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

LIBRARY

An ORS Service

Next Generation Sequencing Data Analysis

Lynn Young, Ph.D.

lynny@mail.nih.gov

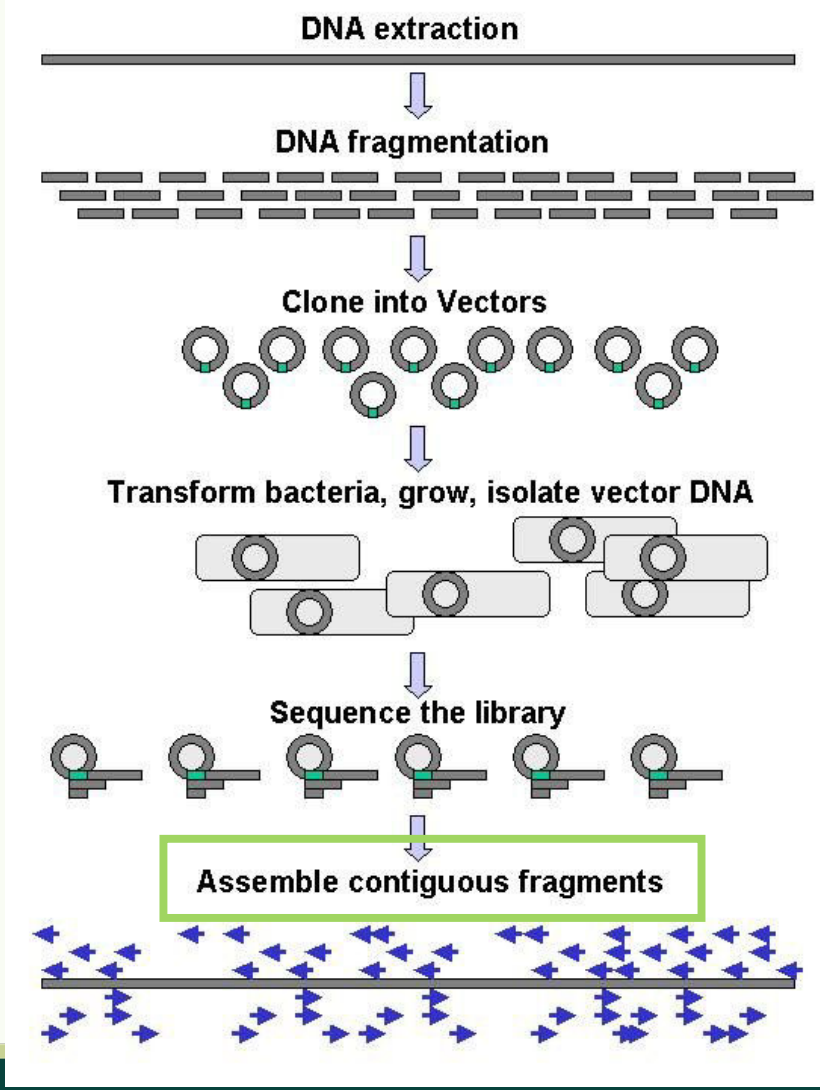
NIH Library Bioinformatics Support Program

20 September 2012

Acknowledgement

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

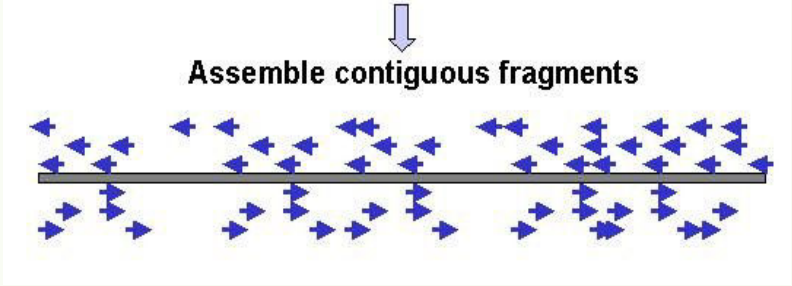
Introduction



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

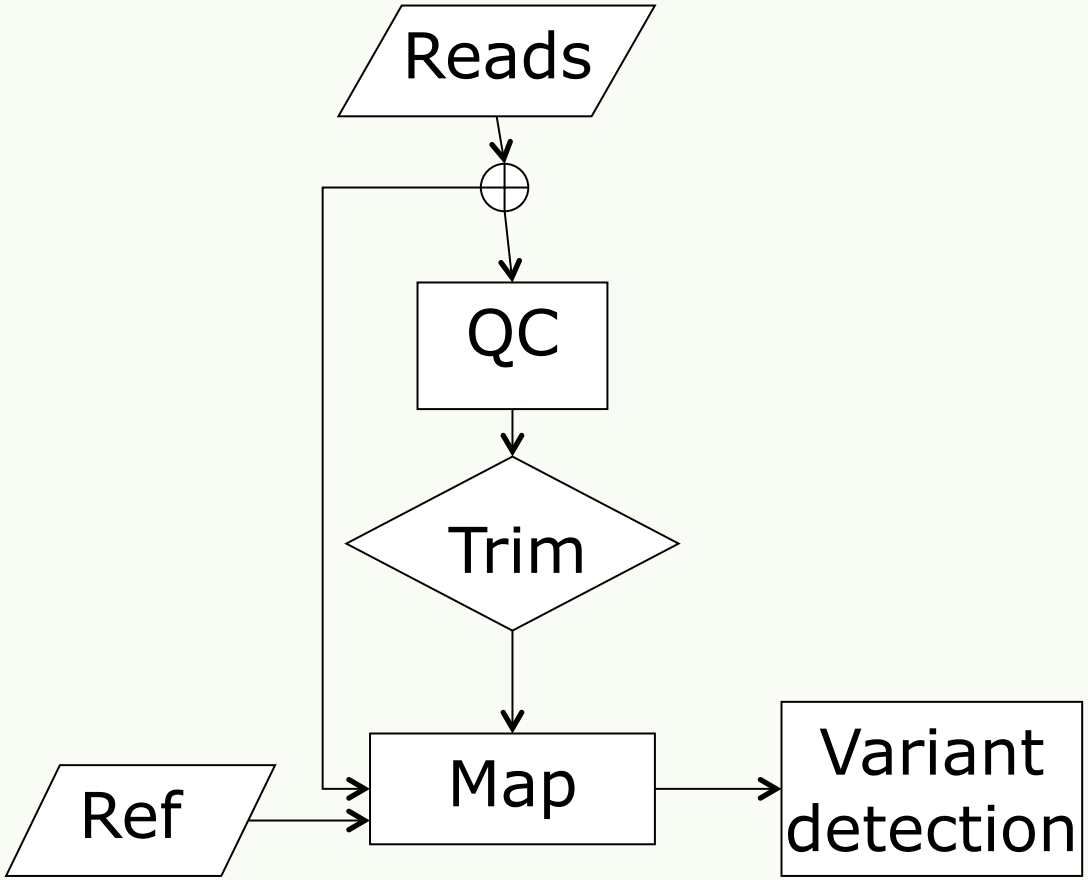
Objectives

- 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
   ATTCATTCCATTCCACTGCACTCCAATCTTCACATAAAATGTAGACAGAAGCTTTCTGAGAAACTTTTCT
   CTGATGTGTGCATTCATCTCACAGATGTGAACCATTCTTTTGTTTGAGCAGTTTGGTAACATTCTTTTTG
   TAGAATCTGCAAAAGGATATTTGTGAGCACTTTGAAGCCTATGGTGAAAAAGGAAATATCTTCAGAGAAA
   AACTAGAAAGAAGGTTTCTGAGAAACTGCTTTGTCATGTGTGAATTAGTCTCACAGATTTGAACCTTTCT
   GTTGATTGAACATATTGGAAACCTTCTTTTTGTAGAATCTGCAAAGGGATATTTGTGAACACTTGGAGGC
   CAATGGTGAAAAAGGAAATATATTCACATGAAAACTAGACAGAATCTTTCTGAGACACTTCTGTGTTTGG
```

Reads – FASTQ Format

```
@SRR016862.16884
ATTTTGAGTGGTACATCTAGGTAGCCGTTTTTGGAAACGGG
+
IIIIII,IIIII?III?I&II9$H+/I>IA%1.$,$%$#$F
@SRR016862.58801
ATTTTGAGTGGTACATCTAGGTAGCCGTTTTTGAAACCAGG
+
IIIIIIIIIIIIIIIII9IIIOII4.II@&?6&$&#%'@.
```

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	ATTTTGAGTGGTACATCTAGGTAGCCGTTTT
4	SRR016862.58801	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGTAGCCGTTTT
5	SRR016862.198085	0	chr21	27002753	37	11M1D30M	*	0	0	ATTTTGAGTGGACATCTAGGTAGCCGTTTCT
6	SRR016862.219598	0	chr21	27002753	37	41M	*	0	0	ATTTTGAGTGGTACATCTAGGTAGCCGTTTC

CCGTTTTTGGAAACGGG	IIIII,IIII?III?I&II9$H+/I>IA%1.$,$%$#$F	XT:A:U	NM:i:3	X0:i:1	X1:i:0	XM:i:3	XO:i:0	XG:i:0	MD:Z:30C2A4A2
CCGTTTTTGAAACCAGG	IIIIIIIIIIIIIIII9III0II4.II@&?6&$&#%'@.	XT:A:U	NM:i:2	X0:i:1	X1:i:0	XM:i:2	XO:i:0	XG:i:0	MD:Z:30C5A4
CGTTTCTTAAAAAAGGT	IIIIIAI@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
CCGTTTCTGAACACAGG	IIIIIIIIIIAE@4I-H6CC8"F&13+:8$I1$($/'$$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
```

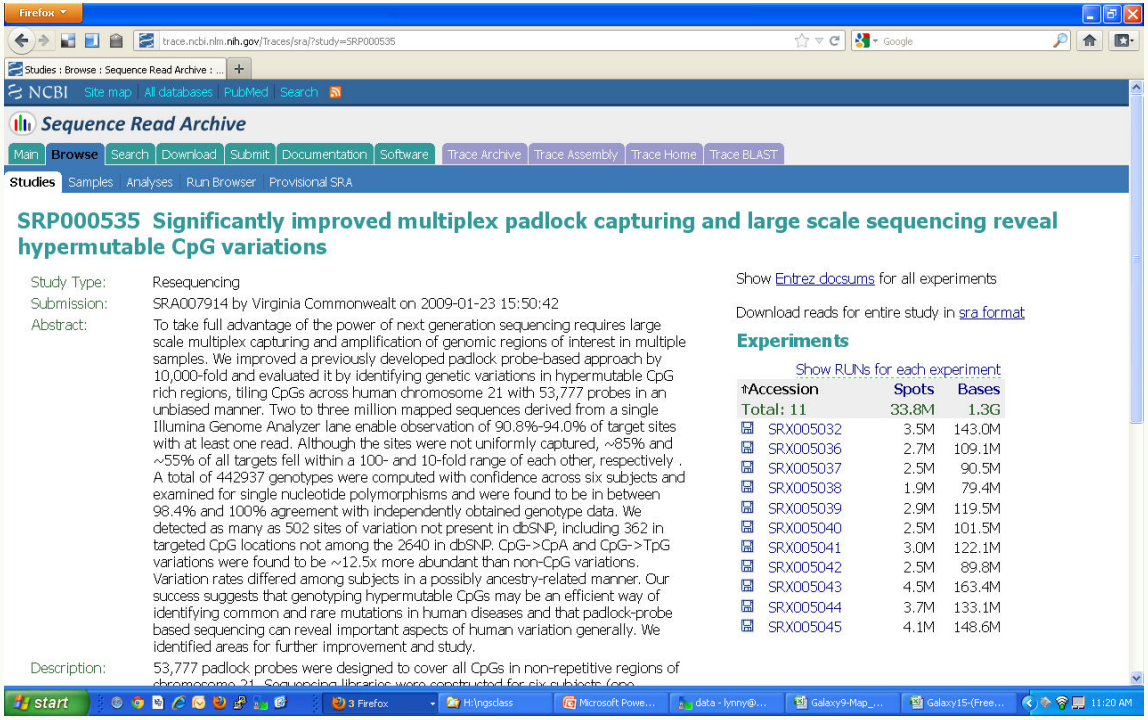
VCF Format - Data

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 observat							
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;AN
55	chr21	27010656	.	TT	AC	256.001	.	AB=0;ABP=0;AC=2;AF=1;AN
56	chr21	27010825	.	CA	AC	5571.43	.	AB=0;ABP=0;AC=2;AF=1;AN
57	chr21	27011565	.	AA	TG	76.0432	.	AB=0;ABP=0;AC=2;AF=1;AN
58	chr21	27011976	.	G	A	7662.29	.	AB=0;ABP=0;AC=2;AF=1;AN

	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

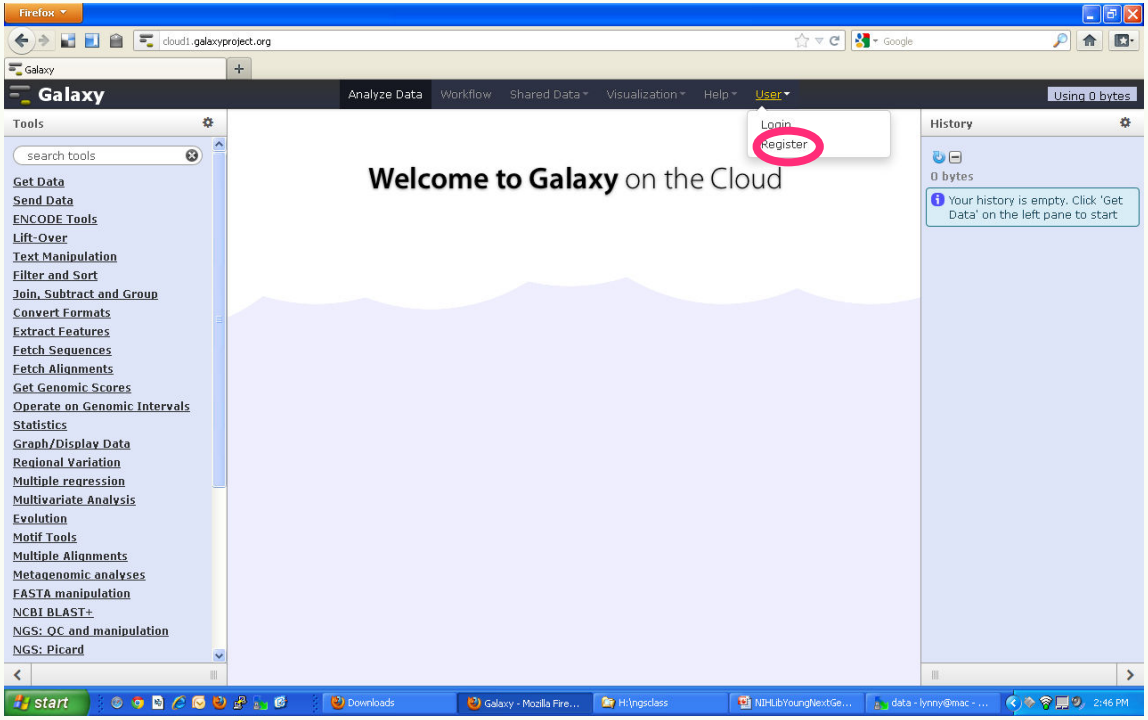
http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP000535



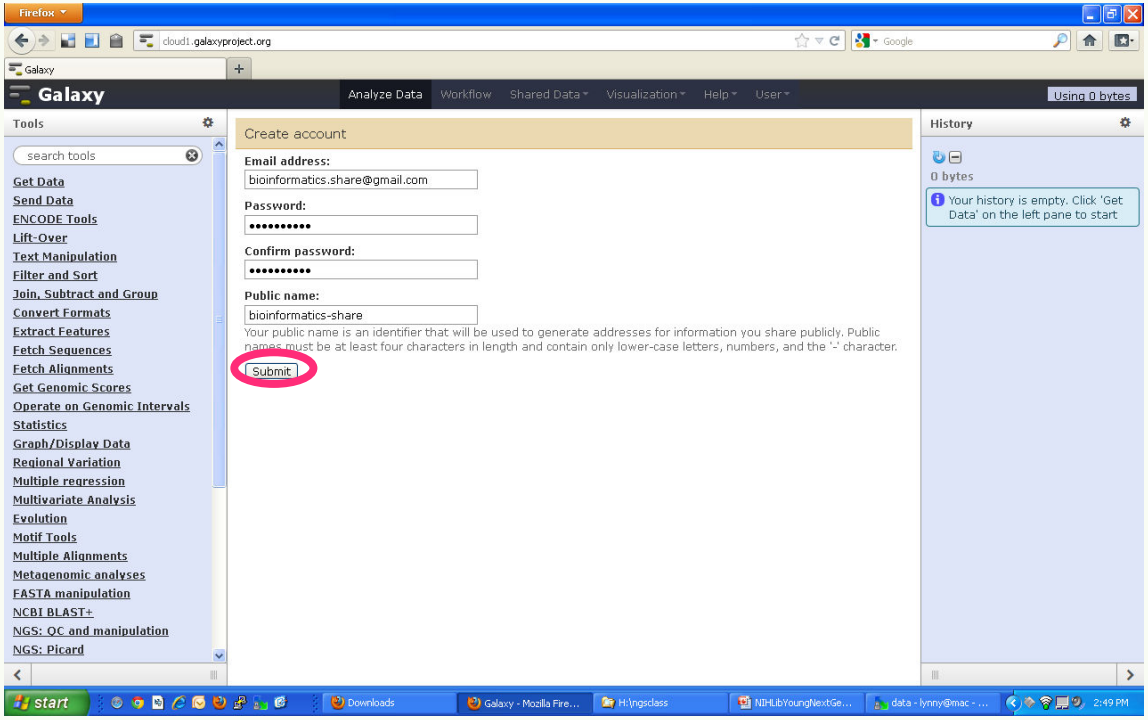
Galaxy

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

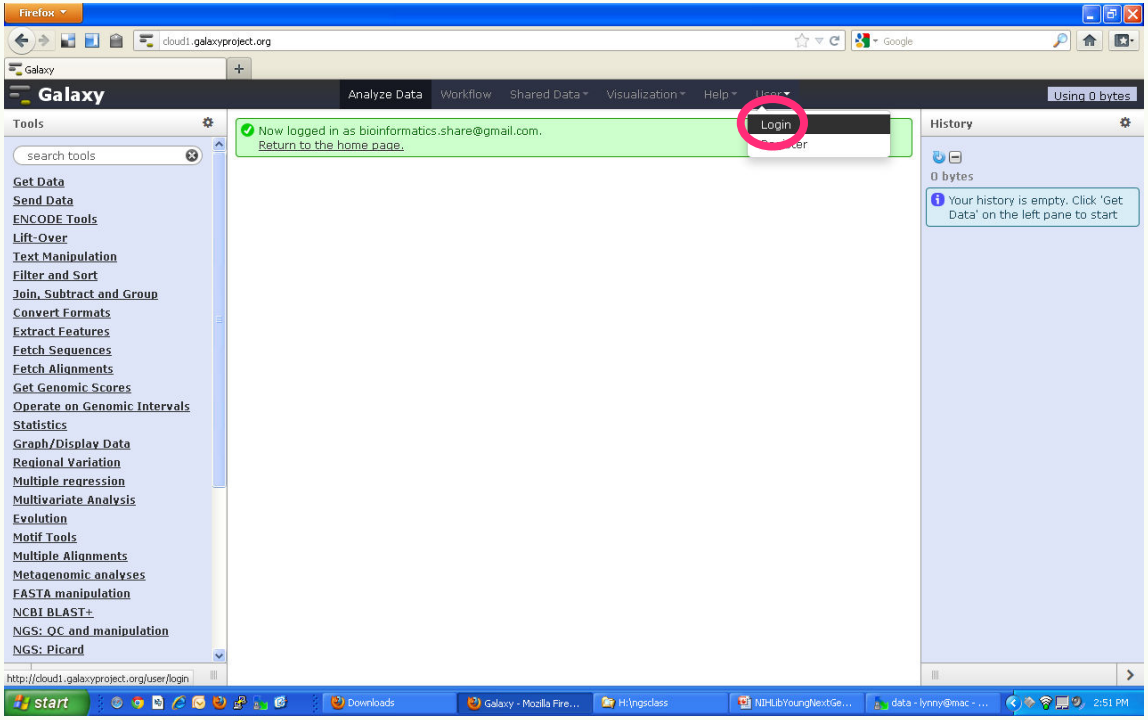
Galaxy Account Registration



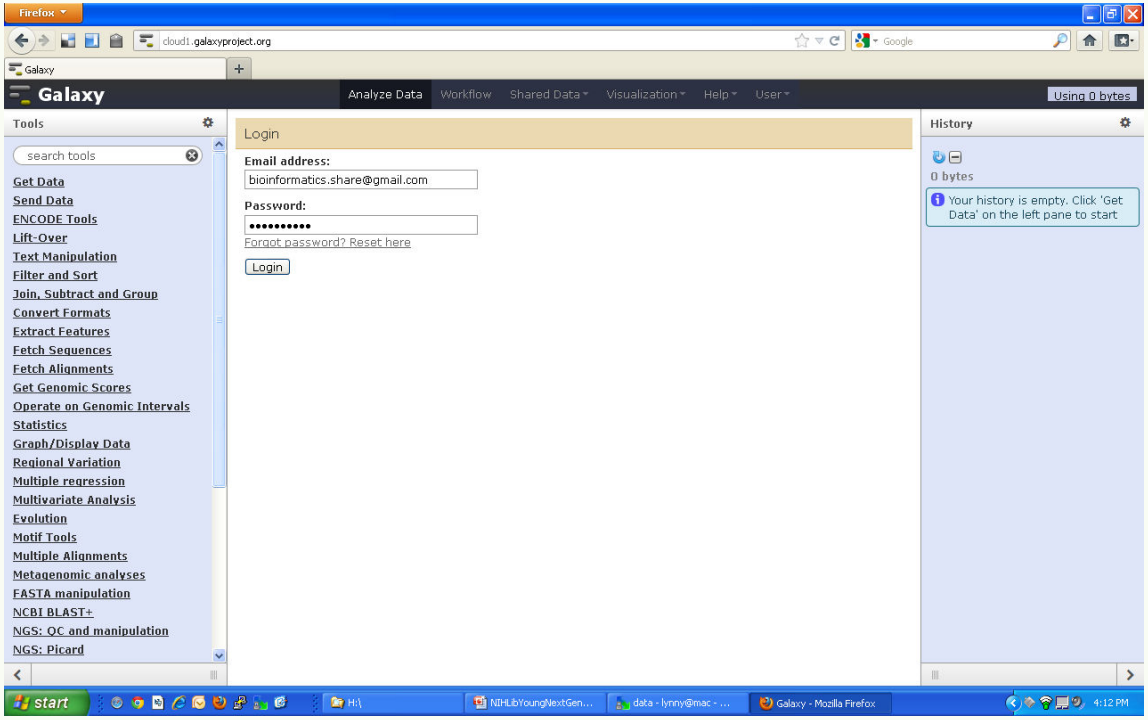
Galaxy Account Registration



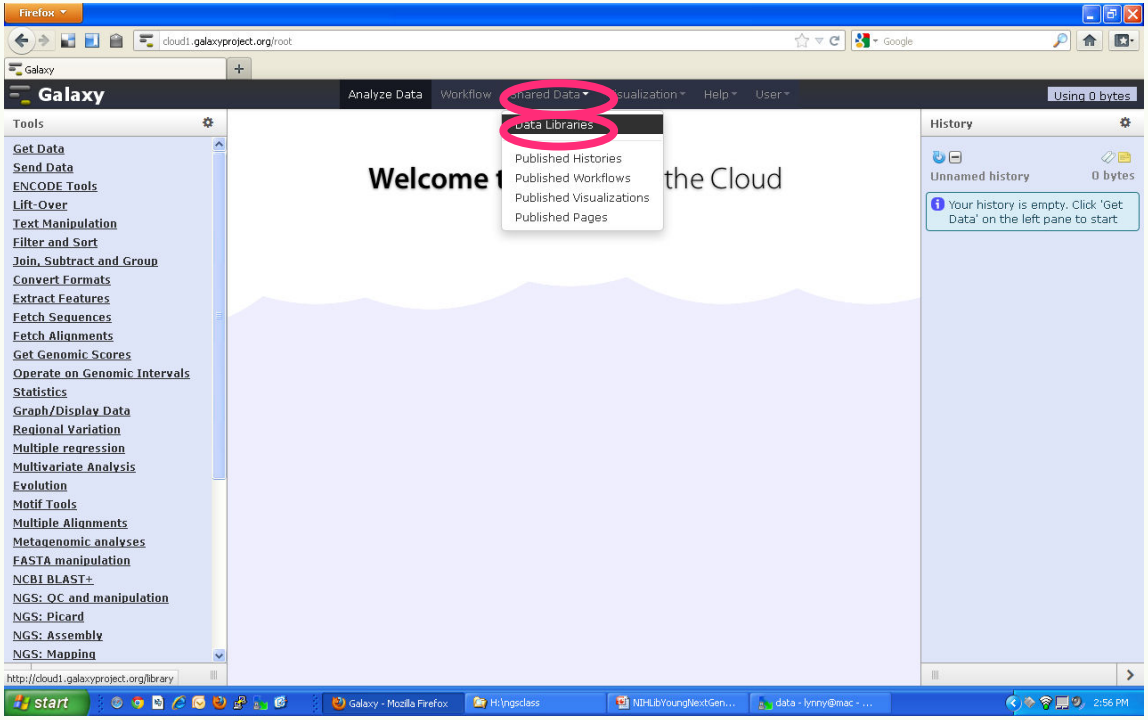
Galaxy Login if Already Have Account



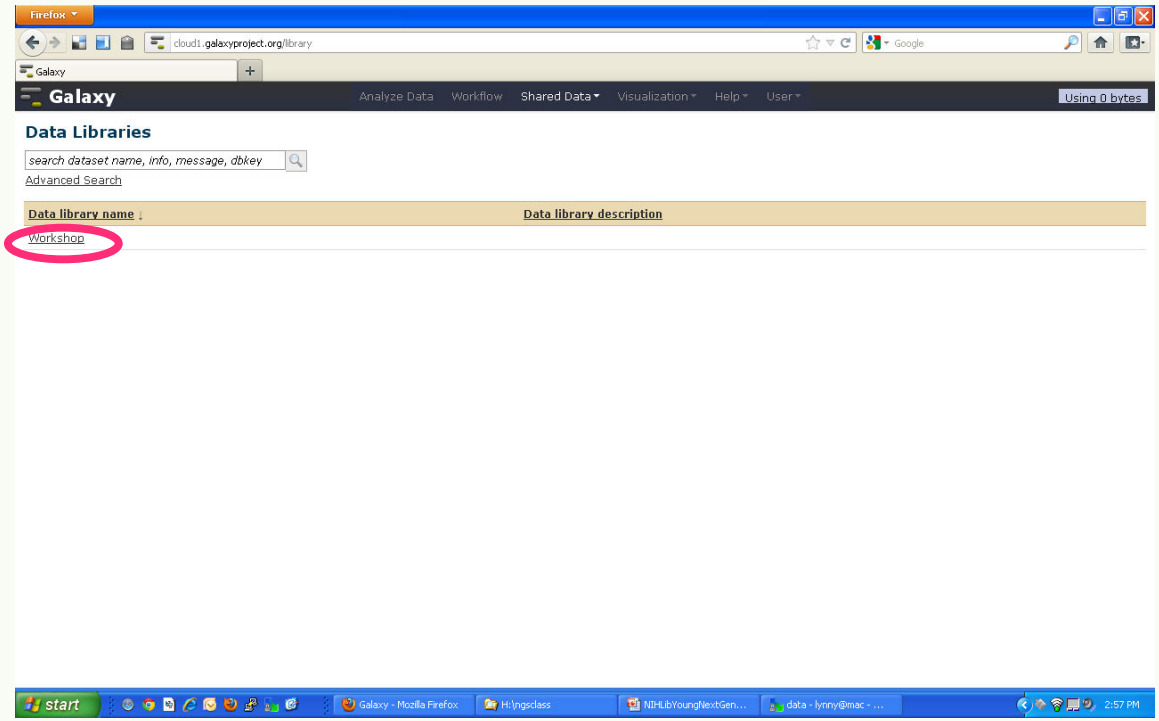
Galaxy Login



Galaxy – Shared Data

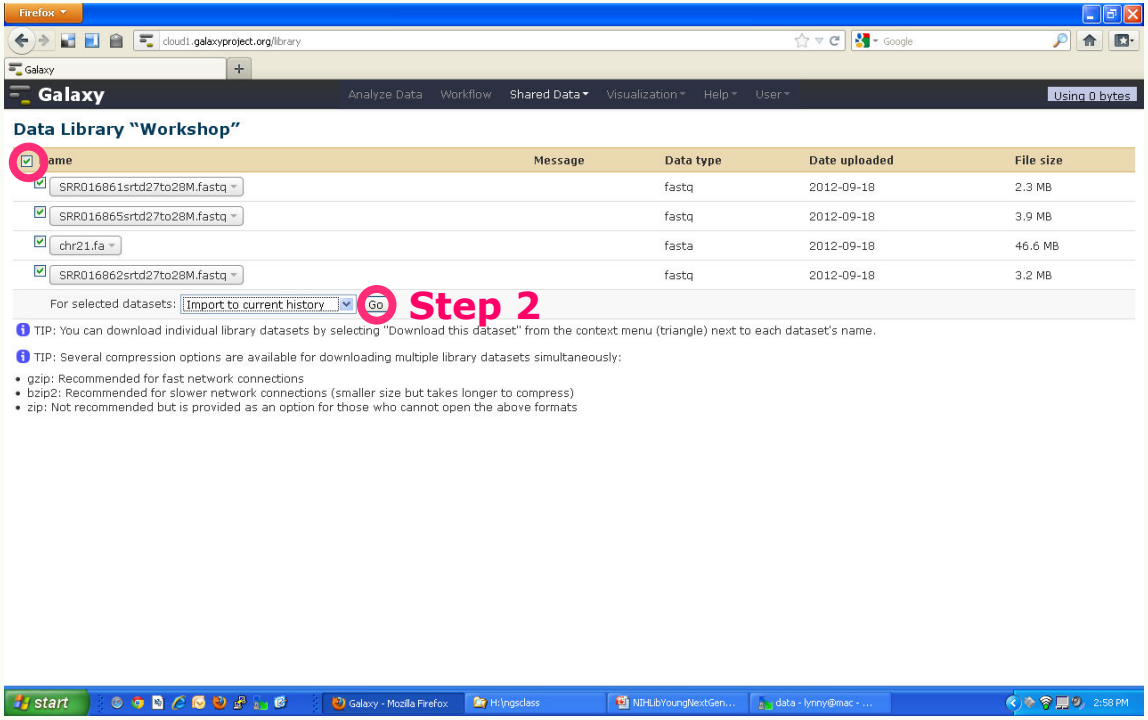


Galaxy – Obtain Shared Data

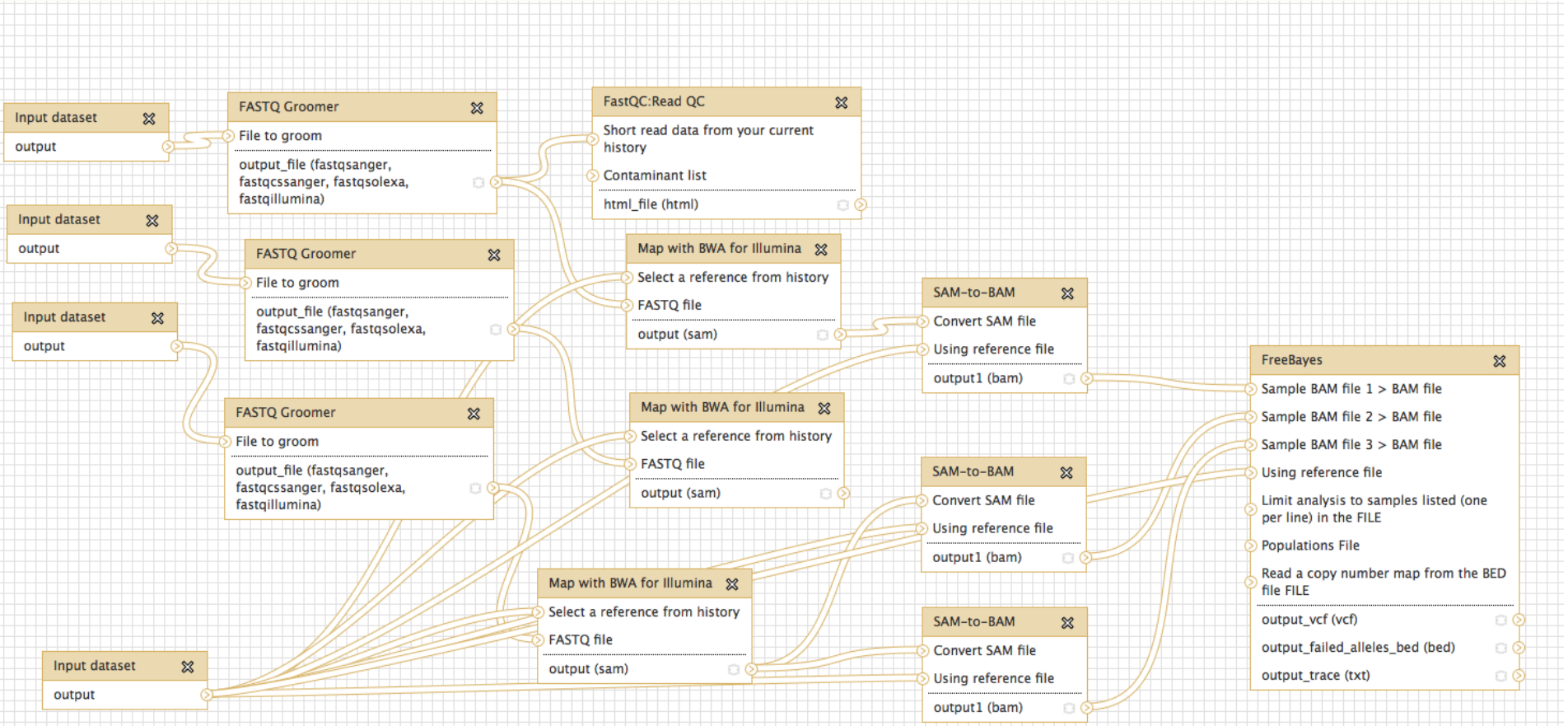


Galaxy – Obtain Share Data for Input Datasets

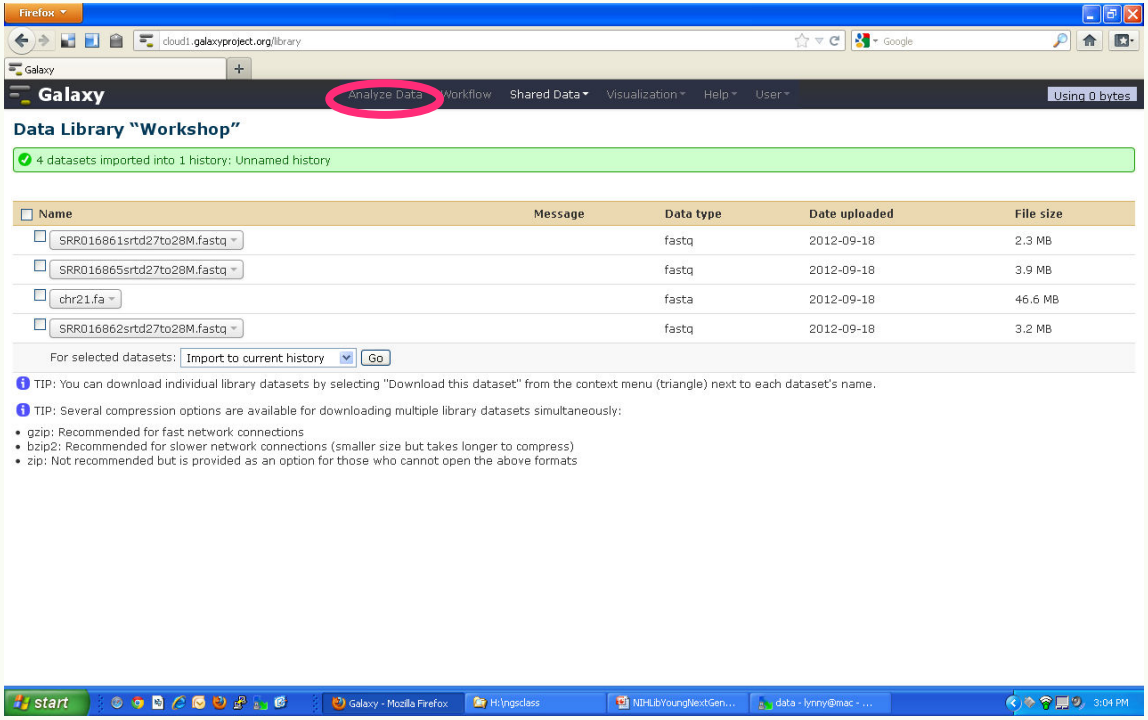
Step 1



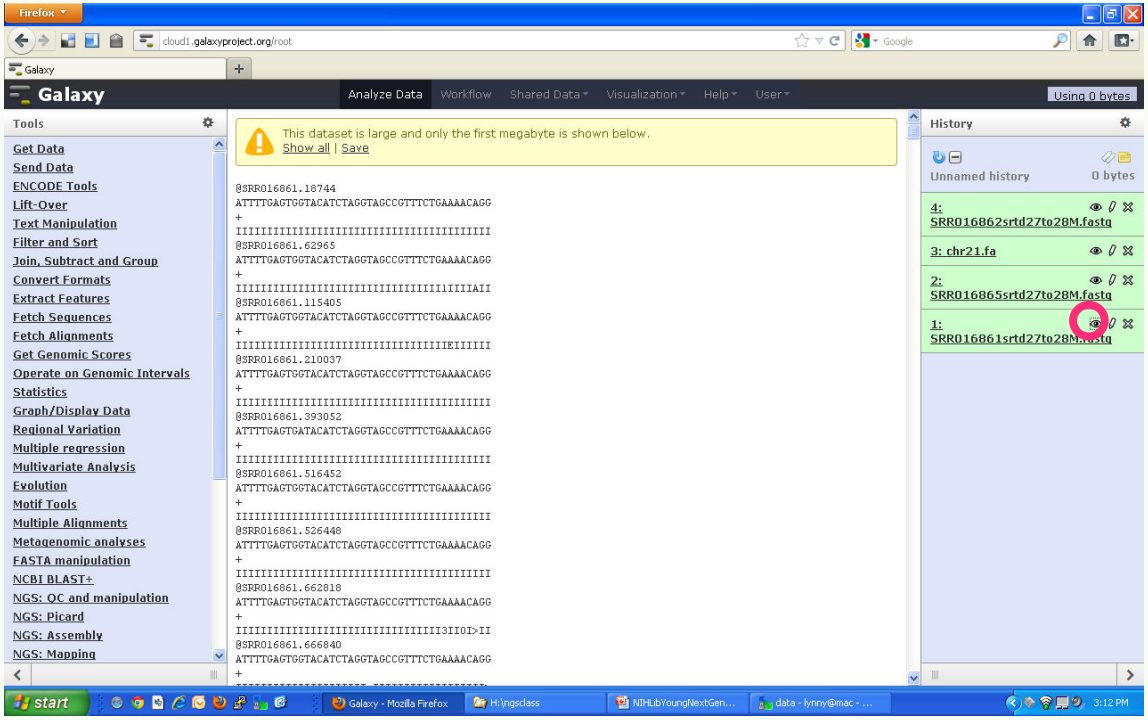
Galaxy Data Analysis Workflow - Details



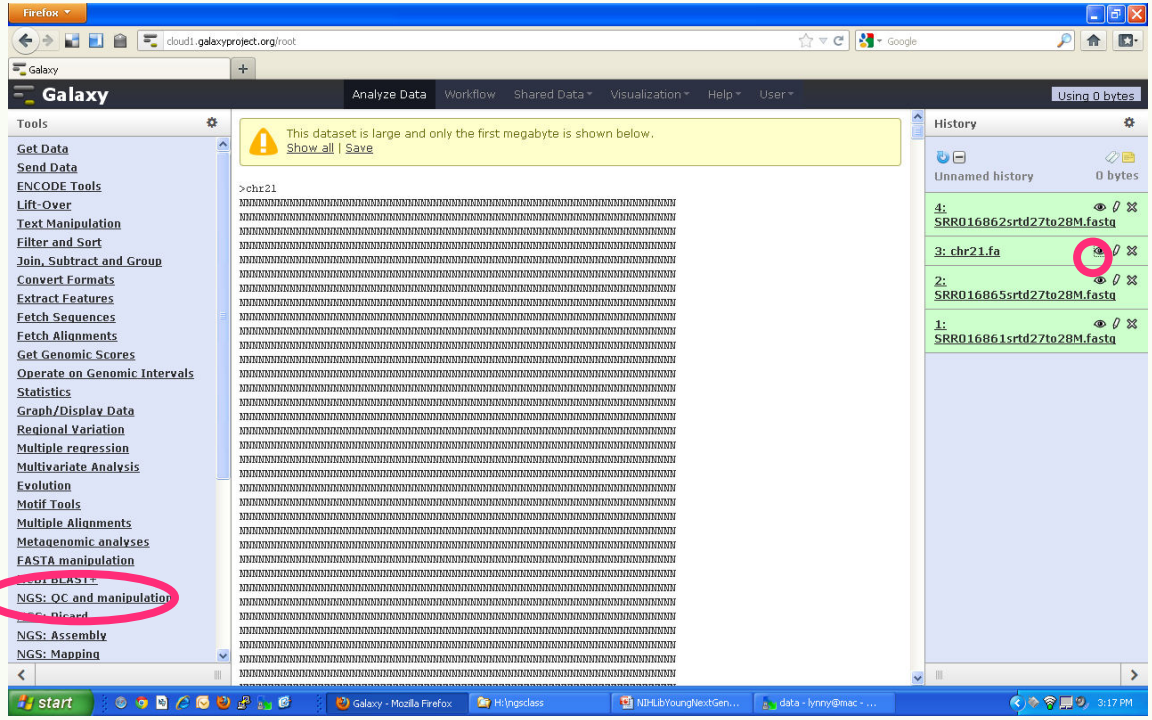
Galaxy – Analyze Data



Galaxy Input Dataset – View Sequence Reads



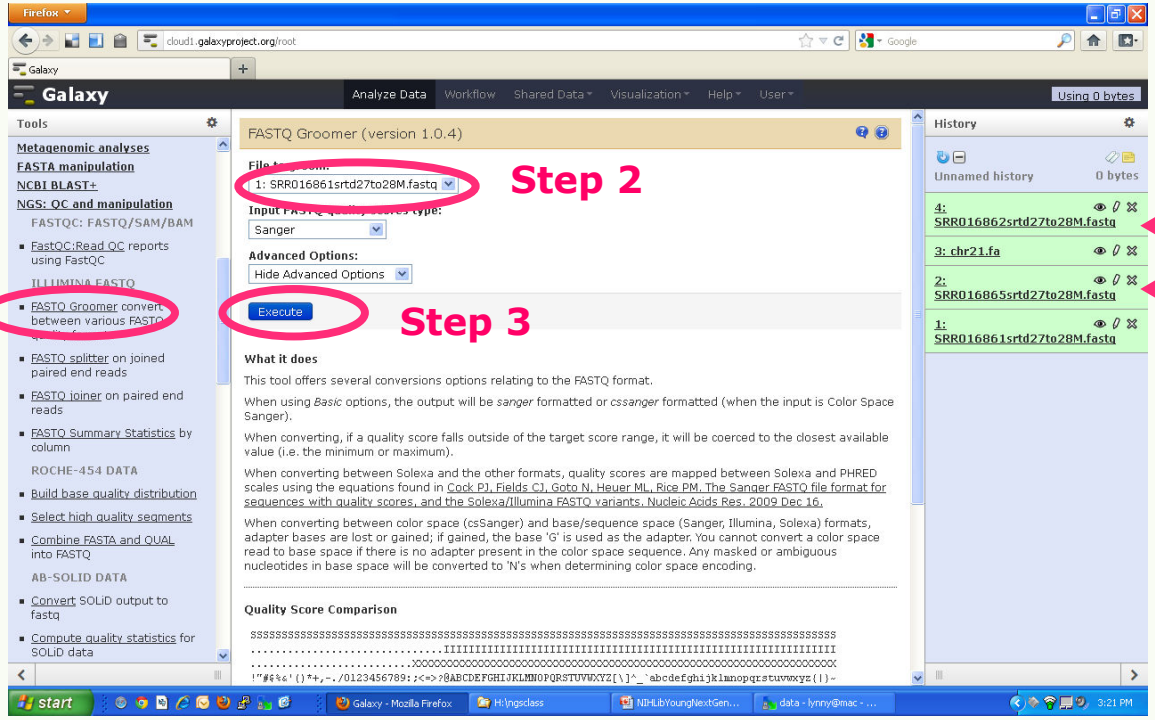
Galaxy Input Dataset – View Reference Sequence



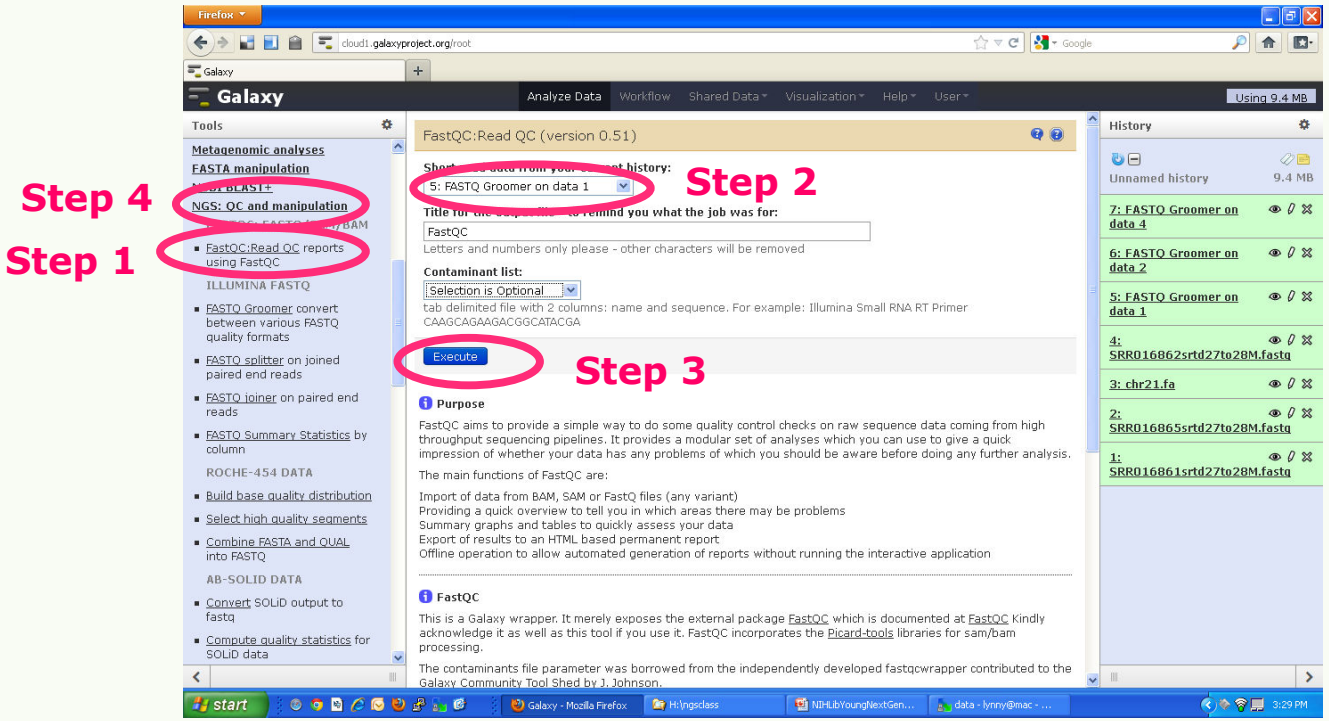
Reference Sequence

For next slide

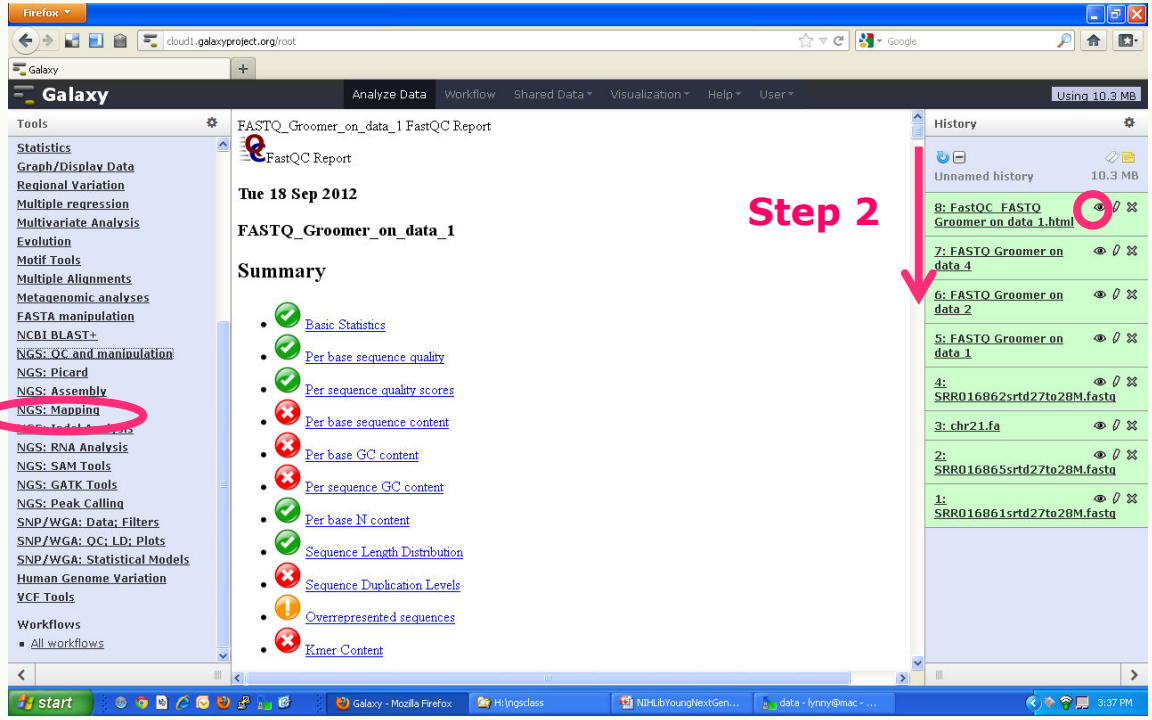
Galaxy - FASTQ Groomer



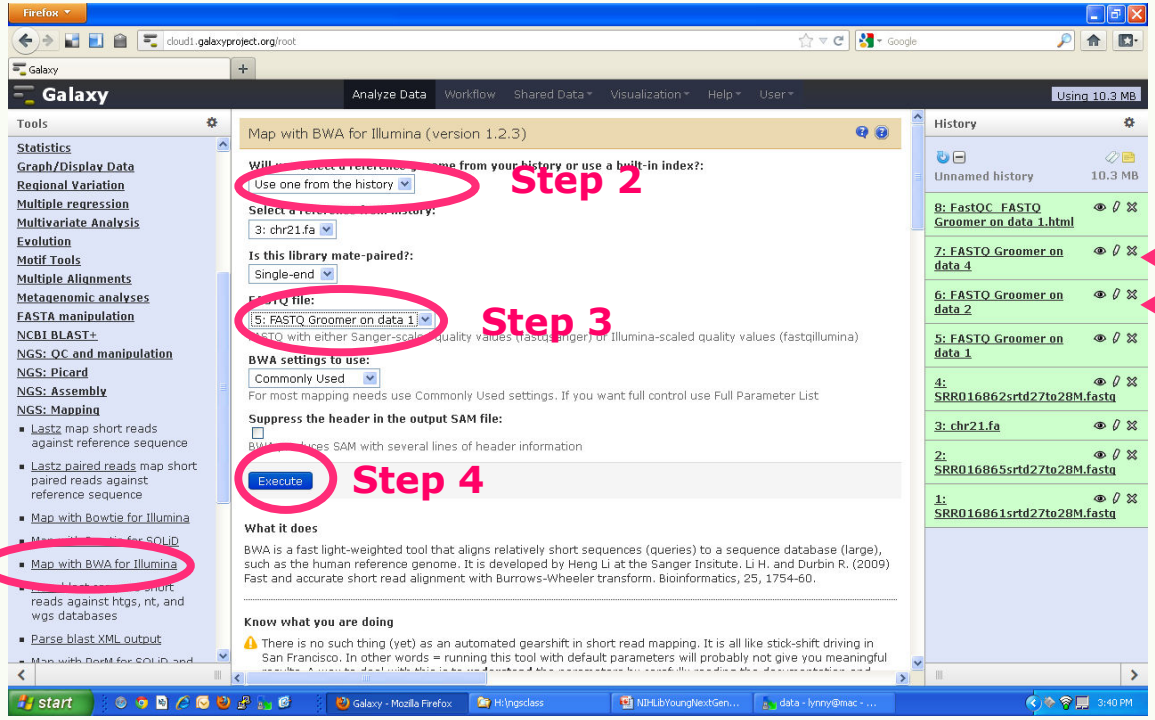
Galaxy - FastQC



Galaxy – FastQC Results

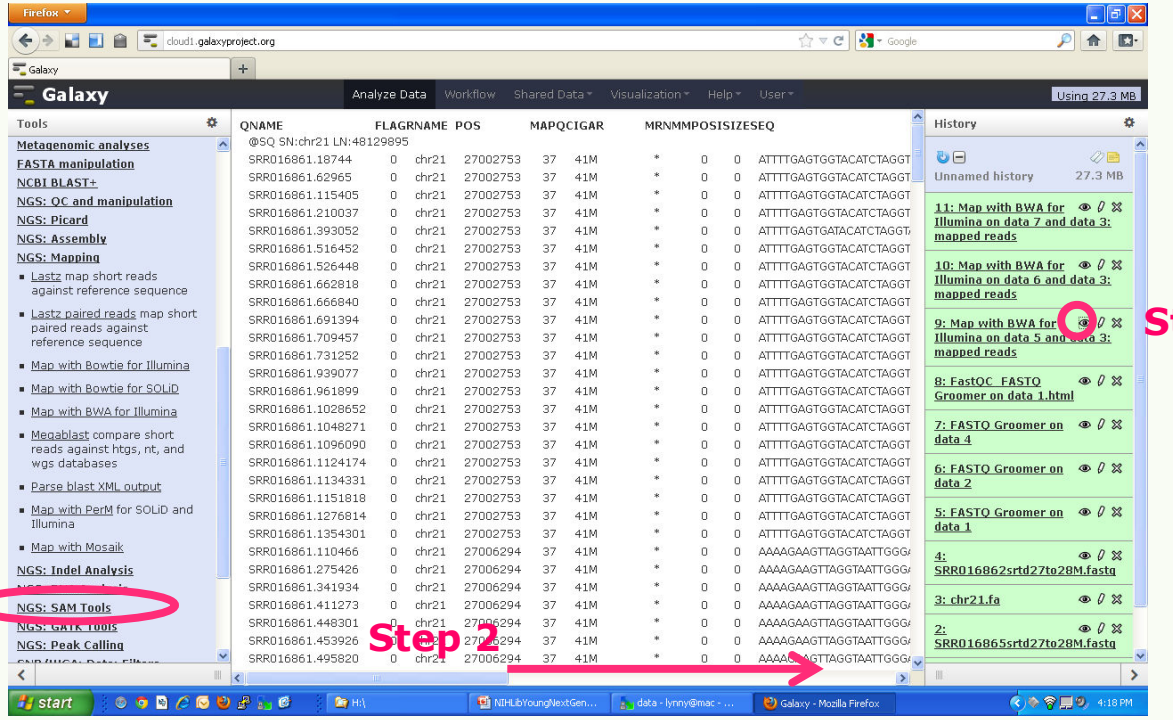


Galaxy Mapping – Burris Wheeler Aligner (BWA)

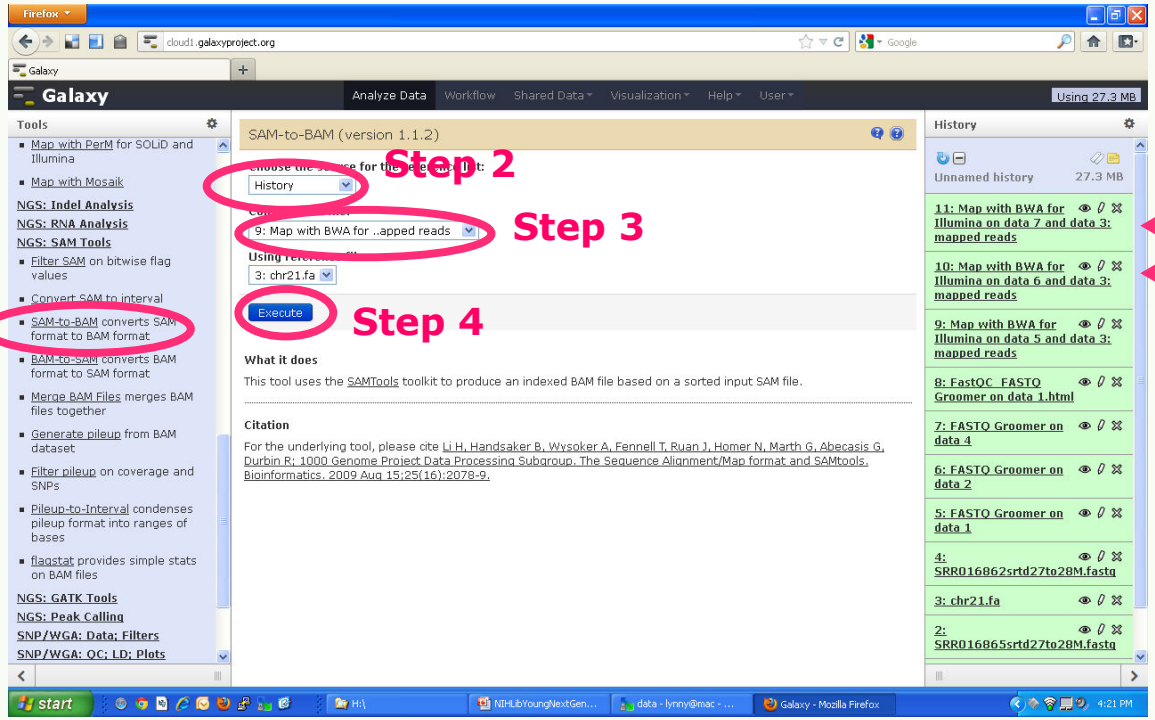


Repeat steps for the other two Groomed files

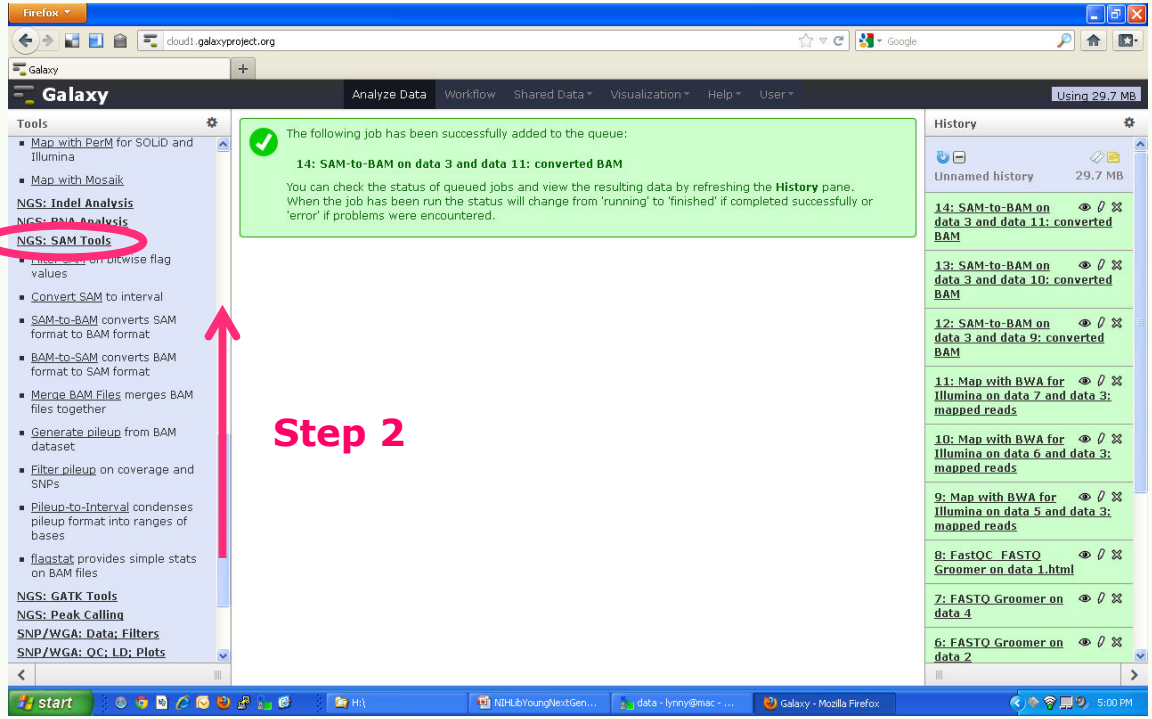
Galaxy – View BWA Results



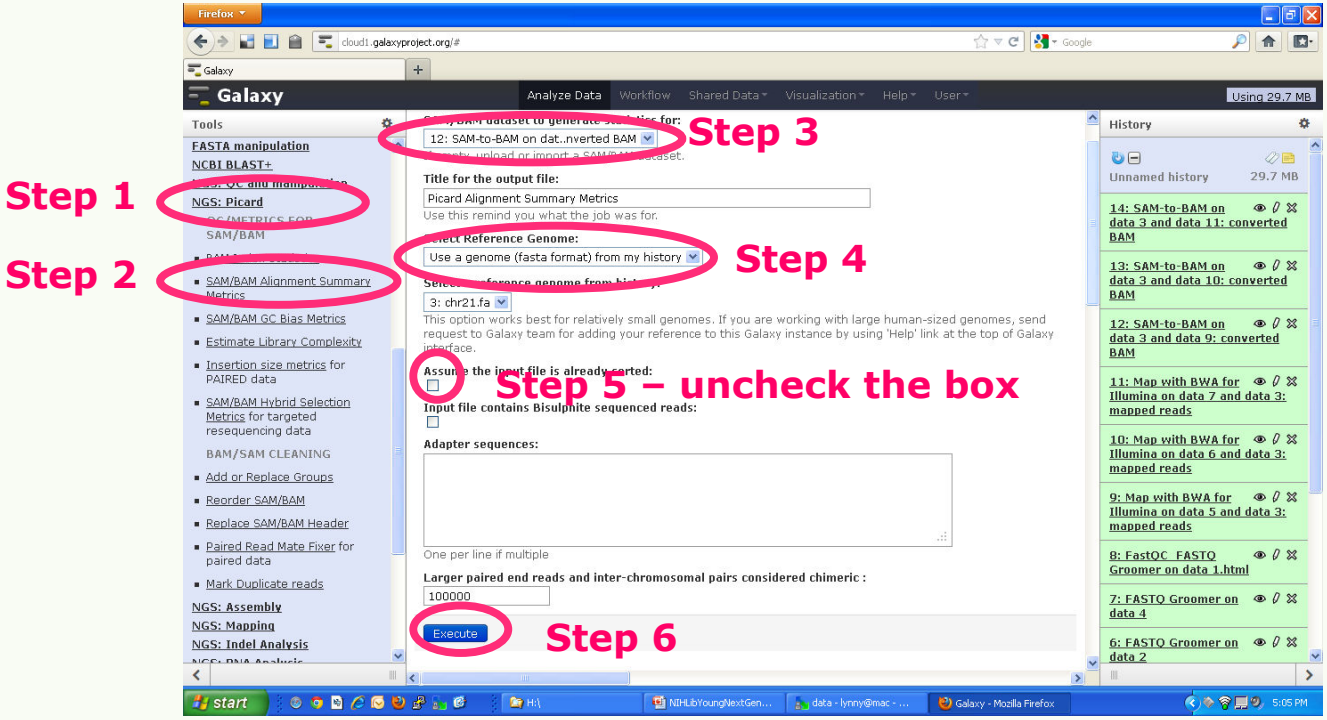
Galaxy – SAM to BAM



Galaxy – Navigation to Picard Alignment Summary Metrics

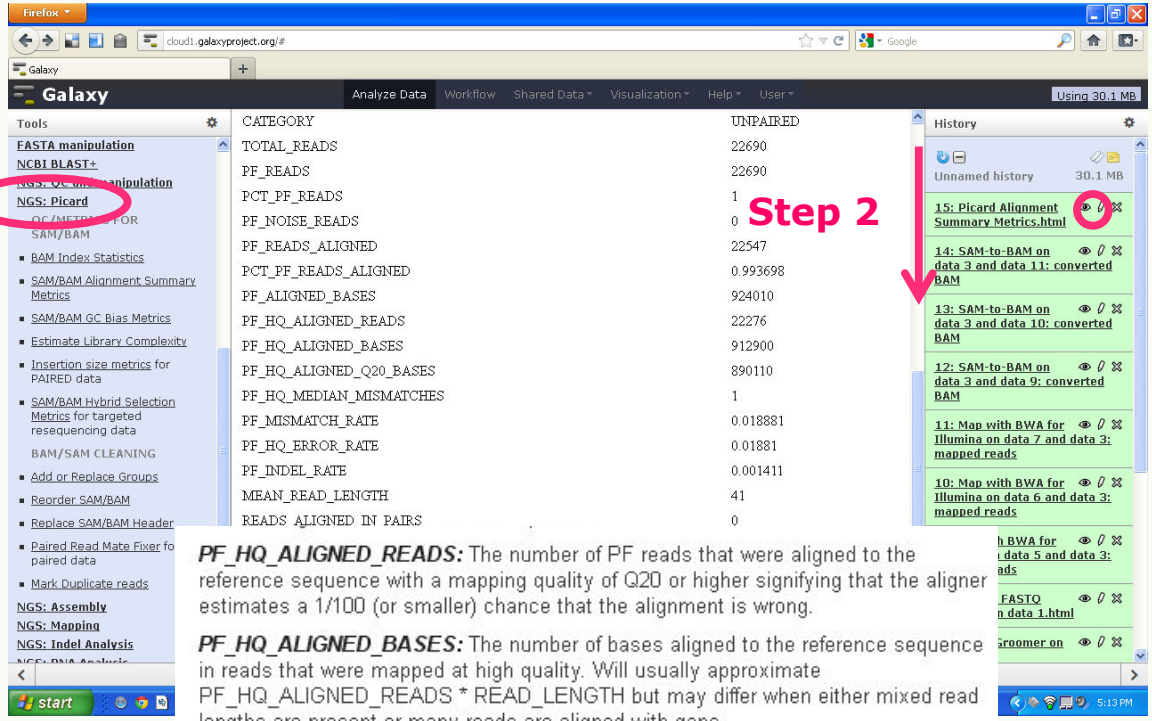


Galaxy – Picard Alignment Summary Metrics



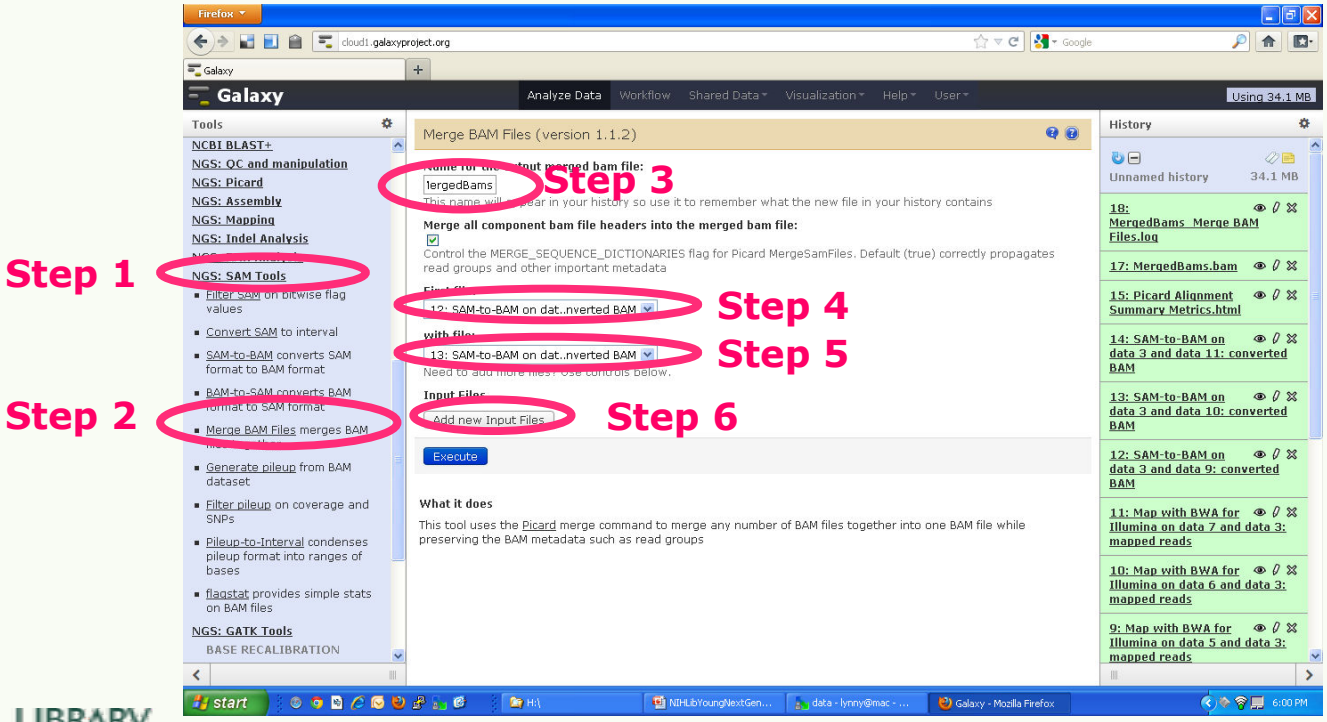
Galaxy – Results of Picard Summary Alignment Metrics

For next slide

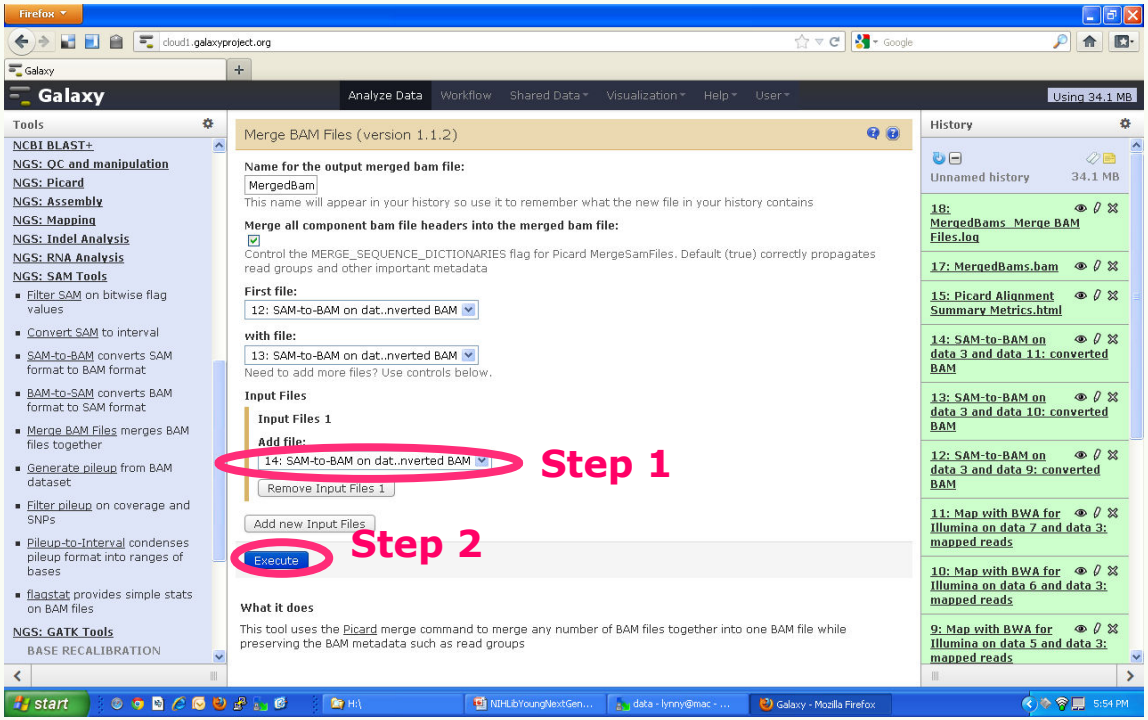


Key - http://picard.sourceforge.net/picard-metric-definitions.shtml

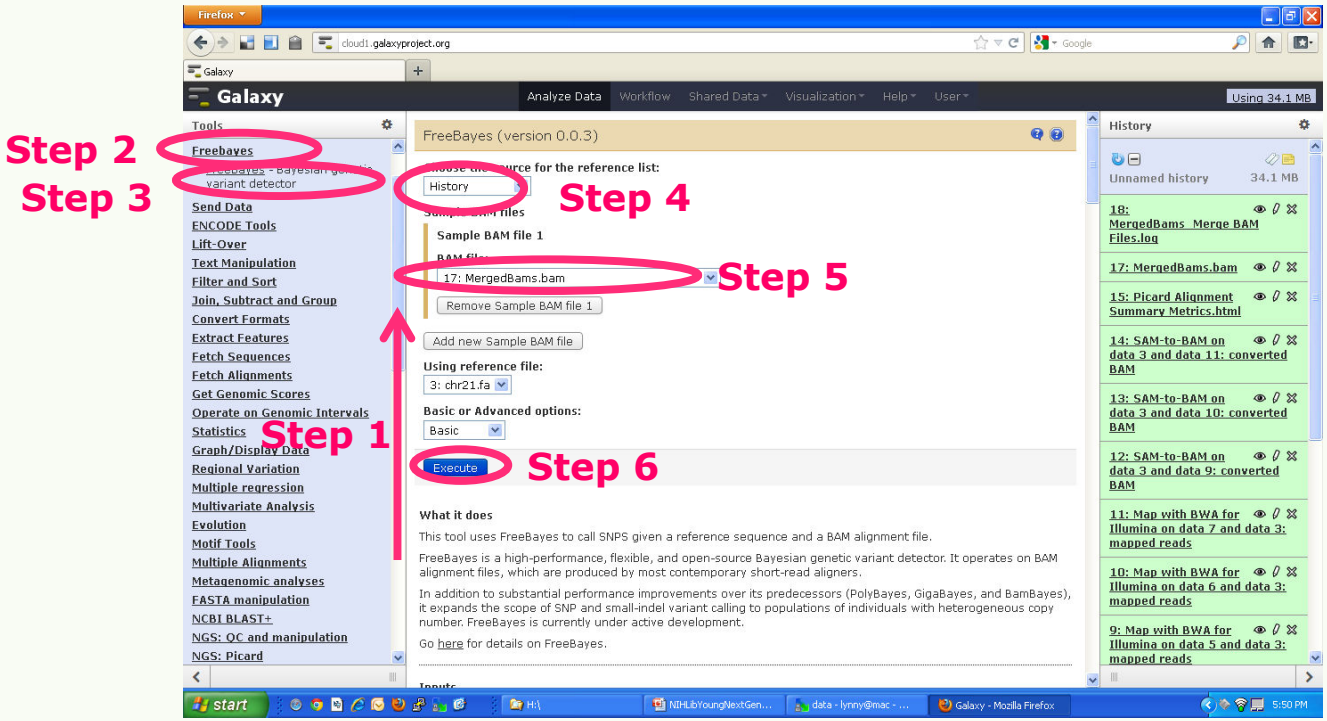
Galaxy Variant Detection – Preparation Merging Bam Files



Galaxy Variant Detection – Preparation Merging Bam Files

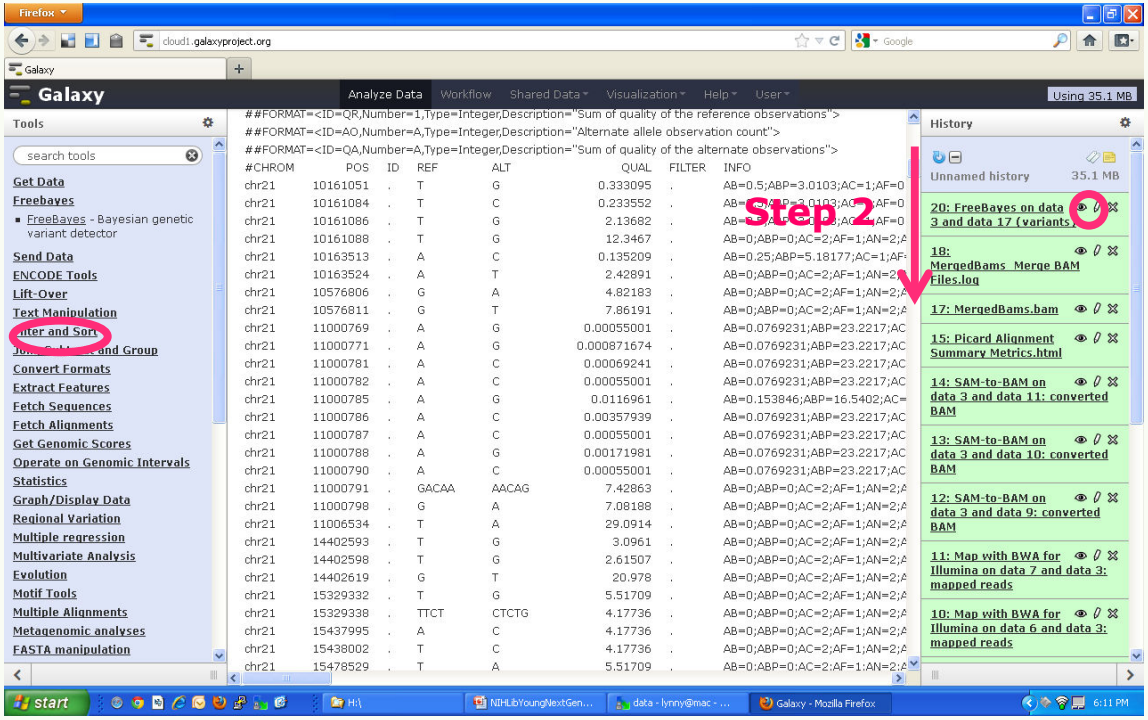


Galaxy Variant Detection - FreeBayes

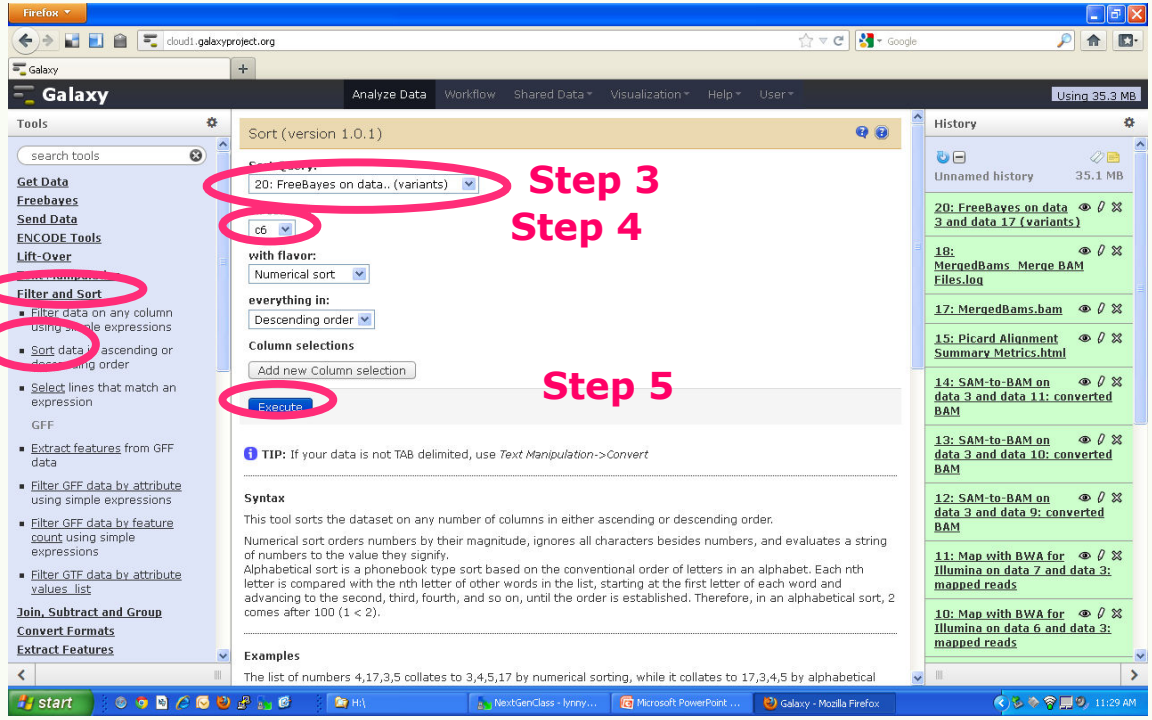


Galaxy Variant Detection – FreeBayes Results

For next slide

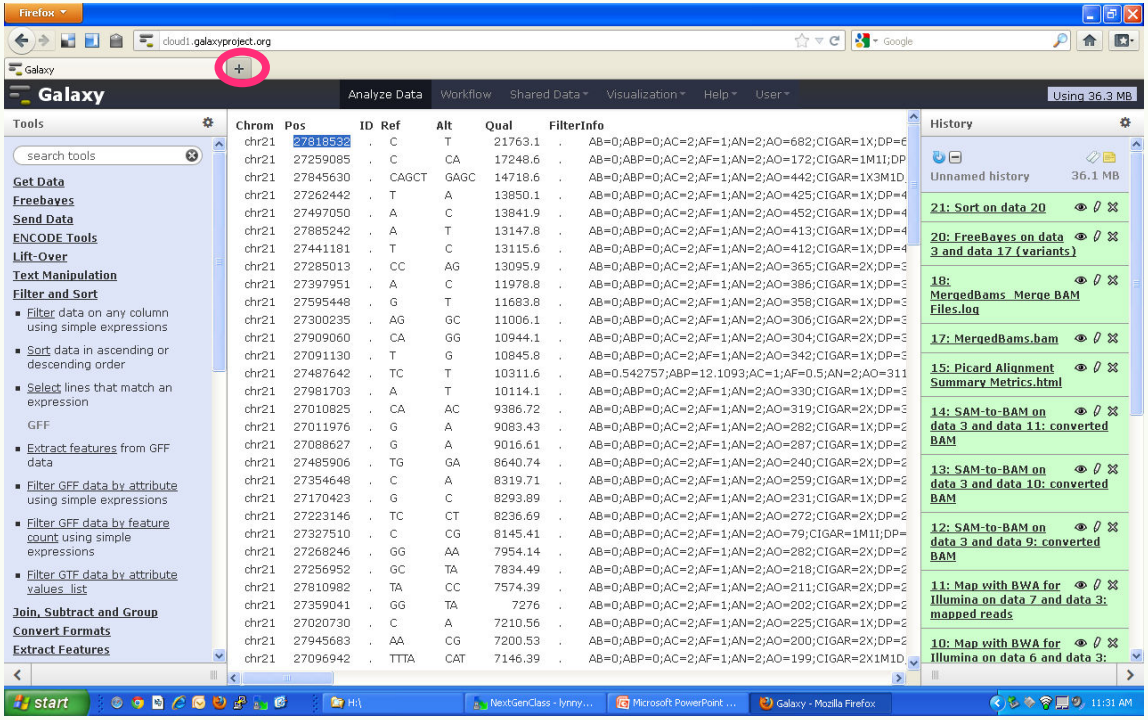


Galaxy – Filter and Sort



Galaxy – Filter and Sort Results

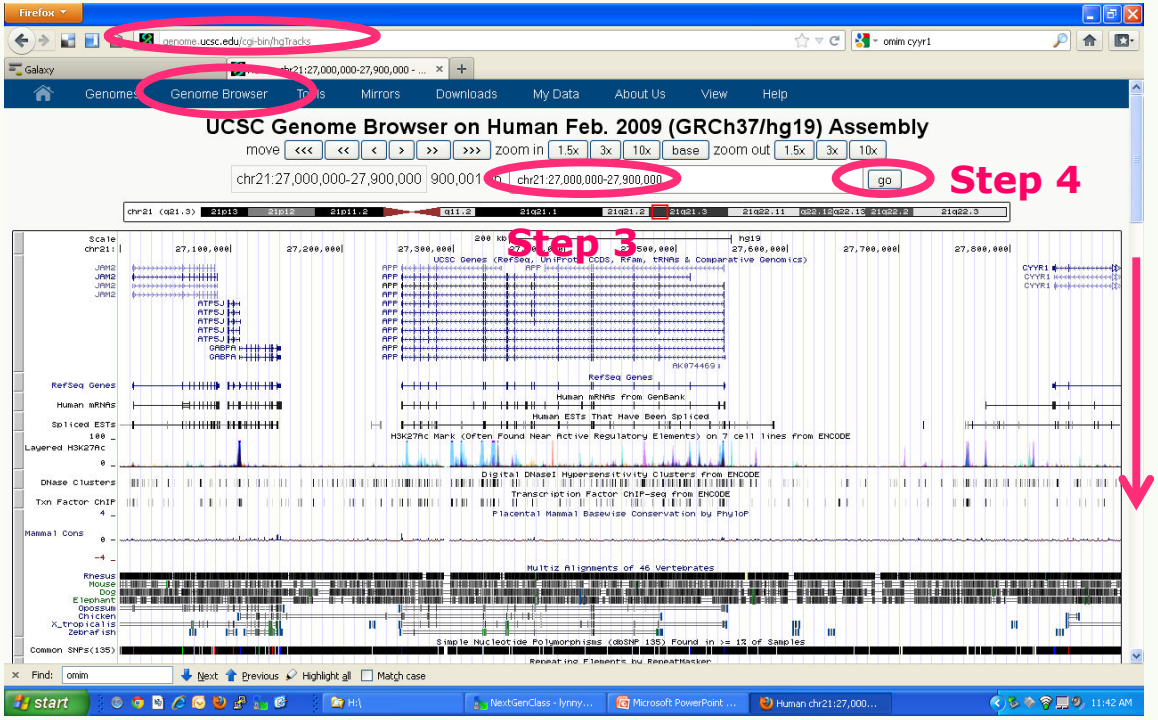
For next slide, open new tab



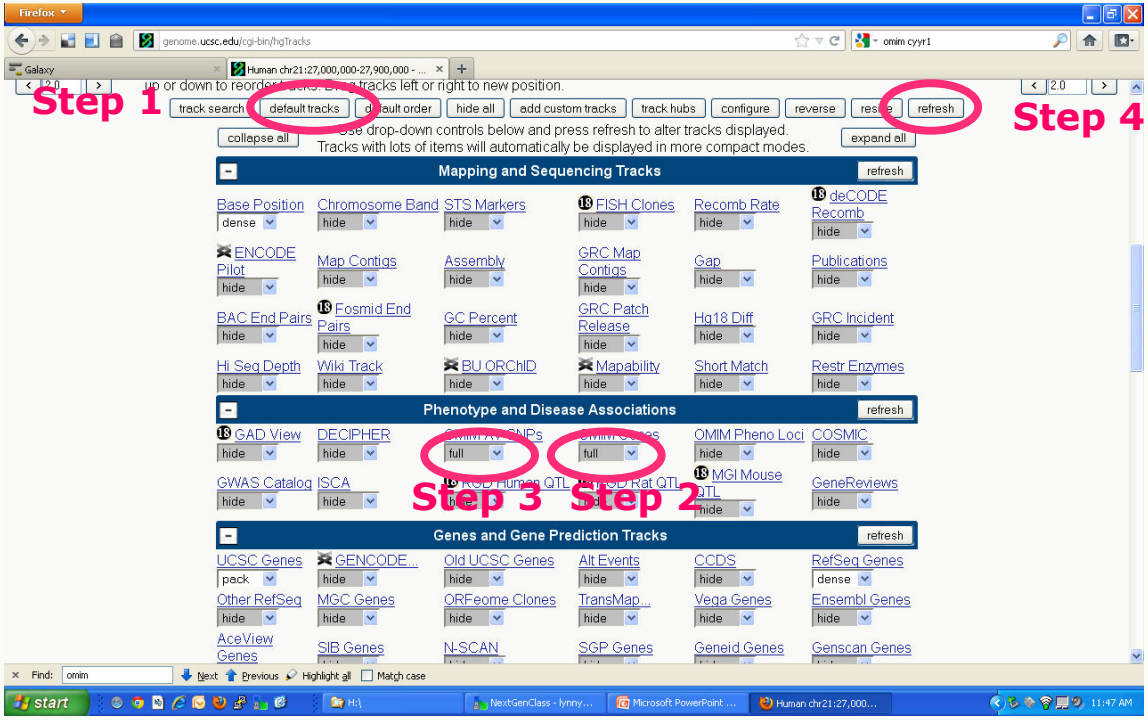
Biological Context
UCSC Genome Browser
http://genome.ucsc.edu

Step 1

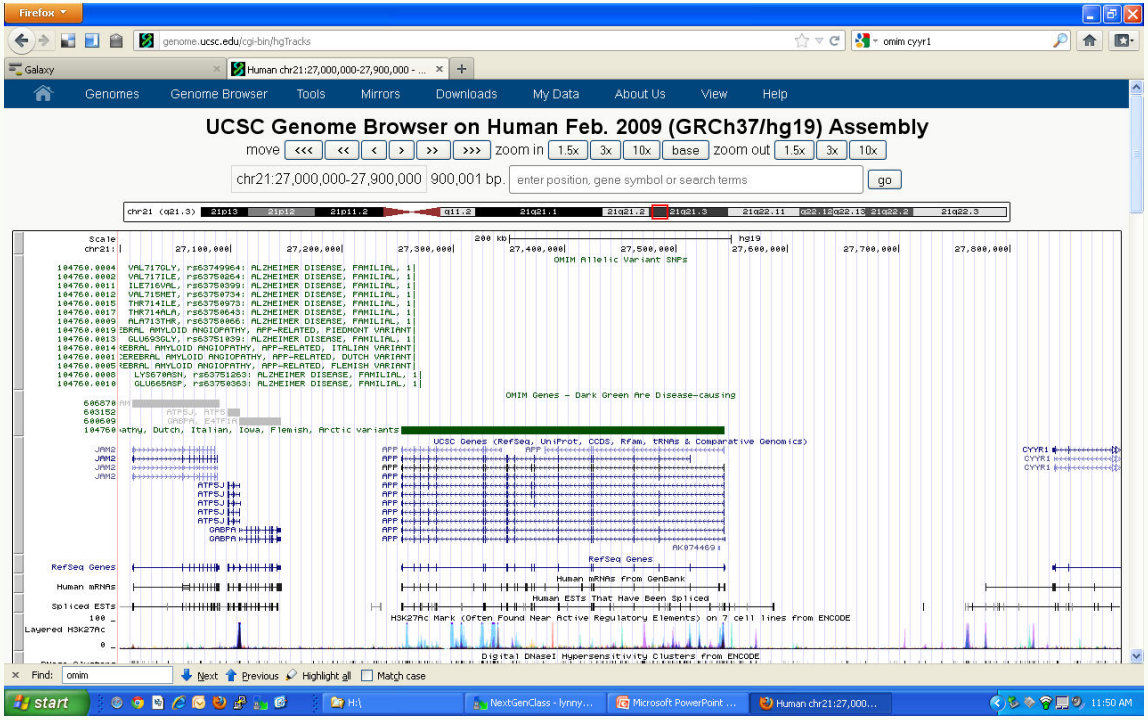
Step 2



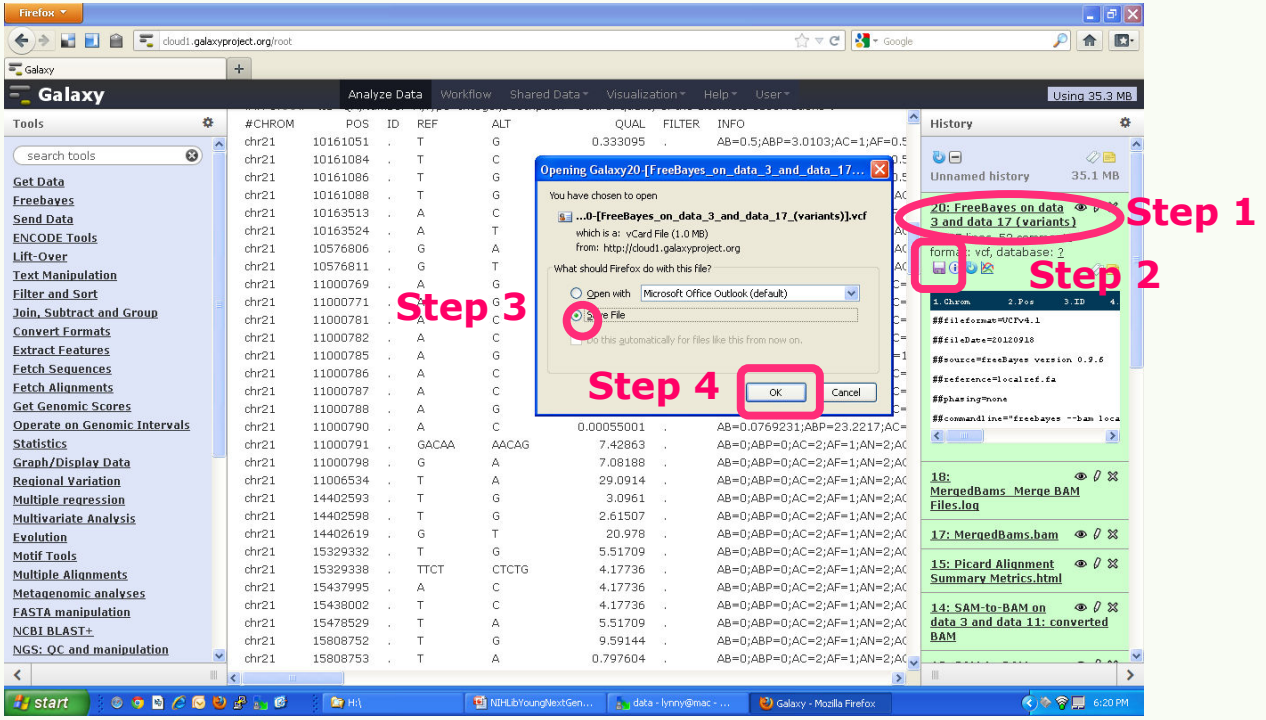
UCSC Genome Browser – OMIM Genes, OMIM AV SNPs



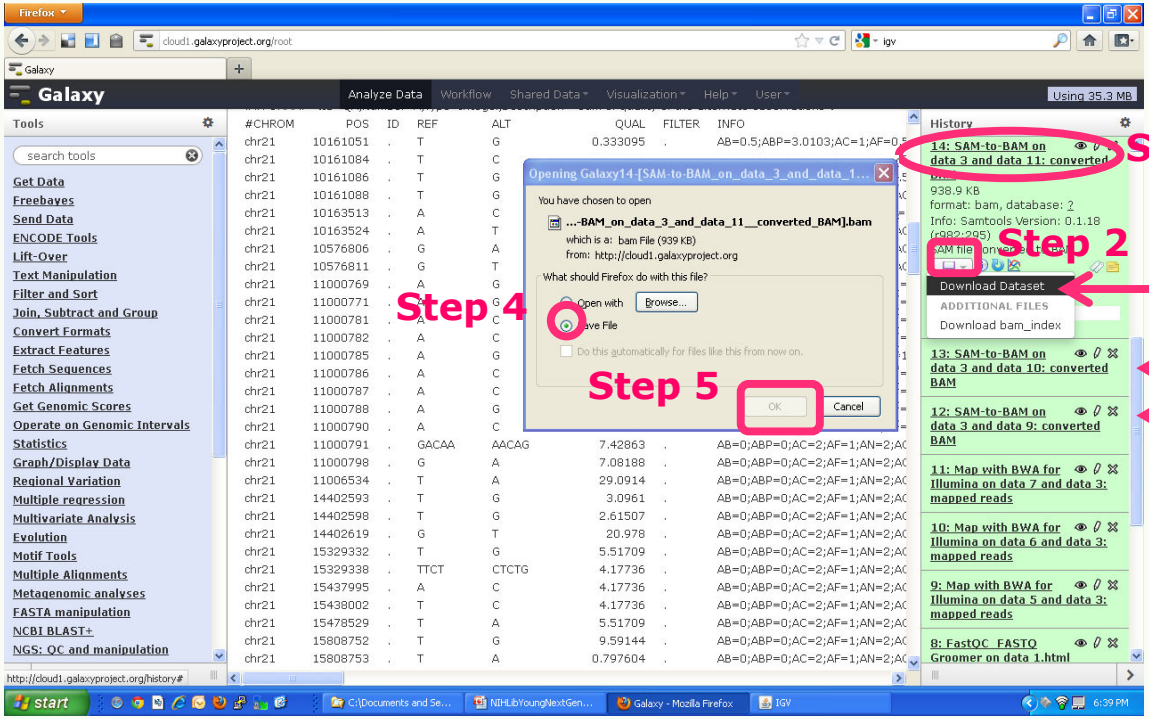
UCSC Genome Browser - Results



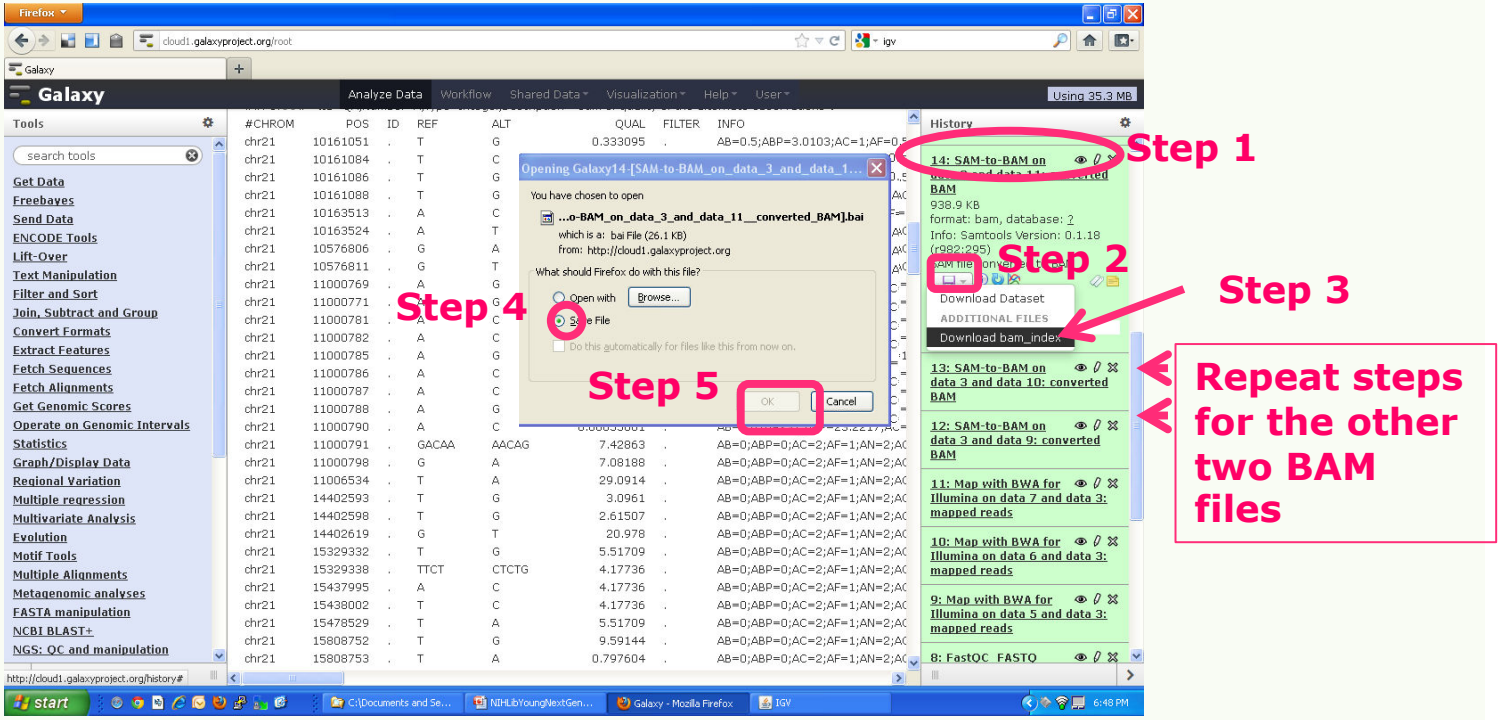
Galaxy – Exporting Data Download VCF File



Galaxy – Exporting Data Download BAM Files



Galaxy - Exporting Data Download BAI Files



Thank you for attending.