source for the reads 1:' dropdown menu, which is set to 'History'.
- Step 3:** Points to the '9: Map with BWA for mapped reads' dropdown menu.
- Step 4:** Points to the 'Execute' button at the bottom of the tool configuration panel.

On the right side, the 'History' panel shows a list of previous jobs. A red box with arrows points to the first two entries, with the text: 'Repeat steps for the other two SAM files'.

What it does
This tool uses the [SAMtools](#) toolkit to produce an indexed BAM file based on a sorted input SAM file.

Citation
For the underlying tool, please cite [Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25\(16\):2078-9.](#)

Galaxy – Navigation to Picard Alignment Summary Metrics

Step 1

Step 2

The screenshot displays the Galaxy web interface in a Firefox browser window. The address bar shows 'cloud1.galaxyproject.org'. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A status bar at the top right indicates 'Using 29.7 MB'. On the left, the 'Tools' panel is expanded, showing a list of tools under various categories. The 'NGS: SAM Tools' category is selected, and the 'Convert SAM to interval' tool is highlighted. A red circle and an arrow point to this tool, labeled 'Step 1'. In the center, a green notification box with a checkmark states: 'The following job has been successfully added to the queue: 14: SAM-to-BAM on data 3 and data 11: converted BAM. You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' A red arrow points from the 'Tools' panel to this notification, labeled 'Step 2'. On the right, the 'History' panel shows a list of previous jobs, including '14: SAM-to-BAM on data 3 and data 11: converted BAM' at the top.



Galaxy – Picard Alignment Summary Metrics

Step 1 NGS: Picard

Step 2 SAM/BAM Alignment Summary Metrics

Step 3 12: SAM-to-BAM on data 11: converted BAM

Step 4 Use a genome (fasta format) from my history

Step 5 – uncheck the box Assume the input file is already sorted

Step 6 Execute



Galaxy – Results of Picard Summary Alignment Metrics

For next slide

Tools

- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard
- NGS: SAM/BAM
- BAM Index Statistics
- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data
- BAM/SAM CLEANING
- Add or Replace Groups
- Reorder SAM/BAM
- Replace SAM/BAM Header
- Paired Read Mate Fixer for paired data
- Mark Duplicate reads
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis

CATEGORY	UNPAIRED
TOTAL_READS	22690
PF_READS	22690
PCT_PF_READS	1
PF_NOISE_READS	0
PF_READS_ALIGNED	22547
PCT_PF_READS_ALIGNED	0.993698
PF_ALIGNED_BASES	924010
PF_HQ_ALIGNED_READS	22276
PF_HQ_ALIGNED_BASES	912900
PF_HQ_ALIGNED_Q20_BASES	890110
PF_HQ_MEDIAN_MISMATCHES	1
PF_MISMATCH_RATE	0.018881
PF_HQ_ERROR_RATE	0.01881
PF_INDEL_RATE	0.001411
MEAN_READ_LENGTH	41
READS_ALIGNED_IN_PAIRS	0

History

- 15: Picard Alignment Summary Metrics.html
- 14: SAM-to-BAM on data 3 and data 11: converted BAM
- 13: SAM-to-BAM on data 3 and data 10: converted BAM
- 12: SAM-to-BAM on data 3 and data 9: converted BAM
- 11: Map with BWA for Illumina on data 7 and data 3: mapped reads
- 10: Map with BWA for Illumina on data 6 and data 3: mapped reads
- 9: Map with BWA for Illumina on data 5 and data 3: mapped reads
- 8: FASTQ to BAM on data 1.html
- 7: Zoomer on

Step 2

Step 1

PF_HQ_ALIGNED_READS: The number of PF reads that were aligned to the reference sequence with a mapping quality of Q20 or higher signifying that the aligner estimates a 1/100 (or smaller) chance that the alignment is wrong.

PF_HQ_ALIGNED_BASES: The number of bases aligned to the reference sequence in reads that were mapped at high quality. Will usually approximate $PF_HQ_ALIGNED_READS * READ_LENGTH$ but may differ when either mixed read lengths are present or many reads are aligned with gaps.

PF_HQ_ALIGNED_Q20_BASES: The subset of PF_HQ_ALIGNED_BASES where the base call quality was Q20 or higher.



Key - <http://picard.sourceforge.net/picard-metric-definitions.shtml>

Galaxy Variant Detection – Preparation Merging Bam Files

Step 1 NGS: SAM Tools

Step 2 Merge BAM Files merges BAM

Step 3 Name for the output merged bam file:
mergedBams

Step 4 Merge all component bam file headers into the merged bam file:

Step 5 Input Files

Step 6 Add new Input Files

History:

- 18: MergedBams Merge BAM Files.log
- 17: MergedBams.bam
- 15: Picard Alignment Summary Metrics.html
- 14: SAM-to-BAM on data 3 and data 11: converted BAM
- 13: SAM-to-BAM on data 3 and data 10: converted BAM
- 12: SAM-to-BAM on data 3 and data 9: converted BAM
- 11: Map with BWA for Illumina on data 7 and data 3: mapped reads
- 10: Map with BWA for Illumina on data 6 and data 3: mapped reads
- 9: Map with BWA for Illumina on data 5 and data 3: mapped reads



Galaxy Variant Detection – Preparation Merging Bam Files

The screenshot shows the Galaxy web interface with the 'Merge BAM Files (version 1.1.2)' tool selected. The tool configuration is as follows:

- Name for the output merged bam file:** MergedBam
- Merge all component bam file headers into the merged bam file:** ☒
- First file:** 12: SAM-to-BAM on dat..nverted BAM
- with file:** 13: SAM-to-BAM on dat..nverted BAM
- Input Files:** 14: SAM-to-BAM on dat..nverted BAM
- Buttons:** 'Add file' (Step 1) and 'Execute' (Step 2) are highlighted with red circles.

The 'History' panel on the right shows a list of previous jobs, including '18: MergedBams Merge BAM Files.log', '17: MergedBams.bam', '15: Picard Alignment Summary Metrics.html', '14: SAM-to-BAM on data 3 and data 11: converted BAM', '13: SAM-to-BAM on data 3 and data 10: converted BAM', '12: SAM-to-BAM on data 3 and data 9: converted BAM', '11: Map with BWA for Illumina on data 7 and data 3: mapped reads', '10: Map with BWA for Illumina on data 6 and data 3: mapped reads', and '9: Map with BWA for Illumina on data 5 and data 3: mapped reads'.



Galaxy Variant Detection - FreeBayes

Step 2
Step 3

Step 4

Step 5

Step 6

Step 1

Galaxy (version 0.0.3)

Tools

FreeBayes

FreeBayes - Bayesian genetic variant detector

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

NCBI BLAST+

NGS: QC and manipulation

NGS: Picard

Choose the source for the reference list:

History

Sample BAM file 1

BAM file:

17: MergedBams.bam

Remove Sample BAM file 1

Add new Sample BAM file

Using reference file:

3: chr21.fa

Basic or Advanced options:

Basic

Execute

What it does

This tool uses FreeBayes to call SNPS given a reference sequence and a BAM alignment file.

FreeBayes is a high-performance, flexible, and open-source Bayesian genetic variant detector. It operates on BAM alignment files, which are produced by most contemporary short-read aligners.

In addition to substantial performance improvements over its predecessors (PolyBayes, GigaBayes, and BamBayes), it expands the scope of SNP and small-indel variant calling to populations of individuals with heterogeneous copy number. FreeBayes is currently under active development.

Go [here](#) for details on FreeBayes.

History

Unnamed history 34.1 MB

18: MergedBams Merge BAM Files.log

17: MergedBams.bam

15: Picard Alignment Summary Metrics.html

14: SAM-to-BAM on data 3 and data 11: converted BAM

13: SAM-to-BAM on data 3 and data 10: converted BAM

12: SAM-to-BAM on data 3 and data 9: converted BAM

11: Map with BWA for Illumina on data 7 and data 3: mapped reads

10: Map with BWA for Illumina on data 6 and data 3: mapped reads

9: Map with BWA for Illumina on data 5 and data 3: mapped reads



Galaxy Variant Detection – FreeBayes Results

For next
slide

Galaxy

Tools

FreeBayes

FreeBayes - Bayesian genetic variant detector

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Split, Merge and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

History

20: FreeBayes on data 3 and data 17 (variants)

18: MergedBams_Merge BAM Files.log

17: MergedBams.bam

15: Picard Alignment Summary Metrics.html

14: SAM-to-BAM on data 3 and data 11: converted BAM

13: SAM-to-BAM on data 3 and data 10: converted BAM

12: SAM-to-BAM on data 3 and data 9: converted BAM

11: Map with BWA for Illumina on data 7 and data 3: mapped reads

10: Map with BWA for Illumina on data 6 and data 3: mapped reads

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr21	10161051	.	T	G	0.333095	.	AB=0.5;ABP=3.0103;AC=1;AF=0
chr21	10161084	.	T	C	0.233552	.	AB=0.5;ABP=3.0103;AC=1;AF=0
chr21	10161086	.	T	G	2.13682	.	AB=0.5;ABP=3.0103;AC=1;AF=0
chr21	10161088	.	T	G	12.3467	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	10163513	.	A	C	0.135209	.	AB=0.25;ABP=5.18177;AC=1;AF
chr21	10163524	.	A	T	2.42891	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	10576806	.	G	A	4.82183	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	10576811	.	G	T	7.86191	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	11000769	.	A	G	0.00055001	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000771	.	A	G	0.000871674	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000781	.	A	C	0.00069241	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000782	.	A	C	0.00055001	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000785	.	A	G	0.0116961	.	AB=0.153846;ABP=16.5402;AC=
chr21	11000786	.	A	C	0.00357939	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000787	.	A	C	0.00055001	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000788	.	A	G	0.00171981	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000790	.	A	C	0.00055001	.	AB=0.0769231;ABP=23.2217;AC
chr21	11000791	.	GACAA	AACAG	7.42863	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	11000798	.	G	A	7.08188	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	11006534	.	T	A	29.0914	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	14402593	.	T	G	3.0961	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	14402598	.	T	G	2.61507	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	14402619	.	G	T	20.978	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	15329332	.	T	G	5.51709	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	15329338	.	TTCT	CTCTG	4.17736	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	15437995	.	A	C	4.17736	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	15438002	.	T	C	4.17736	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A
chr21	15478529	.	T	A	5.51709	.	AB=0;ABP=0;AC=2;AF=1;AN=2;A

Step 1

Step 2



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Filter and Sort

Step 1 Filter and Sort

Step 2 Filter data on any column using simple expressions

Step 3 20: FreeBayes on data.. (variants)

Step 4 c6

Step 5 Execute

Sort (version 1.0.1)

with flavor: Numerical sort

everything in: Descending order

Column selections

Add new Column selection

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

Syntax

This tool sorts the dataset on any number of columns in either ascending or descending order.

Numerical sort orders numbers by their magnitude, ignores all characters besides numbers, and evaluates a string of numbers to the value they signify.

Alphabetical sort is a phonebook type sort based on the conventional order of letters in an alphabet. Each nth letter is compared with the nth letter of other words in the list, starting at the first letter of each word and advancing to the second, third, fourth, and so on, until the order is established. Therefore, in an alphabetical sort, 2 comes after 100 (1 < 2).

Examples

The list of numbers 4,17,3,5 collates to 3,4,5,17 by numerical sorting, while it collates to 17,3,4,5 by alphabetical

History

Unnamed history 35.1 MB

- 20: FreeBayes on data 3 and data 17 (variants)
- 18: MergedBams_Merge BAM Files.log
- 17: MergedBams.bam
- 15: Picard Alignment Summary Metrics.html
- 14: SAM-to-BAM on data 3 and data 11: converted BAM
- 13: SAM-to-BAM on data 3 and data 10: converted BAM
- 12: SAM-to-BAM on data 3 and data 9: converted BAM
- 11: Map with BWA for Illumina on data 7 and data 3: mapped reads
- 10: Map with BWA for Illumina on data 6 and data 3: mapped reads



Galaxy – Filter and Sort Results

For next slide, open new tab

The screenshot shows the Galaxy web interface in a Firefox browser. The address bar shows 'cloud1.galaxyproject.org'. The Galaxy logo is in the top left, and the top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A red circle highlights the '+' icon in the browser's tab bar. The main content area displays a table of genomic data with columns: Chrom, Pos, ID, Ref, Alt, Qual, and FilterInfo. The table lists various genomic features across different chromosomes (chr21, chr2, chr1). The right sidebar shows a 'History' panel with a list of recent jobs, including '21: Sort on data 20', '20: FreeBayes on data 3 and data 17 (variants)', '18: MergedBams Merge BAM Files.log', '17: MergedBams.bam', '15: Picard Alignment Summary Metrics.html', '14: SAM-to-BAM on data 3 and data 11: converted BAM', '13: SAM-to-BAM on data 3 and data 10: converted BAM', '12: SAM-to-BAM on data 3 and data 9: converted BAM', '11: Map with BWA for Illumina on data 7 and data 3: mapped reads', and '10: Map with BWA for Illumina on data 6 and data 3: mapped reads'.

Chrom	Pos	ID	Ref	Alt	Qual	FilterInfo
chr21	27219532	.	C	T	21763.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=682;CIGAR=1X;DP=...
chr21	27259085	.	C	CA	17248.6	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=172;CIGAR=1M11;DP=...
chr21	27845630	.	CAGCT	GAGC	14718.6	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=442;CIGAR=1X3M1D;...
chr21	27262442	.	T	A	13850.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=425;CIGAR=1X;DP=4...
chr21	27497050	.	A	C	13841.9	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=452;CIGAR=1X;DP=4...
chr21	27885242	.	A	T	13147.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=413;CIGAR=1X;DP=4...
chr21	27441181	.	T	C	13115.6	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=412;CIGAR=1X;DP=4...
chr21	27285013	.	CC	AG	13095.9	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=365;CIGAR=2X;DP=3...
chr21	27397951	.	A	C	11978.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=386;CIGAR=1X;DP=3...
chr21	27595448	.	G	T	11683.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=358;CIGAR=1X;DP=3...
chr21	27300235	.	AG	GC	11006.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=306;CIGAR=2X;DP=3...
chr21	27909060	.	CA	GG	10944.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=304;CIGAR=2X;DP=3...
chr21	27091130	.	T	G	10845.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=342;CIGAR=1X;DP=3...
chr21	27487642	.	TC	T	10311.6	AB=0.542757;ABP=12.1093;AC=1;AF=0.5;AN=2;AO=311...
chr21	27981703	.	A	T	10114.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=330;CIGAR=1X;DP=3...
chr21	27010825	.	CA	AC	9386.72	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=319;CIGAR=2X;DP=3...
chr21	27011976	.	G	A	9083.43	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=282;CIGAR=1X;DP=2...
chr21	27088627	.	G	A	9016.61	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=287;CIGAR=1X;DP=2...
chr21	27485906	.	TG	GA	8640.74	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=240;CIGAR=2X;DP=2...
chr21	27354648	.	C	A	8319.71	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=259;CIGAR=1X;DP=2...
chr21	27170423	.	G	C	8293.89	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=231;CIGAR=1X;DP=2...
chr21	27223146	.	TC	CT	8236.69	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=272;CIGAR=2X;DP=2...
chr21	27327510	.	C	CG	8145.41	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=79;CIGAR=1M11;DP=...
chr21	27268246	.	GG	AA	7954.14	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=282;CIGAR=2X;DP=2...
chr21	27256952	.	GC	TA	7834.49	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=218;CIGAR=2X;DP=2...
chr21	27810982	.	TA	CC	7574.39	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=211;CIGAR=2X;DP=2...
chr21	27359041	.	GG	TA	7276	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=202;CIGAR=2X;DP=2...
chr21	272020730	.	C	A	7210.56	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=225;CIGAR=1X;DP=2...
chr21	27945683	.	AA	CG	7200.53	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=200;CIGAR=2X;DP=2...
chr21	27096942	.	TTTA	CAT	7146.39	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=199;CIGAR=2X1M1D;...



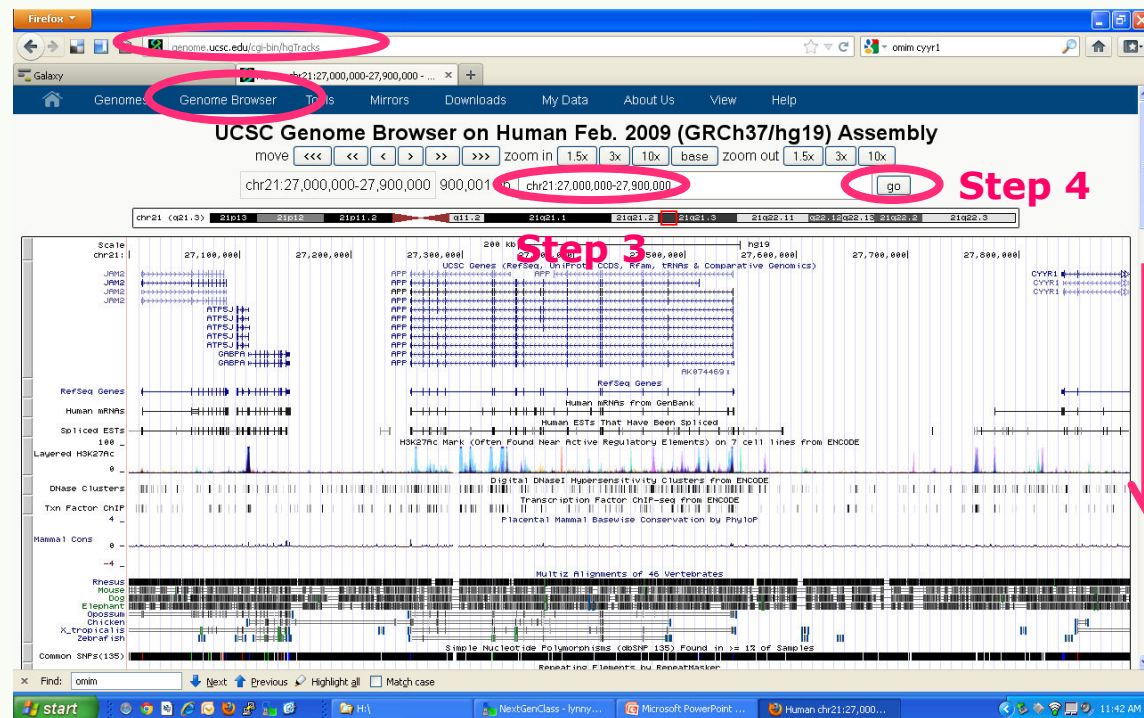
An ORS Service

Next Generation Sequencing Data Analysis

Biological Context UCSC Genome Browser <http://genome.ucsc.edu>

Step 1

Step 2



Step 4

Step 5



An ORS Service

Next Generation Sequencing Data Analysis

UCSC Genome Browser – OMIM Genes, OMIM AV SNPs

Step 1

Step 2

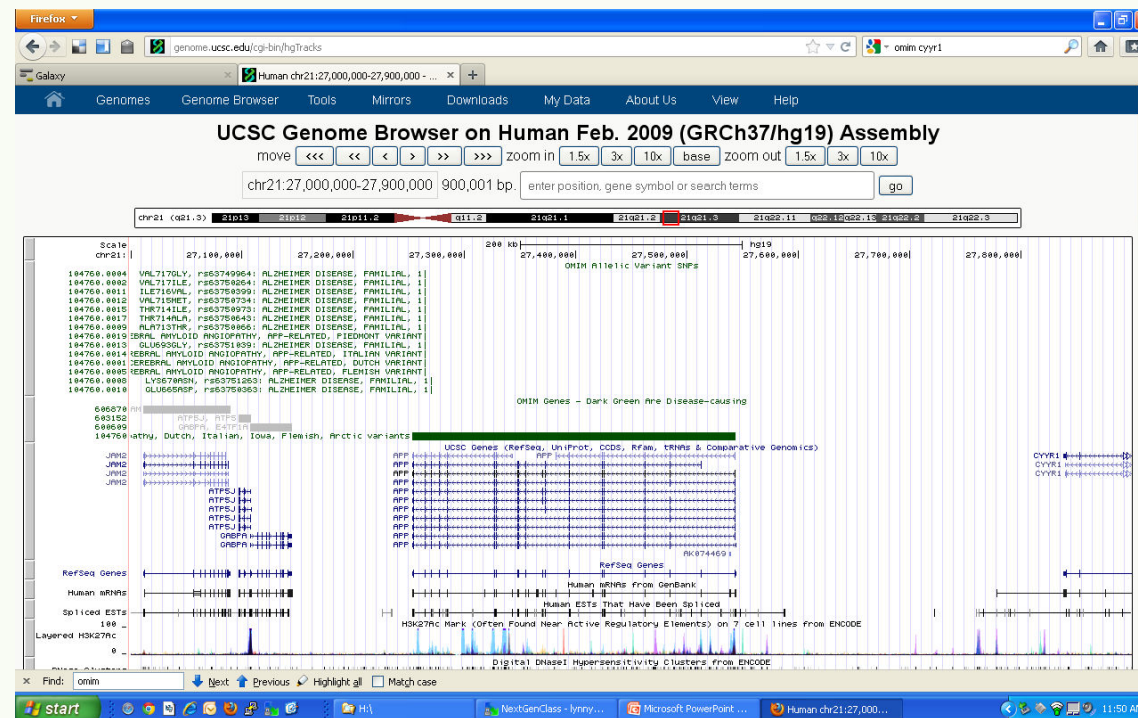
Step 3

Step 4

UCSC Genome Browser interface showing the OMIM track. The interface includes a search bar, track controls, and various genomic tracks. Annotations highlight specific steps: Step 1 points to the 'track search' button; Step 2 points to the 'OMIM AV SNPs' dropdown menu; Step 3 points to the 'full' dropdown menu; and Step 4 points to the 'refresh' button.



UCSC Genome Browser - Results



Galaxy – Exporting Data Download VCF File

The screenshot shows the Galaxy web interface with a table of genomic data. The 'Tools' sidebar on the left lists various analysis tools. The central panel displays a table with columns: #CHROM, POS, ID, REF, ALT, QUAL, FILTER, and INFO. The 'History' panel on the right shows a list of jobs, with '20: FreeBayes on data 3 and data 17 (variants)' highlighted. A dialog box is open in the center, asking 'What should Firefox do with this file?'. The dialog has two options: 'Open with' (selected) and 'Save File'. The 'Open with' option has a dropdown menu showing 'Microsoft Office Outlook (default)'. The 'Save File' option is also visible. The 'OK' button is highlighted with a red circle.

Step 1: Select the file in the History panel.

Step 2: Click the file icon in the History panel.

Step 3: Select the 'Open with' option in the dialog box.

Step 4: Click the 'OK' button in the dialog box.



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Exporting Data Download BAM Files

Galaxy interface showing the process of downloading BAM files. The interface includes a 'Tools' sidebar, a central data table, and a 'History' panel on the right. A dialog box is open in the center, asking 'What should Firefox do with this file?' with options to 'Open with', 'Save File', or 'Do this automatically'. Red annotations highlight specific steps: Step 1 points to the 'Download Dataset' button in the History panel; Step 2 points to the 'Download bam_index' button; Step 3 points to the 'Repeat steps for the other two BAM files' instruction; Step 4 points to the 'Open with' button in the dialog box; and Step 5 points to the 'OK' button in the dialog box.



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy - Exporting Data Download BAI Files

The screenshot shows the Galaxy web interface in a Firefox browser. The main panel displays a workflow history table with columns for job ID, name, status, and output. The workflow history table is as follows:

Job ID	Job Name	Status	Output
14	SAM-to-BAM on data 3 and data 11 converted BAM	Completed	938.9 KB format: bam, database: 2 Info: Samtools Version: 0.1.18 (r982:295) SAM file converted to BAM
13	SAM-to-BAM on data 3 and data 10 converted BAM	Completed	
12	SAM-to-BAM on data 3 and data 9 converted BAM	Completed	
11	Map with BWA for Illumina on data 7 and data 3 mapped reads	Completed	
10	Map with BWA for Illumina on data 6 and data 3 mapped reads	Completed	
9	Map with BWA for Illumina on data 5 and data 3 mapped reads	Completed	
8	FastQC FASTQ	Completed	

A dialog box titled "Opening Galaxy14 [SAM-to-BAM on data 3 and data 11 converted BAM].bai" is open, showing the file path and a "Download Dataset" button. The dialog box has a "Download Dataset" button and a "Download bam_index" button. The "Download Dataset" button is highlighted with a red circle and labeled "Step 2". The "Download bam_index" button is highlighted with a red circle and labeled "Step 3". The "Download Dataset" button is also labeled "Step 1". The "Download bam_index" button is also labeled "Step 4". The "Download Dataset" button is also labeled "Step 5".

Step 1

Step 2

Step 3

Step 4

Step 5

Repeat steps for the other two BAM files



An ORS Service

Next Generation Sequencing Data Analysis

Thank you for attending.

