# Sample Size Does Matter: Scaling Up Analysis in Galaxy with Metagenomics

Daniel Blankenberg dan@bx.psu.edu Post Doctoral Research Associate Galaxy Team / Penn State

#### **Metagenomics Powerhouse**



## Sarah J. Craig, PhD sarah@bx.psu.edu

Center for Medical Genomics, Penn State University, University Park Hershey Medical School, Pennsylvania, USA Department of Biology, Penn State University, University Park, Pennsylvania

~500 Samples: Buccal & Stool from Mother-Child Pairs

# The **– Galaxy** Team





Enis Afgan



Dannon Baker



Dan Blankenberg



**Dave Bouvier** 



Marten Čech



John Chilton



**Dave Clements** 



Nate Coraor



**Carl Eberhard** 



Jeremy Goecks



Sam Guerler



Rémi Marenco



Jen Jackson



**Ross Lazarus** 



**Nick Stoler** 



Anton Nekrutenko



James Taylor



Nitesh Turaga

http://wiki.galaxyproject.org/GalaxyTeam

#### Overview

Brief Introduction to Metagenomics
Doing It
Challenges, Solutions, and Future Work

#### Overview

- Brief Introduction to Metagenomics
- Doing It
- Challenges, Solutions, and Future Work

#### Metagenomics

 Study of genetic material recovered directly from environmental samples

isolation and lab cultivation not required

- High-throughput sequencing
  - 16S rRNA targeted
  - Whole-genome shotgun



Grice E.A. & Segre J.A. (2012) The Human Microbiome: Our Second Genome, Annu. Rev. Genomics Human Genet. 13, 151-170



Grice E.A. & Segre J.A. (2012) The Human Microbiome: Our Second Genome, Annu. Rev. Genomics Human Genet. 13, 151-170

#### Whole-Genome Shotgun



Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. Biol. Procedures Online (2009).

#### 16S rRNA



Julien Tremblay. JGI.

1400

800

1000

1200

V9

Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples.Nat Rev Genet. 2005 Nov;6(11):805-14. Review. PubMed PMID: 16304596.

#### Overview

- Brief Introduction to Metagenomics
  Doing It
- Challenges, Solutions, and Future Work

### Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### Importing FTP uploaded files into Galaxy

= Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -
Tools	Developed from web an unload from diek
search tools	Download from web or upload from disk
Get Data	Regular Composite
Send Data	
Text Manipulation	
Filter and Sort	
Join, Subtract and Group	
Convert Formats	ETP files
Extract Features	
Fetch Sequences	This Galaxy server allows you to upload files via FTP. To upload some files, log in to
Fetch Alignments	the FTP server at 54.211.185.247 using your Galaxy credentials (email address and password)
Statistics	
Graph/Display Data	Die files:
Metagenomics: Mothur	Name Size Created
Operate on Genomic Intervals	F3D0_S188_L001_R1_001.fastq 4.2 MB 06/25/2016 11:01:21 PM
Phenotype Association	F3D0_S188_L001_R2_001.fastq 4.2 MB 06/25/2016 11:01:22 PM
snpEff	
BEDtools	F3D141_S207_L001_R1_001.fastq 3.2 MB 06/25/2016 11:01:23 PM
EMBOSS Regional Variation	Type (set a F3D141_S207_L001_R2_001.fastq 3.2 MB 06/25/2016 11:01:24 PM
FASTA manipulation	
Evolution	Contraction of the Contraction o
Multiple Alignments	Choose local file Choose FTP file & Paste/Fetch data Pause Reset Start Close
Metagenomic analyses	
_001	fastq
► F3D142_S2	08_L001_R2 1.7 MB fastosanger 🗶 📿 Additional Sp 💌 🏟 🔿 🖄
Туре (	set all): fastqsanger v C Genome (set all): Additional Species A v
	□ Choose local file 🕞 Choose FTP file 🕼 Paste/Fetch data Pause Reset Start Close

History	2≎⊡				History	24
search datasets MiSeq Data 38 shown 157 95 MB All None	For all selected	Create Datase	History < Back to 434 pairs no fa 'QC and Join paired end ( uploaded file)'	C Copy of (imported from	K Back to Microbiome - 4 Metagenomic Pairs (Dan Load) 102C1Buccal_S13_L001 a pair of datasets	<u>34</u> Dec. 18, 2 _ <b>001</b>
Hide datasets		1: Microbiome - 434 Metageno	Microbiome - 434 Meta	agenomic	<u>forward</u>	۲
Unhide datase	ts	mic Pairs (Dan Dec. 18, 2015 Lo ad)	a list of paired datasets	.5 Load)	reverse	۲
Undelete dataset	sets	a list of 434 dataset pairs	102C1Buccal S13 L001	001	132.7 MB format: <b>fastqsanger</b> , dat	abase: <u>?</u>
Build Dataset I	List		102C1Stool S1 L001 0	01	uploaded fastqsanger fi	е
Build Dataset F	Pair		a pair of datasets		E 0 III	•
$\frac{34: F3D7 S195}{q}$ $\frac{33: F3D7 S195}{q}$	Create a collection o	f paired datasets atasets have been successfully paired				د. در ×
<u>     32: F3D6 S194</u> <u>     q     </u>	19 unpaired forward	(19 filtered out) Ch	oose filters Clear filters Auto-pair		19 unpaired reverse - (19 filtered	l out)
<u> <u> </u></u>	F3D0_S188_L001_R1_0	01.fastq	Pair these datasets		F3D0_S188_L001_R2_001.fa	istq S
- 30: F3D5 S193	F3D1_S189_L001_R1_0	01.fastq	Pair these datasets		F3D1_S189_L001_R2_001.fa	istq s
đ	F3D2_S190_L001_R1_0	01.fastq	Pair these datasets		F3D2_S190_L001_R2_001.fa	istq
29: F3D5 5193	F3D3_S191_L001_R1_0	01.fastq	Pair these datasets		F3D3_S191_L001_R2_001.fa	istq
д	F3D5_S193_L001_R1_0	01.fastq	Pair these datasets		F3D5_S193_L001_R2_001.fa	istq S
	F3D6_S194_L001_R1_0	01.fastq	Pair these datasets		F3D6_S194_L001_R2_001.fa	stq
	F3D7_S195_L001_R1_0	01.fastq	Pair these datasets		F3D7_S195_L001_R2_001.fa	stq
			0 paired			

(no paired datasets yet)

#### QC, Prepare, manipulate, filter Reads



Input: a List of paired sequencing Reads Outputs: QC'd Joined paired reads, filtered for chimeras (FASTA) Quality Reports (html and single summary table)

#### Create Workflow for Easy-to-Read FastQC Report

The second state of the se	User - Using 5.6 GB
Workflow Canvas       Name:         Single FASTQ Files       Select lines from         Collection Type:       Select lines from         list       Details         Output       Select lines that match an expression (Galaxy Version 1.0.1)         Select lines from       Configure Output: 'out	Column Join × Tabular files tabular_output (tabular) • script_output (txt)
✓ FastQC × ▲ that	(Galaxy Version 0.0.1)
Short read data from your current Matching	this must be Tabular files
Institution       Image: Contaminant list       Image: Contaminant list <td< td=""><td>e the output more</td></td<>	e the output more
Change datatype	item
	• 0
Add an annotation or note for this step. It will be shown with the workflow.	Fill character
Email notification         Yes       No         An email notification will be sent when the job has completed.       This action will set tags dataset.	s for the Additional datasets to create
USE AS a SUDWORKIOW Treat an entire workflow as a single Tool Module Output cleanup Yes No Delete intermediate outputs if they are not used as input for another job. TIP: If your data is no delimited, use Text M >Convert	ot TAB Manipulation-

#### Single Summary FastQC Report

🔿 🔿 🖉 🗮 Galaxy / Da	an's Brubeck Pla	a × 📃 🕻	Galaxy		×																								Gmail 🛆	n <sub>M</sub>
← → C 🗋 brubec	ck.bx.psu.ed	<b>u</b> :8051																											2	≣
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now														marks																
🔁 Galaxy / Da	an's Bru	ıbeck	Playg	round	d					A	Analyze Dat	ta Workf	low Shar	red Data <del>-</del>	Visualiza	ation <del>-</del> A	dmin He	elp <del>+</del> Use	er-										Using 9.4	4 TB
Tools	1	buccal 254C1-	_S20_L001 buccal S3	_001_2 30 L001 (	253C1- 001 3	buccal_S2 254C1-s	20_L001_ stool S1	001_3 2 L001 00	253C1-4 01 2	stool_S1 254C1-	7_L001_0 stool S1	01_2 2 L001 00	253C1-4 01 3	stool_S17 254M-bu	7_L001_0	01_3 0 L001 0	253M-bu 01 2	uccal_S40 254M-bi	0_L001_0 uccal S4	01_2 0 L001 0	253M-bi 01 3	1ccal_S40 255C1-b	0_L001_0	01_3 2 L001 00	254C1-1 01 2	buccal_S 255C1-	30_L001_(	001_2	History 24	<b>¢</b> 🗆
search tools		buccal	_S2_L001_	001_3	255M-b	uccal_S4_	_L001_00	1_2 5_1001_0	255M-bi	uccal_S4	_L001_00	1_3 44 T001 (	256C1-1	buccal_S	33_L001_0	001_2	256C1-1	buccal_S	33_L001_	001_3	256C1-s	stool_S2_	_L001_00	1_2 1	256C1-	stool_S2	_L001_00	1_3	search datasets	
search tools		buccal	_S26_L001	_001_2	257M-b	ouccal_S26	6_L001_0	01_3	258C1-1	buccal_S	39_L001_	001_2	258C1-1	buccal_S	39_L001_	001_3	258C1-s	stool_S9	_L001_00	1_2	258C1-	stool_S9	_L001_00	1_3	258M-b	uccal_S2	7_L001_0	01_2	Scarch datasets	<u> </u>
Stampy Genotyping		258M-bi buccal	uccal_S27 S45 L001	_1001_00 001 3	01_3 260C1-	259C1-1 stool S8	L001 00	32_L001_( 1 2	001_2 260C1-4	259C1-1 stool S8	L001 00	32_L001_0 1 3	001_3 260M-bi	259M-bu accal S29	uccal_S1 9 L001 00	5_L001_0 01 2	01_2 260M-bi	259M-bi accal S2	uccal_S1 9 L001 0	5_L001_0	262C1-1	260C1-E Duccal SI	buccal_S 13 L001 (	45_1001_0 001 2	262C1-1	260C1- buccal Si	13 L001 (	001 3	434 pairs with QC and Join pair end	red
Get Data		262M-b	uccal_S23	_L001_0	01_2	262M-bu	iccal_S2	3_L001_0	01_3	263C1-	buccal_S	19_L001_0	001_2	263C1-1	buccal_S	19_L001_	001_3	263C1-	stool_S1	6_L001_0	01_2	263C1-s	stool_S1	6_L001_00	01_3	263M-		_	6010 shown, 7908 hidden	
Send Data		264M-bi	_S34_L001 uccal S37	L_001_2 L001_0	263M-b 01 3	uccal_S34 265C1-b	4_L001_0	01_3 18 L001 (	264C1-1 001 2	buccal_S 265C1-l	31_L001_ buccal S	001_2 18 L001 (	264C1-1 001 3	265C1-8	31_L001_( stool S4(	001_3 6 L001 0	264C1-s 01 2	265C1-1	_L001_00 stool S4	1_2 6 L001 0	264C1-s 01 3	265M-bu 265M-bu	_L001_00	1_3 8 L001 00	264M-bi 01 2	1cca1_S3 265M-	7_L001_0	01_2	457.13 GB	•
Lift-Over		buccal	S28_L001	_001_3	266C1-	buccal_S4	47_L001_	001_2	266C1-1	buccal_S	47_L001_	001_3	266C1-	stool_S7	_L001_00	1_2	266C1-s	stool_S7	_L001_00	1_3	266M-bu	iccal_S21	1_L001_0	01_2	266M-b	uccal_S2	1_L001_0	01_3		
Text Manipulation		267C1-l buccal	buccal_S3 S36 L001	88_L001_0	001_2 268C1-	267C1-buccal S3	buccal_S 36 L001	38_L001_0 001 3	001_3 268C1-	267C1- sttool S	stoo1_S1 5 L001 0	0_L001_00 01 2	01_2 268C1-	267C1-s sttool S	stool_S10 5 L001 00	0_L001_0 01 3	01_3 268M-bi	267M-bi lccal S22	uccal_S3 2 L001 0	_L001_00	1_2 268M-bi	267M-bu iccal S22	1ccal_S3 2 L001 0	_L001_001 01 3	1_3 270C1-1	268C1- buccal Si	17 L001 (	001 2	13918: Column Join o	×
Filter and Sort		270C1-	buccal_S1	7_L001_0	001_3	270C1-s	stool_S1	4_L001_0	01_2	270C1-	stool_S1	4_L001_0	01_3	270M-bu	uccal_S2	0_L001_0	01_2	270M-b	uccal_S2	0_L001_0	01_3	271C1-4	buccal_S	24_L001_0	001_2	271C1-			3915, and others	
Join, Subtract and Group		272C1-	_S24_L001 stool S46	L_001_3	271C1-	stool_S4_ 272C1=9	_L001_00	1_2 6 1.001 00	271C1-4	stool_S4 272M-b	_L001_00	1_3 0_1.001_0	271M-bi 01 2	1ccal_S25 272M-bi	5_L001_0	01_2 0 T.001 0	271M-bu 01 3	274C1-1	5_LOO1_O	01_3 22 T.001 (	272C1-1	274C1-k	25_L001_0	001_2 22 T.001 (	272C1-1	274C1-	25_L001_0	001_3		_
Convert Formats		stool_	S42_L001_	001_2	274C1-	stool_S42	2_L001_0	01_3	276C1-1	buccal_S	15_L001_	001_2	276C1-1	buccal_S	15_L001_	001_3	276C1-	stool_S4	3_L001_0	01_2	276C1-	stool_S43	3_L001_0	01_3	276M-b	uccal_S2	7_L001_0	01_2	13917: Cut on collection 134 81	×
Extract Features		276M-bi	uccal_S27 S7 1.001	L001_0	01_3 278C1-	277C1-h	buccal_S	23_L001_( 001_2	001_2 278C1_1	277C1-l	buccal_S	23_L001_( 001_3	278C1-4	277C1-s	stool_S4	5_L001_0	01_2 278C1_4	277C1-	stool_S4	5_L001_0	278M-bi	277M-bu	100al_S7	_L001_001	1_2 278M-bi	277M-	6 1.001 0	01 3	a list of datasets	
Fetch Sequences		280C1-1	buccal_S2	4_L001_0	001_2	280C1-h	ouccal_S	24_L001_0	001_3	280C1-	stool_S3	7_L001_0	01_2	280C1-s	stool_S3	7_L001_0	01_3	280M-b	uccal_S2	8_L001_0	01_2	280M-bu	uccal_S2	B_L001_00	01_3	281C1-	0_1001_0		12482: Column Join o	
Fetch Alignments		buccal	_S9_L001_ stool S40	001_2	281C1-	-buccal_S9	9_L001_0	01_3	281C1-4	282M-b	6_L001_0	01_2	281C1-	stool_S36	6_L001_0	01_3 13_1001	281M-bu	1ccal_S2	6_L001_0	01_2	281M-bi	1ccal_S26	6_L001_0	01_3 9 T001 00	282C1-	stool_S4(	0_L001_0	01_2	n data 11305. data 11	×
Get Genomic Scores		stool_	S39_L001_	001_3	283M-b	uccal_S3_	_L001_00	00100. 1_2	283M-bi	uccal_S3	_L001_00	001001 13	284C1-1	buccal_Si	14_L001_0	001_2	284C1-1	buccal_S	14_L001_	001_3	284C1-s	stool_S30	0_L001_0	01_2	284C1-	stool_S3	0_L001_0	01_3	304, and others	
Operate on Genomic Intel	ervais	284M-b	uccal_S12	2_L001_0	01_2	284M-bu	iccal_S1	2_L001_0	01_3	285C1-1	buccal_S	17_L001_0	001_2 28601 V	285C1-1	buccal_S	17_L001_	001_3	285C1-	stool_S3	3_L001_0	22601	285C1-s	stool_S3	3_L001_00	01_3	285M-	1 7 001 0	01.2	13481: Paste on collection 10	
<u>Statistics</u>		286M-b	uccal_S21	0010	01_3	289C1-b	buccal_S	8_L001_0	01_2	289C1-1	buccal_S	8_L001_0	01_3	289C1-s	stool_S4	4_L001_0	01_2	289C1-	stool_S4	4_L001_0	01_3	289M-bu	accal_S6	_L001_001	1_2	289M-	1_1001_0	°1_2	871 and collection 9131	^
Granh (Display Data		buccal	_S6_L001_	001_3	290C1-	buccal_S4	48_L001_	001_2	290C1-1	buccal_S	48_L001_	001_3	290M-b	accal_S41	1_L001_0	01_2 2 T001 0	290M-bu	accal_S4	1_L001_0	01_3	294C1-1	ouccal_SI	18_L001_	001_2	294C1-1	buccal_Si	18_L001_(	001_3	a list of datasets	
Graph/Display Data		stool	S38_L001	001_2	302C1-	stool_S38	B_L001_0	4_L001_00 01_3	302M-bi	uccal_S2	9_L001_0	2_L001_00 01_2	302M-bi	iccal_S29	9_L001_0	2_L001_0 01_3	01_3	30201-1	buccal_s	10_1001_0	001_2	30201-0	buccar_s	10_1001_0	501_5	30201-			13046: VSearch chimera dete	~
Nultiple regression		>>Adap	ter Conte	ent	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	ction on collection 10436: Ch	-
Multiveriete Analysis		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	imera Alignments	
Evolution		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	a list of datasets	
Evolution Motif Tools		pass	pass	pass pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	13045: VSearch chimera dete	×
Multiple Alignments		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	ction on collection 10436: No	
Mataganomic analyses		pass	pass	pass pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	a list of datasets	
EASTA manipulation		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	12044 Manual allowers date	
NGS: OC and manipulation	n	pass	pass	pass pass	pass	pass	pass	pass	pass pass	pass	pass pass	pass	pass	pass	pass	pass pass	pass	pass	pass	pass	pass pass	pass	pass	pass	pass	pass	pass	pass pass	ction on collection 10436	×
NGS: Picard (heta)	<u>///</u>	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	a list of datasets	
NGS: Mapping		pass	pass	pass pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	13043: VSearch chimera dete	
NGS: RNA Analysis		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	ction on collection 10436: Ch	*
NGS: SAM Tools		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	imera Information	
NGS: GATK Tools (beta)		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	a list of datasets	
NGS: Variant Detection		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	11469: VSearch chime 💿 🖋	×
NGS: Peak Calling		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	ra detection on data 1	
NGS: Simulation		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	0042: Chimera Alignments	
SNP/WCA: Data: Filters		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	11468: VSearch chime 💿 🖋	×
Human Cenome Variation	n	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	ra detection on data 1	
		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	0042: Non Chimera	
(		pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass		>

### Doing It

# QC and Preparation Classify Reads

- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### **Classify Reads**

#### • "Binning"

"best effort" to identify reads or contigs with certain groups

of organisms designated as OTUs

Operational Taxonomic Units (OTUs)

•operational definition of a species or group of species

Algorithms

- Taxonomy Dependent
  - alignment to known sequences and species
  - misses reads that are absent in database
  - obtain estimates of the profile/abundance of 'known' taxonomic groups

Taxonomy Independent

- group/bin reads in a given dataset based on their mutual similarity
- considers content of reads only
- no database

#### **Kraken Classifier**

- assign taxonomic labels to short DNA sequences
- exact alignment of k-mers



#### **Classify Reads**



Input: a List of sequencing Reads (FASTA or FASTQ) Outputs: Summary Table of counts (filtered or non-filtered joined)

#### **Abundance Counts**

000 / 🗮 Workflow home / Dan's	Gmail 🛆 🖉 Workflow home / Dan's Bru × 🚍 Galaxy / Dan's Brubeck Pla × 🛒 Galaxy × 🗐 brubeck.bx.psu.edu:8051/ ×																													
← → C 🗋 brubeck.bx.psu.e	edu:	8051																												☆≡
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now																														
🔁 Galaxy / Dan's Br	rub	beck P	laygı	round						A	nalyze Data	a Workfl	ow Shar	red Data <del>-</del>	Visualiza	tion <del>-</del> A	Admin H	lelp <del>+</del> Use	er-	•										Using 9.4 TB
Tools		buccal_S3	_L001_0	001_2	284C1-b	ouccal_SI	4_L001_0	001_2	284C1-	stool_S3	0_L001_00	01_2 T 001_0	284M-b	uccal_S1	2_L001_00	01_2	285C1-	buccal_SI	17_L001_	001_2	285C1-s	stool_S3	3_L001_0	01_2	285M-b	uccal_S31	1_L001_00	01_2	History	2≎⊡
search tools	I	buccal_S4	8_L001 al S29	001_2 001_00	290M-bu	iccal_S41	_L001_00	)1_2	294C1-1	buccal_S	18_L001_0	001_2	294C1-	stool_S3	4_L001_00	)1_2	294M-b	uccal_S32	2_L001_0	01_2	302C1-1	puccal_S	10_L001_	001001001001001001001	302C1-	stool_S38	B_L001_00	01_2	search datasets	0
Stampy Genotyping Get Data Send Data Lift-Over Text Manipulation Filter and Sort Noin, Subtract and Group Convert Formats Extract Features Fetch Sequences Fetch Alignments Get Genomic Scores Operate on Genomic Intervals Statistics Wavelet Analysis Graph/Display Data Regional Variation Multiple regression Multivariate Analysis		ACI         C           8         2           9         1           0         0           1         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           1         0           1         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0           0         0		-1 - -1 - 	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 1 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 1 0 0 1 3 0 0 0 0 0 0 0 0 0 0 0 0 0	0 1 0 1 1 0 0 0 4 2 1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 2 0 0 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 1 8 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 1 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 1 1 0 0 0 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0	0 4 1 0 3 1 0 1 0 0 0 0 0 1 4 2 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 2 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 4 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 1 0 0 1 4 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0	2 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0	0 1 0 2 0 0 0 0 4 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	4 3 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	gg on 434 pairs krak and join on single rea database 3162 shown, 760 hidden 68.16 GB <b>Q</b> 3922: Select on da a 3920 <b>Q</b> 3921: Column Join on data 3917, data 39 16, and others <b>Q</b> 3920: Column Join on data 3917, data 39 16, and others 3919: Select on data 3482 3918: Cut on collection a list of datasets <b>Q</b> 3917: Cut on data 3480	en and filter d - set
<u>Evolution</u> Motif Tools Multiple Alignments		0 0 0 0		0 0 0	0 1 0	0 0 1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 1 0	0 0 1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 1 1	0 0 1	0 0 0	<u>3916: Cut on data</u> <u>3479</u>	(1)
<u>Multiple Alignments</u> Metagenomic analyses		0 ( 0 1 0 (		1 0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	<u>3915: Cut on data</u> 3478	A X
ASTA manipulation <u>VGS: QC and manipulation</u> <u>VGS: Picard (beta)</u> <u>VGS: RNA Analysis</u> <u>VGS: SAM Tools</u> <u>VGS: GATK Tools (beta)</u> <u>NGS: Variant Detection</u> <u>NGS: Peak Calling</u> <u>NGS: Simulation</u> <u>SNP/WGA: Data; Filters</u> Human Genome Variation		13     10       13     1       23     3       7     1       0     1       76     4       17     6       18     1       16     7       5     1       4     1       8     6       9     1       37     2	5 5 2 7 3 3 6 6 1 1 4 4 0 0	75 590 23 4 55 1 0 4 9 29 36 1 23 9 37 5 5	16 39 7 5 33 5 20 18 14 3 6 8 9 22 17 23 16759	37 33 69 5 182 9 2 6 4 9 2 2 1 8 20 7 14	22 12 7 37 50 4 3 20 3 9 2 3 13 3 16	6 19 9 113 124 4 3 11 13 21 1 11 15 9 14	14 14 33 26 196 11 2 1 3 73 7 2 8 13 10	9 8 12 17 155 9 2 20 5 11 3 3 20 2 18	28 20 19 39 3 5 32 29 15 7 5 14 15 24	12 12 10 2 124 4 0 15 19 16 2 2 5 7 33	20 4 6 15 5 5 2 18 11 6 1 10 4 12 27	25 14 12 50 57 1 11 21 6 17 7 7 10 17 12	152 9 1 109 69 3 5 15 5 25 0 111 16 3 26	11 7 7 66 508 5 25 6 22 4 4 2 1 10 13	37 2 7 127 4 3 1 4 33 20 4 6 6 24 3	14 13 10 84 100 7 24 15 5 49 0 5 6 0 35	19 19 3 25 87 7 24 13 17 13 5 2 10 14 15	60 12 16 65 3 29 12 18 6 13 6 20 39 11	20 2 8 84 217 16 2 3 4 13 5 2 13 5 12	19 3 6 45 164 5 10 9 10 19 2 1 8 14 11	38 7 83 67 39 2 9 11 29 6 11 1 16 10 24	14 13 6 30 168 8 18 15 11 6 2 3 13 2 27	28 23 13 86 93 4 11 15 14 4 6 13 28 25	28 24 9 41 3 2 5 6 25 9 2 4 22 5 16	12 14 5 47 34 15 4 20 1 13 5 4 12 780	29 2 3 127 3 11 1 1 17 26 1 36 6 43 11	<ul> <li>3914: Cut on data 3477</li> <li>3913: Cut on data 3476</li> <li>3912: Cut on data 3475</li> <li>3911: Cut on data 3474</li> <li>3910: Cut on data 3473</li> </ul>	• / X • / X • / X • / X
VCE Tools		Actir 716 1 547 1	087 4732	69999 45796	16758 5339 5158	1443 2097 15158 42005	4717 7755	4773 963	2284 7783 2542	3464 7341 8895 2004	6413 62951	4581 31618	2865 2054 7831	8268 1063 182	4147 17782	9672 3846 9495	2484 668 6606	34206 9477 1407 31402	2195 7265	6182 22055 3561	4511 18371 839 21024	6135 346 2410	3985 536 16410	16768 3855 4544	4511 2883 8400	29882 5604 807	13431 12421	17438 219	<b>3909: Cut on data</b>	<pre></pre>

### Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### Normalizing

#### Rarefy

- Total sum scaling
- Cumulative Sum Scaling
- DeSeq2 process
- •

#### Rarefaction



Mohd Shaufi MA, Sieo CC, Chong CW, Gan HM, Ho YW. Deciphering chicken gut microbial dynamics based on highthroughput 16S rRNA metagenomics analyses. Gut Pathog. 2015 Feb 26;7:4. doi: 10.1186/s13099-015-0051-7.

• Have we sequenced enough? Normalize counts across samples



1	2
"x"	
"X102C1Buccal_S13_L001_001_2"	0.0439628348495116
"X102C1Stool_S1_L001_001_2"	0.0436719139786513
"X102M.buccal_2"	0.0441353284015752
"X103C1.buccal_S26_L001_001_2"	0.0412752117548852
"X103C1.stool_S15_L001_001_2"	0.0375047214380159
"X103M.buccal_S37_L001_001_2"	0.0468916650084611
"X105C1Buccal_S14_L001_001_2"	0.0355364524551467
"X105C1Stool_S2_L001_001_2"	0.0475142673724111
"X105M.buccal_S1_L001_001_2"	0.0470150093788334
"X106C1Buccal_S15_L001_001_2"	0.0467482587333604
"X106C1Stool_S3_L001_001_2"	0.0534846235423753
"X106M.buccal_S2_L001_001_2"	0.040421620256698
"X107C1Buccal_S16_L001_001_2"	0.0471970891841178
"X107C1Stool_S4_L001_001_2"	0.0533971575341763
"X107M.buccal_S3_L001_001_2"	0.0472890304994955
"X108C1Buccal_S17_L001_001_2"	0.0481609441445347
"X108C1Stool_S5_L001_001_2"	0.045170380186428
"X108M.buccal_S4_L001_001_2"	0.0378253483991975
"X109C1Buccal_S18_L001_001_2"	0.0310756229168307
"X109C1Stool_S6_L001_001_2"	0.0364931311092052
"X109M.buccal_S5_L001_001_2"	0.0375674715083014
"X110C1Buccal_S19_L001_001_2"	0.035845121476847
"X110C1Stool_S7_L001_001_2"	0.0492491404316483
"X110M.buccal_S6_L001_001_2"	0.039821866382552
"X112C1.buccal_S27_L001_001_2"	0.0325941364457868
"X112C1.stool_S16_L001_001_2"	0.0524261115683378
"X112M.buccal_S38_L001_001_2"	0.0359530450432949
"X113C1.buccal_S28_L001_001_2"	0.0360305077235005
"X113C1.stool_S17_L001_001_2"	0.0517149011158896

#### Vegan Tool Suite - Rarefaction Slopes

### Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### a-diversity

## biodiversity in a defined

- habitat
- ecosystem
- sample
- Community Richness
  - how many organisms are really there

## Alpha Diversity

<b>=</b> Galaxy	Analyze Data Workflow Shared Data <del>-</del> Visualiz	ation <del>-</del> Admin Help	p∓ User∓		Us	ing 5.8 GB
Fools	Vegan Diversity index (Galaxy Version 0.0.3)			▼ Options	History	€\$□
Regional Variation	File with abundance values for community				search datasets	8
FASTA manipulation Evolution	C       507: Vegan Rarefaction on data 500 (Random rarefied comr         Rows are samples; columns are species/phyla/community classifier	nunity matrix)		•	MiSeq Data 130 shown, 378 <u>hidden</u>	
Multiple Alignments Metagenomic analyses	Group name column				812.52 MB	2 > >
Beta Diversity using scikit-bio	Column: 1			•	508: Column frequencies	• # ×
Vegan Fisher Alpha index	Species, phylum, etc				on data 500	
Vegan Diversity index	Sample count columns				507: Vegan Rarefaction	• # ×
Vegan Rarefaction curve and statistics	Select/Unselect all Column: 2 × Column: 3 × Column: 4 × Column: 5 × Column: 6	× Column: 7 × Colu	ımn: 8 🗙 Column: 9 🗶 Colu	mn: 10	on data 500 (Random rar efied community matrix)	
VSearch alignment	× Column: 11 × Column: 12 × Column: 13 × Column: 14 × Column	n: 15 🗶 Column: 16	× Column: 17 × Column: 18	8	506: Vegan Rarefaction on data 500 (plot)	• / ×
VSearch dereplication	× Column: 19 × Column: 20				505: Vegan Parefaction	
VSearch sorting	Select each column that contains counts				on data 500 (slope of cu	• # X
VSearch shuffling	Innut has a header line				<u>rve)</u>	
VSearch search	Yes No	1	2		504: Vegan Rarefaction	• 🖋 🗙
VSearch masking	Diversity index to use	5350	X		babilities)	
VSearch clustering		F3D0	2.21245501532285		503: Vegan Rarefaction	• # X
VSearch chimera detection		F3D2	2.01951605149471		on data 500 (estimated r	
BIOM metadata add-metadata	Margin for which the index is computed	F3D3	1.64741465806235		icnness)	
Convert BIOM formats	1	F3D5	2.21811786026141		502: Vegan Rarefaction on data 500 (frequency	• 🖋 🗙
Kraken taxonomic report view	The logarithm base	F3D6	2.01124753861779		of species)	
report of classification for multiple	Natural Logarithm: exp(1)	F3D7	1.56781184051996	•	501: Vegan Rarefaction	@ # ¥
samples		F3D8	2.25470249485439		on data 500 (number of	
Kraken-translate convert	✓ Execute	F3D9	2.16635907862611		species)	
taxonomy IDs to names		F3D141	1.88938647101194		500: Column substitutio	• / ×
Kraken-report view a sample	Calculate Diversity index using vegan and selected method.	F3D142	1.69697928769152		<u>n on data 498</u>	
, f 1 16 ,		F3D143	1.86853892207042			
		F3D144	1.71506115634714			
		F3D145	1.55429517220742			
		F3D146	2.1131686074353			
		F3D147	1.6036205003031			
		F3D148	1.73860419546772			
		F3D149	2.02799055973392			
		F3D150	1.99456407827727			

### Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### **β-diversity**

compares species diversity between

- habitats
- samples
- How similar are two samples?

#### **Beta Diversity**

<b>-</b> Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -		Usi	ing 5.8 GB
ools	Vegan Beta Diversity curve and statistics (Galaxy Version 0.0.2)	ptions	History	<i>℃</i> � □
Kraken taxonomic report view report of classification for multiple	File with abundance values for community		search datasets	0
samples	C       507: Vegan Rarefaction on data 500 (Random rarefied community matrix)	-	MiSeq Data	
<u>Kraken-translate</u> convert taxonomy IDs to names	Group name column		812.52 MB	<b>•</b>
Kraken-report view a sample report of your classification	Column: 1	•	509: Vegan Diversity on data 507	• / ×
Kraken-mpa-report view report of classification for multiple samples	Sample count columns		508: Column frequencies on data 500	• / ×
Kraken-filter filter classification by confidence score	x Column: 2       x Column: 3       x Column: 4       x Column: 5       x Column: 6       x Column: 7       x Column: 8       x Column: 9       x Column: 10		507: Vegan Rarefaction	• / ×
Kraken assign taxonomic labels to sequencing reads	x Column: 11 x Column: 12 x Column: 13 x Column: 14 x Column: 15 x Column: 16 x Column: 17 x Column: 18		efied community matrix)	
Summarize taxonomy	Select each column that contains counts		506: Vegan Rarefaction on data 500 (plot)	● # ×
Draw phylogeny	Input has a header line		505: Vegan Rarefaction	• / ×
Poisson two-sample test	Yes No		on data 500 (slope of cu rve)	
Fetch taxonomic representation	X-axis label		504: Vegan Rarefaction	• / ×
Find diagnostic hits	Sample Size		<u>on data 500 (species pro</u> <u>babilities)</u>	
<u>Krona pie chart</u> from taxonomic profile	Y-axis label Group		503: Vegan Rarefaction on data 500 (estimated r	• / ×
Vegan Beta Diversity curve and	Label beta_diversity curves by rownames of X		ichness)	
Iotif Tools	Yes No		on data 500 (frequency	• / ×
IGS: QC and manipulation	Diversity index to compute			
IGS: Mapping	1  "w" = (b+c)/(2*a+b+c)	-	on data 500 (number of	• / ×
IGS: Picard	Order sites by increasing number of species		species)	

### Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization

#### Metadata Handling

- Additional information about each sample
  - Parallel tabular files (e.g. LEfSe: cat on top of input)
  - BIOM format (not handled by many tools)

Use converters between tabular files and BIOM, various ways of pulling in and out metadata values

#### **Combine Metadata with Relative Abundances**

<b>=</b> Galaxy		Analyze Data Wo	orkflow Shared Data <del>-</del> N	/isualization <del>-</del> Admin	Help∓ User∓				Usi	ng 5.8 GB
Tools	Concatenate datasets	tail-to-head (Galaxy	Version 1.0.0)			▼ Optio	ons	History		€‡□
search tools	Concatenate Dataset							search datase	ets	8
Get Data	C 2 516	: MiSeq Early / Late M	letadata				•	MiSeq Data	hidden	
Send Data Text Manipulation	Dataset						劶	812.56 MB		<b>V &gt; </b>
Compute an expression on every row	Select							516: MiSeq Earl	ly / Late	• / ×
Add column to an existing dataset		08: Column frequenci	ies on data 500				•	515: Vegan Bet	a Diversit	• / ×
Concatenate datasets tail-to-head	+ Insert Dataset							<u>y on data 507 (</u>	plot)	
Cut columns from a table Merge Columns together	✓ Execute							514: Vegan Beta y on data 507 (	<u>a Diversit</u> output b	• / ×
y	<u> </u>							eta diversity se	cores no tr	iangular)
1	2	3	4	5	6	7	8	ç	)	1
group	F3D0	F3D1	F3D2	F3D3	F3D5	F3D6	F3D7	F	3D8	F
time	Early	Early	Early	Early	Early	Early	Early	E	arly	E
#ID	F3D0	F3D1	F3D2	F3D3	F3D5	F3D6	F3D7	F	3D8	F
Thiotrichales	0.0	0.0	0.0	0.000156690692573	0.0	0.0	0.0	C	).0	(
Pseudomonadales	0.00130684788291	0.00141317788377	0.000736338294851	0.000156690692573	0.000455684666211	0.000639959042621	0.00039	4866732478 0	0.000964320	)154291 (
Enterobacteriales	0.000784108729744	0.00035329447094	2 0.000315573554936	0.000156690692573	0.000911369332422	0.000511967234097	0.00059	2300098717 0	0.000771456	5123433 (
Chromatiales	0.0	0.00017664723547	1 5.25955924893e-05	0.0	0.0	0.000127991808524	0.0	C	).0	(
Xanthomonadales	0.0	0.00017664723547	1 0.0	0.0	0.0	0.000127991808524	0.0	C	).0	(
Alteromonadales	0.000130684788291	0.0	0.0	0.0	0.0	0.0	0.0	C	).0	(
Aeromonadales	0.000130684788291	0.0	0.0	0.0	0.0	0.0	0.0	C	).0	(
Pasteurellales	0.0	0.00017664723547	1 0.0	0.0	0.0	0.0	0.0	C	).0	(
Bickettsiales	0.0	0.0	0.0	0.0	0.0	0.0	0.0		00010286	1020050 (
<b>=</b> Galaxy		Analyze Data Wo	orkflow Shared Data <del>-</del> \	Visualization <del>-</del> Admin	Help- User-				Usi	ng 5.8 GB
Tools								History		C & ∏
	Remove beginning of	a file (Galaxy Version	1.0.0)			✓ Optio	ons			
search tools	Remove first						_	search datase	ets	8
Get Data	lines							MiSeq Data 139 shown, 378	hidden	
Send Data	intes							812 57 MR		
Text Manipulation	from							012.37 MB		
Compute an expression on every row	C 2 517	: Concatenate datase	ts on data 508 and data 516	5			•	517: Concaten ets on data 50	<u>ate datas</u> 8 and dat	• / ×
Add column to an existing dataset	✓ Execute							<u>a 516</u>		
Concatenate datasets tail-to-head	What it does							516: MiSeq Earl Metadata	ly / Late	• / ×
Cut columns from a table	This tool removes a spec	ified number of lines	from the beginning of a dat	aset.				515: Vegan Bet	a Diversit	• / ×
merge columns together								y on data 507 (	plot)	

#### **LEfSe - LDA Effect Size**



https://bytebucket.org/biobakery/galaxy\_lefse/wiki/lefse\_ove.png

#### Format Data for LEfSe

<b>=</b> Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Us	ing 5.8 GB
Tools	A) Format Data for LEfSe (Galaxy Version 1.0)	History	€‡⊡
BEDtools EMBOSS	Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe - See samples below -	search datasets	8
Regional Variation	Image: State of the second and the	MiSeq Data 140 shown, 378 <u>hidden</u>	
Evolution	Select whether the vectors (features and meta-data information) are listed in rows or columns	812.58 MB	•
Multiple Alignments	Rows	518: Remove beginning	• / ×
Motif Tools	Select which row to use as class	517: Concatenate datase	• / ×
NGS: QC and manipulation NGS: Mapping	#1:time  Select which row to use as subclass	<u>ts on data 508 and data</u> 516	
NGS: BAM Tools	no subclass	516: MiSeq Early / Late	• / ×
NGS: VCF Manipulation	Select which row to use as subject	<u>Metadata</u> 515: Vegan Beta Diversit	@ # ¥
NGS: SAM Tools NGS: Peak Calling	▼2:#ID	y on data 507 (plot)	
NGS: Variant Detection	Per-sample normalization of the sum of the values to 1M (recommended when very low values are present)	514: Vegan Beta Diversit y on data 507 (output b	• / ×
NGS: RNA Analysis NGS: GATK Tools	Yes	eta diversity scores no tr	riangular)
LEfSe: LDA Effect Size	✓ Execute	<u>y on data 507 (output b</u>	• / ×
F) Plot Differential Features	What it does	<u>eta diversity scores plot)</u> 512: Vegan Reta Diversit	
D) Plot Cladogram	LDA Effect Size (LEfSe) (Segata et. al 2010) is an algorithm for high-dimensional biomarker discovery and explanation that identifies genomic features (genes, pathways, or taxa) characterizing the differences between two or more biological conditions (or classes, see figure below). It emphasizes both	y on data 507 (output b eta diversity scores trian	gular)
C) Plot LEfSe Results	statistical significance and biological relevance, allowing researchers to identify differentially abundant features that are also consistent with biologically meaningful categories (subclasses). LEfSe first robustly identifies features that are statistically different among biological classes. It then performs additional	511: Vegan Beta Diversit	• / ×
A) Format Data for LEISe B) LDA Effect Size (LEISe)	tests to assess whether these differences are consistent with respect to expected biological behavior. Specifically, we first use the non-parametric factorial Kruskal-Wallis (KW) sum-rank test to detect features with significant differential abundance with	y on data 507 (Mean out put mean beta diversity	index)

#### **Plot LEfSe Results**



#### Visualization: Phinch

#### In IPHINCH



Biological Observation Matrix v1 format: **biom1**, database: <u>?</u>

B 0 2 III



#### view biom at Phinch



https://github.com/PitchInteractiveInc/Phinch

https://github.com/blankenberg/Phinch http://www.bx.psu.edu/~dan/Phinch/

#### Overview

# Brief Introduction to Metagenomics Doing It Challenges, Solutions, and Future Work

#### Decreasing load times for list of saved histories

# Operations that can be done in the Database should be done in the Database

		gg on – 3–17 join on single	–2015 – kraken and fi e read – set database	ilter and	-	57754		<u>(</u>	<u>) Ta</u>	<u>gs</u>	<u>Accessible</u>	60.5 GB	Mar 17, 2016	Mar 23, 2016	i	
		HMR16STR c filter and join	himera search then kr n on single end	aken and	•	47217	3372 70	26	<u>0 Ta</u>	ags		109.6 GB	Jan 15, 2016	Jan 22, 2016	i	
> Code	(!)	Issues 451	ີ່ງ Pull requests 42		11	Graphs	Settings	Showing	change	s from all	I commits - 1 change xy/webapps/galaxy/cor	ed file - htrollers/history.py			+16 –11	Diff option
Se c atas <sup>Merged</sup>	data et c	1base 0 #19	<b>Query rathe</b> 48 commits into galaxypro	r than	da om jg	taset	<b>iterati</b>	2 38 39 40 41 41 42 43 44 45 46 47 48 40	38 39 40	() () () () () () () () () () () () () (	<pre>8,18 +38,23 @@ class # Custom column type class DatasetsByStat def get_value( s state_counts 'ok' : 0 'running 'queued' 'error' } for hda in h if hda.v</pre>	<pre>HistoryListGrid( grids.() s eColumn( grids.GridColumn elf, trans, grid, history = { , ' : 0, : 0, istory.datasets: isible and not hda.delete </pre>	ed and hda.state in	n state_counts.keys():		
Conve	ersation	2 -0- Co	ommits 5 🗄 Files o	changed 1			Colour Proj	50	41 42 43 44 45 46 47 48	- + + + + + + + + + +	# States to states_to_sh # Get datase state_counts	<pre>show in column. ow = ( 'ok', 'running', ' it counts for each state i</pre>	'queued', 'error' in a state-count d ) for state, count n.query( model.Data el.HistoryDatasetA: model.Dataset.sta	) intionary. aset.state, func.count(mod ssociation ) te )	del.Dataset.sta	ite))
	our	nts in saved his	stories grid. In my testin	g, speedup	is 4-1(	0X.	Galaxy FIOJE		49 50 51 52 53 54 55	+ + + + +	# Create HTM	.filter( mc mc mc )	odel.HistoryDatase odel.HistoryDatase odel.HistoryDatase odel.Dataset.state	<pre>tAssociation.history_id == tAssociation.visible == tr tAssociation.deleted == fa .in_(states_to_show) )</pre>	= history.id, rue(), alse(),	
	<u>)</u> 1							51 52 53 54	56 57 58 59	+	for state in for state in for state in count = if count	<pre>state_counts.keys(): states_to_show: state_counts.get( state ) :</pre>	)			

#### **Multiple History View**

03	ing 5.5 16
History	<b>℃</b>
search datasets	8
HMR16STR upload fasta 5035 shown	
37.8 GB	
5035: SRS066188.fsa	• / ×

lising 0 5 T

# Fast view across many histories Drag and drop to copy Datasets

- Galaxy / Dan's	Brudeck P	Taygroundiyze Data Worktin	ow Shared Da	ata - Visualization - Admin F	leip <del>-</del> User -	:::		Using 9.5 TB
Done search histories		search all datasets						Create new
Current History	•	Switch to	•	Switch to	•	Switch to	•	Switch to
HMR16STR upload fasta 5035 shown		Copy of 'HMR16STR upload fasta' 4076 shown		Copy of 'vegan on 434 pairs' 79 shown		Copy of '434 Pairs - GG - Abun	dances @ P'	434 Pairs - GG - /
37.8 GB		29.74 GB	2 > >	11.67 MB	<b>S</b>	94.99 KB		94.99 KB
search datasets	8	search datasets	8	search datasets	8	search datasets	8	search datasets
Drag datasets here to copy them to the o	current history	4076: SRS054236.fsa	@ # ¥	79: Vegan Rarefaction on data 1	(R 💿 🖋 🗶	🚹 🔅 loading datasets		🚹 🔅 loading dat
5035: SRS066188.fsa	● / ×	4075: SBS054222 fc2		script)				
5034: SRS066144.fsa	• / ×	4074: SRS054219.fsa	• / ×	78: Vegan Rarefaction on data 1 andom rarefied community mate	( <u>R</u> () (R) (R) (R) (R) (R) (R) (R) (R) (R)			
5033: SRS065725.fsa	• # ×	4073: SRS054126.fsa	• / ×	77: Vegan Rarefaction on data 1	(p () / x			
5032: SRS065719.fsa	• / ×	4072: SRS054121.fsa	• / ×	76: Vegan Rarefaction on data 1	<u>(s</u> 🕑 🖋 🗙			
5031: SRS065718.fsa	● # ×	4071: SRS054112.fsa	• / ×	lope of curve)	(5 <b>a b y</b>			
<u>5030: SRS065711.fsa</u>	• / ×	4070: SRS054108.fsa	• / ×	pecies probabilities)				
5029: SRS065701.tsa		4069: SRS054094.fsa	• / ×	74: Vegan Rarefaction on data 1 stimated richness)	<u>(e</u> 🕐 🖋 🗙			
5028: SR5065676 fca	• • ×	4068: SRS054072.fsa	• / ×	73: Vegan Rarefaction on data 1	(f 💿 🖋 🗙			
<u>3027, 383003070.13a</u>	• / ×	4067: SRS054065.fsa	• / ×	72: Vegan Parefaction on data 1				
<u>5026: SRS065675.fsa</u>	• / ×	4066: SRS054043.fsa	• / ×	umber of species)	<u></u>			
<u>5025: SRS065669.fsa</u>	• / ×	4065: SRS054039.fsa	• / ×	71: Vegan Rarefaction on data 1	<u>(R</u> 👁 🖋 🗙			
5024: SRS065666.fsa	• / ×	4064: SRS054007.fsa	• / ×	70: Vegan Rarefaction on data 1	<u>(R</u> 👁 🖋 🗙			
5023: SRS065665.fsa	● / ×	4063: SRS053956.fsa		andom rarefied community mate	rix)			
5022: SRS065656.fsa	• / ×	4062: SBS052025 fca		69: Vegan Rarefaction on data 1	(p 🕘 🖋 🗙			
		4002: 5K5055925.15d	• / ×	····/				

#### **Dataset Collection Operations**

#### Collection Operations (Limited) #2434

ាំ Merged	martenso	n merged 10 commits in	galaxyproject:dev	from j	jmchilton:fewer_collection_opts	15 days ago			
다 Conver	rsation 13	-O- Commits 10	E Files changed 3	0					
	jmchilton co	mmented 28 days ago			Galaxy Project member	+ 💼 🥒			
	Overview								
	This PR introduces Tool-derived framework-level plumbing for dealing collections at the model level instead of at the file level, allowing operations that generate new HDAs and collections without duplicating Dataset objects. Together operations vastly expand the expressiveness of Galaxy workflows.								

#### **The Collection Operations:**

- Zip (two datasets -> paired collection). Like all these tools, it can be mapped over so it can easily be
  used to take two dataset lists and build a list of pairs for instance.
- Unzip (paired collection -> two datasets).
- Filter failed datasets (list -> list). Given a list it produces another list without any failed datasets. (Most commonly requested of these operations.)
- Flatten collection (\* -> list). Produces a flat list from any collection, joining identifiers on user supplied character.

#### Naming of Datasets

#### Concern

 Default Galaxy names lose Sample names (Tool xyz on Dataset 1,2,3,4)

# Solution

- Dataset Collections have an Element Identifier that is maintained throughout jobs
- Tools that make use of \${dataset.name} should now use \${dataset.element\_identifier}

#### Future Work: Scaling Up

- Viewing Large Histories
  - Pagination
  - Infinite Scrolling
  - Dataset Bundling



have that interface scale well

- Uploading Large Collections
  - Better handling of selecting many files
  - Importing multiple files from an archive
    - Provide a manifest file that instructs Galaxy to build a Collection

#### **Additional Options Available: Mother**

galaxyproject / tools-iuc	tch <b>-</b> 25 ★ Star 25 % Fork 74								
<> Code (!) Issues 78 (?) Pull requests (41 III) Wiki -4 Pulse III Graphs (Settings									
mothur: update make.contigs to use paired collections #853									
Conversation 0 -O- Commits 1 E Files changed 1	+15 -8								
shiltemann commented 3 days ago       Galaxy Project member       + (a)         No description provided.	Labels Or None yet								
-o- 🛐 update make.contigs to use paired collections Verified 🗙 8	Milestone 🌣								
Shiltemann referenced this pull request 3 days ago IUC Contribution Fest - Metagenomics and mothur #419 0 of 8 tasks complete	Assignees     Assignees     No one-assign yourself     1 participant								
Add more commits by pushing to the mohtur-updates branch on galaxyproject/tools-iuc.									

#### **Additional Options Available: Qiime**

#### WIP: Add Qiime wrappers #431

1) Open bgruening wants to merge 76 commits into master from gime

Conversation 39

-O- Commits 76 🗄 Files changed 72



Edit

+8,002 -1

#### **Additional Options Available: FROGS**



#### Take Home

- You can do metagenomics with Galaxy
  - Many different modules to swap in and out
- You can do large-scale multiple sample analysis with Galaxy
  - 500 Samples are no problem
  - 5,000+ Samples through API
    - Client-side is under active development

#### Acknowledgements

- Everyone Here
- Tool Developers
- Galaxy Committers and Contributors
- Galaxy Community
- Funding: grant number HG005542 from the National Human Genome Research Institute, National Institutes of Health as well as grants HG005133, HG004909 and HG006620 and NSF grant DBI 0543285. Additional funding is provided by Huck Institutes for the Life Sciences at Penn State and, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

#### Watch out for Drop Gators!



http://www.winknews.com/2016/07/06/spotted-in-cape-coral-an-alligator-in-a-tree/