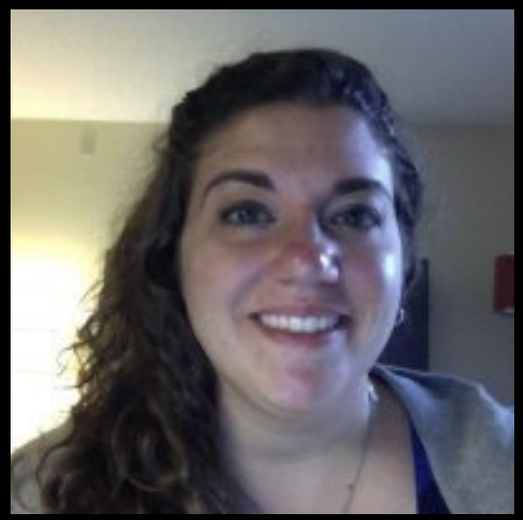
Sample Size Does Matter: Scaling Up Analysis in Galaxy with Metagenomics

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

Metagenomics Powerhouse



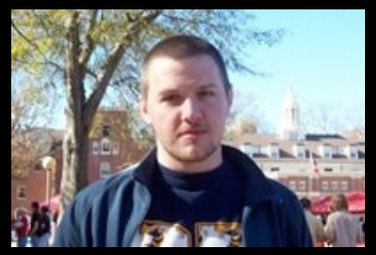
Sarah J. Craig, PhD
sarah@bx.psu.edu

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

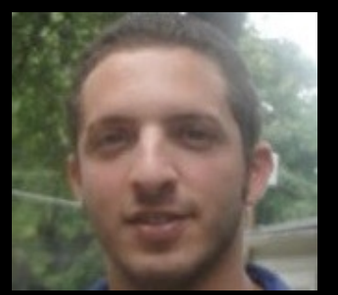
John Chilton

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**

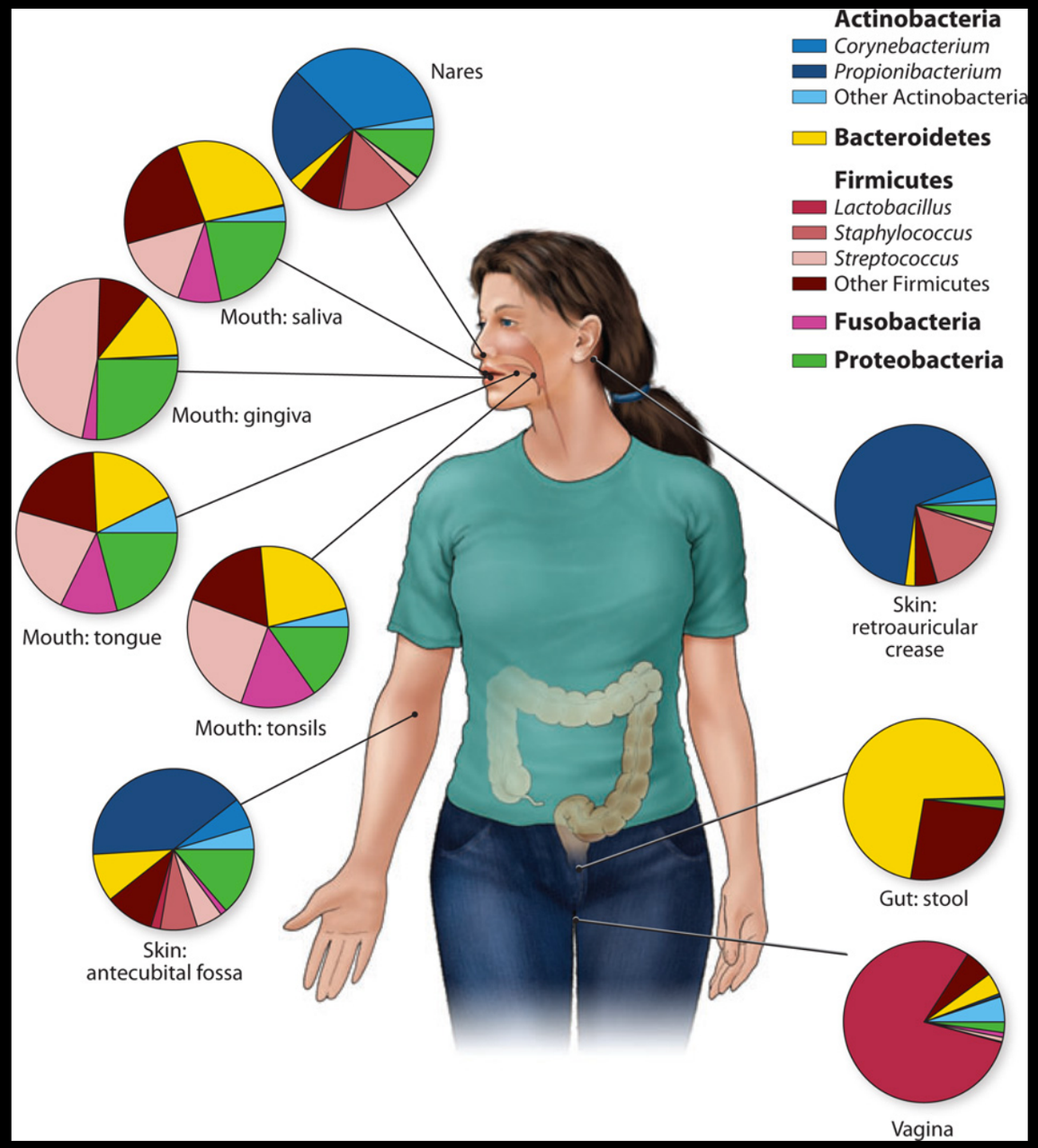
Overview

- **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

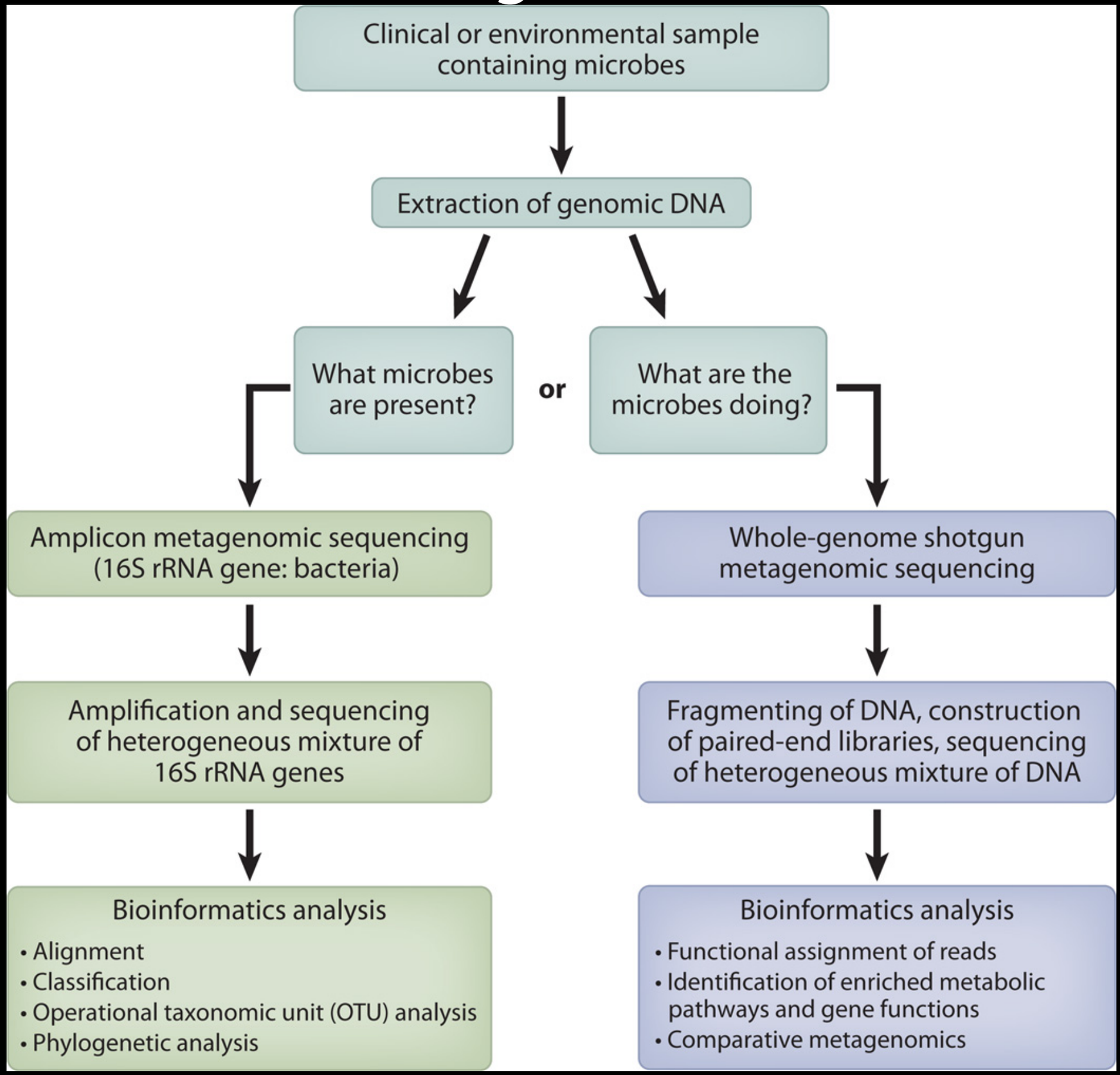


Grice E.A. & Segre J.A. (2012) *The Human Microbiome: Our Second Genome*, Annu. Rev. Genomics Human Genet. 13, 151-170

Metagenomics

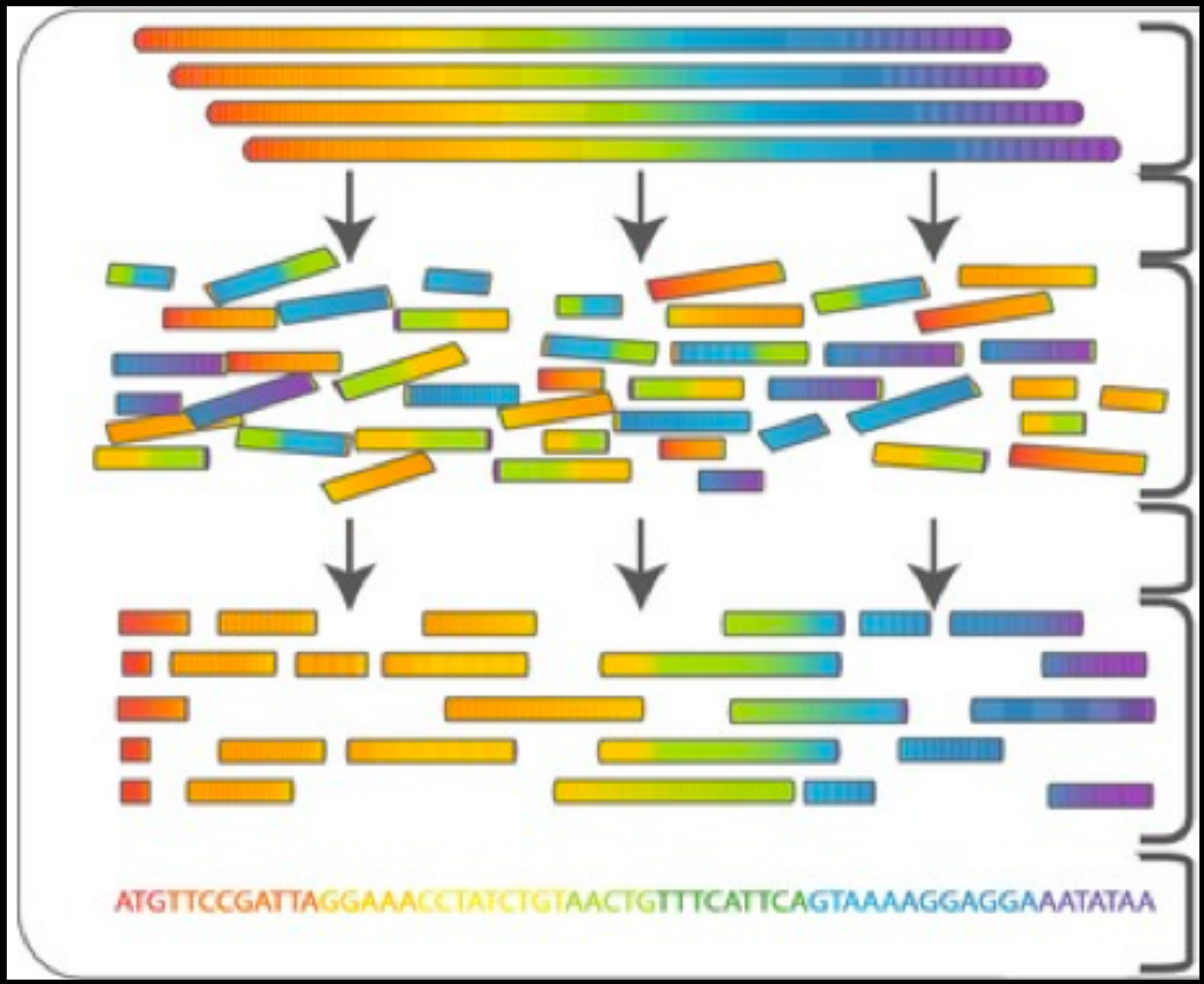
16S rRNA

Whole-Genome



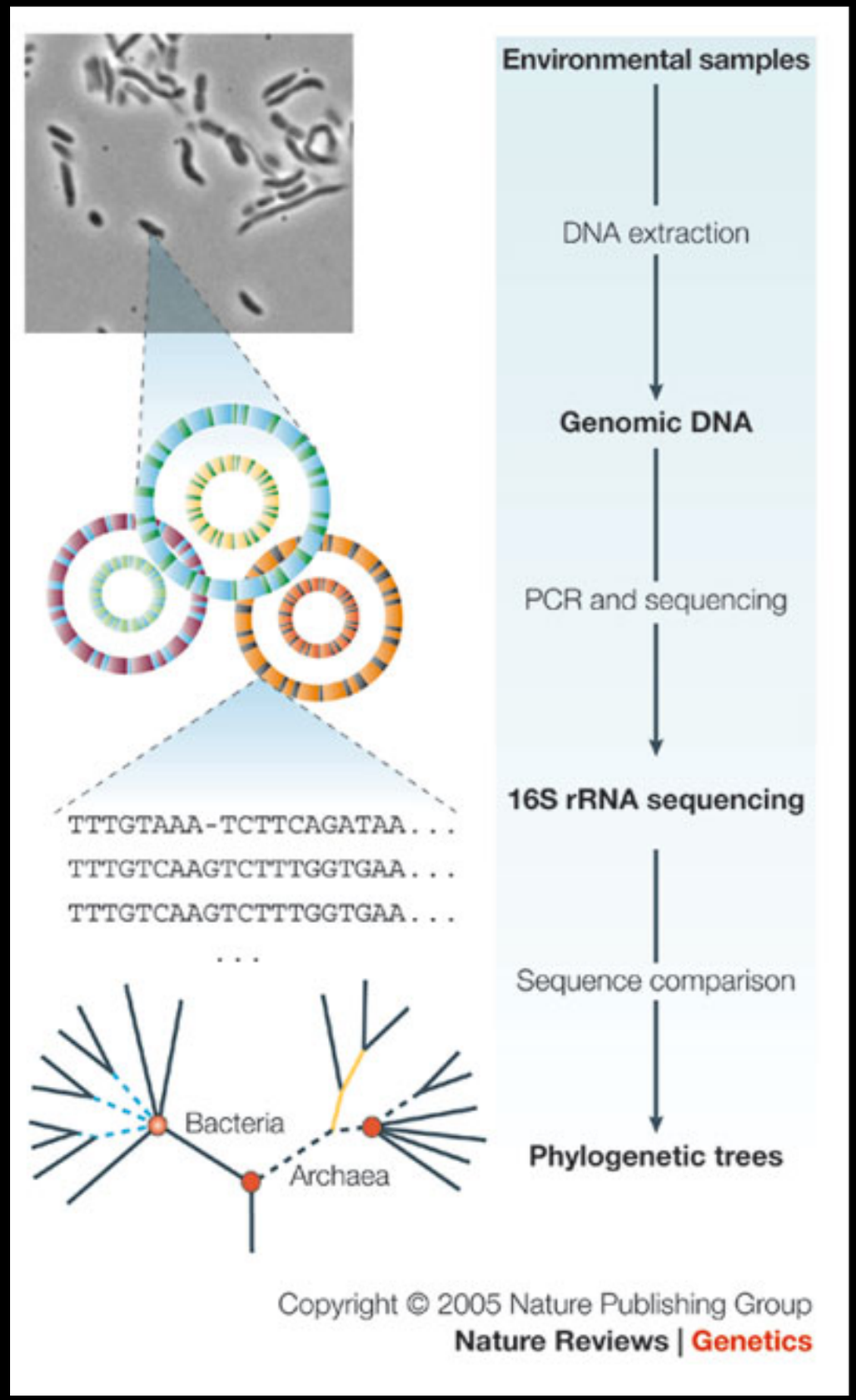
Grice E.A. & Segre J.A. (2012) *The Human Microbiome: Our Second Genome*, Annu. Rev. Genomics Human Genet. 13, 151-170

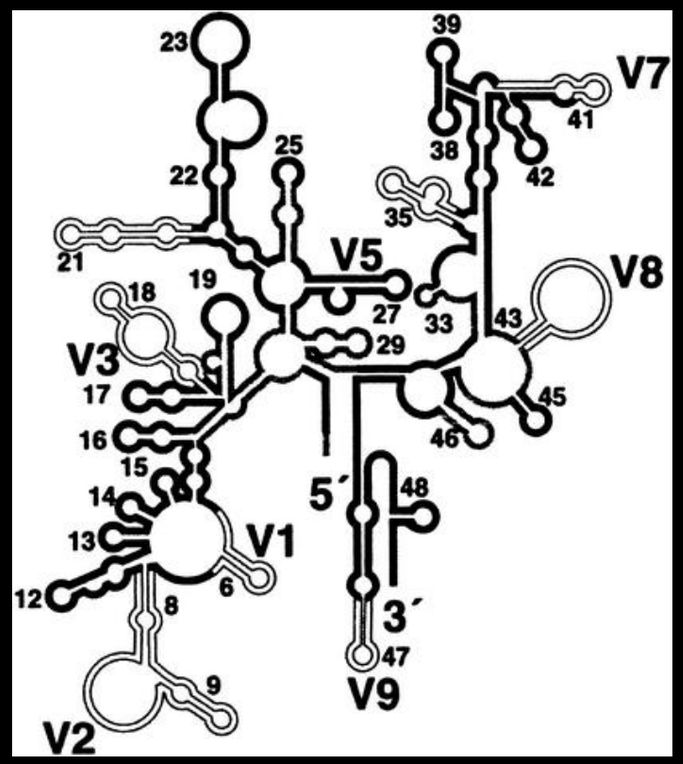
Whole-Genome Shotgun



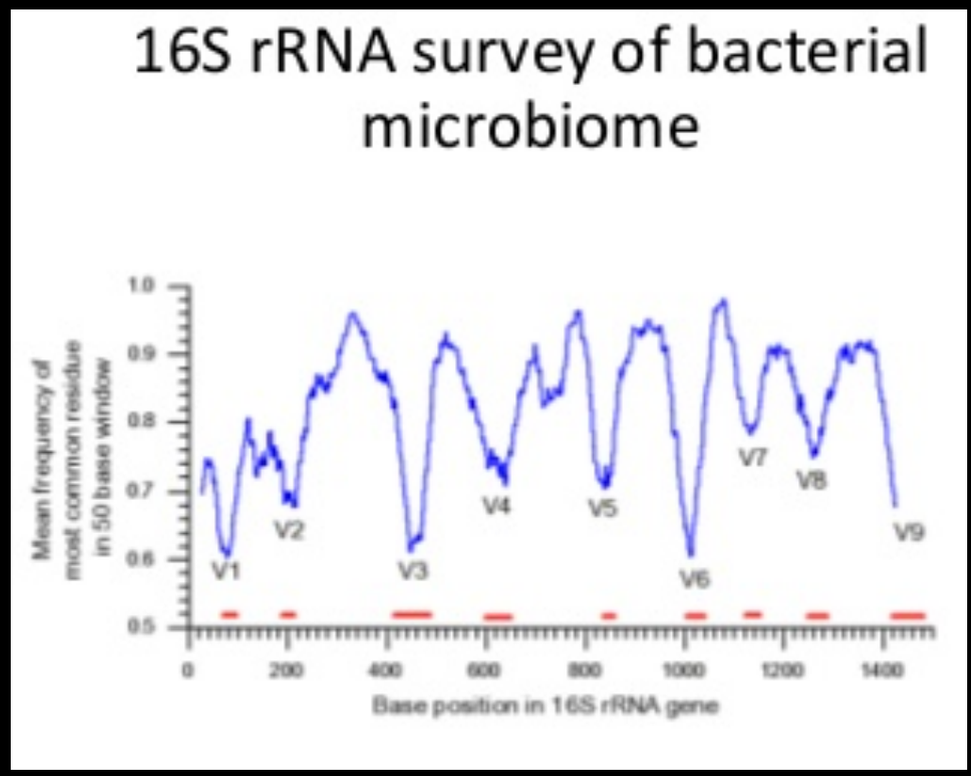
Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. Biol. Procedures Online (2009).

16S rRNA





Tortoli E. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. Clin Microbiol Rev. 2003 Apr;16(2):319-54. Review.



Julien Tremblay. JGI.

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**

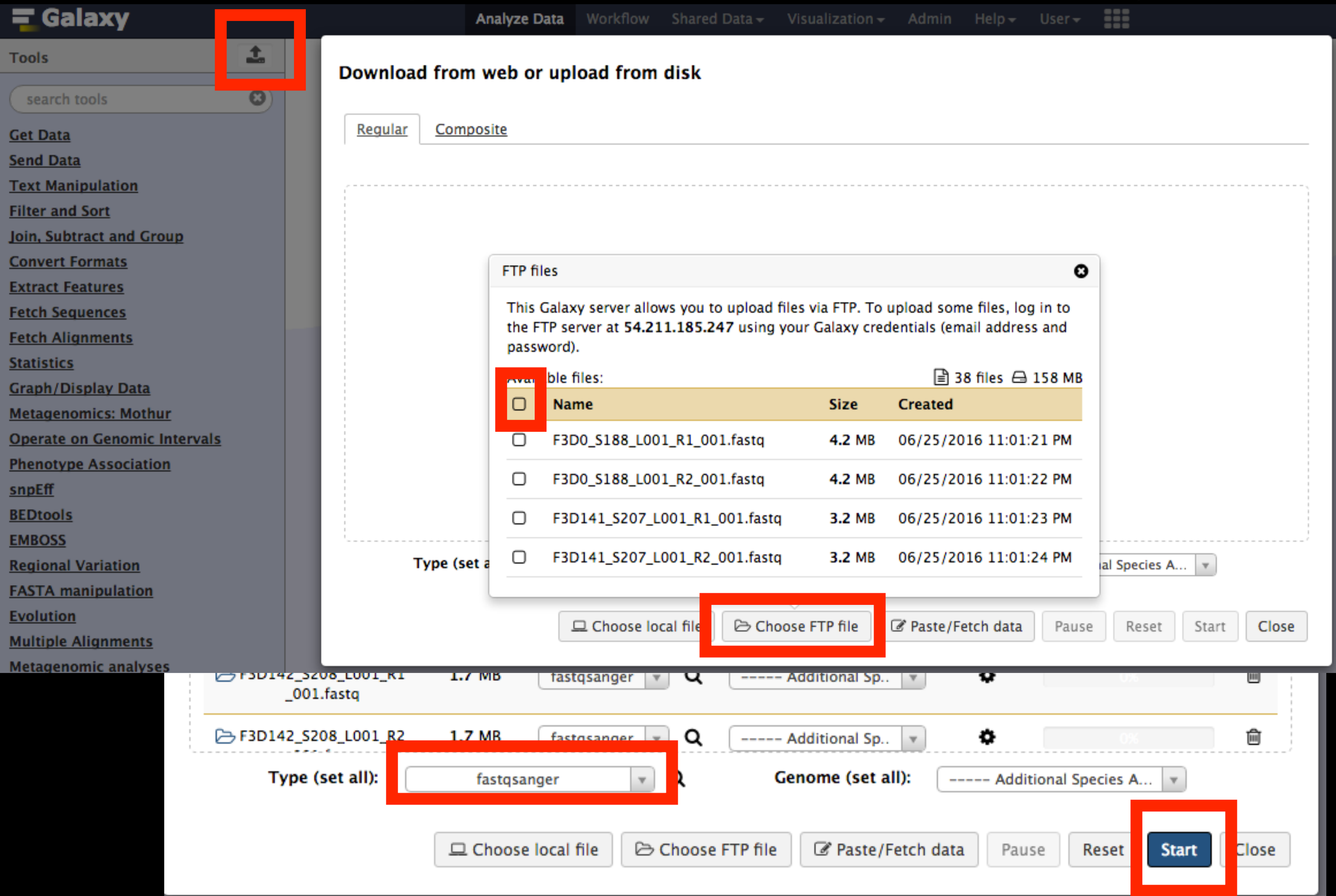
Doing It

- **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



Create Dataset Collection

History

search datasets

MiSeq Data
38 shown

157.95 MB

All | None | For all selected...

- Hide datasets
- Unhide datasets
- Delete datasets
- Undelete datasets
- Permanently delete datasets
- Build Dataset List
- Build Dataset Pair
- Build List of Dataset Pairs

1: Microbiome – 434 Metagenomic Pairs (Dan Dec. 18, 2015 Load)
a list of 434 dataset pairs

34: F3D7 S195 g
33: F3D7 S195 g
32: F3D6 S194 g
31: F3D6 S19 g
30: F3D5 S193 g
29: F3D5 S193 g

History

Back to 434 pairs no fastqc Copy of 'QC and Join paired end (imported from uploaded file)'

Microbiome – 434 Metagenomic Pairs (Dan Dec. 18, 2015 Load)
a list of paired datasets

102C1Buccal_S13_L001_001
a pair of datasets

102C1Stool_S1_L001_001
a pair of datasets

103C1-buccal_S26_L001_001

History

Back to Microbiome – 434 Metagenomic Pairs (Dan Dec. 18, 2 Load)

102C1Buccal_S13_L001_001
a pair of datasets

forward

reverse

132.7 MB
format: **fastqsanger**, database: ?

uploaded fastqsanger file

@M02286:19:000000000-AA549:1:1101:

Create a collection of paired datasets

19 pairs created: all datasets have been successfully paired

19 unpaired forward (19 filtered out)

_R1_

Choose filters Clear filters
Auto-pair

19 unpaired reverse – (19 filtered out)

_R2_

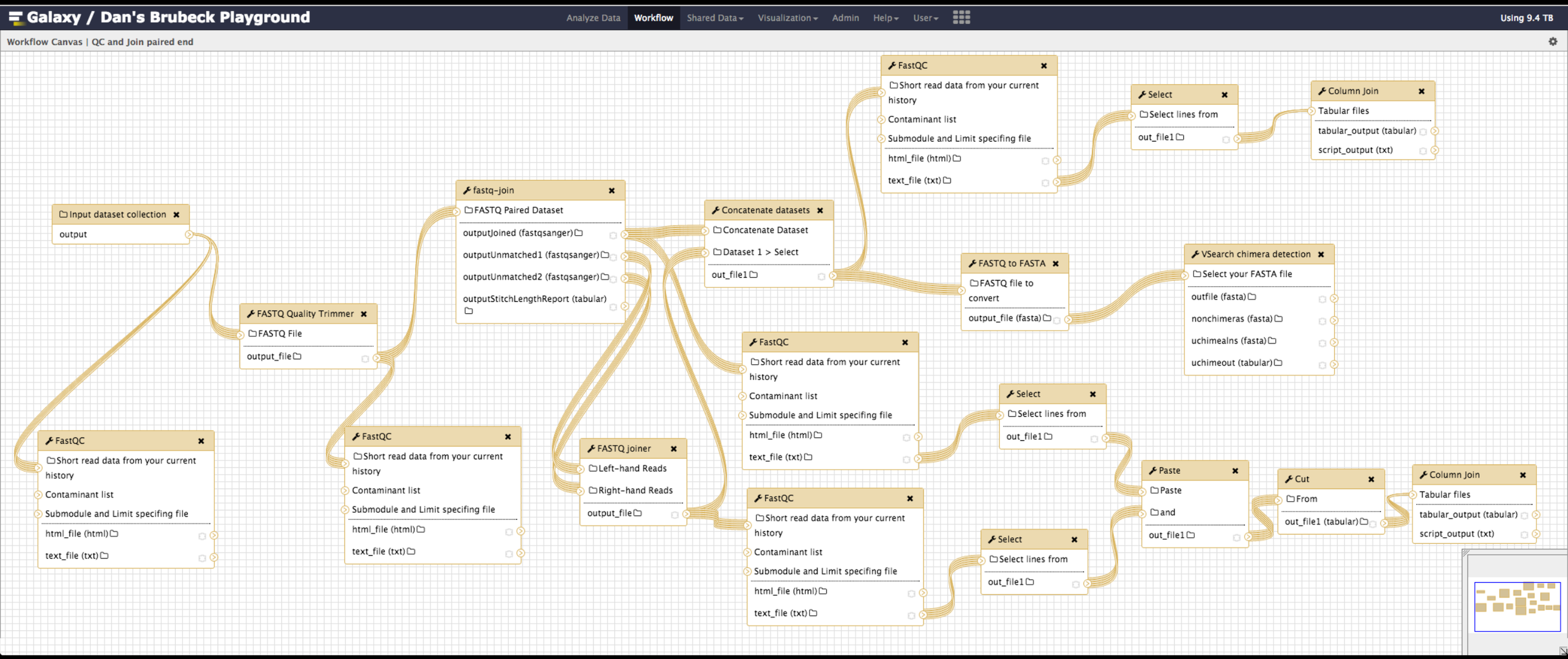
F3D0_S188_L001_R1_001.fastq	Pair these datasets	F3D0_S188_L001_R2_001.fastq
F3D1_S189_L001_R1_001.fastq	Pair these datasets	F3D1_S189_L001_R2_001.fastq
F3D2_S190_L001_R1_001.fastq	Pair these datasets	F3D2_S190_L001_R2_001.fastq
F3D3_S191_L001_R1_001.fastq	Pair these datasets	F3D3_S191_L001_R2_001.fastq
F3D5_S193_L001_R1_001.fastq	Pair these datasets	F3D5_S193_L001_R2_001.fastq
F3D6_S194_L001_R1_001.fastq	Pair these datasets	F3D6_S194_L001_R2_001.fastq
F3D7_S195_L001_R1_001.fastq	Pair these datasets	F3D7_S195_L001_R2_001.fastq

...

0 paired

(no paired datasets yet)

QC, Prepare, manipulate, filter Reads

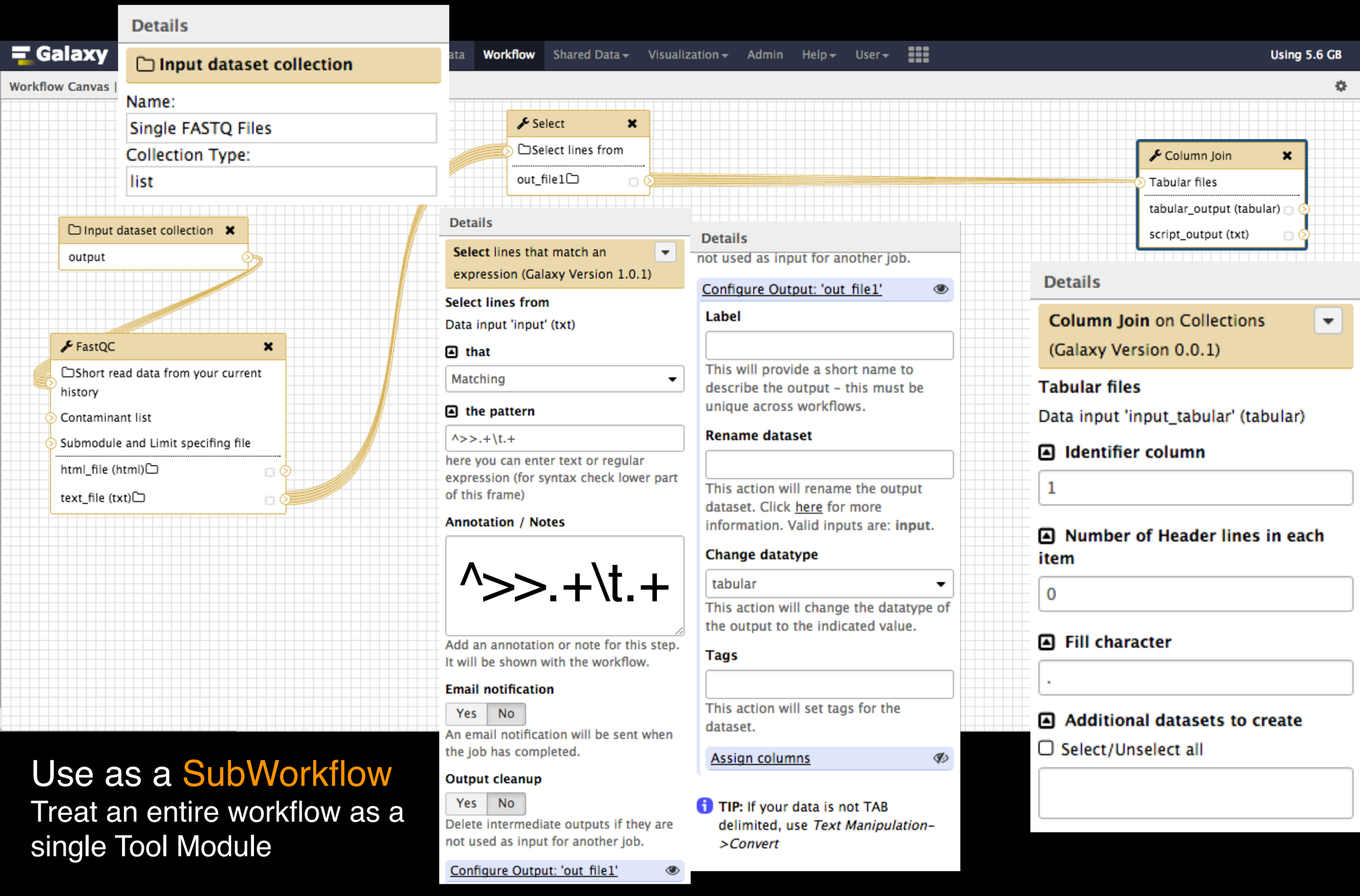


Input: a List of paired sequencing Reads

Outputs: QC'd Joined paired reads, filtered for chimeras (FASTA)

Quality Reports (html and single summary table)

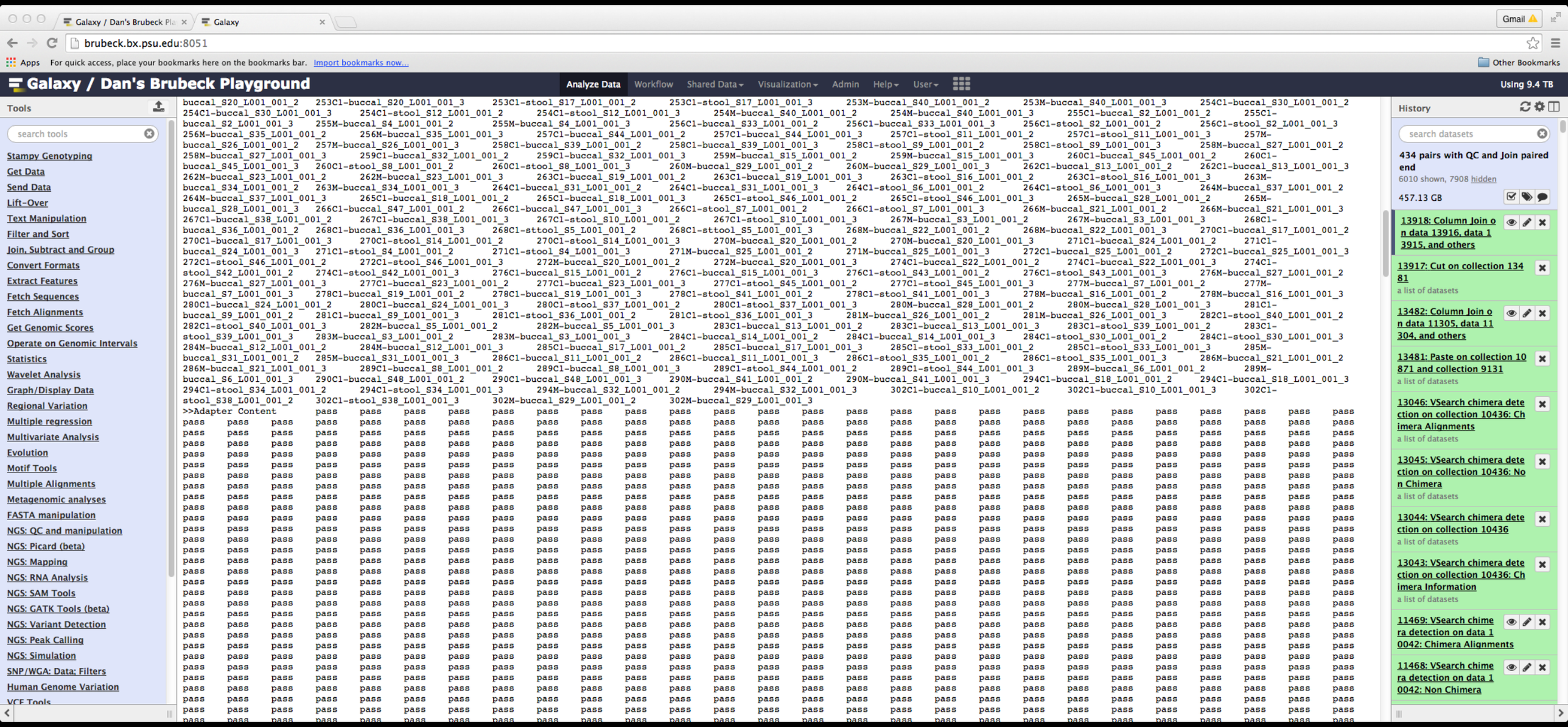
Create Workflow for Easy-to-Read FastQC Report



Use as a SubWorkflow

Treat an entire workflow as a single Tool Module

Single Summary FastQC Report



Doing It

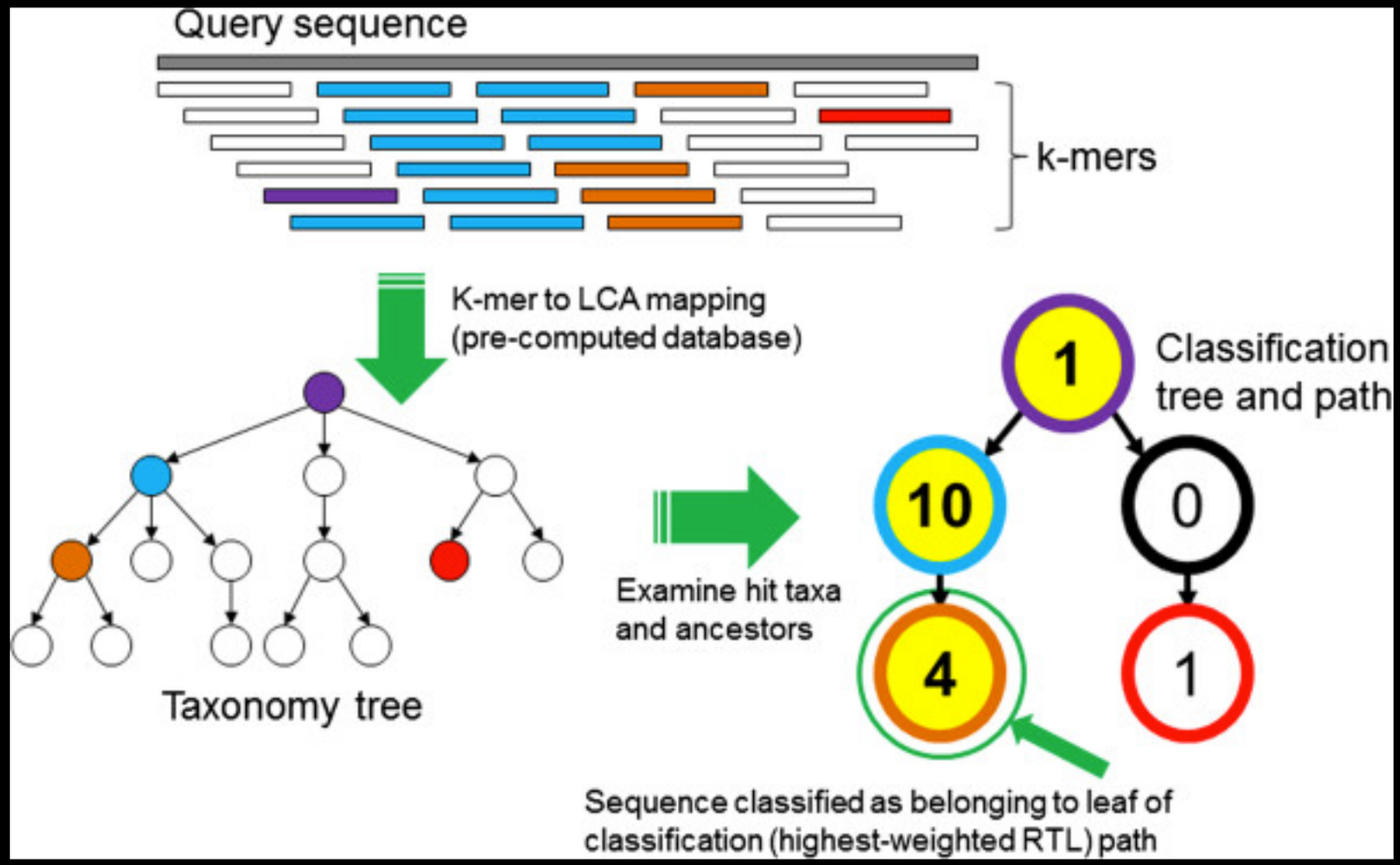
- **QC and Preparation**
- **Classify Reads**
- **Normalize**
- **alpha diversity**
- **beta diversity**
- **Downstream Analysis and Visualization**

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

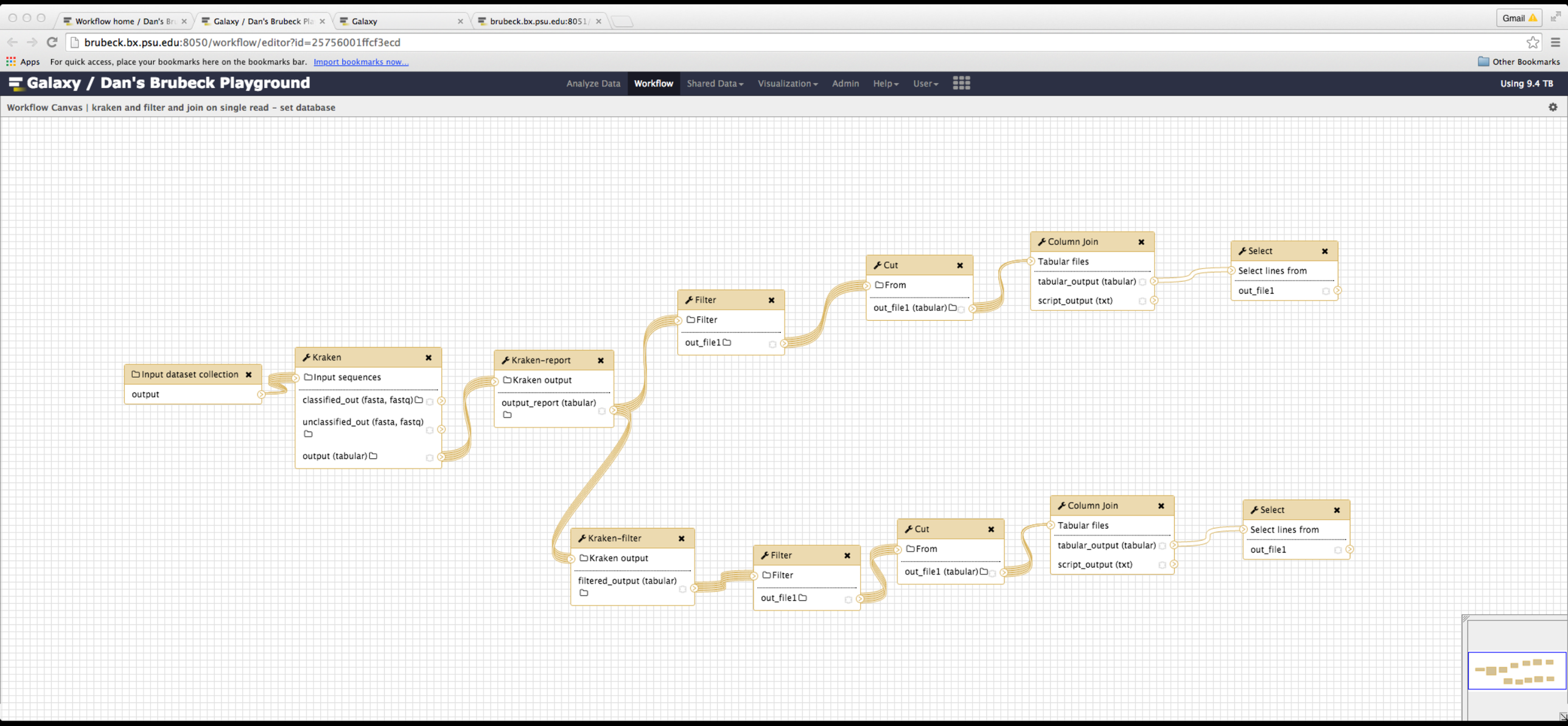
Kraken Classifier

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



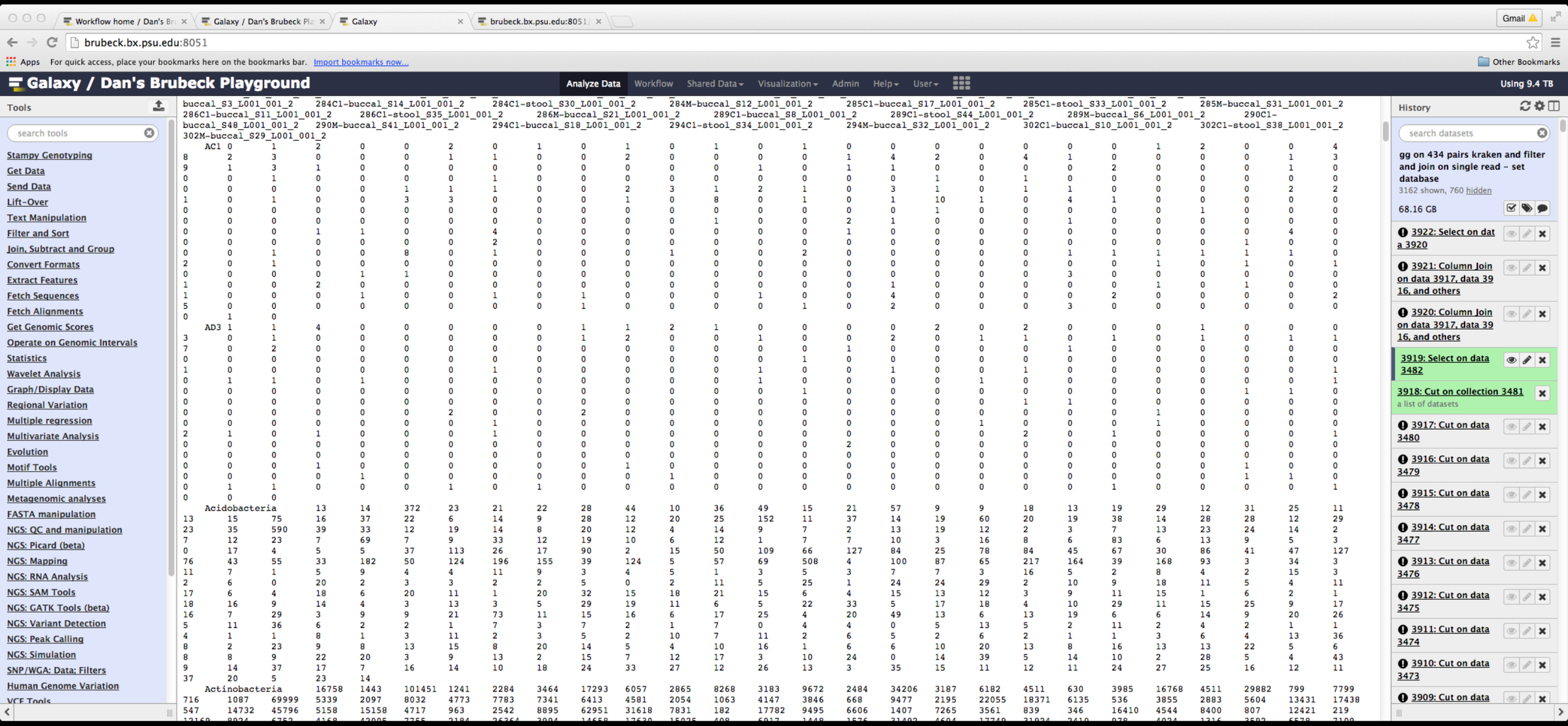
Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 2014, 15:R46.

Classify Reads



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

Abundance Counts



Doing It

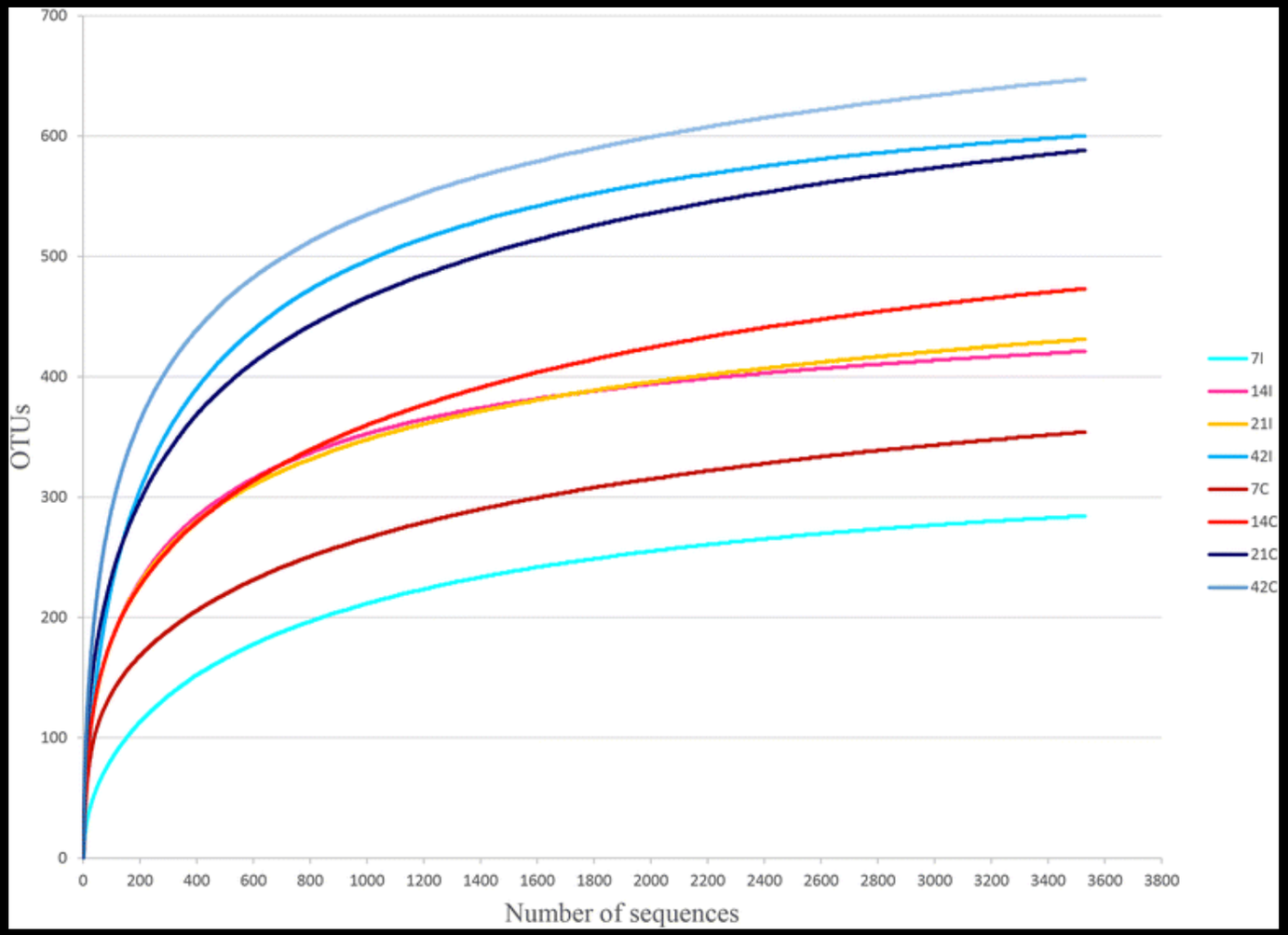
- **QC and Preparation**
- **Classify Reads**
- **Normalize**
- **alpha diversity**
- **beta diversity**
- **Downstream Analysis and Visualization**

Normalizing

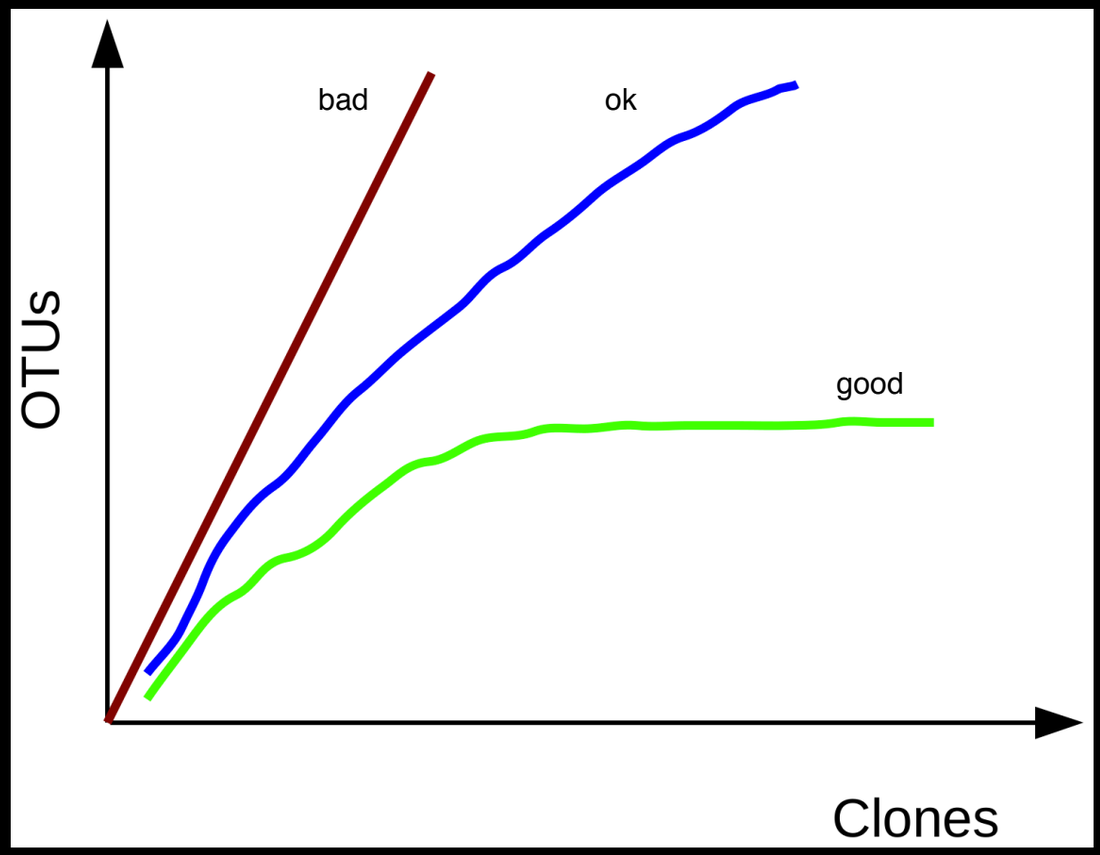
- Rarefy
- Total sum scaling

- Cumulative Sum Scaling
- DeSeq2 process
- ...

Rarefaction

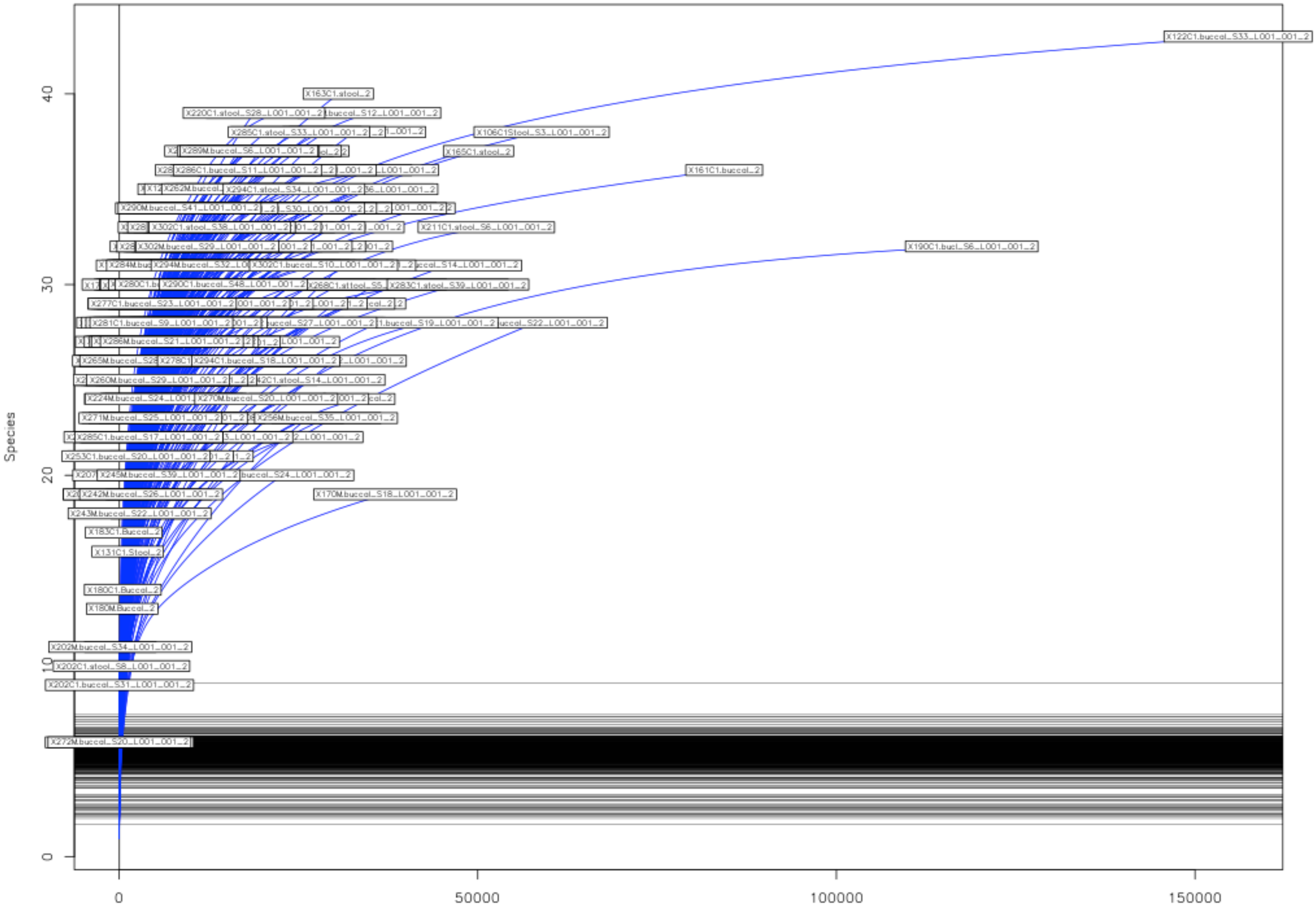


Mohd Shaufi MA, Sieo CC, Chong CW, Gan HM, Ho YW. Deciphering chicken gut microbial dynamics based on high-throughput 16S rRNA metagenomics analyses. Gut Pathog. 2015 Feb 26;7:4. doi: 10.1186/s13099-015-0051-7.



Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010 Feb 26;6(2):e1000667.

- Have we sequenced enough?
- Normalize counts across samples



Vegan Tool Suite - Rarefaction Plot

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**

α-diversity

- biodiversity in a defined
  - habitat
  - ecosystem
  - sample
- Community Richness
  - how many organisms are really there

Alpha Diversity

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 5.8 GB

Tools

- Regional Variation
- FASTA manipulation
- Evolution
- Multiple Alignments
- Metagenomic analyses
- Beta Diversity using scikit-bio
- Vegan Fisher Alpha index
- Vegan Diversity index
- Vegan Rarefaction curve and statistics
- VSearch alignment
- VSearch dereplication
- VSearch sorting
- VSearch shuffling
- VSearch search
- VSearch masking
- VSearch clustering
- VSearch chimera detection
- BIOM metadata add-metadata
- Convert BIOM formats
- Kraken taxonomic report view report of classification for multiple samples
- Kraken-translate convert taxonomy IDs to names
- Kraken-report view a sample

Vegan Diversity index (Galaxy Version 0.0.3) — Options

File with abundance values for community

507: Vegan Rarefaction on data 500 (Random rarefied community matrix)

Rows are samples; columns are species/phyla/community classifier

Group name column

Column: 1

Species, phylum, etc

Sample count columns

Select/Unselect all

× Column: 2 × Column: 3 × Column: 4 × Column: 5 × Column: 6 × Column: 7 × Column: 8 × Column: 9 × Column: 10 × Column: 11 × Column: 12 × Column: 13 × Column: 14 × Column: 15 × Column: 16 × Column: 17 × Column: 18 × Column: 19 × Column: 20

Select each column that contains counts

Input has a header line

Yes | No

Diversity index to use

Inverse Simpson

Margin for which the index is computed

1

The logarithm base

Natural Logarithm: exp(1)

✔ Execute

Calculate Diversity index using vegan and selected method.

1	2
	x
F3D0	2.21245501532285
F3D1	2.37118203363653
F3D2	2.01951605149471
F3D3	1.64741465806235
F3D5	2.21811786026141
F3D6	2.01124753861779
F3D7	1.56781184051996
F3D8	2.25470249485439
F3D9	2.16635907862611
F3D141	1.88938647101194
F3D142	1.69697928769152
F3D143	1.86853892207042
F3D144	1.71506115634714
F3D145	1.55429517220742
F3D146	2.1131686074353
F3D147	1.6036205003031
F3D148	1.73860419546772
F3D149	2.02799055973392
F3D150	1.99456407827727

History

search datasets

MiSeq Data
130 shown, 378 hidden
812.52 MB

- 508: Column frequencies on data 500
- 507: Vegan Rarefaction on data 500 (Random rarefied community matrix)
- 506: Vegan Rarefaction on data 500 (plot)
- 505: Vegan Rarefaction on data 500 (slope of curve)
- 504: Vegan Rarefaction on data 500 (species probabilities)
- 503: Vegan Rarefaction on data 500 (estimated richness)
- 502: Vegan Rarefaction on data 500 (frequency of species)
- 501: Vegan Rarefaction on data 500 (number of species)
- 500: Column substitution on data 498

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

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 5.8 GB

Tools

- Kraken taxonomic report view report of classification for multiple samples
- Kraken-translate convert taxonomy IDs to names
- Kraken-report view a sample report of your classification
- Kraken-mpa-report view report of classification for multiple samples
- Kraken-filter filter classification by confidence score
- Kraken assign taxonomic labels to sequencing reads
- Summarize taxonomy
- Draw phylogeny
- Poisson two-sample test
- Find lowest diagnostic rank
- Fetch taxonomic representation
- Find diagnostic hits
- Krona pie chart from taxonomic profile
- Vegan Beta Diversity curve and statistics

Motif Tools
NGS: QC and manipulation
NGS: Mapping
NGS: BAM Tools
NGS: Picard

Vegan Beta Diversity curve and statistics (Galaxy Version 0.0.2) — Options

File with abundance values for community

507: Vegan Rarefaction on data 500 (Random rarefied community matrix)

Rows are samples; columns are species/phyla/community classifier

Group name column

Column: 1

Species, phylum, etc

Sample count columns

Select/Unselect all

× Column: 2 × Column: 3 × Column: 4 × Column: 5 × Column: 6 × Column: 7 × Column: 8 × Column: 9 × Column: 10 × Column: 11 × Column: 12 × Column: 13 × Column: 14 × Column: 15 × Column: 16 × Column: 17 × Column: 18 × Column: 19 × Column: 20

Select each column that contains counts

Input has a header line

Yes | No

X-axis label

Sample Size

Y-axis label

Group

Label beta_diversity curves by rownames of X

Yes | No

Diversity index to compute

1 "w" = (b+c)/(2*a+b+c)

Order sites by increasing number of species

History

search datasets

MiSeq Data
131 shown, 378 hidden
812.52 MB

509: Vegan Diversity on data 507
508: Column frequencies on data 500
507: Vegan Rarefaction on data 500 (Random rarefied community matrix)
506: Vegan Rarefaction on data 500 (plot)
505: Vegan Rarefaction on data 500 (slope of curve)
504: Vegan Rarefaction on data 500 (species probabilities)
503: Vegan Rarefaction on data 500 (estimated richness)
502: Vegan Rarefaction on data 500 (frequency of species)
501: Vegan Rarefaction on data 500 (number of species)

Doing It

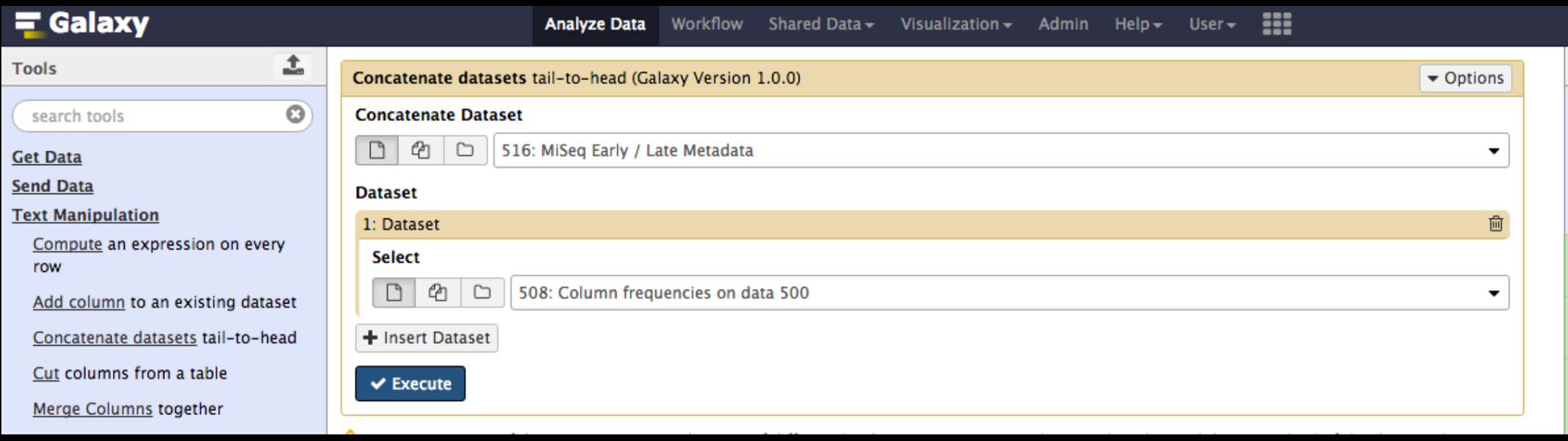
- **QC and Preparation**
- **Classify Reads**
- **Normalize**
- **alpha diversity**
- **beta diversity**
- **Downstream Analysis and Visualization**

Metadata Handling

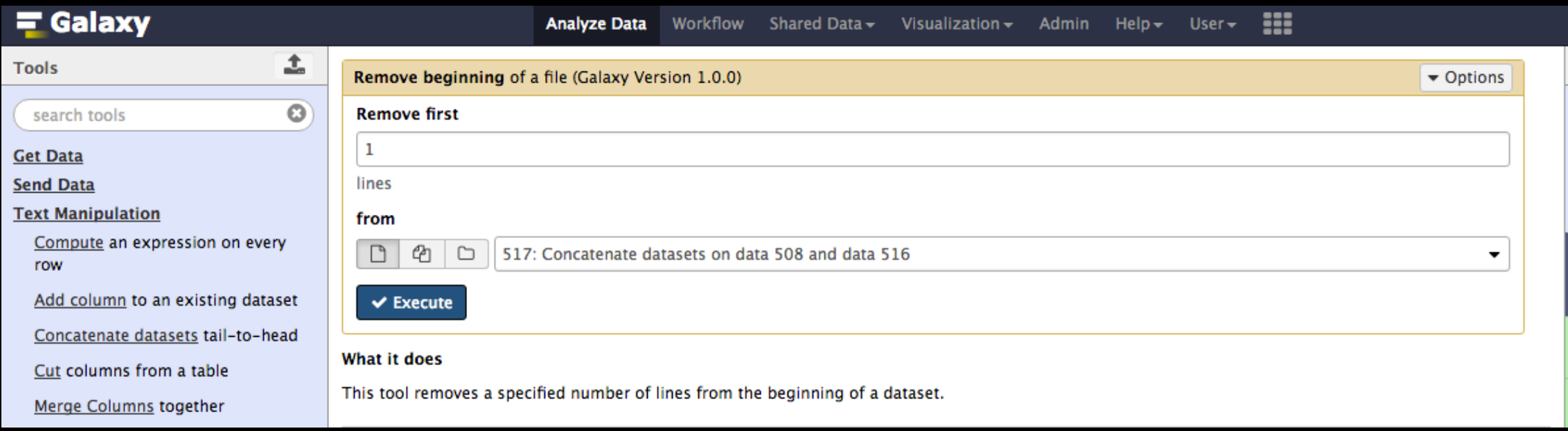
- Additional information about each sample
  - Parallel tabular files (e.g. LEfSe: cat on top of input)
  - BIOM format (not handled by many tools)

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

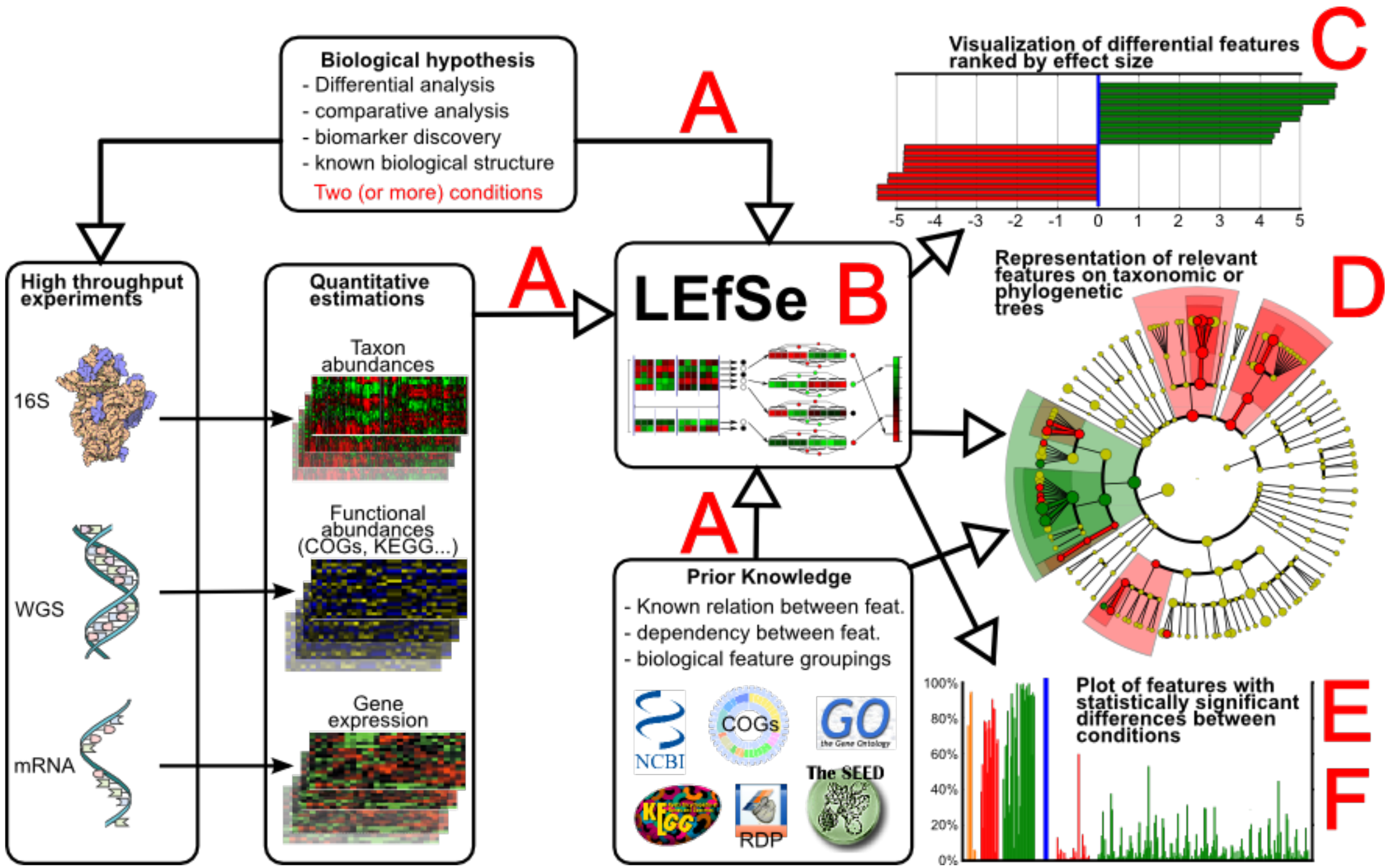
Combine Metadata with Relative Abundances



1	2	3	4	5	6	7	8	9
group	F3D0	F3D1	F3D2	F3D3	F3D5	F3D6	F3D7	F3D8
time	Early	Early	Early	Early	Early	Early	Early	Early
#ID	F3D0	F3D1	F3D2	F3D3	F3D5	F3D6	F3D7	F3D8
Thiotrichales	0.0	0.0	0.0	0.000156690692573	0.0	0.0	0.0	0.0
Pseudomonadales	0.00130684788291	0.00141317788377	0.000736338294851	0.000156690692573	0.000455684666211	0.000639959042621	0.000394866732478	0.000964320154291
Enterobacteriales	0.000784108729744	0.000353294470942	0.000315573554936	0.000156690692573	0.000911369332422	0.000511967234097	0.000592300098717	0.000771456123433
Chromatiales	0.0	0.000176647235471	5.25955924893e-05	0.0	0.0	0.000127991808524	0.0	0.0
Xanthomonadales	0.0	0.000176647235471	0.0	0.0	0.0	0.000127991808524	0.0	0.0
Alteromonadales	0.000130684788291	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aeromonadales	0.000130684788291	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pasteurellales	0.0	0.000176647235471	0.0	0.0	0.0	0.0	0.0	0.0



LEfSe - LDA Effect Size



https://bytebucket.org/biobakery/galaxy_lefse/wiki/lefse_ove.png

Format Data for LEfSe

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 5.8 GB

Tools

- BEDtools
- EMBOSS
- Regional Variation
- FASTA manipulation
- Evolution
- Multiple Alignments
- Metagenomic analyses
- Motif Tools
- NGS: QC and manipulation
- NGS: Mapping
- NGS: BAM Tools
- NGS: Picard
- NGS: VCF Manipulation
- NGS: SAM Tools
- NGS: Peak Calling
- NGS: Variant Detection
- NGS: RNA Analysis
- NGS: GATK Tools
- LEfSe: LDA Effect Size
  - F) Plot Differential Features
  - E) Plot One Feature
  - D) Plot Cladogram
  - C) Plot LEfSe Results
  - A) Format Data for LEfSe
  - B) LDA Effect Size (LEfSe)

A) Format Data for LEfSe (Galaxy Version 1.0) — Options

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe - See samples below - Please use Galaxy Get-Data/Upload-File. Use File-Type = tabular

518: Remove beginning on data 517

Select whether the vectors (features and meta-data information) are listed in rows or columns

Rows

Select which row to use as class

#1:time

Select which row to use as subclass

no subclass

Select which row to use as subject

#2:#ID

Per-sample normalization of the sum of the values to 1M (recommended when very low values are present)

Yes

Execute

What it does

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 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 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 respect to the class of interest; biological significance is subsequently investigated using a set of pairwise tests among subclasses using the (unpaired)

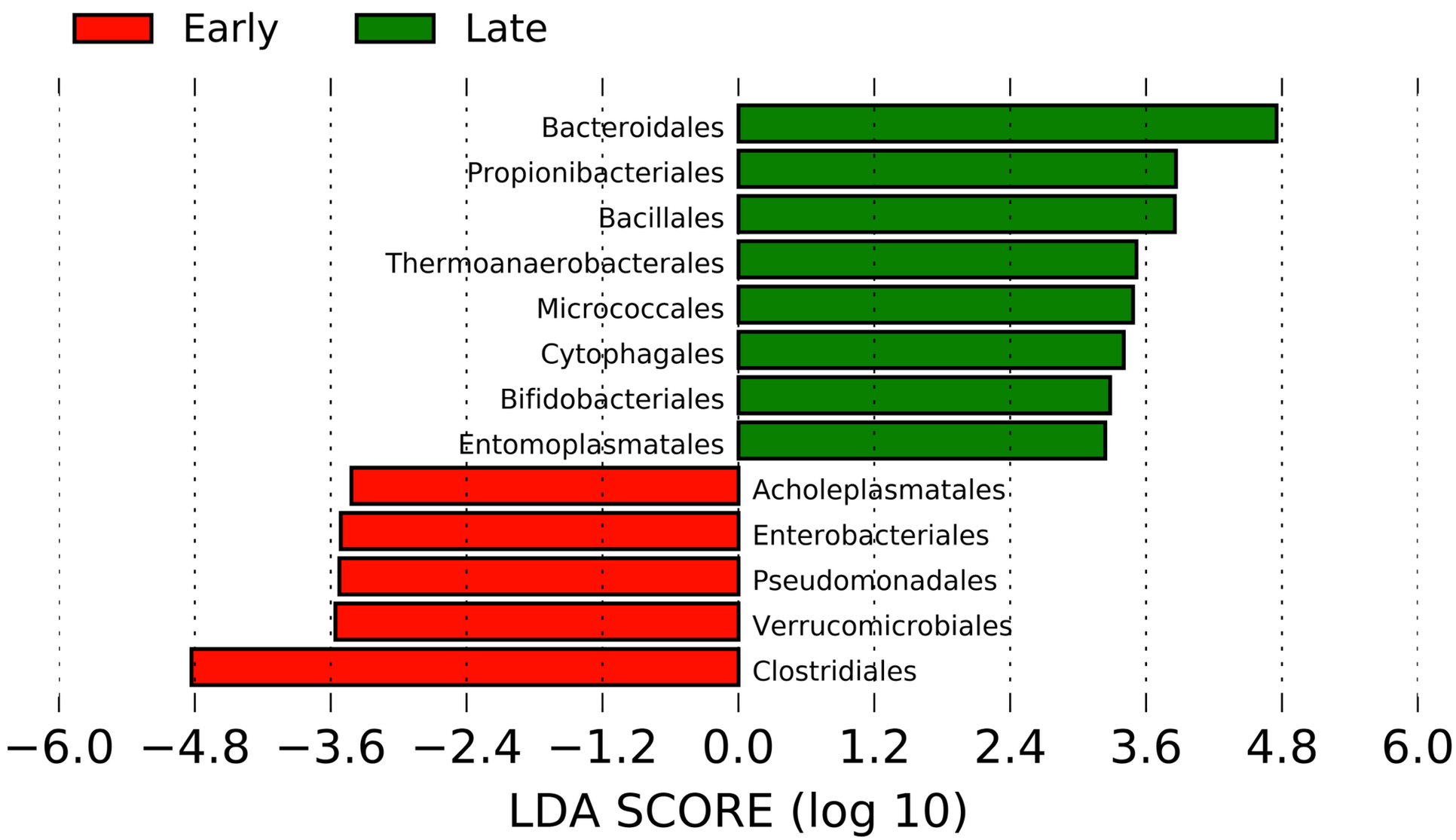
History

search datasets

MiSeq Data
140 shown, 378 hidden
812.58 MB

- 518: Remove beginning on data 517
- 517: Concatenate datasets on data 508 and data 516
- 516: MiSeq Early / Late Metadata
- 515: Vegan Beta Diversity on data 507 (plot)
- 514: Vegan Beta Diversity on data 507 (output_beta_diversity_scores_no_triangular)
- 513: Vegan Beta Diversity on data 507 (output_beta_diversity_scores_plot)
- 512: Vegan Beta Diversity on data 507 (output_beta_diversity_scores_triangular)
- 511: Vegan Beta Diversity on data 507 (Mean output_mean_beta_diversity_index)
- 510: Vegan Beta Diversit

Plot LEfSe Results



Visualization: Phinch

PHINCH

36: Convert BIOM on data 1 and data 35

Biological Observation Matrix v1

format: **biom1**, database: ?

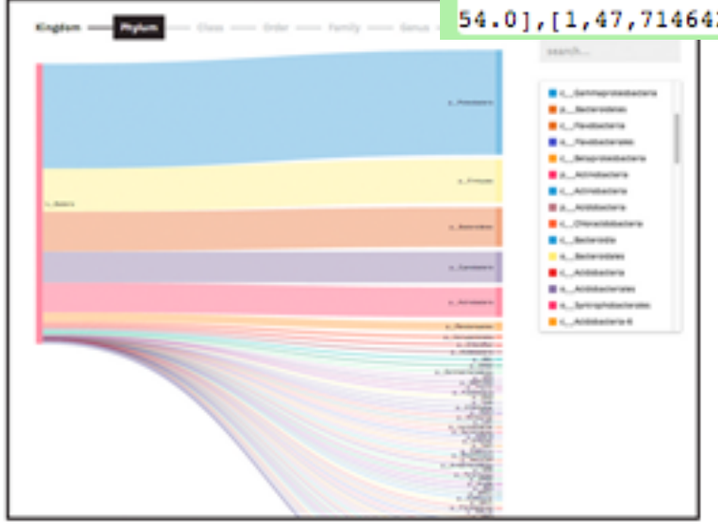
view biom at Phinch

{"id": "None","format": "Biological Obs
: [2681, 431],"data": [[0,1,1.0],[0,2,1
,1.0],[0,394,13.0],[0,397,1.0],[0,430,2
[1,14,88953.0],[1,15,305717.0],[1,16,30
0,92012.0],[1,31,286741.0],[1,32,92374.
54.0],[1,47,714642.0],[1,48,225806.0],[

Taxonomy Bar Chart

Bubble Chart

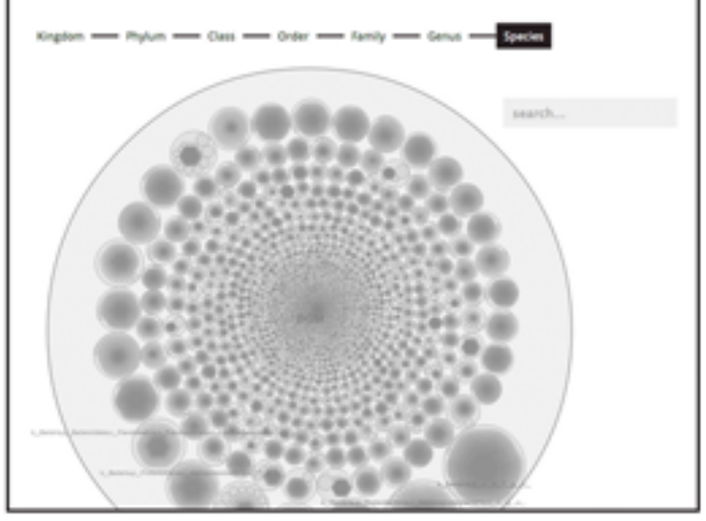
Sankey Diagram



Donut Partition

Attributes Column Chart

Bubble Chart By OTU



https://github.com/PitchInteractiveInc/Phinch

https://github.com/blankenberg/Phinch

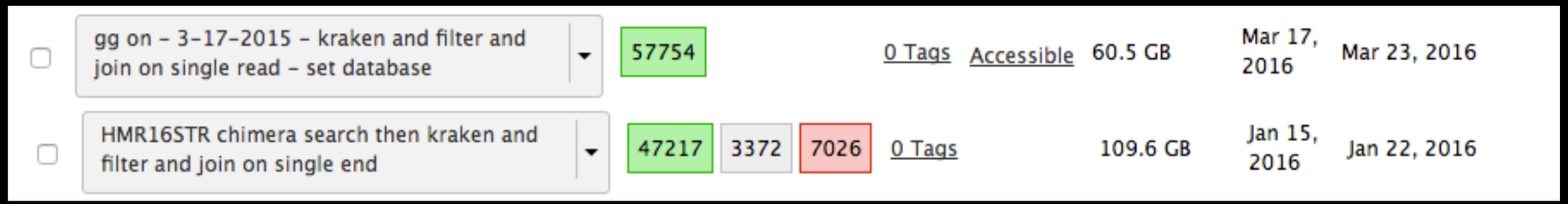
http://www.bx.psu.edu/~dan/Phinch/

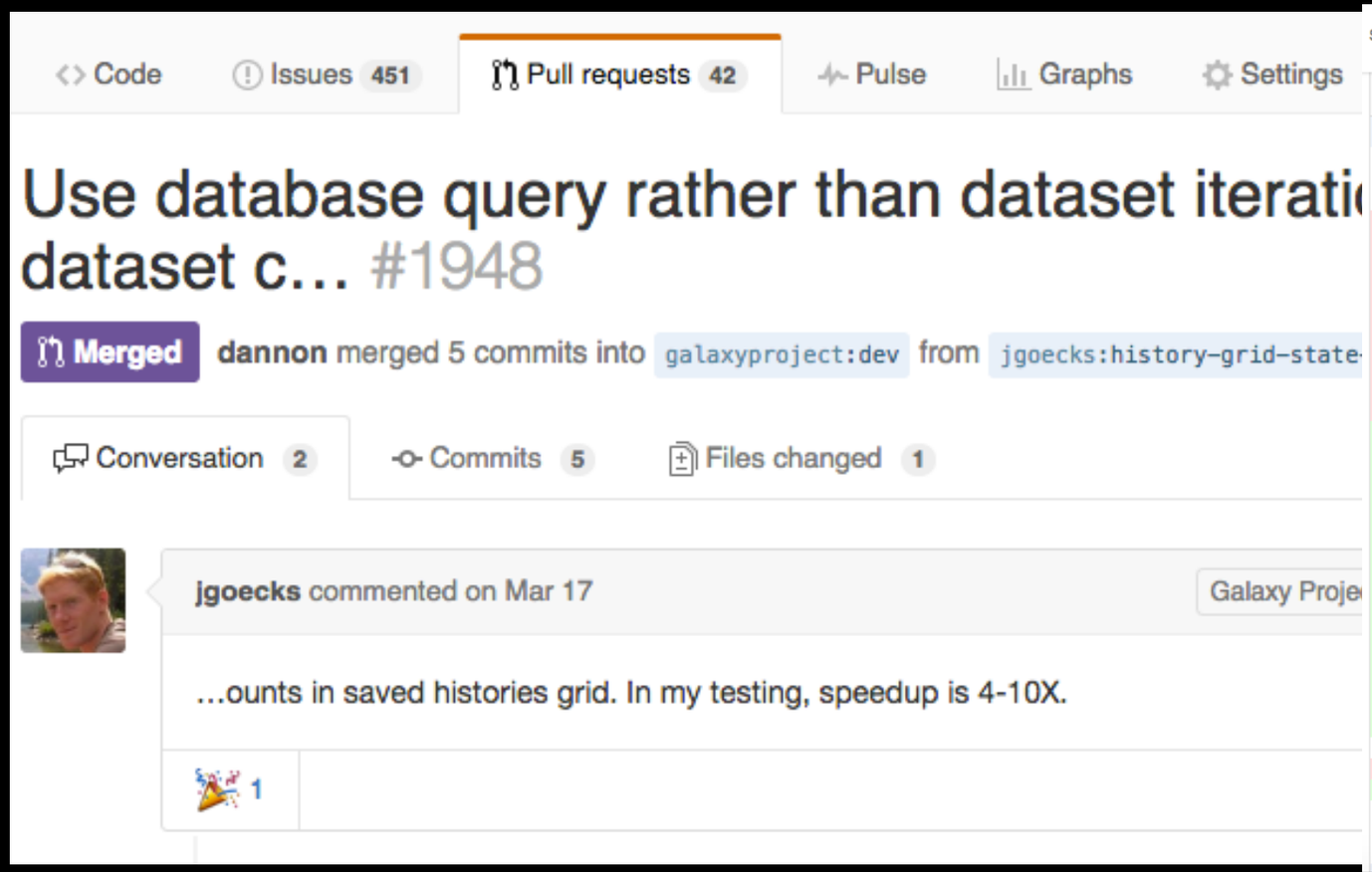
Overview

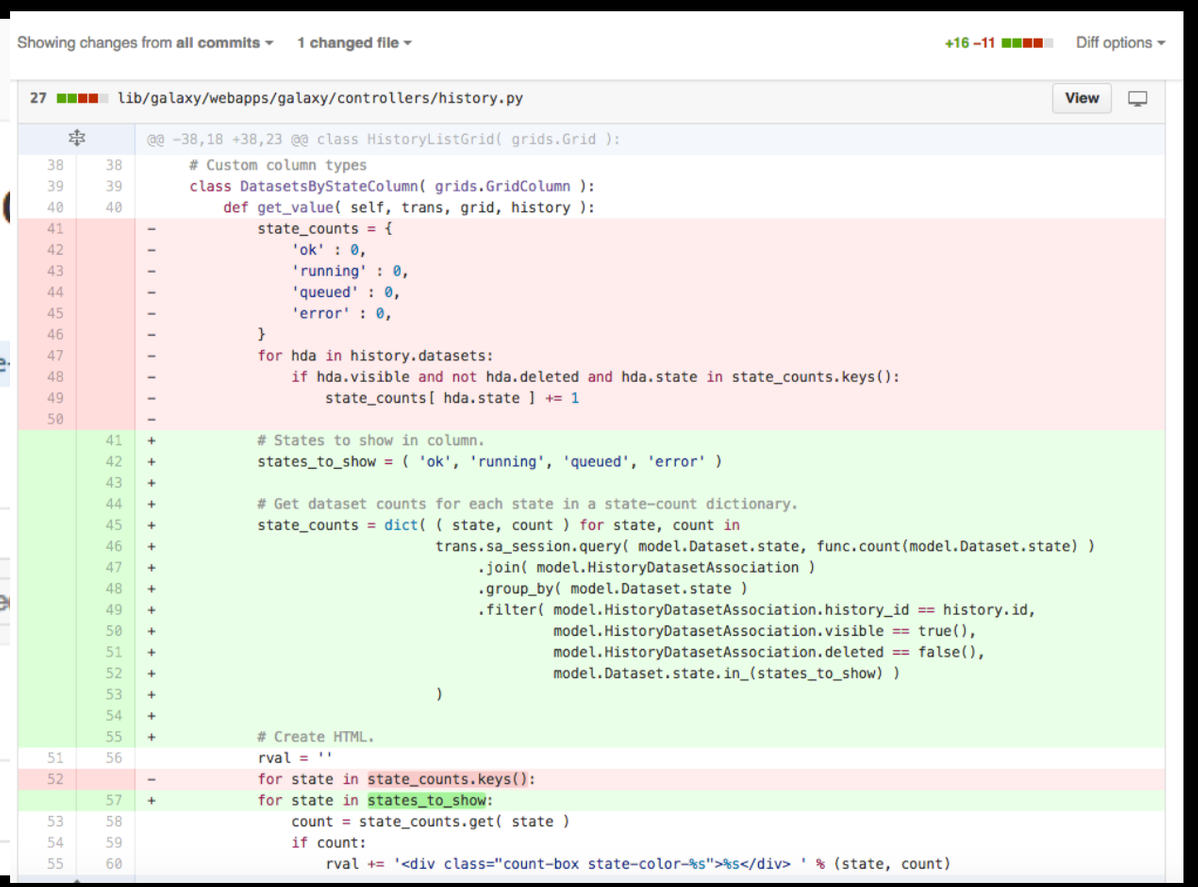
- **Brief Introduction to Metagenomics**
- **Doing It**
- **Challenges, Solutions, and Future Work**

Decreasing load times for list of saved histories

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

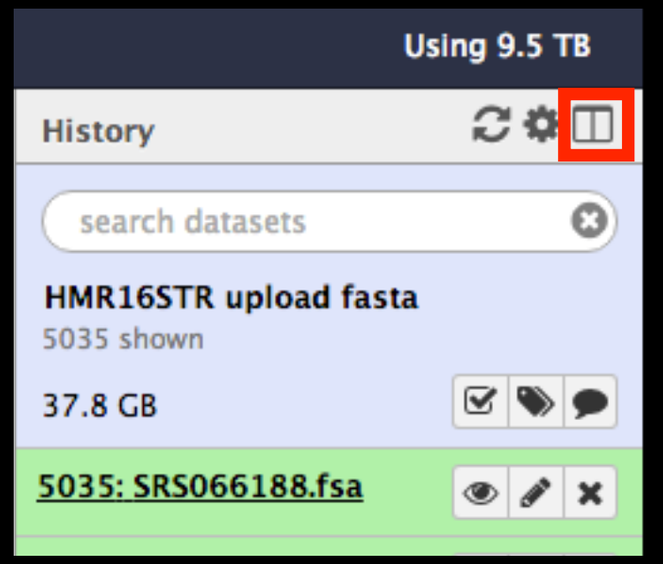






Multiple History View

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



Galaxy / Dan's Brubeck Playground lyze Data Workflow Shared Data Visualization Admin Help User Using 9.5 TB

Done | search histories | search all datasets | Create new

Current History

HMR16STR upload fasta
5035 shown
37.8 GB
search datasets
Drag datasets here to copy them to the current history

- 5035: SRS066188.fsa
- 5034: SRS066144.fsa
- 5033: SRS065725.fsa
- 5032: SRS065719.fsa
- 5031: SRS065718.fsa
- 5030: SRS065711.fsa
- 5029: SRS065701.fsa
- 5028: SRS065685.fsa
- 5027: SRS065676.fsa
- 5026: SRS065675.fsa
- 5025: SRS065669.fsa
- 5024: SRS065666.fsa
- 5023: SRS065665.fsa
- 5022: SRS065656.fsa

Switch to

Copy of 'HMR16STR upload fasta'
4076 shown
29.74 GB
search datasets

- 4076: SRS054236.fsa
- 4075: SRS054232.fsa
- 4074: SRS054219.fsa
- 4073: SRS054126.fsa
- 4072: SRS054121.fsa
- 4071: SRS054112.fsa
- 4070: SRS054108.fsa
- 4069: SRS054094.fsa
- 4068: SRS054072.fsa
- 4067: SRS054065.fsa
- 4066: SRS054043.fsa
- 4065: SRS054039.fsa
- 4064: SRS054007.fsa
- 4063: SRS053956.fsa
- 4062: SRS053925.fsa

Switch to

Copy of 'vegan on 434 pairs'
79 shown
11.67 MB
search datasets

- 79: Vegan Rarefaction on data 1 (R script)
- 78: Vegan Rarefaction on data 1 (Random rarefied community matrix)
- 77: Vegan Rarefaction on data 1 (plot)
- 76: Vegan Rarefaction on data 1 (slope of curve)
- 75: Vegan Rarefaction on data 1 (species probabilities)
- 74: Vegan Rarefaction on data 1 (estimated richness)
- 73: Vegan Rarefaction on data 1 (frequency of species)
- 72: Vegan Rarefaction on data 1 (number of species)
- 71: Vegan Rarefaction on data 1 (R script)
- 70: Vegan Rarefaction on data 1 (Random rarefied community matrix)
- 69: Vegan Rarefaction on data 1 (plot)

Switch to

Copy of '434 Pairs - GG - Abundances @ P'
94.99 KB
search datasets
loading datasets...

Switch to

434 Pairs - GG -
94.99 KB
search datasets
loading dat

Dataset Collection Operations

Collection Operations (Limited) #2434

Merged **martenson** merged 10 commits into `galaxyproject:dev` from `jmchilton:fewer_collection_opts` 15 days ago

Conversation 13 | Commits 10 | Files changed 31

jmchilton commented 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

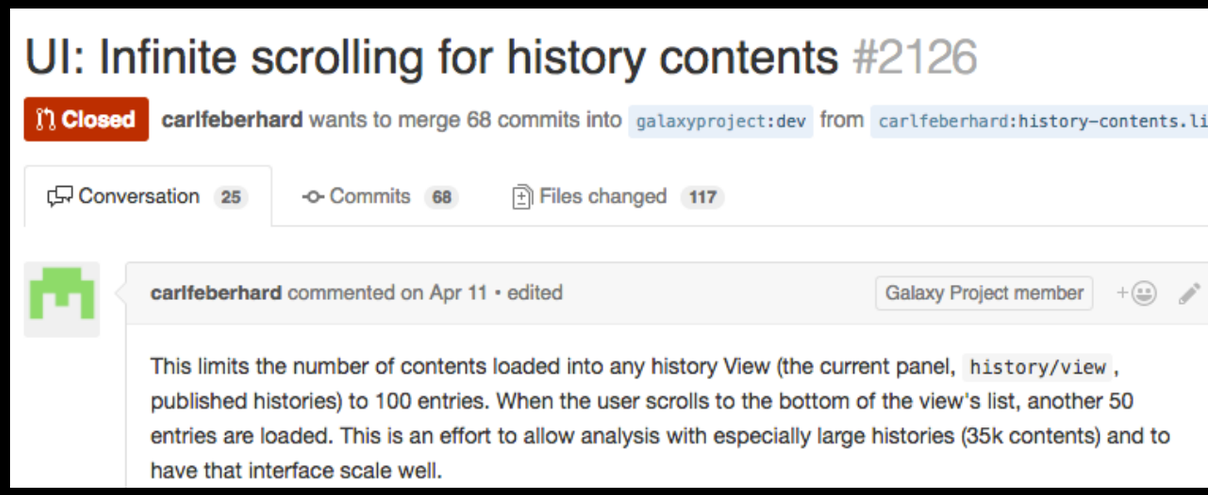
- Default Galaxy names lose Sample names (Tool xyz on Dataset 1,2,3,4)

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



- 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

Unwatch 25 | Star 25 | Fork 74

Code | Issues 78 | Pull requests 41 | Wiki | Pulse | Graphs | Settings

mothur: update make.contigs to use paired collections #853

Edit

Open **shiltemann** wants to merge 1 commit into `master` from `mohtur-updates`

Conversation 0 | Commits 1 | Files changed 1 | +15 −8

shiltemann commented 3 days ago — Galaxy Project member

No description provided.

`update make.contigs to use paired collections` Verified 832267a

shiltemann referenced this pull request 3 days ago

IUC Contribution Fest - Metagenomics and mothur #419

0 of 8 tasks complete — Open

Add more commits by pushing to the **`mohtur-updates`** branch on **galaxyproject/tools-iuc**.

Labels

None yet

Milestone

No milestone

Assignees

No one—assign yourself

1 participant

Additional Options Available: Qiime

WIP: Add Qiime wrappers #431

Edit

Open **bgruening** wants to merge 76 commits into `master` from `qiime`

Conversation 39 · Commits 76 · Files changed 72 · +8,002 −1

bgruening commented on Nov 30, 2015 • edited — Galaxy Project member

This is our WIP PR during the metagenomics hackathon

edit Saskia:
..and the GCC2016 hackathon.

Resources:
Qiime script documentation: http://qiime.org/scripts/index.html
Qiime github repo: https://github.com/biocore/qiime
Qiime test data in github repo (folder qiime_test_data)

Labels

missing tests

wip

Milestone

No milestone

Assignees

No one—assign yourself

Additional Options Available: FROGS

FROGS: Find Rapidly OTUs with Galaxy[5] Solution

Frédéric ESCUDIE*, Lucas AUER*, Maria BERNARD, Laurent CAUQUIL, Katia VIDAL, Sarah MAMAN, Mahendra MARIADASSOU, Guillermina HERNANDEZ-RAQUET, Géraldine PASCAL

*These authors have contributed equally to the present work.

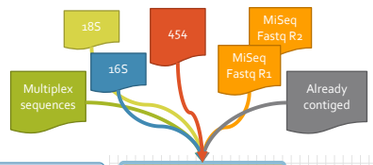
Contact: geraldine.pascal@toulouse.inra.fr

Analyses of bacterial communities → High-throughput sequencing, 16S/18S RNA amplicons → Illumina data, sequenced, great depth → Bioinformatics data processing (FROGS)

ID	Taxonomy	Sample A	Sample B
OTU_1	Bacteria;chlorobi	12 895	54
OTU_2	Bacteria;firmicutes	0	6 812

User-friendly + Accuracy + Speed + Scalability = FROGS

Core workflow



Pre-process

FROGS Pre-process
- Archive file
- dereplicated_file (fasta)
- count_file (tabular)
- summary_file (html)

To merge paired-end reads : Flash
Trimming: Cutadapt

Clustering

FROGS Clustering swarm
- Sequences file
- Count file
- abundance_biom (txt)
- seed_file (fasta)
- swarms_composition (tabular)

Denoising and Clustering with <u>local</u> threshold: SWARM [1]

Chimera

FROGS Remove chimera
- Sequences file
- Abundance file
- non_chimera_fasta (fasta)
- out_abundance_biom (txt)
- out_abundance_count (tabular)
- summary_file (html)

Efficient Chimera Removal : VSEARCH [2]

Filters

FROGS Filters
- Sequence file (Fasta)
- Abundance file (BIOM)
- excluded (txt)
- fasta_output (fasta)
- web (html)
- biom_output (txt)
- krona (html)

<u>A wide choice of filters.</u>

We advise to filter OTU abundances at 0.005% [7].

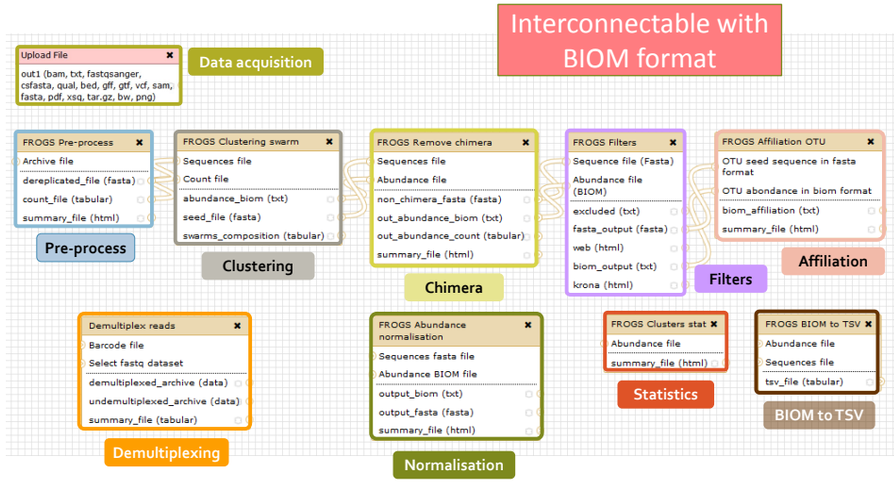
Affiliation

FROGS Affiliation OTU
- OTU seed sequence in fasta format
- OTU abondance in biom format

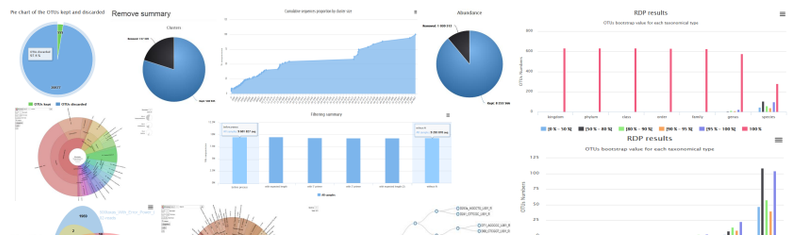
<u>Double taxonomics affiliation</u>
RDP classifier [3] until species.
Blast+ [4] with multi species

FROGS = complete analyze

Interconnectable with BIOM format



Great graphics outputs



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/