FungiDB: Data analysis via EuPathDB Galaxy

Variant Calling, Part I: Uploading data and starting the workflow (Group Exercise)

In this exercise we will work in groups to retrieve DNA sequence data from the sequence repository and analyze it for variants using a workflow in EuPathDB Galaxy.

There are different ways to get data into Galaxy. Here we will use the sample ID and get the data using the "Get Data via Globus from the EBI server using your unique file identifier" link as well direct urls, as we have done this in the Galaxy RNA-Seq section. Follow these steps to "Get Data via Globus from the EBI server":

- 1. Click on the Get Data link.
- 2. Click on the Get Data via Globus from the EBI server link.
- 3. The next window allows you to enter the sample ID. This ID starts with the letters 'SAM'. Choose the sample ID for your group from the list below and use it in this form.

Note: it is very important that you select whether the data is single or paired-ended.

4. Once the form is properly filled, click on the *Execute* button to start the data transfer process.

🌀 🏷 globus Geno	mics		Analyze Data	Workflow Shared Data - Visualization - Help - User -		Using 804.4 GB	
Tools	<u>±</u>	With EuPat	hDB Gal	axy you can:	History	€*□	
search tools		1. Start anal	zing your o	data now. All EuPathDB genomes are pre-loaded. Pre-configured workflows are	search datasets	0	
Get Data	🌀 🦄 glo	bus Genomics	e-scale o	data analysis with no prior programming or bioinformatics experience.	Unnamed history		
NGS APPLICATIONS	Tools	±	pm workf	flows using an interactive workflow editor. Learn how	0 b	•	
NGS: QC and manipulation	search too	ls 😵	rivate, c	share it with colleagues or the community.	1 This history is en	npty. You can load	
NGS: Mapping	Get Data		abo 🔰	Get Data via Globus from the EBI server using your unique file identifier	Galaxy Tool Versio	on 1.0.0)	 Options
NGS: Mapping QC NGS: RNA Analysis	Get Data v upload	ia Globus High speed file	vith E	Enter your ENA Sample id			
NGS: DNAse NGS: Peak Calling	Get Data via Globus from the EBI server using your unique file		flow	SAMEA35659918			
NGS: SAM Tools	identifier		pton i.	e. SAMN00189025			
NGS: BAM Tools NGS: SNPiR Tools	Send Data	via Globus Transfers data	flow	Data type to be transferred			
NGS: Picard	via Globus.		pton	fastq			•
NGS: Indel Analysis NGS: GATK Tools	NGS APPLIC	ATIONS	flow	Single or Paired-Ended			
NGS: GATK2 Tools NGS: GATK3 Tools	NGS: Assembly		lopHa	Paired			•
NGS: FermiKit Suite NGS: Variant Detection	NGS: Mappin	Pronie a trans Tools: FastQC	cripton C, Trimn	✓ Execute			
Consensus Genotyper for Ex	come Variants	EuPathDB We	orkflow				
NGS: Interval Tools Profile and an NGS: VCF Tools Tools: Bowtie		alyse S 2, FreeBayes, orkflow for Va	and SnpEff ariant Calling, paired-end sequencing				
NGS: EMBOSS Profile and a NGS: PECALLER Tools: Bowt		alyse SNPs. 2, FreeBayes,	and SnpEff				
<		EuPathDB Galaxy workspaces are provi	ied free of charge. V	We encrypt data transfers and storage but ultimately we cannot guarantee the security of data transmissions between EuPathDB, Globus and menufibility to back on one of the and ability and which many long form your study, and for both the mine to use a data for ability of a back on a format of the security of the sec		>	

If you click on the "Upload File from your computer" you will be able to use url links to initiate file download.

g globus Genomics	Analyze Data Workflow Shared Data - Visualization - Help - User -	u u	lsing 710.1 GB
Tools		History	<i>₽</i> ¢⊡
Search tools	1 Job has been successfully added to the queue - resulting in the following datasets: 1: ERR1767828.fastq.gz	search datasets Unnamed history	8
Get Data via Globus High speed file upload	2: ERR1767828_1.fastq.gz	2 shown 0 b	S
Get Data via Globus from the EBI server using your unique file identifier	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.	2: ERR1767828_1.fastq.gz	• / ×
Upload File from your computer		1: ERR1767828.fastq.gz	
Send Data via Globus Transfers data via Globus.			
NGS APPLICATIONS			
NGS: QC and manipulation			
NGS: Assembly			
NGS: Mapping			
NGS: Mapping QC			
NGS: RNA Analysis			
NGS: DNAse			
NGS: Peak Calling			
NGS: SAM Tools			
NGS: BAM Tools			
NGS: SNPiR Tools			

Group assignments:

Group assignments will be given in class and will be also available after the course online.

Running a variant calling workflow:

- Once the data files have been transferred into your galaxy history you need to choose an appropriate workflow. EuPathDB provides some preconfigured workflows on the EuPathDB Galaxy instance home page.
- Remember to choose the appropriate workflow Single-read or paired ended.

g blobus Genomics	Analyze Data Workflow Shared Data + Visualization + Help + User+	Using 88.1 GB	
Tools	EuPathDB system status	History C 🗘 🗔	
search tools	Lakarjone Penseges Daleven Krownen	search datasets	
Get Data	Welcome to the EuPathDB Galaxy Site	SNPCalling	
EUPATHDB APPLICATIONS	A free, interactive, web-based platform for large-scale data analysis	5.91 GB	
EuPathDB Export Tools		This history has been purged and	
NGS APPLICATIONS	With EuPathDB Galaxy you can:	deleted	
NGS: QC and manipulation	 Start analyzing your data now. All EuPathDB genomes are pre-loaded. Pre-configured workflows are available. 	50: the lifte are able to uklual (the state	
NGS: Assembly	2. Perform large-scale data analysis with no prior programming or bioinformatics experience.		
NGS: Mapping OC	Create custom worklows using an interactive worklow editor. Learn now A. Visualize your results (BioWioh GRowse)	fastq	
NGS: RNA Analysis	5. Keep data private, or share it with colleagues or the community.	57: ftp://ftp.sra.ebi.ac.uk/vol 🗶 🖋 🗙	
NGS: DNAse	To learn more shout Galavy check out nublic Galavy resources: Learn Galavy	1/fastq/SRR538/SRR538771/SRR538771_1.	
NGS: Peak Calling	to learn more about Galaxy crieck out public Galaxy resources, Learn Galaxy	fastq	
NGS: SAM Tools			
NGS: BAM Tools	Get started with pre-configured workflows:		
NGS: SNPiR Tools	(additional workflows will be added soon)		
NGS: Picard	EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates Profile a transcriptome and analyze differential gene expression		
NGS: Indel Analysis	Tools: FastQC, Sickle, CSNAP, CuffLinks, CuffDiff.		
NGS: GATK Tools	EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates		
NGS: GATK3 Tools	Profile a transcriptome and analyze differential gene expression.		
NGS: EermiKit Suite	loois: rastud, inimiomatic, lophatz, Culturins, Luituin.		
NGS: Variant Detection	EUPathDB Workflow for Illumina paired-end KNA-sea, biological replicates Profile a transcriptome and analyze differential eene expression.		
Consensus Genotyper for Exome Variants	Tools: FastOC, Trimmomatic, TopHat2, HTseq, DESeq2.		
NGS: Interval Tools	EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates		
NGS: VCF Tools	Profile a transcriptome and analyze differential gene expression.		
NGS: EMBOSS	5. Public D Media and a Marine Callera diselected and an analysis		
NGS: PECALLER	EDPathob Worklow for Variant Cauling, single-read sequencing Profile and analyse SNPs.		
NGS: SOAP	Tools: Sickle, Bowtie2, FreeBayes, and SnpEff		
NGS: Simulation	EuPathDB Workflow for Variant Calling, paired-end sequencing		
SNP/WGA: Data; Filters	Profile and analyse SNPs. Tools: Sirkle Bowle? EreaBayes and SnpEff		
SNP/WGA: GC; ED; Plots SNP/WGA: Statistical Models	rous, siene, downey, ricelayes, and argues		
Deserve Association	EVENTIBE Galaxy involvages are provided free of many. We encycle data transfers and disciple to ultimately we cancel guarded the sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and Williams, Amazon Cloud Services, and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy of data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy data transmission. Between 6, 540008, Galaxy and the ultimate in a sourchy data transmission. Betw		
<	information in the files you analyze. SuPathDB, Globus and attilates University of Pennsylvania, University of Everpool, and Amazon Cloud Services do not take any responsibility and are not liable for the loss and/or release of	II >	

The pre-configured workflows follow these steps:

- Determine quality of the reads in your files and generates FASTQC reports
- Trim reads based on their quality scores
- Align reads to a reference genome using Bowtie2 and generating coverage plots
- Sort alignments with respect to their chromosomal positions
- Detect variants using FreeBayes
- Filter SNP candidates
- Analyze and annotate of variants, and calculation of the effects via SnpEff



- Next, set workflow parameters.
 - Make sure that the input steps for paired-end are set to the $xxxx_1.fastq.gz$ and $xxxx_2.fastq.gz$ as by default both have the same one selected.

Step 1: Input dataset 1	
Input Dataset 🍈	
53: SRR834923_1.fastq.gz	٥
type to filter	
Step 2: Input dataset	
8	
Input Dataset 🍈	
54: SRR834923_2.fastq.gz	٥
type to filter	

- Select the correct reference genome (for steps: Bowtie2, FreeBayes, SnpEff)
- Click on the *Run Workflow* button.