

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

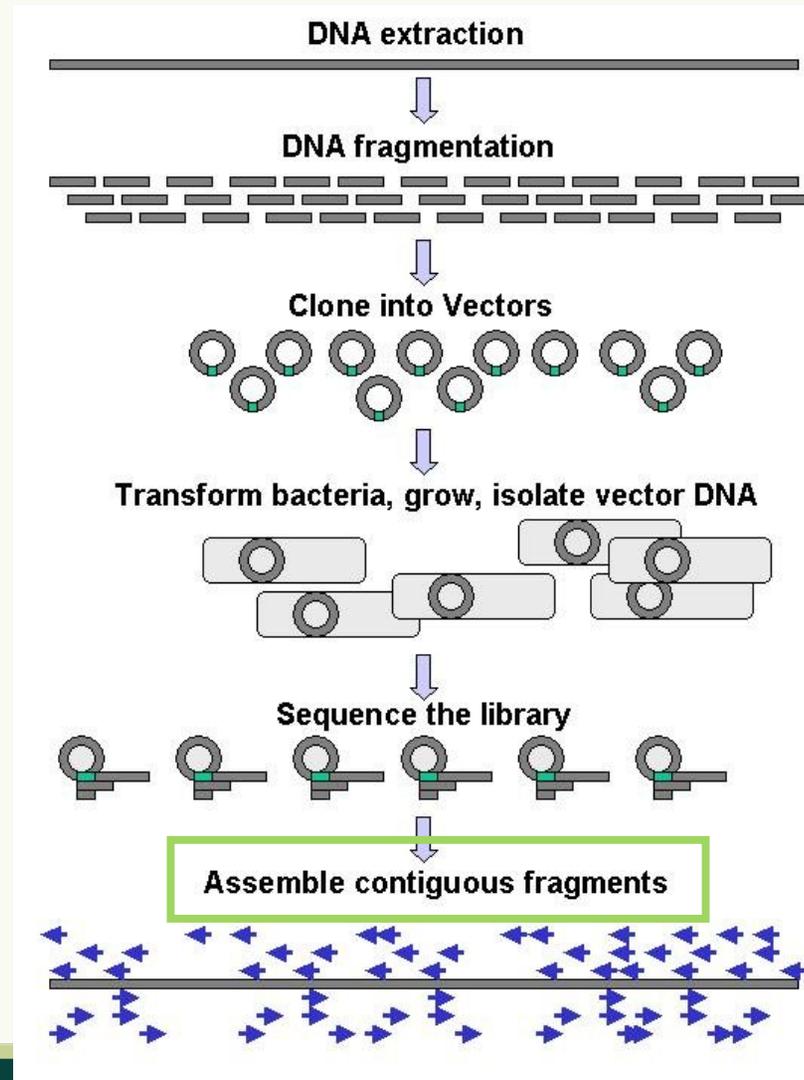
20 September 2012

Acknowledgement

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



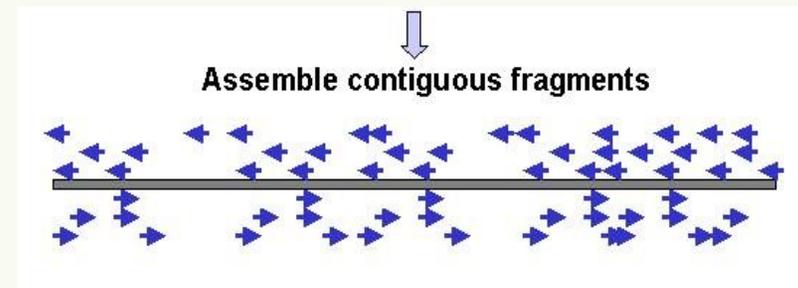
Introduction



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

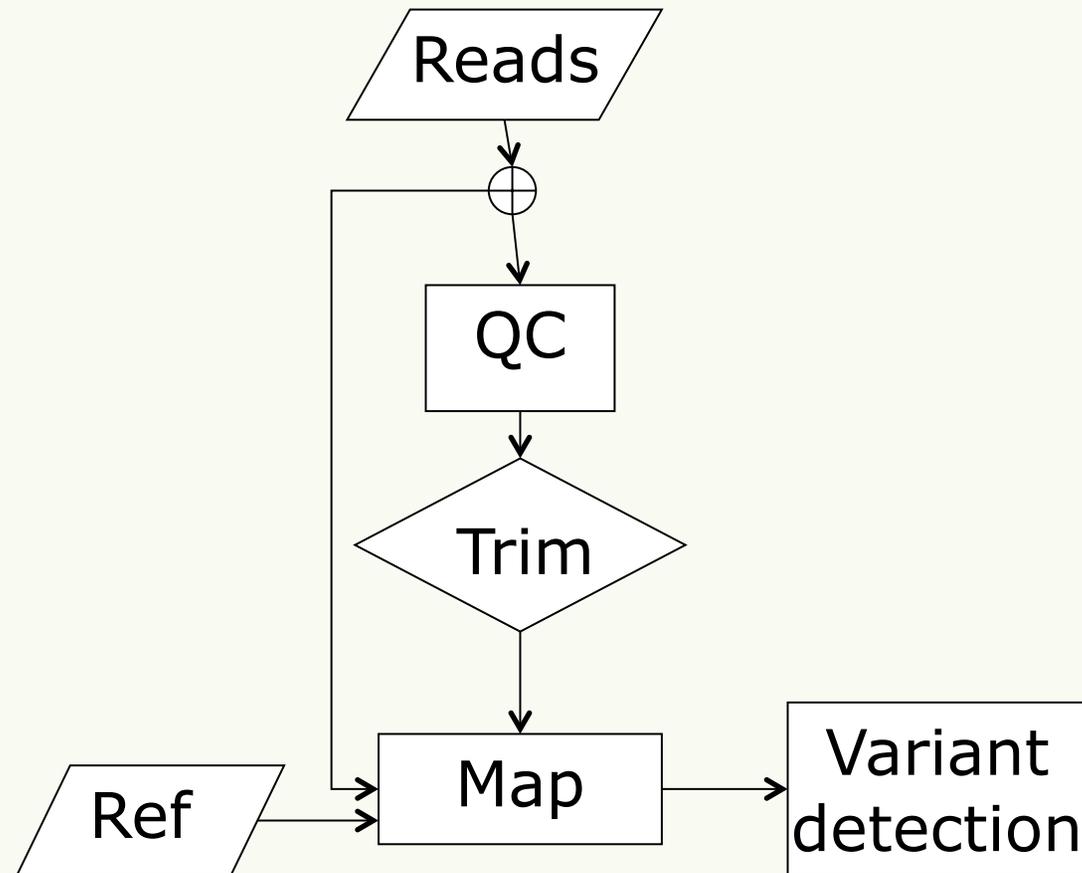
Objectives

- 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  
CTGATGTGTGCATTCATCTCACAGATGTGAACCATTCTTTTGTGGTGGTAAACATTCTTTTGTG  
TAGAATCTGCAAAGGATATTTGTGAGCACTTTGAAGCCTATGGTGAAAAAGGAAATATCTTCAGAGAAA  
AACTAGAAAGAAGGTTTCTGAGAACTGCTTTGTCATGTGTGAATTAGTCTCACAGATTTGAACCTTTCT  
GTTGATTGAACATATTGGAAACCTTCTTTTGTAGAATCTGCAAAGGGATATTTGTGAACACTTGGAGGC  
CAATGGTGAAAAAGGAAATATATTCACATGAAAACCTAGACAGAATCTTTCTGAGACACTTCTGTGTTTGG
```



Reads – FASTQ Format

@SRR016862.16884

ATTTGAGTGGTACATCTAGGTAGCCGTTTTTGGAAACGGG

+

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

@SRR016862.58801

ATTTGAGTGGTACATCTAGGTAGCCGTTTTTGAACCAGG

+

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	ATTTTGAGTGGACATCTAGGTAGCCGTTTCT				
6	SRR016862.219598	0 chr21 27002753 37 41M	*	0	0	ATTTTGAGTGGTACATCTAGGTAGCCGTTTCT				

CGTTTTTGGAAACGGG	, ? ?I& 9\$H+ > A%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
CGTTTTTGAACCCGGG	9 0 4. @&?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	A @ ; % C & 8 1(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
CGTTTCTGAACACGGG	AE@4 -H6CC*F@13+;:8\$1#(\$/'\$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 devia"
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 observat							
52	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
53	chr21	27006332	.	A	G	3.36835	.	AB=0.0833333;ABP=21.1059
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 main heading is "SRP000535 Significantly improved multiplex padlock capturing and large scale sequencing reveal hypermutable CpG variations". Below this, there are sections for "Study Type", "Submission", and "Abstract". The "Abstract" section contains a detailed description of the study. To the right, there are links for "Show Entrez docs" and "Download reads for entire study". Below these is an "Experiments" section with a table listing accession numbers, spots, and bases.

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

Experiments

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

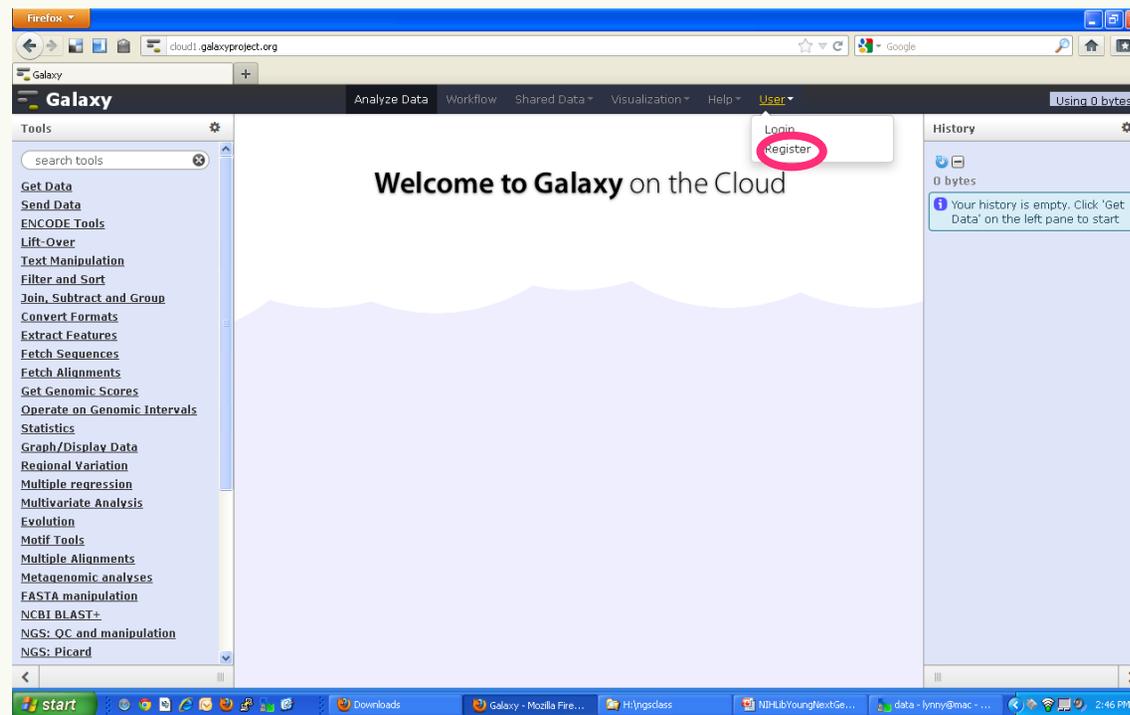
Next Generation Sequencing Data Analysis

Galaxy

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



Galaxy Account Registration



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy Account Registration

The screenshot shows the Galaxy web interface in a Firefox browser window. The address bar displays 'cloud1.galaxyproject.org'. The main content area is titled 'Create account' and contains the following form fields:

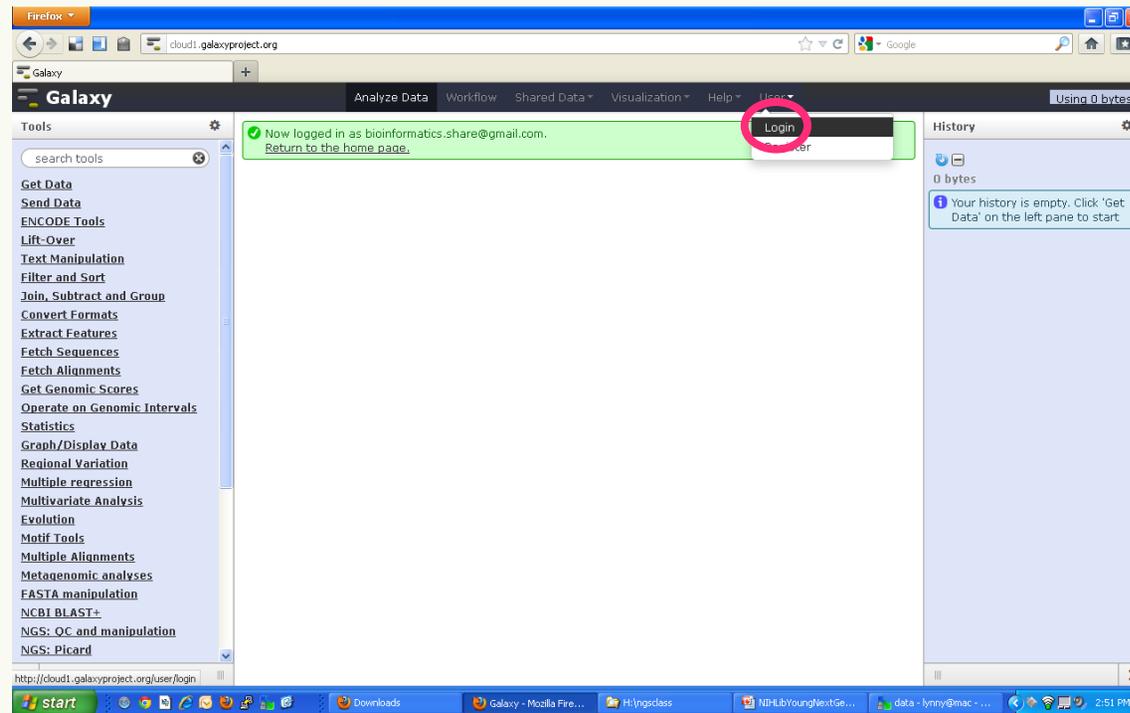
- Email address:** bioinformatics.share@gmail.com
- Password:** [masked with dots]
- Confirm password:** [masked with dots]
- Public name:** bioinformatics-share

Below the 'Public name' field, there 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.

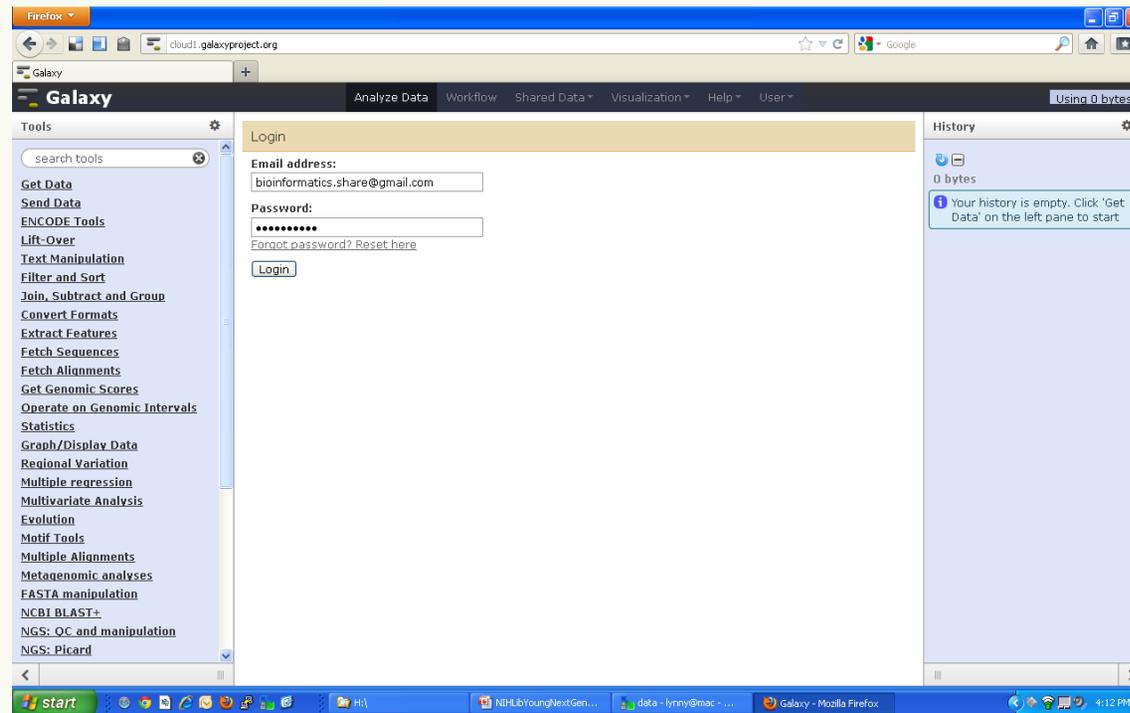
The left sidebar shows a 'Tools' menu with categories like 'Get Data', 'Send Data', 'ENCODE Tools', etc. The right sidebar shows a 'History' section with '0 bytes' and a message: "Your history is empty. Click 'Get Data' on the left pane to start."



Galaxy Login if Already Have Account



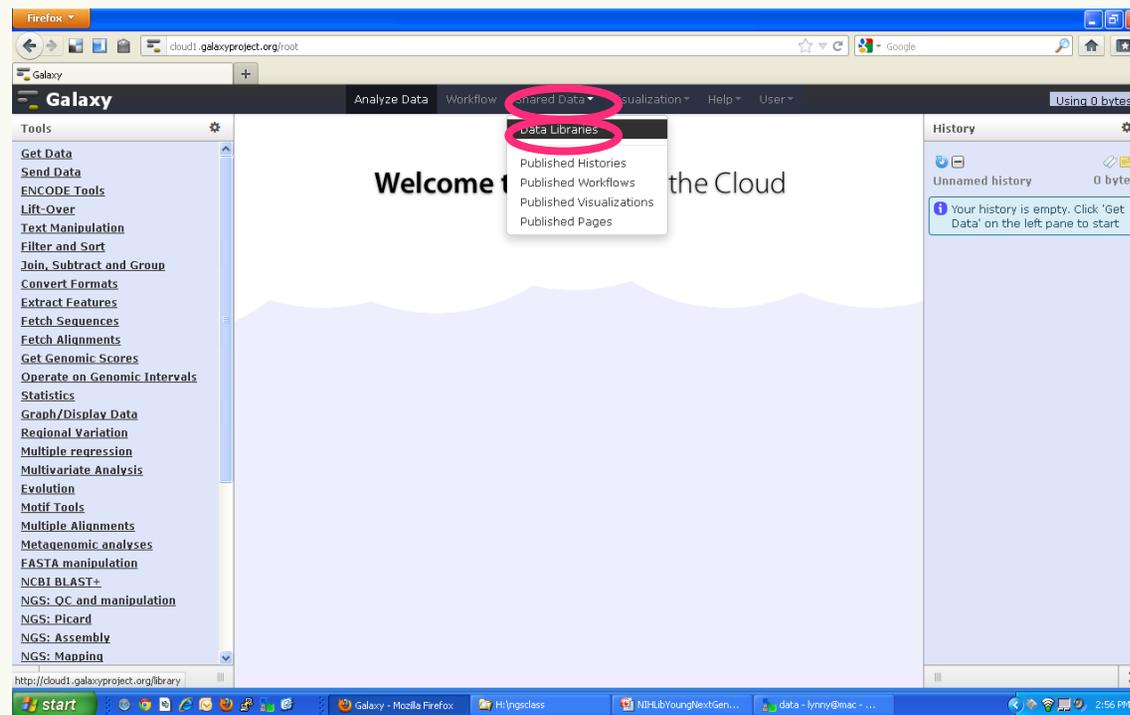
Galaxy Login



An ORS Service

Next Generation Sequencing Data Analysis

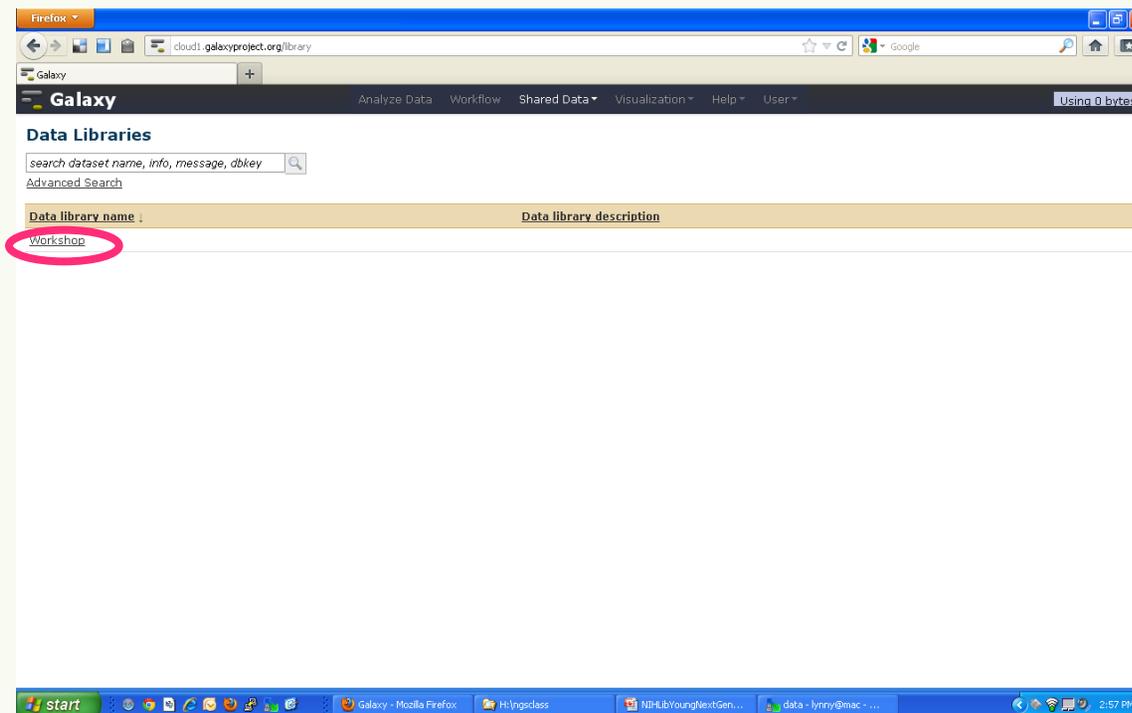
Galaxy – Shared Data



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Obtain Shared Data



The screenshot shows a web browser window displaying the Galaxy project's Data Libraries page. The browser's address bar shows the URL `cloud1.galaxyproject.org/library`. The page title is "Galaxy" and the navigation menu includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". A search bar is present with the placeholder text "search dataset name, info, message, dbkey". Below the search bar, there is a table with the following structure:

Data library name	Data library description
Workshop	

The word "Workshop" in the first row of the table is circled in red.



Galaxy – Obtain Share Data for Input Datasets

Step 1

<input checked="" type="checkbox"/>	name	Message	Data type	Date uploaded	File size
<input checked="" type="checkbox"/>	SRR016861srt27to28M.fastq		fastq	2012-09-18	2.3 MB
<input checked="" type="checkbox"/>	SRR016865srt27to28M.fastq		fastq	2012-09-18	3.9 MB
<input checked="" type="checkbox"/>	chr21.fa		fasta	2012-09-18	46.6 MB
<input checked="" type="checkbox"/>	SRR016862srt27to28M.fastq		fastq	2012-09-18	3.2 MB

For selected datasets: **Step 2**

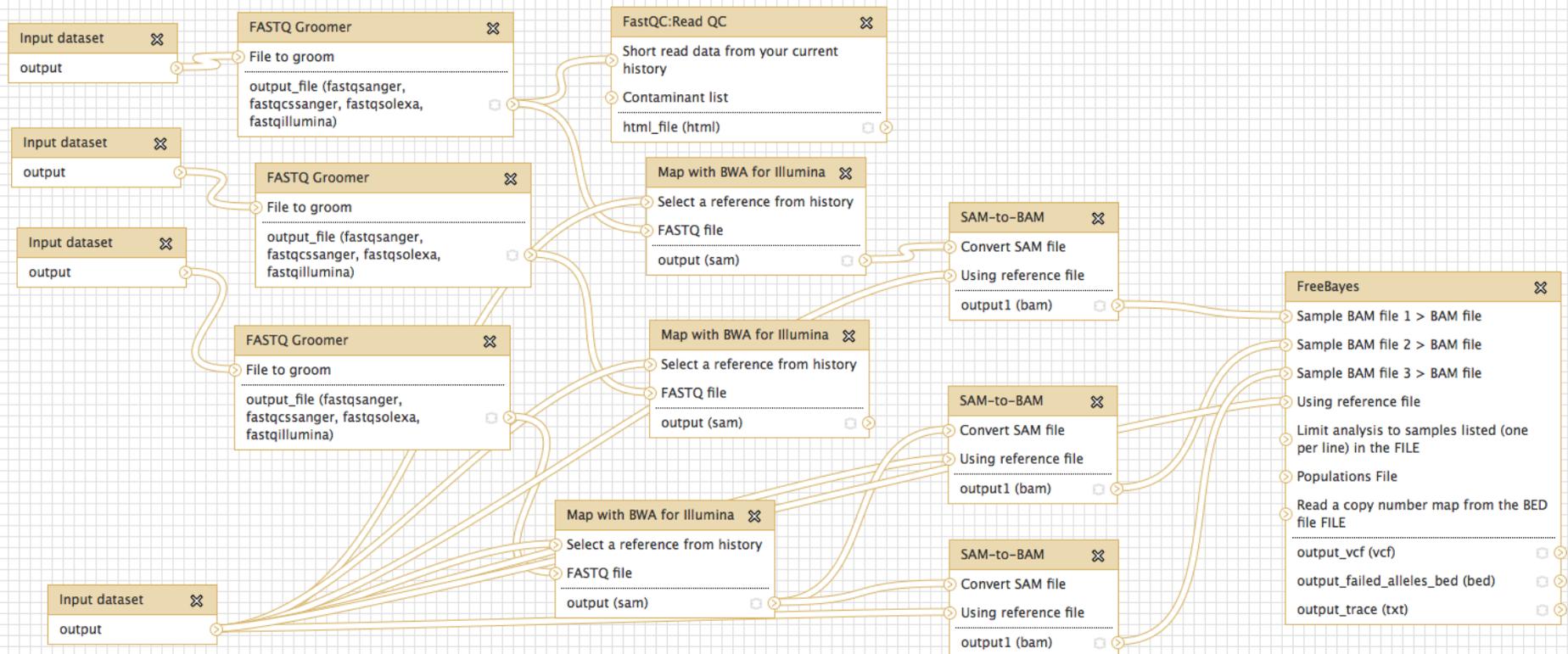
TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

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

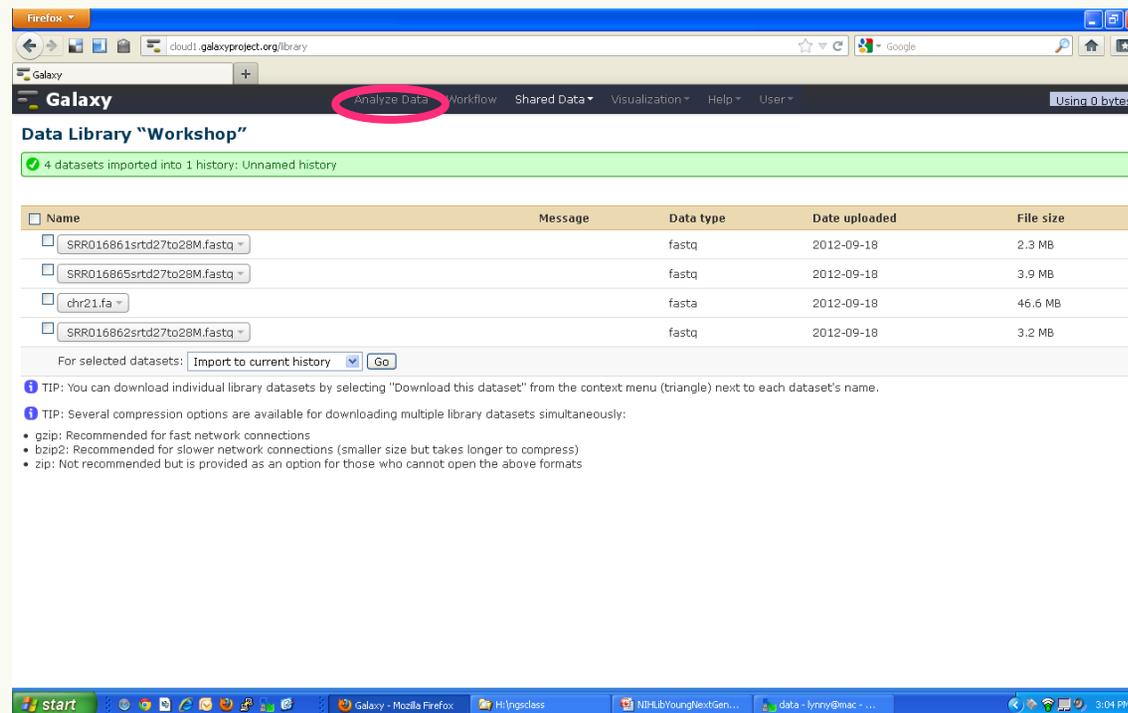
- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats



Galaxy 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
<input type="checkbox"/> SRR016061srt27to28M.fastq		fastq	2012-09-18	2.3 MB
<input type="checkbox"/> SRR016065srt27to28M.fastq		fastq	2012-09-18	3.9 MB
<input type="checkbox"/> chr21.fa		fasta	2012-09-18	46.6 MB
<input type="checkbox"/> SRR016062srt27to28M.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

start Galaxy - Mozilla Firefox H:Ingclass NDHLBYoungNextGen... data - lynny@mac - ... 3:04 PM



Galaxy - FASTQ Groomer

The screenshot shows the Galaxy web interface for the FASTQ Groomer tool (version 1.0.4). The interface is annotated with three steps:

- Step 1:** A red circle highlights the "FASTQ Groomer" tool in the "Tools" sidebar under "ILLUMINA-FASTQ".
- Step 2:** A red circle highlights the file selection dropdown menu, which shows "1: SRR016861srtdd27to28M.fastq" selected.
- Step 3:** A red circle highlights the "Execute" button.

Additional annotations include:

- A red box on the right side of the interface contains the text: "Repeat steps for the other two FASTQ files".
- The "History" panel on the right shows a list of files: "4: SRR016862srtdd27to28M.fastq", "3: chr21.fa", "2: SRR016865srtdd27to28M.fastq", and "1: SRR016861srtdd27to28M.fastq".



Galaxy - FastQC

Step 4

Step 1

Step 2

Step 3

FastQC:Read QC (version 0.51)

Short description from your recent history:
5: FASTQ Groomer on data 1

Title for the report:
Letters and numbers only please - other characters will be removed

Contaminant list:
Selection is Optional

Purpose
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.
The main functions of FastQC are:
Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

FastQC
This is a Galaxy wrapper. It merely exposes the external package `FastQC`, which is documented at `FastQC`. Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the `Picard-tools` libraries for sam/bam processing.
The contaminants file parameter was borrowed from the independently developed `fastqcwrapper` contributed to the Galaxy Community Tool Shed by J. Johnson.

History
Unnamed history 9.4 MB
7: FASTQ Groomer on data 4
6: FASTQ Groomer on data 2
5: FASTQ Groomer on data 1
4: SRR016862srtd27to28M.fastq
3: chr21.fa
2: SRR016865srtd27to28M.fastq
1: SRR016861srtd27to28M.fastq



Galaxy – FastQC Results

The screenshot displays the Galaxy web interface. On the left, the 'Tools' sidebar lists various analysis tools, with 'NGS: Mapping' circled in red. The main content area shows the 'FASTQ_Groomer_on_data_1' report, including a 'Summary' section with a list of metrics and their status (e.g., 'Basic Statistics' is successful, while 'Per base sequence content' is failed). On the right, the 'History' panel shows a list of recent jobs, with the top entry circled in red. A red arrow points from the History panel towards the main report area, and another red arrow points from the History panel towards the 'Step 1' label.

For next slide

Step 2

Step 1



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy Mapping – Burris Wheeler Aligner (BWA)

Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-in index?:
Use one from the history

Select a reference genome from history:
3: chr21.fa

Is this library mate-paired?:
Single-end

FASTQ file:
5: FASTQ Groomer on data 1

BWA settings to use:
Commonly Used

Execute

History

- 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: SRR016862srtid27to28M.fastq
- 3: chr21.fa
- 2: SRR016865srtid27to28M.fastq
- 1: SRR016861srtid27to28M.fastq

Step 1

Step 2

Step 3

Step 4

Repeat steps for the other two Groomed files



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – View BWA Results

The screenshot displays the Galaxy web interface with the following components:

- Tools Sidebar (Left):** Lists various NGS tools. 'NGS: SAM Tools' is circled in red.
- Central Table:** A table with columns: QNAME, FLAGNAME, POS, MAPQCIGAR, and MRNMMPOSIZSEEQ. It contains multiple rows of sequencing data.
- History Sidebar (Right):** Shows a list of workflow steps. 'Step 1' is circled in red.

Annotations on the slide:

- A red circle highlights 'NGS: SAM Tools' in the Tools sidebar.
- A red arrow points from the bottom of the table to the text 'Step 2'.
- A red circle highlights 'Step 1' in the History sidebar.

For next slide

Step 1

Step 2



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – SAM to BAM

Step 1 (circled) points to the **SAM-to-BAM** tool in the Tools sidebar.

Step 2 (circled) points to the **History** dropdown menu.

Step 3 (circled) points to the **9: Map with BWA for mapped reads** dropdown menu.

Step 4 (circled) points to the **Execute** button.

Repeat steps for the other two SAM files (pink callout box) with arrows pointing to the history entries for **data 7** and **data 3**.



Galaxy – Navigation to Picard Alignment Summary Metrics

Step 1

Step 2

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.

History

- 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: FastQC FASTQ Groomer on data 1.html
- 7: FASTQ Groomer on data 4
- 6: FASTQ Groomer on data 2



Galaxy – Picard Alignment Summary Metrics

Step 1 NGS: Picard

Step 2 SAM/BAM Alignment Summary Metrics

Step 3 12: SAM-to-BAM on dat...nverted 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

History:

- 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: FastQC FASTQ Groomer on data 1.html
- 7: FASTQ Groomer on data 4
- 6: FASTQ Groomer on data 2



Galaxy – Results of Picard Summary Alignment Metrics

For next slide

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

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 First file:
12: SAM-to-BAM on dat..nverted BAM

Step 5 with file:
13: SAM-to-BAM on dat..nverted BAM

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 displays the Galaxy web interface for the 'Merge BAM Files' tool. 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

Red annotations highlight the 'Add file:' button (Step 1) and the 'Execute' button (Step 2). The history panel on the right shows a list of previous jobs, including 'MergedBams_Merge BAM Files.log' and 'Picard Alignment Summary Metrics.html'.



Galaxy Variant Detection - FreeBayes

Step 2
Step 3

Step 4

Step 5

Step 6

Step 1

The screenshot shows the Galaxy web interface with the FreeBayes tool configuration page. The interface includes a left-hand navigation menu, a central tool configuration area, and a right-hand history panel. Red circles and arrows highlight specific steps: Step 1 points to the 'Tools' menu; Step 2 and Step 3 highlight 'FreeBayes' and 'FreeBayes - Bayesian genetic variant detector' respectively; Step 4 highlights the 'History' dropdown; Step 5 highlights the '17: MergedBams.bam' file selection; and Step 6 highlights the 'Execute' button.



Galaxy Variant Detection – FreeBayes Results

For next slide

Step 1

#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



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Filter and Sort

Step 1 Filter and Sort

Step 2 Sort data

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

Step 4 c6

Step 5 Execute

The screenshot shows the Galaxy web interface with the 'Sort (version 1.0.1)' tool selected. The 'Tools' sidebar on the left lists various categories like 'Get Data', 'Send Data', and 'Filter and Sort'. The main configuration area shows the tool's settings, including the input data source, column selection, and sorting options. The 'History' panel on the right shows a list of previous jobs. Red annotations highlight the steps for using the tool.



Galaxy – Filter and Sort Results

For next slide, open new tab

The screenshot shows the Galaxy web interface in a Firefox browser window. The main content area displays a table of genomic data with columns for Chromosome, Position, ID, Reference, Alternative, Quality, and FilterInfo. A sidebar on the left contains various tools, with the 'Filter and Sort' tool selected. The 'Filter and Sort' tool options include: 'Filter data on any column using simple expressions', 'Sort data in ascending or descending order', 'Select lines that match an expression', 'Extract features from GFF data', 'Filter GFF data by attribute using simple expressions', 'Filter GFF data by feature count using simple expressions', and 'Filter GTF data by attribute values list'. The right sidebar shows a history panel with a list of recent jobs, including 'Sort on data 20', 'FreeBayes on data 3 and data 17 (variants)', 'MergedBams Merge BAM Files.log', 'MergedBams.bam', 'Picard Alignment Summary Metrics.html', 'SAM-to-BAM on data 3 and data 11: converted BAM', 'SAM-to-BAM on data 3 and data 10: converted BAM', 'SAM-to-BAM on data 3 and data 9: converted BAM', 'Map with BWA for Illumina on data 7 and data 3: mapped reads', and 'Map with BWA for Illumina on data 6 and data 3: mapped reads'. The browser address bar shows 'cloud1.galaxyproject.org' and the page title is 'Galaxy'.

Chrom	Pos	ID	Ref	Alt	Qual	FilterInfo
chr21	27319532	.	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;DP=...
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=2
chr21	27397951	.	A	C	11978.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=386;CIGAR=1X;DP=2
chr21	27595448	.	G	T	11683.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=358;CIGAR=1X;DP=2
chr21	27300235	.	AG	GC	11006.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=306;CIGAR=2X;DP=2
chr21	27909060	.	CA	GG	10944.1	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=304;CIGAR=2X;DP=2
chr21	27091130	.	T	G	10845.8	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=342;CIGAR=1X;DP=2
chr21	27487642	.	TC	T	10311.6	AB=0.54275;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=2
chr21	27010825	.	CA	AC	9386.72	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=319;CIGAR=2X;DP=2
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	27020730	.	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;DP=...



An ORS Service

Next Generation Sequencing Data Analysis

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

Step 1

Step 2

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

chr21:27,000,000-27,900,000 900,001 chr21:27,000,000-27,900,000 go

UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)

RefSeq Genes

Human mRNAs

Spliced ESTs

Layered H3K27ac

DNase Clusters

Txn Factor ChIP

Human Cons

Multiz Alignments of 46 Vertebrates

Common SNPs (135)

Step 4

Step 3

Step 5



An ORS Service

Next Generation Sequencing Data Analysis

UCSC Genome Browser – OMIM Genes, OMIM AV SNPs

The screenshot shows the UCSC Genome Browser interface with the following tracks and annotations:

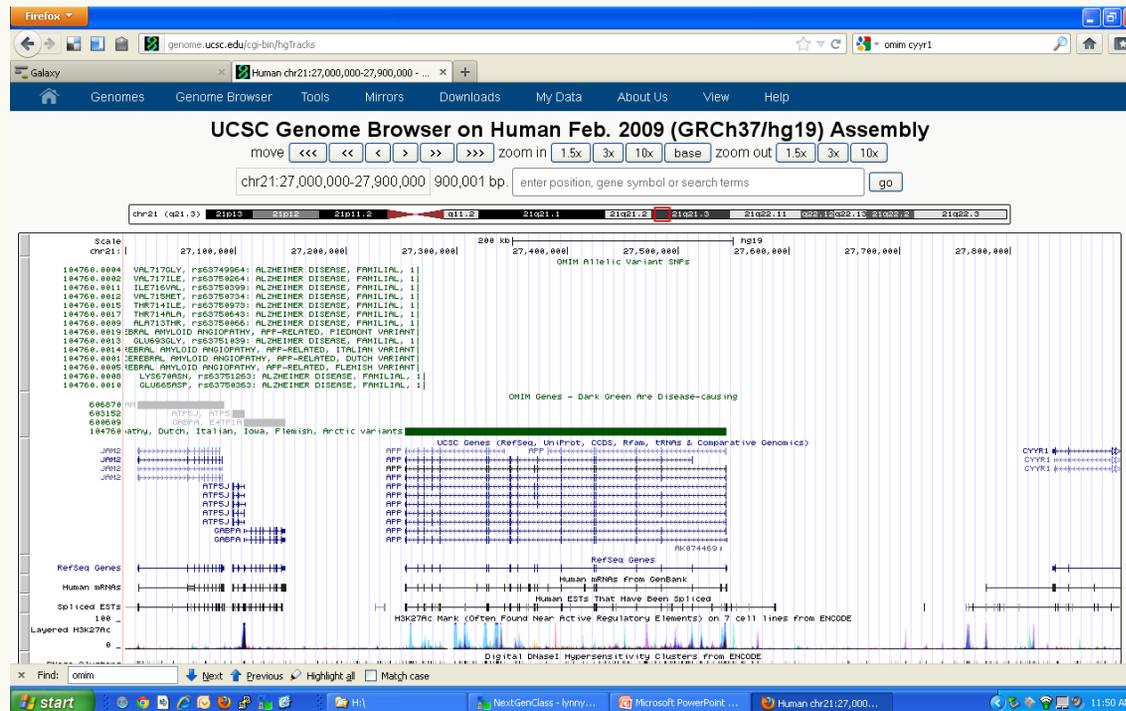
- Step 1:** A red circle highlights the "default tracks" button in the top navigation bar.
- Step 2:** A red circle highlights the "OMIM AV SNPs" dropdown menu in the "Phenotype and Disease Associations" section.
- Step 3:** A red circle highlights the "full" option in the dropdown menu.
- Step 4:** A red circle highlights the "refresh" button in the top right corner of the browser window.

The browser window displays the following tracks:

- Mapping and Sequencing Tracks:** Base Position, Chromosome Band, STS Markers, FISH Clones, Recomb Rate, deCODE Recomb, ENCODE Pilot, Map Contigs, Assembly, GRC Map Contigs, Gap, Publications, BAC End Pairs, Fosmid End Pairs, GC Percent, GRC Patch Release, Hg18 Diff, GRC Incident, Hi Seq Depth, Wiki Track, BU ORChID, Mapability, Short Match, Restr Enzymes.
- Phenotype and Disease Associations:** GAD View, DECIPHER, OMIM AV SNPs, OMIM Genes, OMIM Pheno Loci, COSMIC, GWAS Catalog, ISCA, MGI Mouse, GeneReviews.
- Genes and Gene Prediction Tracks:** UCSC Genes, GENCODE, Old UCSC Genes, Alt Events, CCDS, RefSeq Genes, Other RefSeq, MGC Genes, ORFeome Clones, TransMap, Vega Genes, Ensembl Genes, AceView Genes, SIB Genes, N-SCAN, SGP Genes, Geneid Genes, Genscan Genes.



UCSC Genome Browser - Results



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Exporting Data Download VCF File

The screenshot shows the Galaxy web interface with a VCF file export dialog box open. The dialog box contains the following text:

Opening Galaxy20 [FreeBayes_on_data_3_and_data_17...]
You have chosen to open
...0-[FreeBayes_on_data_3_and_data_17_(variants)].vcf
which is a: vCard File (1.0 MB)
from: http://cloud1.galaxyproject.org
What should Firefox do with this file?
 Open with Microsoft Office Outlook (default)
 Save File
Do this automatically for files like this from now on.
OK Cancel

Four steps are highlighted with red circles and labels:

- Step 1:** The file name in the dialog box: "...0-[FreeBayes_on_data_3_and_data_17_(variants)].vcf".
- Step 2:** The file type in the dialog box: "vCard File (1.0 MB)".
- Step 3:** The "Save File" radio button.
- Step 4:** The "OK" button.

The background shows a table of genomic data with columns: #CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO. The INFO column contains values like "AB=0.5;ABP=3.0103;AC=1;AF=0.5".



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy – Exporting Data Download BAM Files

The screenshot shows the Galaxy web interface with a workflow history on the right and a file download dialog box in the center. The workflow history lists several steps, with the top one highlighted. The dialog box asks for the action to take with a downloaded file. Red annotations and arrows indicate the steps for downloading a BAM file.

Step 1: Click on the workflow history entry: "14: SAM-to-BAM on data 3 and data 11: converted".

Step 2: Click on the "BAM file" link in the download options.

Step 3: Click on the "Download Dataset" button.

Step 4: Click on the "Open with" button in the dialog box.

Step 5: Click on the "Save File" button in the dialog box.

Repeat steps for the other two BAM files.



An ORS Service

Next Generation Sequencing Data Analysis

Galaxy - Exporting Data Download BAI Files

The screenshot shows the Galaxy web interface with a workflow history on the right and a file download dialog box in the center. The dialog box is titled "Opening Galaxy14-[SAM-to-BAM_on_data_3_and_data_11...]" and contains the following text:

You have chosen to open
...o-BAM_on_data_3_and_data_11_converted_BAM].bai
which is a: bai File (26.1 KB)
from: http://cloud1.galaxyproject.org

What should Firefox do with this file?

Open with Browse...
 Save File
 Do this automatically for files like this from now on.

OK Cancel

Annotations on the screenshot:

- Step 1:** Points to the workflow item "14: SAM-to-BAM on data 3 and data 11: converted BAM" in the history panel.
- Step 2:** Points to the "Download Dataset" button for the selected workflow item.
- Step 3:** Points to the "Download bam_index" button in the "ADDITIONAL FILES" section.
- Step 4:** Points to the "Save File" radio button in the dialog box.
- Step 5:** Points to the "OK" button in the dialog box.

A large red box on the right contains the text: "Repeat steps for the other two BAM files".



Thank you for attending.



An ORS Service

Next Generation Sequencing Data Analysis