

MG-RAST

metagenomics analysis server

v3.6 Tutorial

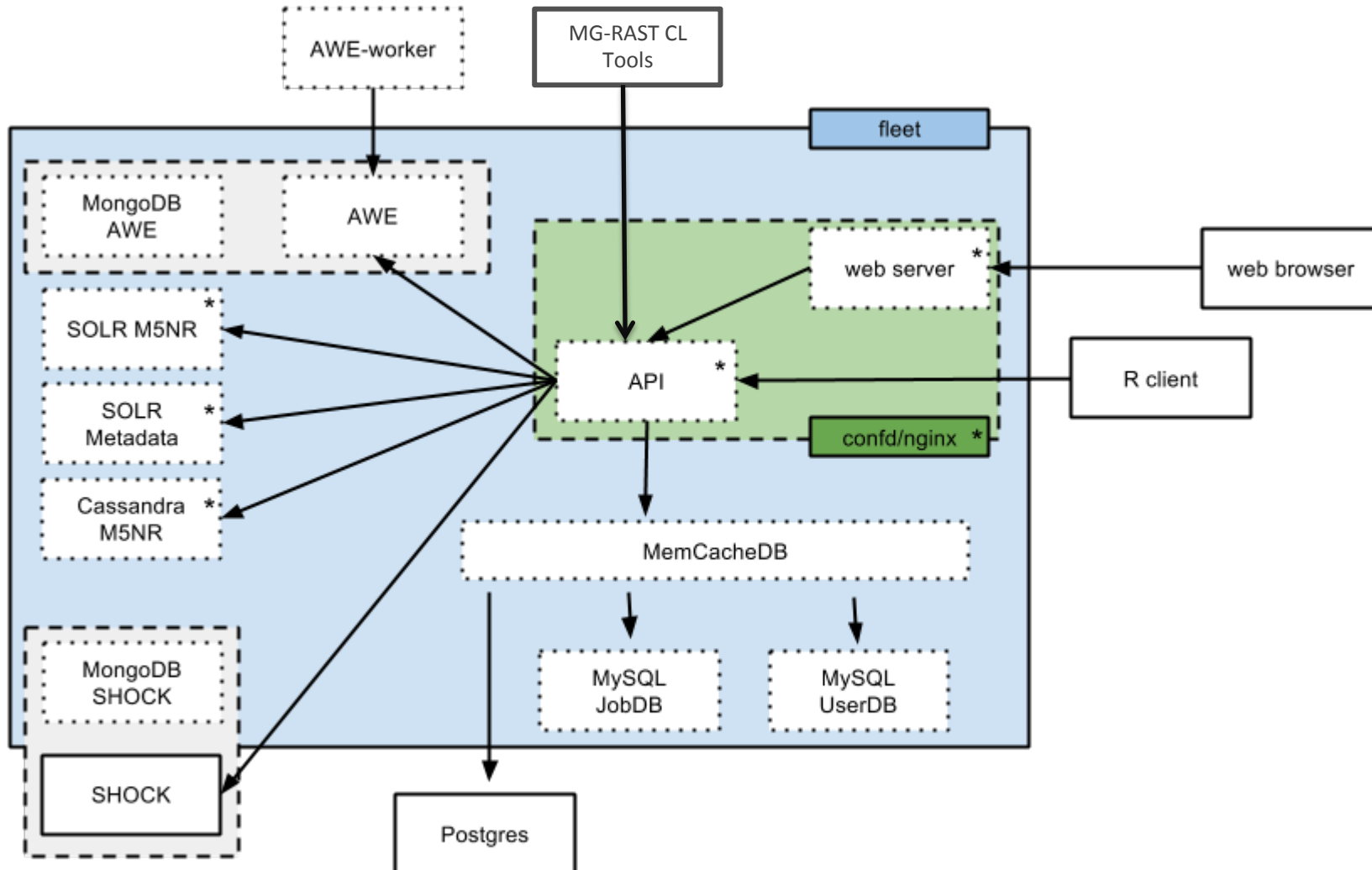
Folker Meyer

Argonne National Laboratory

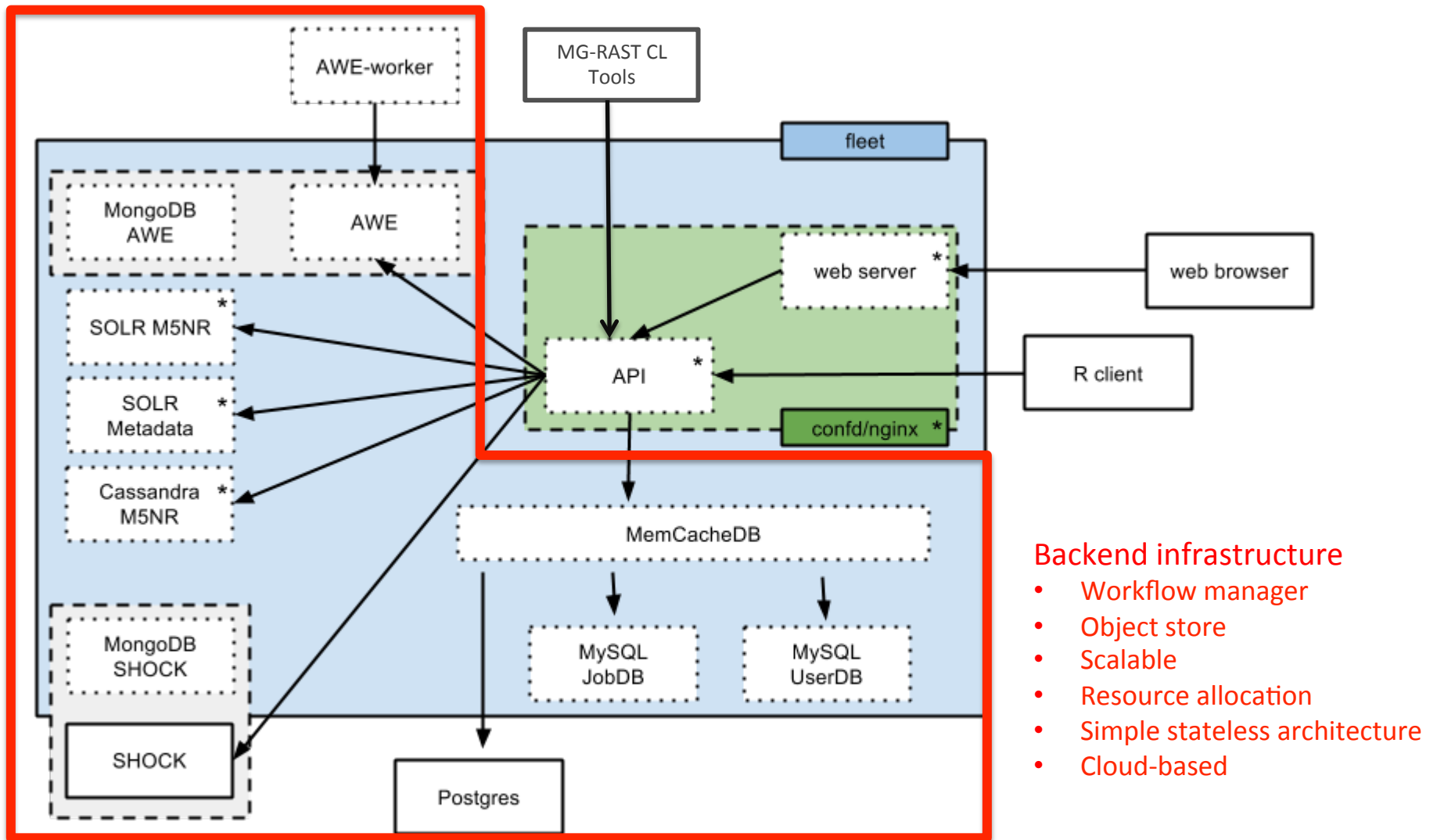
MG-RAST is

- A system to support user-driven analysis of metagenomic data;
- It offers automated quality control, annotation, comparative analysis, and archiving services;
- MG-RAST provides several methods to access data and tools;
- MG-RAST for data analyses and discovery!
 - phylogenetic reconstructions,
 - metabolic reconstructions,
 - explore annotation
 - compare metagenomes

The Big Picture

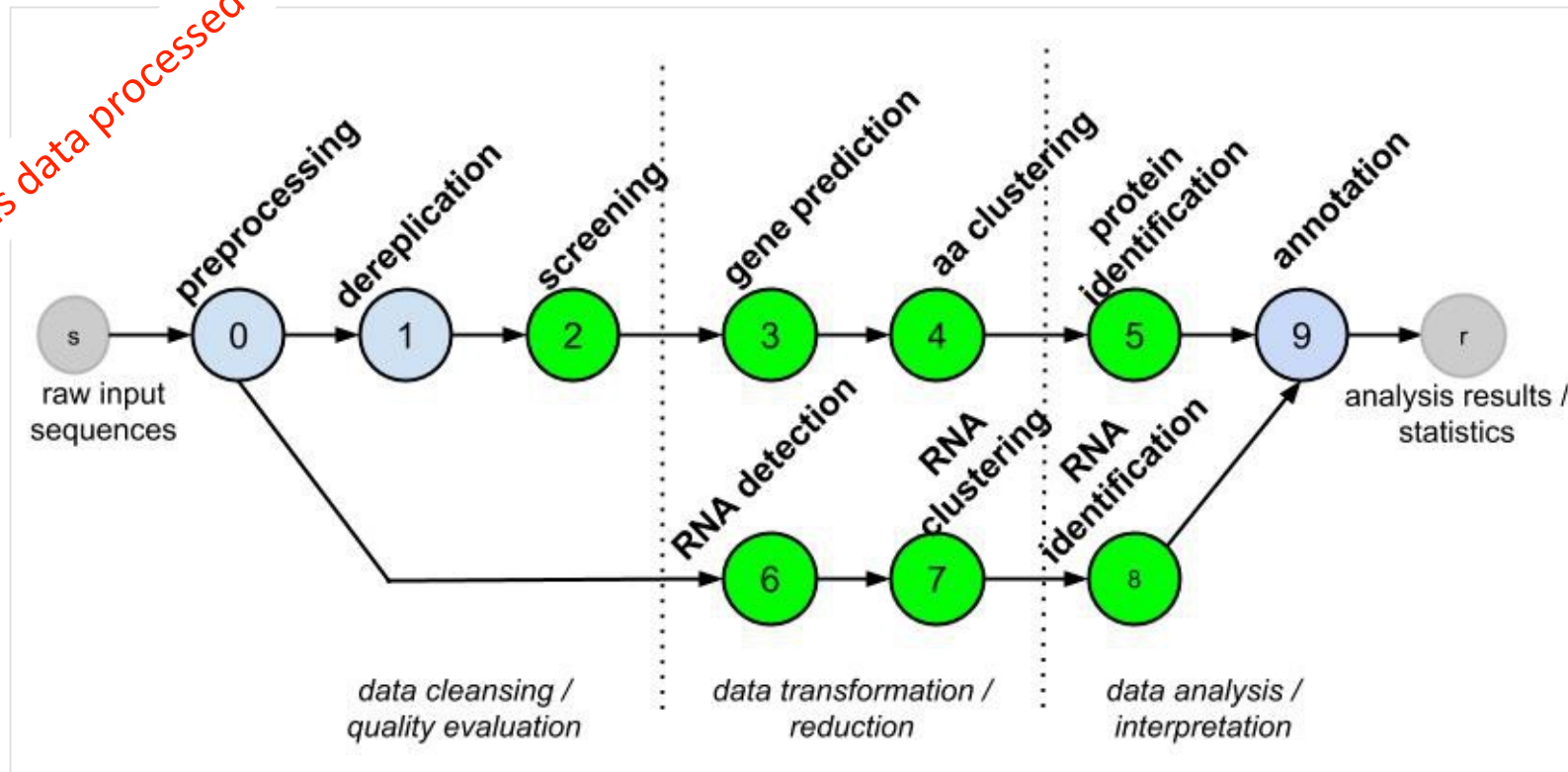


The Big Picture



How is data processed and annotated?

The Big Picture



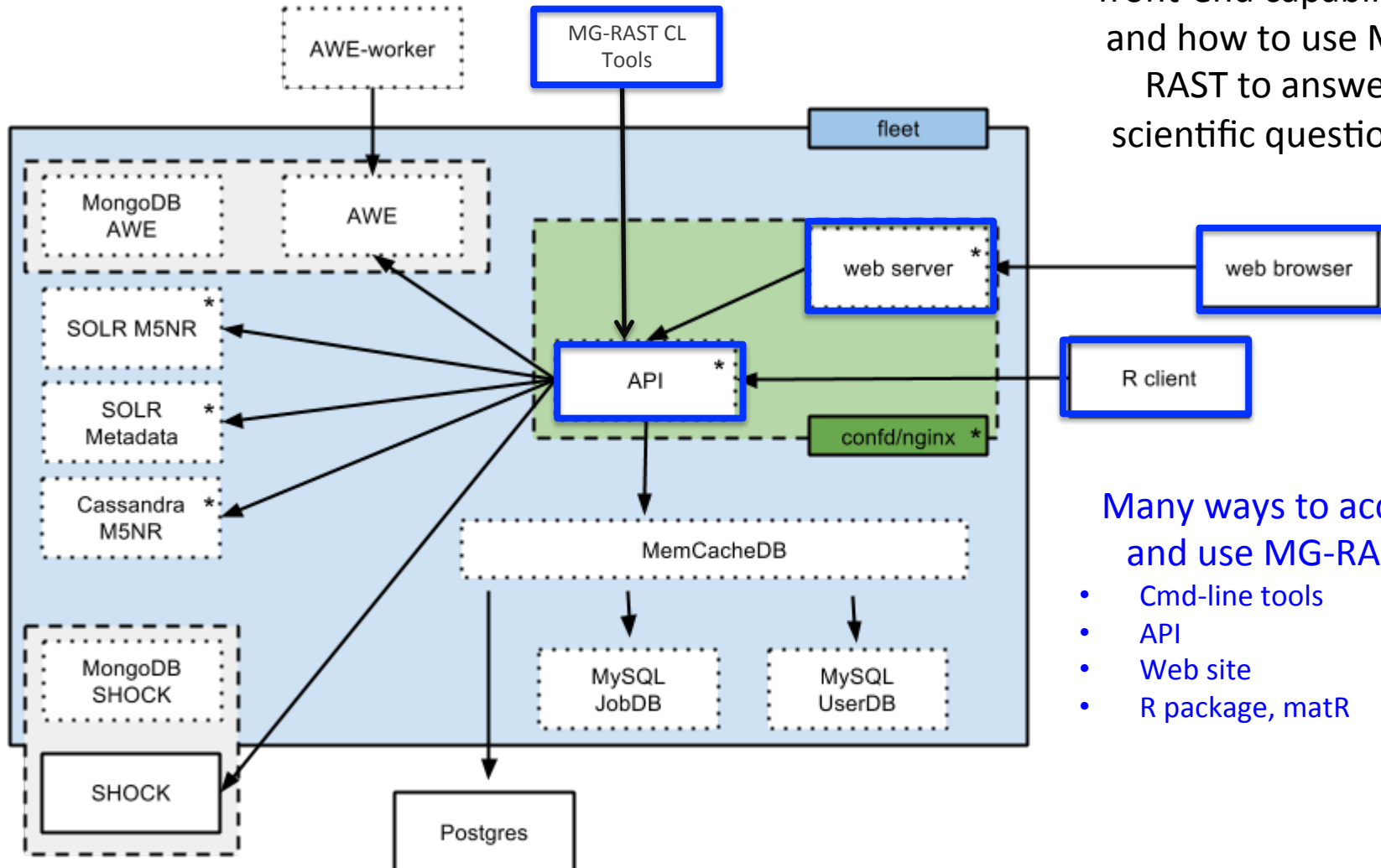
Data submitted to MG-RAST goes through our analysis pipeline.

Composed of three conceptual steps:

1. quality control,
2. data reduction,
3. and analysis

The Big Picture

Focus on MG-RAST front-end capabilities and how to use MG-RAST to answer scientific questions.



Ways to interact with MG-RAST

- Web interface
 - <http://www.metagenomics.anl.gov>
 - Most popular entry point
 - Rich in features for upload, search, data analysis
 - Limitation in number of data sets that can be compared
- API
 - Makes all data created by the pipeline accessible.
 - Complements the existing MG-RAST web interface
 - Compatible with most programming environments
- Cmd-line interface (mg-rast tools)
- R statistical package (matR)

Now for some practical applications!

What you need to get started:

- A MG-RAST account (<http://metagenomics.anl.gov/?page=Register>)
- Web
 - Latest version of Firefox browser
- CMD-Line
 - Download python and install libraries (<https://pip.pypa.io/en/latest/installing.html>)
 - Download MG-RAST command-line tools (<ftp://ftp.metagenomics.anl.gov/tools/upload/>)
 - OS Shell
- R
 - Download and install R package (<https://www.r-project.org/>) AND
 - matR (<https://github.com/MG-RAST/matR/>)
- MD5 sum
 - Need to create MD5sums for files
 - Download <http://www.freebsd.org/> for md5sum command

MG-RAST fundamental concepts

- service side processing
 - stats
- late binding to parameters
 - Filtering data after pipeline completion
- smart data structures
 - enable queries and fast data retrieval
- controlled vocabularies
 - “de-noising” via CVs vs. genbank (dnaA...)

MG-RAST fundamental concepts

- service side processing

- Uses SHOCK storage system
- Enables dynamic storage for server side processing.
- Makes for faster generation of user defined data “views”.
- A more flexible and yet powerful approach.

MG-RAST fundamental concepts

- late binding to parameters

- There is no one size fits all to annotation
- We need multiple ways of viewing the similarities to protein databases.
- Changing from one database (annotation source) to another or changing cutoffs - changes “the picture”.

Therefore we came up with the M5nr and provide “late binding to parameters” to allow users to make their own choices.

MG-RAST fundamental concepts

- late binding to parameters

The M5nr

- MD5-based non-redundant protein database
- Common reference for sharing similarity results.

- Based on databases from:

EBI
GO
JGI

KEGG
NCBI
SEED

Phantome
UniProt
VBI
eggNOG

Source	Source ID	Functional Assignment	Organism
GenBank	EFM52775.1	hypothetical protein ECNC101_19021	Escherichia coli NC101
RefSeq	ZP_07448331.1	hypothetical protein	Escherichia coli NC101
PATRIC	VBIscColi50923_2327	Invasin	Escherichia coli NC101 Unclassified
TrEMBL	EOR0K8	Putative uncharacterized protein	Escherichia coli NC101

>edd13c644cbd9ba3c79ba57ce6d7d09d
MQQTVNYVPNVTNAEITLAASKDPVIADNNDLTTLTAP
SLIQRAMR

+



MG-RAST fundamental concepts

- smart data structures

- MG-RAST has “smart data products” enabling the user -- **at the time of analysis** -- to determine the best parameters.
- No need to recompute data!

MG-RAST fundamental concepts

- controlled vocabularies

- CV's for metadata
- CVs/Ontologies changing over time (<http://bioportal.bioontology.org/>)
- Import of existing CVs into MG-RAST
 - api.metagenomics.anl.gov/metadata
- Defined project wide
- Support multiple versions
 - Default latest
 - Defined within a project

Example 1. Downloading data

Various types of data for download is available.

Dedicated Download page for:

- Metadata
- Submitted data – the original user submission
- Analysis results -- results from EACH STEP of the MG-RAST pipeline.
- Derived data – data based on annotation source and type (taxa or function)

Download from Workbench is also available

Search for metagenome 4472164.3 and download the sequence file that was provided at submission and the metadata file.

Metadata

- Open the metadata file you just downloaded.
 - Here is an example of what format the metadata is in.
 - Can open and edit in a text editor or excel
 - Data is GSC compliant
- Metadata critical for analysis!
- When submitting metadata - you can use an excel template we provide or use our online editor (Metazen).
 - Minimizes effort and reduces complexities of navigating standards.

Exercise 2. Upload and Submission

- Typically, users will come in with their own data and/or looking to analyze public data.
 - *You can use the data you just downloaded to try out the upload if you don't have data of your own.*
- In order to analyze your data – follow some easy steps to upload and submit data.
- Time needed to process data varies considerably (average 2 weeks for WGS), depending on:
 - the size of your data set(s),
 - whether you provide metadata and
 - if you plan to publish data on MG-RAST.

Uploading to MG-RAST

- Green “up” arrow takes you to upload

- Sequences can be in FASTA, FASTQ, or SFF format.

- Can be compressed files.

The screenshot displays the MG-RAST metagenomics analysis server interface. At the top, the logo "MG-RAST metagenomics analysis server" is visible. Below the logo, there are navigation tabs: "upload" (highlighted with a green dot), "submit", and "progress".

The main content area contains instructions for data submission. It states: "Data submission is a two step process. As the first step data is uploaded into your private inbox on the MG-RAST server, this area is write only and only accessible to you. From the inbox data can then be submitted. Use the [upload](#) function or use [our API](#) to upload your data." It also mentions that submission of multiple files requires metadata and provides links to an Excel template and the MetaZen tool. The second step involves submitting data for processing, with a note that data is removed from the inbox upon successful submission and that users can monitor progress via the job status.

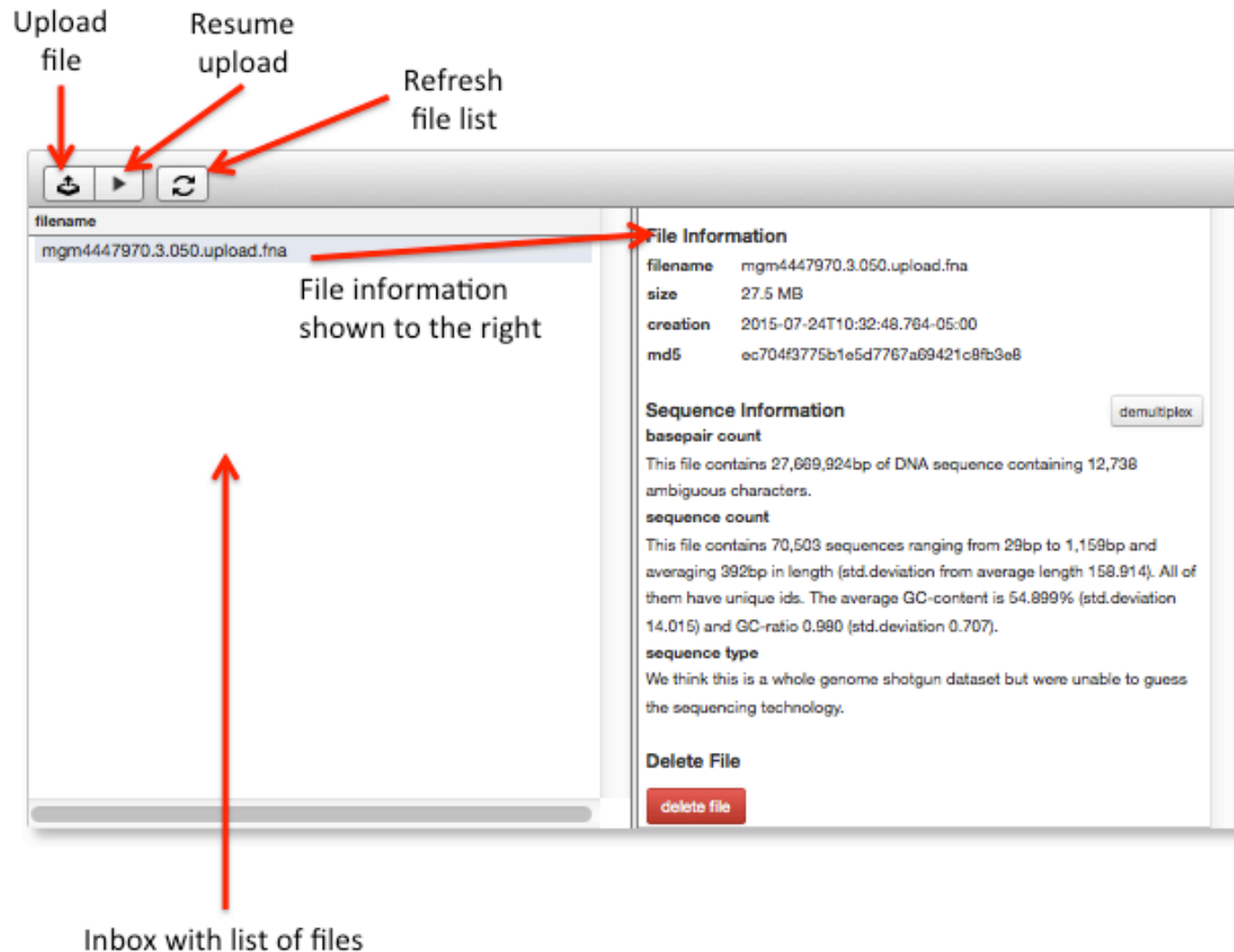
Below the instructions, there is a file upload interface. On the left, a list of files to be uploaded is shown, including "seqs.fasta", "metadata_spreadsheet_F7q1hez.xls", and several "Hiseq-PE-overlapping-small_R1.fastq" and "Hiseq-PE-overlapping-small_R2.fastq" files. A red arrow points to this list with the label "Files to be uploaded".

On the right, a file details panel shows "seqs.gz" with a modified date of "Tue Jun 10 2014 03:43:57 GMT-0500 (CDT)", a size of "19.0 MB", and a type of "application/x-gzip". A green "start upload" button is present, with a red arrow pointing to it and the text "To begin upload – click here". An "advanced" button is also visible, with a red arrow pointing to it and the text "Advanced options let you change chunk size.".

At the bottom right, a green "next" button is shown, with a red arrow pointing to it and the text "Once all files are uploaded, proceed to submission".

On the far right, there are two sidebars: "frequent questions" with links to "File Formats", "Metadata", and "Studies and Projects"; and "running actions" with a note that no actions are currently running on files in the inbox.

Elements of the file browser



Upload progress

The screenshot shows a web-based file upload interface. On the left, a list of files is displayed under the heading 'filename'. The files include 'seqs.fasta', 'metadata_spreadsheet_F7q1hez.xls', and several fastq files. On the right, the upload progress for 'Miseq_PE_joinable_diatom_small_R1.fastq' is shown. The progress bar indicates that 3.5 MB of the 26.6 MB file has been uploaded. A red arrow points from the text 'View upload progress' to the progress bar. Below the progress bar are buttons for 'pause upload' and 'abort'.

filename

- seqs.fasta
- metadata_spreadsheet_F7q1hez.xls
- Hiseq-PE-overlapping-small_R.fastq.fastq
- Hiseq-amplicon-overlapping-SRR1798050small_2.fastq
- Hiseq-amplicon-overlapping-SRR1798050small_1.fastq
- Miseq-PE-amplicon-overlapping_R2.fastq
- Miseq-PE-amplicon-overlapping_R1.fastq
- Hiseq-PE-overlapping-small_R1.fastq
- Hiseq-PE-overlapping-small_R2.fastq
- clovr_acinetobacter_example.sff

uploading
Miseq_PE_joinable_diatom_small_R1.fastq

filename Miseq_PE_joinable_diatom_small_R1.fastq
modified Mon Mar 02 2015 06:41:03 GMT-0600 (CST)
size 26.6 MB
type -

3.5 MB complete

View upload progress

|| pause upload

■ abort

Now to submit your data

1. select metadata file

metadata_spreadsheet_F7q1hez.xls

Select a spreadsheet with metadata for the project you want to submit.

In order to map sequence files to metadata libraries, the names of the sequence files must exactly match the library *file_name* fields or match the library *metagenome_name* fields minus the file extension.

Note: While metadata is not required at submission, the priority for processing data without metadata is lower.

Metadata is important!
Suggested, but not required.

☐ I do not want to supply metadata

select

Select files from your upload to add to a project.

2. select project

3. select sequence file(s)

4. choose pipeline options

5. submit

Submission is done ...

when the
completed sections
turn green.

Test Submission (200706)

[Status](#) [Pipeline](#) [Settings](#)

Below are all tasks that are part of the MG-RAST pipeline for your submission.

- **Green bars**, indicating completed tasks, can be expanded via mouseclick
- **Blue bars** indicate tasks currently being computed on
- **Orange bars** represent the next tasks to be queued
- **Gray tasks** are waiting for completion of another task they depend on
- **Red bars** indicate an error

✓ qc_stats	15/05/2015 12:14:20
started	15/05/2015 11:10:30
completed	15/05/2015 12:14:20
duration	1 hours 3 minutes
inputs	0 (42.4 MB)
outputs	0 (86.6 KB) <i>temporary</i> 1 (7.2 KB) <i>temporary</i> 2 (0.0 B)

✓ preprocess	15/05/2015 10:44:51
✓ dereplication	15/05/2015 10:45:11
✓ screen	15/05/2015 12:53:38
✓ rna detection	15/05/2015 12:14:27
✓ rna clustering	15/05/2015 12:14:37

How far along has my data progressed?

Job ID	Stage	Status	Progress bar
210616	complete	✓ completed	
214843	qc_stats	✖ in-progress	
214844	qc_stats	✖ in-progress	

Progress Bar Legend

Green = completed successfully

Blue = in progress

Orange = queued stage

Red = error

Gray = waiting for other stages to complete

Using MG-RAST to answer scientific questions: An example study

- Feces samples from HMP (Human Microbiome Project) consortium.
- Study citation:
 - Vital M, Howe AC, Tiedje JM. *Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data*. MBio. 2014 Apr 22;5(2):e00889. PubMed PMID: 24757212
- They screened for butyrate synthesis pathways in 15 metagenomes from stool samples of healthy individuals.
 - Why? To understanding the role of butyrate producers in health and disease.
 - Used “expensive” methods to annotate – Use MG-RAST to screen to reduce compute. (e.g. Run pipeline, search and download subset and run through other tools (like Pfam).

Example 3. Finding samples

The research team used 15 random stool samples from HMP healthy patients.

- Go to browse page and filter table by:
 - Project: Human Microbiome Project
 - Material: Feces
 - Disease state: Healthy

Browse Metagenomes

Your Data Summary

Available for analysis? 2
In Progress? 0
Shared with you? 149661
Collections? 6
Projects? 7

Click on the blue links above to browse just your data. For more information visit the Support page.

Public Data Summary

MG-RAST has large number of datasets that users have provided for public use.

of Metagenomes 28198
of Projects 927
Base pairs 9981 Gbp
Sequences 94038 million
Environments 15
PI's 1675

To make your dataset(s) public, select one of your datasets and click the Make Public link.

ALL METAGENOMES
group by project

Current table counts

public (1606) private (0) shared (11)

metagenomes 3 projects 2 biomes 2 features 5 materials 0 altitudes 0 depths 1 locations 0 ph's 2 countries 1 temperatures 1 pi's 1

clear table filters

display 25 items per page

displaying 1 - 25 of 1617

project	name	bps	sequences	biome	feature	material	sequencing type	drisee	select	job #
Human Microbiome Project (HMP)	SRR062417	6650150500	66501505	terrestrial biome	human-associated habitat	feces	WGS	46.102	public	<input type="checkbox"/> id
Human Microbiome Project (HMP)	SRR062287	6371064800	63710648	terrestrial biome	human-associated habitat	mucus	WGS	45.699	public	<input checked="" type="checkbox"/> project
Human Microbiome Project (HMP)	SRR062418	6711513000	67115130	terrestrial biome	human-associated habitat	feces	WGS	45.689	public	<input checked="" type="checkbox"/> name
Human Microbiome Project (HMP)	SRR061345	2852305000	2852305	terrestrial biome	human-associated habitat	mucus	WGS	45.535	public	<input checked="" type="checkbox"/> bps
Human Microbiome Project (HMP)	SRR061908	405671954	4016554	terrestrial biome	human-associated habitat	organic material	WGS	45.497	public	<input checked="" type="checkbox"/> sequences
Human Microbiome Project (HMP)	SRR063501	247123200	2471232	terrestrial biome	human-associated habitat	mucus	WGS	45.443	public	<input checked="" type="checkbox"/> biome
Human Microbiome Project (HMP)	SRR060002	67464400	674644	terrestrial biome	human-associated habitat	mucus	WGS	45.218	public	<input checked="" type="checkbox"/> feature

next» last»

☐ environment package
☒ sequencing type
☐ altitude
☐ depth
☐ location
☐ ph
☐ country
☐ temperature
☐ sequencing method
☐ pi
☐ avg seq length
☒ drisee
☒ a-diversity

You can use the table to filter the HMP samples based on metadata

Finding samples

You can add columns to the Browse table to sort or filter by.

Browse Metagenomes

Your Data Summary

Available for analysis^[?] 2
In Progress^[?] 0
Shared with you^[?] 149661
Collections^[?] 6
Projects^[?] 7

Click on the blue links above to browse just your data. For more information visit the [Support](#) page.

Public Data Summary

MG-RAST has large number of datasets that users have provided for public use.

of Metagenomes 28198
of Projects 927
Base pairs 9981 Gbp
Sequences 94038 million
Enviroments 15
PI's 1675

To make your dataset(s) public, select one of your datasets and click the [Make Public](#) link.

ALL METAGENOMES
group by project

Current table counts

public (1606) private (0) shared (11)

metagenomes 1617 projects 3 biomes 2 features 2 materials 5 altitudes 0 depths 0 locations 1 ph's 0 countries 2 temperatures 1 pi's 1

clear table filters

display 25 items per page

add selected to a collection

next» last»

displaying 1 - 25 of 1617

project	name	bps	sequences	biome	feature	material	sequencing type	drisee	select	
Human IV							WGS			
Human Microbiome Project (HMP)	SRR062417	6650150500	66501505	terrestrial biome	human-associated habitat	feces	WGS	46.102	public	
Human Microbiome Project (HMP)	SRR062282	6371064800	63710648	terrestrial biome	human-associated habitat	mucus	WGS	45.699	public	
Human Microbiome Project (HMP)	SRR062418	6711513000	67115130	terrestrial biome	human-associated habitat	feces	WGS	45.689	public	
Human Microbiome Project (HMP)	SRR061345	285230500	2852305	terrestrial biome	human-associated habitat	mucus	WGS	45.535	public	
Human Microbiome Project (HMP)	SRR061908	405671954	4016554	terrestrial biome	human-associated habitat	organic material	WGS	45.497	public	
Human Microbiome Project (HMP)	SRR063501	247123200	2471232	terrestrial biome	human-associated habitat	mucus	WGS	45.443	public	
Human Microbiome Project (HMP)	SRR060002	67464400	674644	terrestrial biome	human-associated habitat	mucus	WGS	45.218	public	

Filters:

- ☐ job #
- ☐ id
- ☒ project
- ☒ name
- ☒ bps
- ☒ sequences
- ☒ biome
- ☒ feature
- ☒ material
- ☐ enviroment package
- ☒ sequencing type
- ☐ altitude
- ☐ depth
- ☐ location
- ☐ ph
- ☐ country
- ☐ temperature
- ☐ sequencing method
- ☐ pi
- ☐ avg seq length
- ☒ drisee
- ☐ α-diversity

Add in quality score: DRISEE

Example 4: Data Quality - Examine Samples

Sort your filtered samples by DRISSE score and see the range of quality.








Look at the one with best quality (lowest score) and the one with the worst (highest score)

Note: A DRISSE score of zero means the metagenome did not meet standards for calculation.

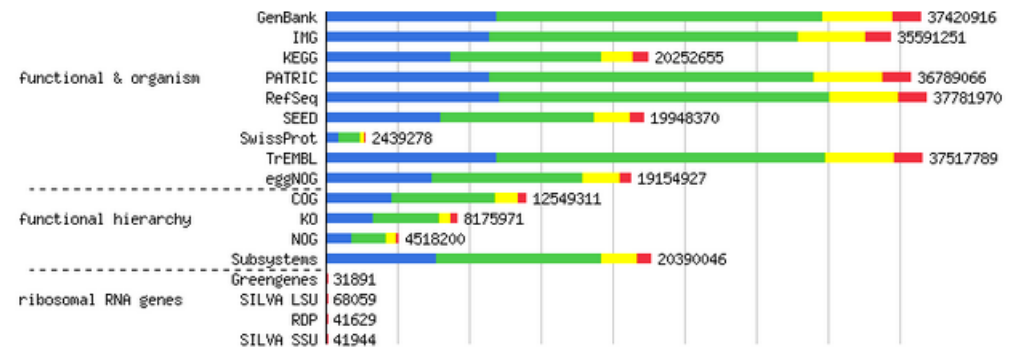
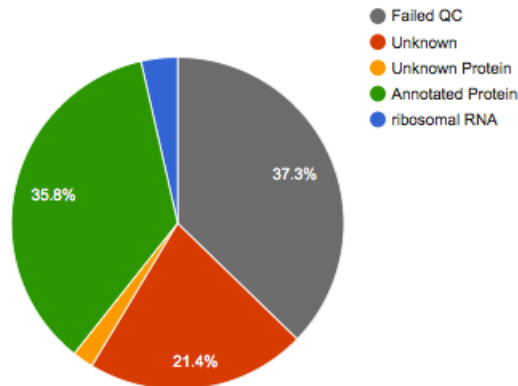
Clicking on the metagenome IDs takes you to the Overview page with a summary of the annotation and analysis.

Metagenome overviews

- Provide information on:
 - Summary of the annotation run (annotated, unannotated)
 - QC  
 - Taxonomic breakdown
 - (for WGS) Functional breakdown
 - Technical data
 - Links to →  Download  Analyze  Search

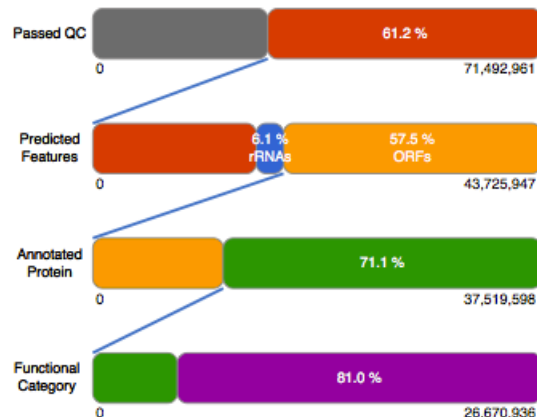
Example Sections of Overview

Sequence Breakdown



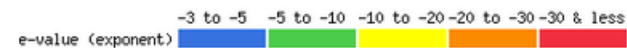
ANALYSIS FLOWCHART

27,767,014 sequences failed quality control. Of those, dereplication identified 23,940,065 sequences (33.5% of total) as artificial duplicate reads (ADRs). Of the 43,725,947 sequences (totaling 4,131,507,015 bps) that passed quality control, 25,150,115 (57.5%) produced a total of 37,519,598 predicted protein coding regions. Of these 37,519,598 predicted protein features, 26,670,936 (71.1% of features) have been assigned an annotation using at least one of our protein databases (M5NR) and 10,848,662 (28.9% of features) have no significant similarities to the protein database (orfans). 21,605,397 features (81.0% of annotated features) were assigned to functional categories.

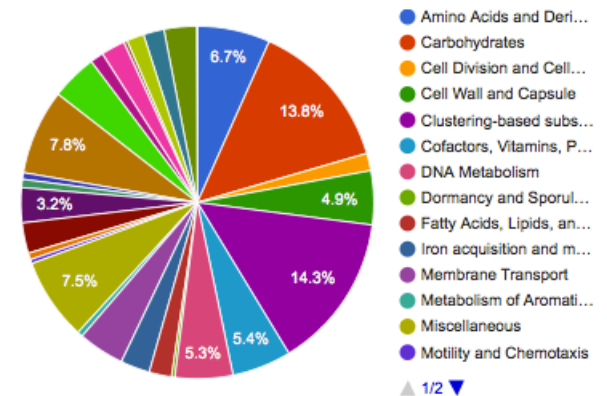


ANALYSIS STATISTICS

Upload: bp Count	7,149,296,100 bp
Upload: Sequences Count	71,492,961
Upload: Mean Sequence Length	100 ± 0 bp
Upload: Mean GC percent	42 ± 9 %
Artificial Duplicate Reads: Sequence Count	23,940,065
Post QC: bp Count	4,131,507,015 bp
Post QC: Sequences Count	43,725,947
Post QC: Mean Sequence Length	94 ± 11 bp
Post QC: Mean GC percent	43 ± 8 %
Processed: Predicted Protein Features	37,519,598
Processed: Predicted rRNA Features	5,998,478
Alignment: Identified Protein Features	26,670,936
Alignment: Identified rRNA Features	110,012
Annotation: Identified Functional Categories	21,605,397



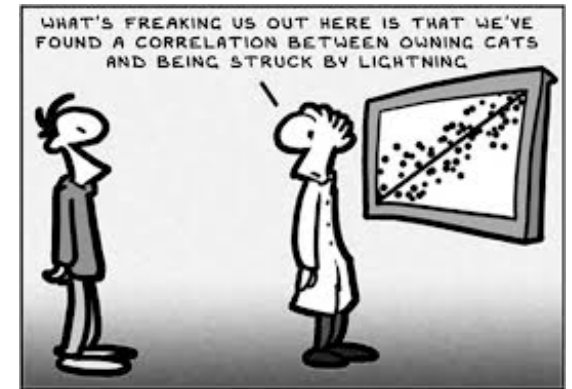
Subsystems [Download chart data](#)
has 14,037,314 predicted functions
37.4% of predicted proteins
52.6% of annotated proteins
[View Subsystems interactive chart](#)



How trustworthy is my data?

The Overview provides insight into the quality of the sequence data.

Why is it so important to know?

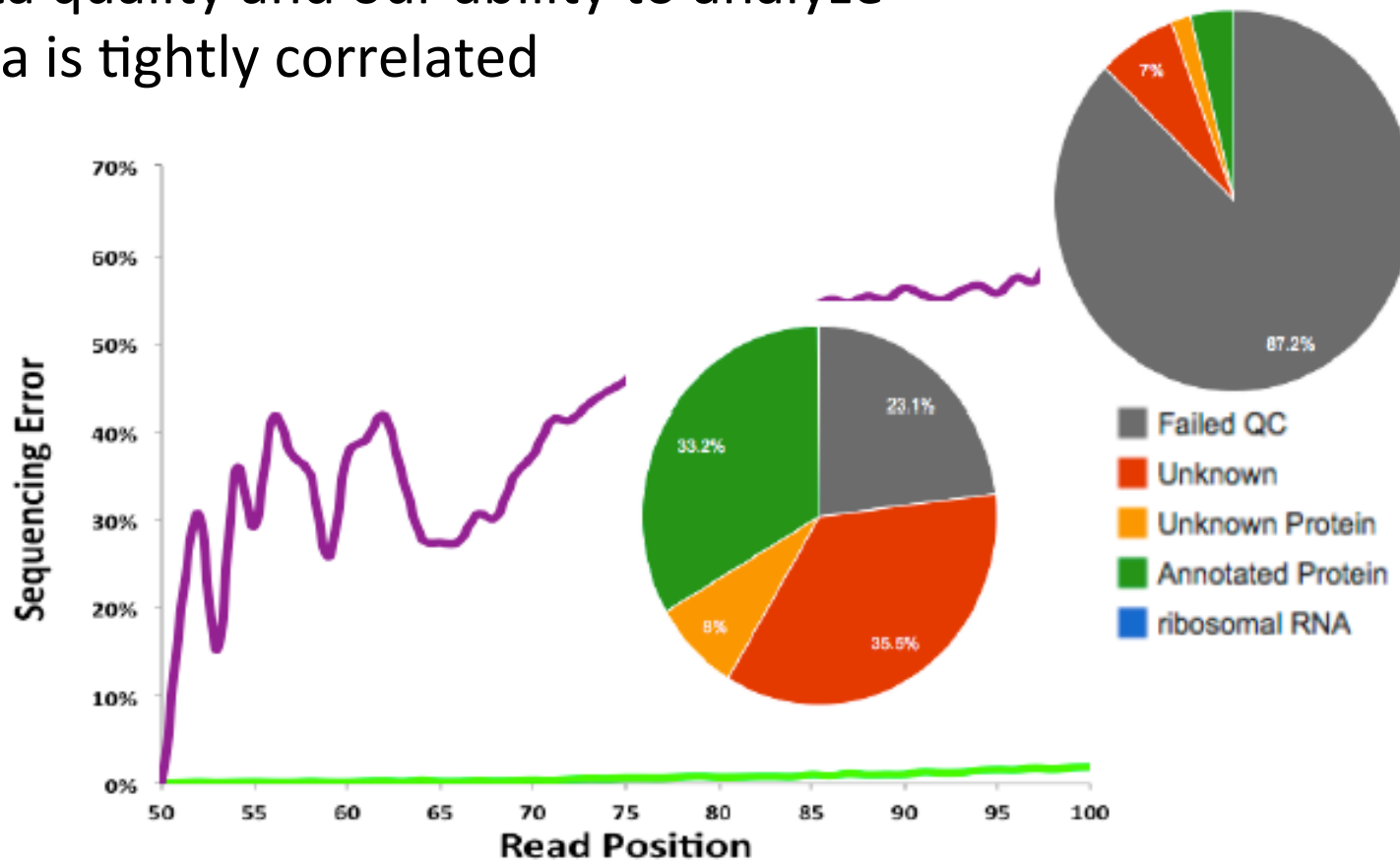


Summaries of technical aspects of the sequence quality to enable sequence data triage:

- DRISSEE for estimating sequence error
- Kmer spectra
- Visualizations of the base caller output

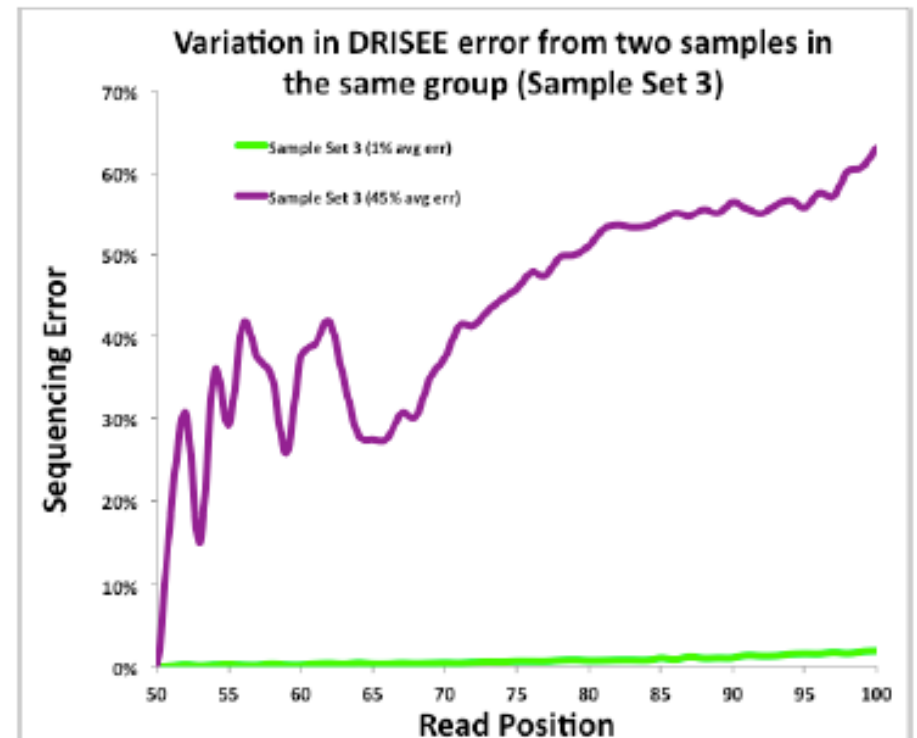
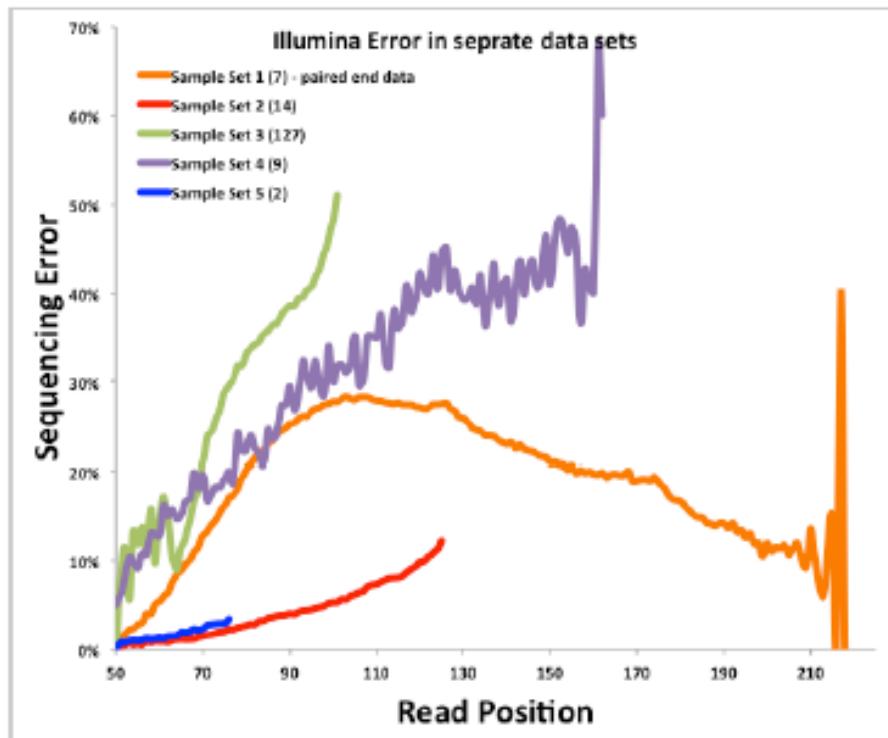
Data Quality Challenge

Data quality and our ability to analyze data is tightly correlated



Duplicate Read Inferred Sequencing Error Estimation --- (DRISEE)

Experiments and even individual samples from a single experiment exhibit unique error profiles.

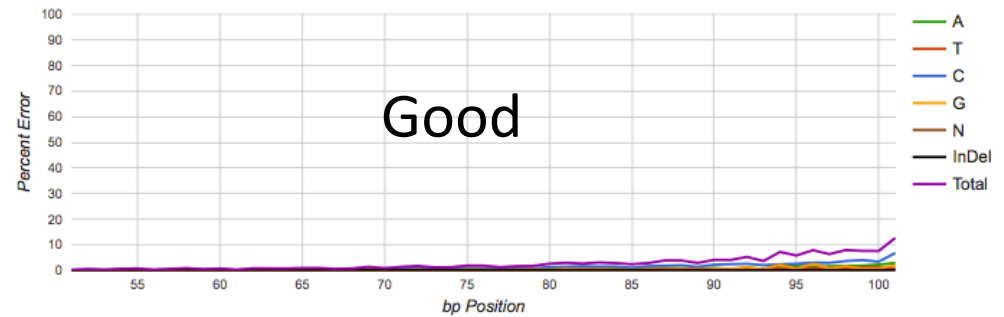


Extreme cases of quality

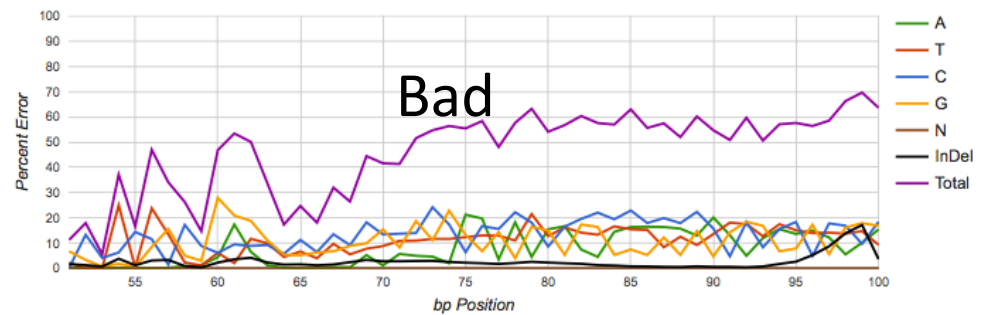
DRISEE [?] [HIDE](#)

Duplicate Read Inferred Sequencing Error Estimation (Keegan et al., PLoS Computational Biology, 2012)

Total DRISEE Error = 2.700 %

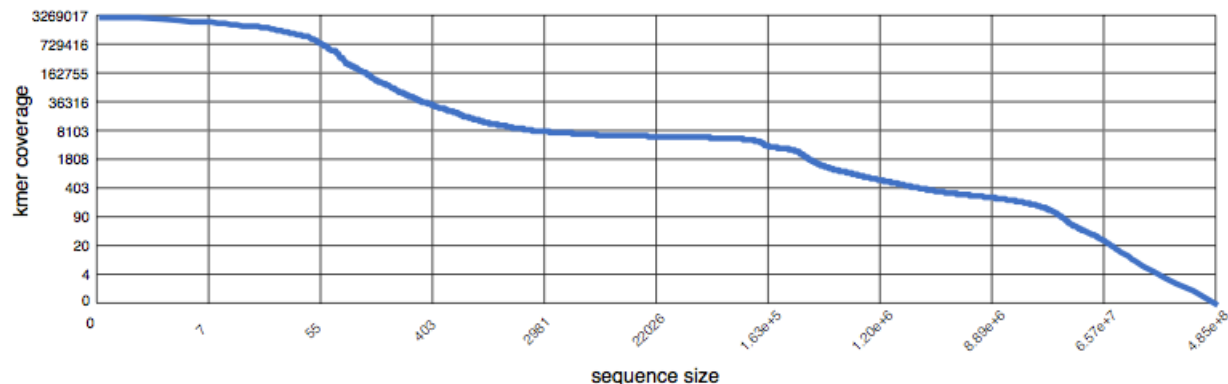


Total DRISEE Error = 46.102 %



Kmer profiles

- The kmer rank abundance plots the relationship between kmer coverage.
- Summarizes the redundancy of sequence datasets.

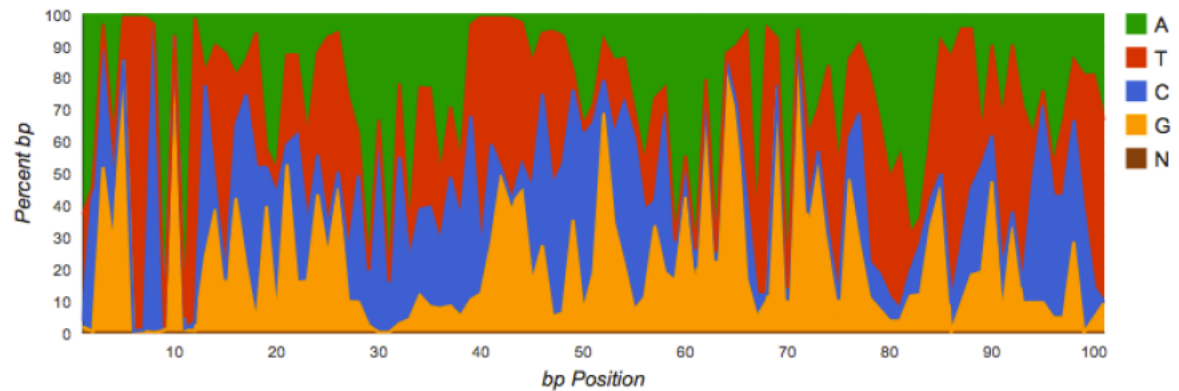


Answers the question “What is the coverage of the n th most-abundant kmer?”.

Nucleotide histogram

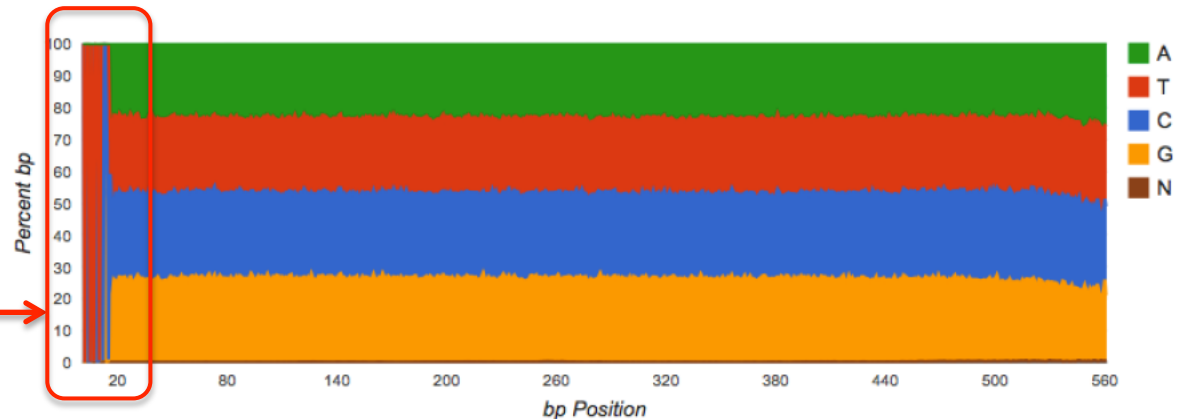
- Amplicon datasets should show biased distributions of bases at each position.

Reflects conservation and variability in the recovered sequences:



- WGS datasets should have roughly equal proportions of basecalls.

An example of
untrimmed barcodes.

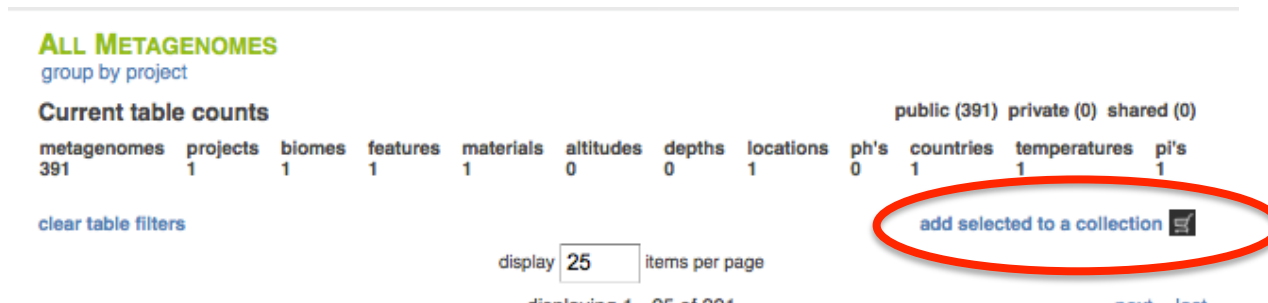


What do the samples look like?

- What characteristics do two extreme HMP samples have?
- What does the nucleotide profile show for the sample with greatest error?
 - If this was your sample, what would you do?
- What data sets would you choose for your comparative analysis? How many?

Example 5. How taxonomically diverse are your samples?

- In Browse table, **select some HMP fecal samples** from healthy patients and add to a collection.
 - Select from table and add create a collection.



- Go to the Analysis page
 - Create rarefaction curve for your collection
 - Create heatmap
 - Create table
 - Select on order Clostridiaceae
 - Subselect and add to workbench
 - Oops – if you can't use MG-RAST web UI with that many features - move over to cmd-line!

Array of analyses

Metagenome Analysis

1 Data Type

ORGANISM ABUNDANCE

Representative Hit Classification

» Best Hit Classification

Lowest Common Ancestor

FUNCTIONAL ABUNDANCE

Hierarchical Classification

All Annotations

2 Data Selection

Metagenomes

☒ compare individually ☐ compare as groups

collections ☒ Amplicon ☒ AmpliconGene ☒ MT ☒ Unknown ☒ WGS ☒

available metagenomes

Fierer-Soil [10]
Fierer-Soil2 [15]
Fierer-Soil3 [16]
Freshwater [15]
HMP feces healthy 6 [6]
HotSpring [9]
Oral Metagenomes [8]

selected metagenomes

SRR059461
SRR059379
SRR059425
SRR059417
SRR059343
SRR063545

Annotation Sources

Max. e-Value Cutoff 1e-5

Min. % Identity Cutoff 60 %

Min. Alignment Length Cutoff 15

Workbench ☐ use features from workbench

3 Data Visualization

alpha

beta

gamma

delta

table

heatmap

PCoA

rarefaction

generate

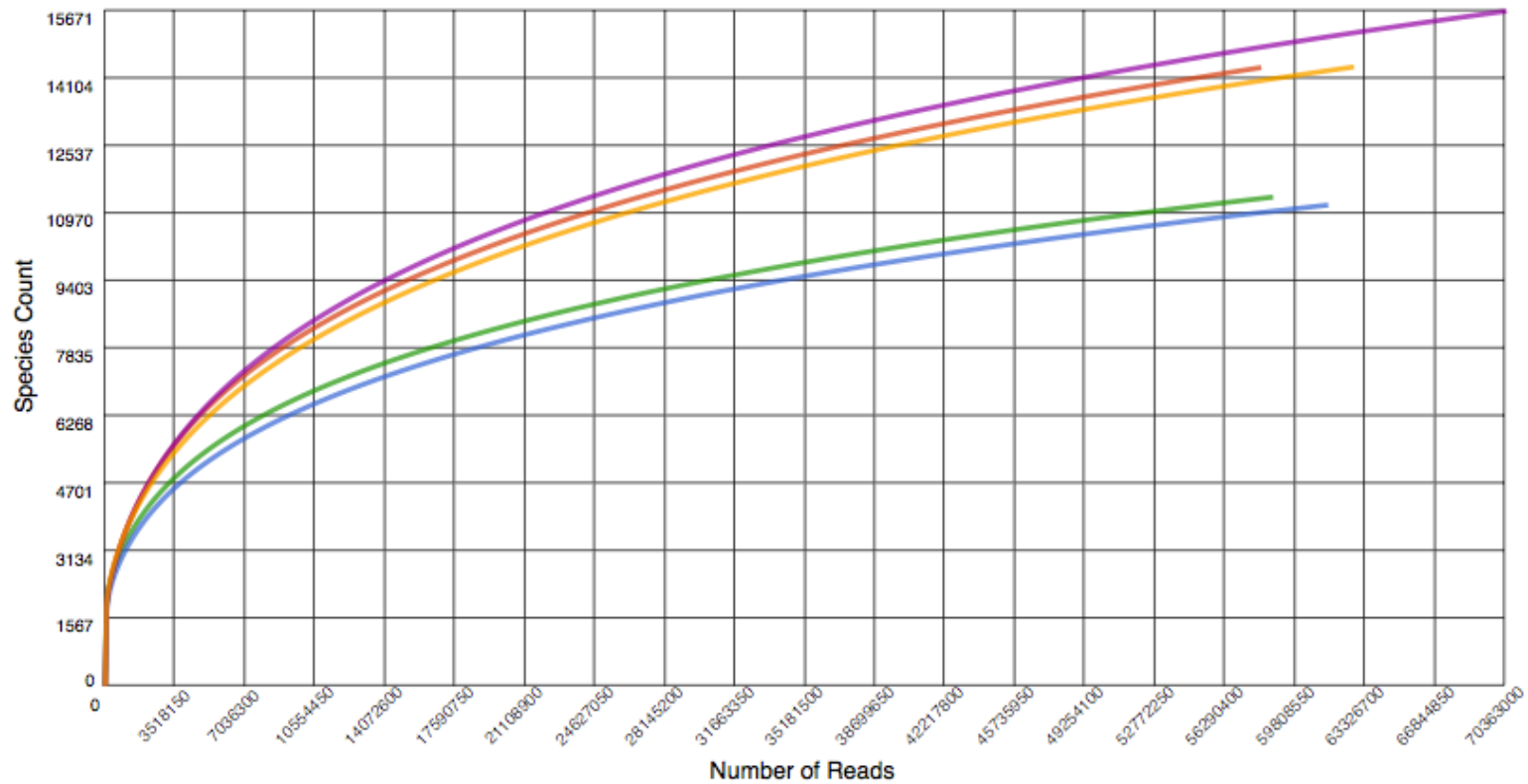
Select your data

What kind of data will you analyze?

Set your parameters and annotation source

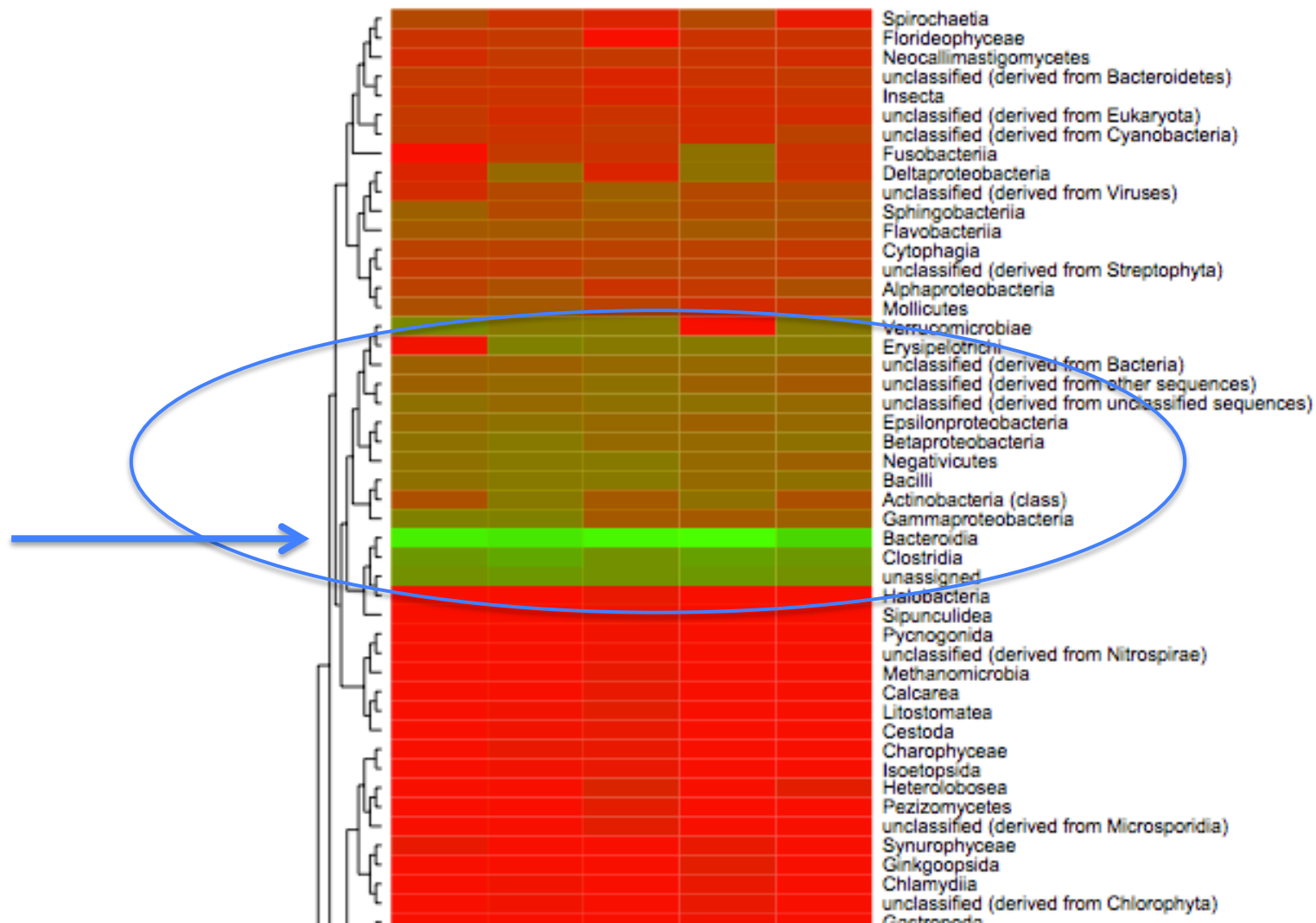
Choose your tool

Sample diversity



rarefaction curve	metagenome ▲▼	alpha diversity ▲▼
	<input type="text"/>	< <input type="text"/>
	44722223.3	79.30
	44721655.3	80.61
	44721999.3	133.28
	44721919.3	161.97
	44721299.3	184.78

What taxa are in common?



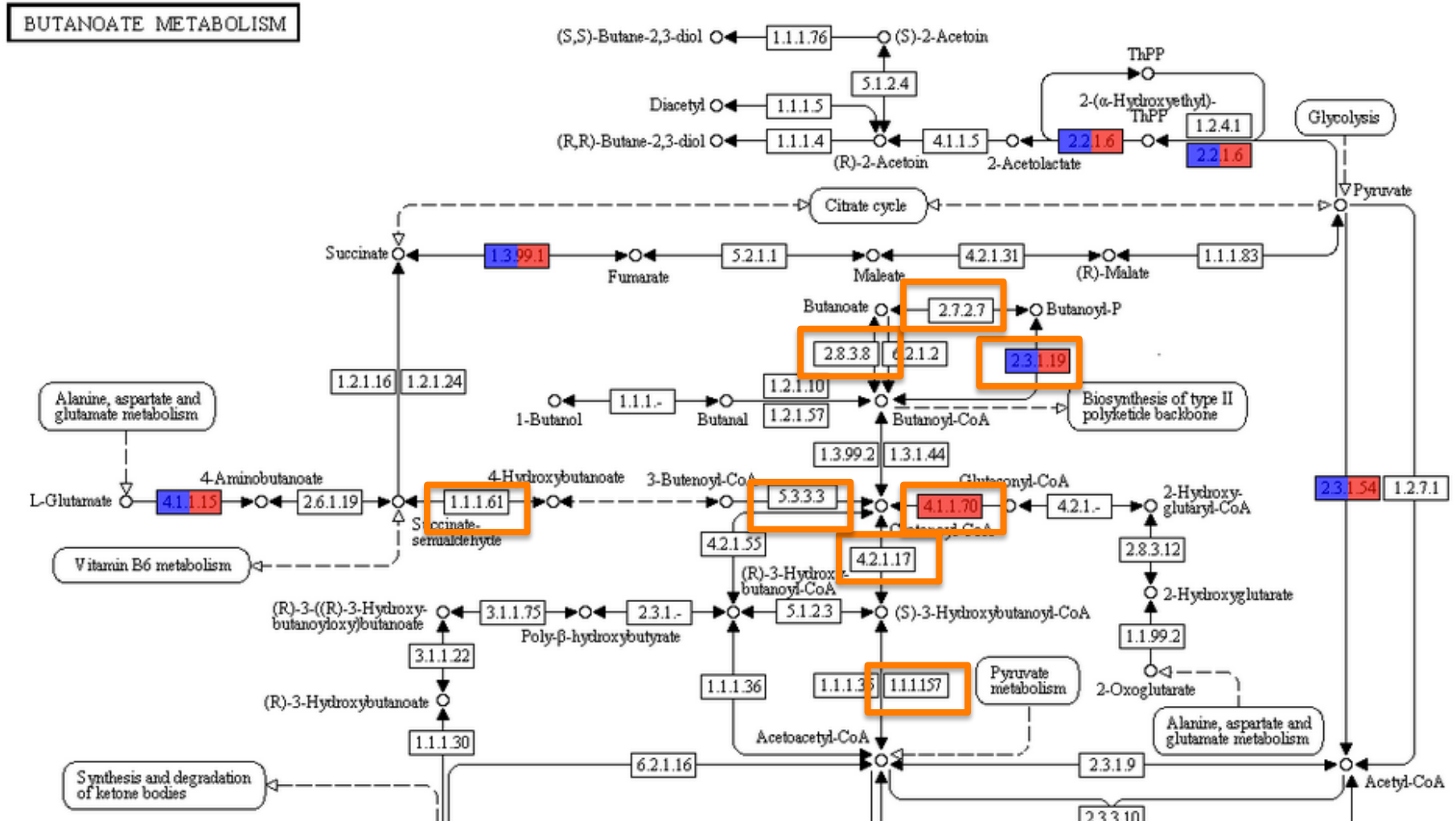
Are there major differences in the samples?

- Comparisons of many large metagenomes can be challenging.
- Out of the hundred or so HMP healthy-patient fecal samples in MG-RAST, how many did you choose?
 - Are we seeing different patterns based on the ones chosen?
- How does subsampling impact results?

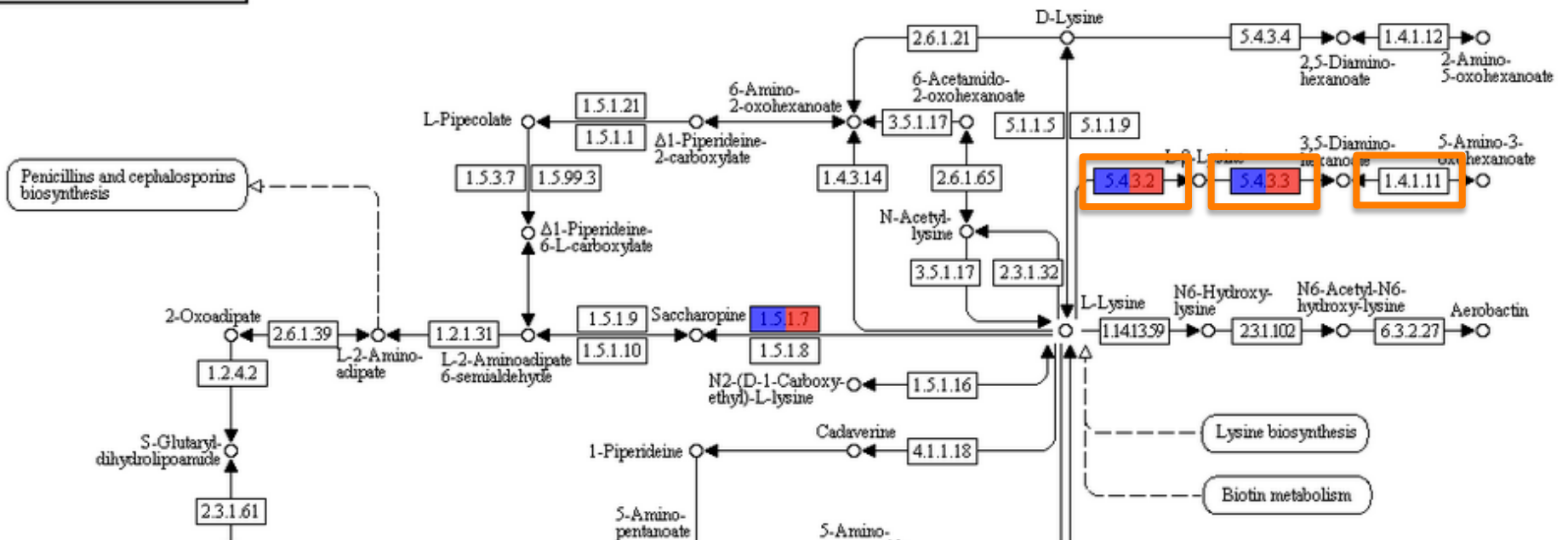
Example 6. Functional diversity

- Switch over to using the Functional Abundance analysis selections.
- Use your HMP collection
 - Search for enzymes involved in butyrate synthesis, like: lysine 2,3-aminomutase (EC 5.4.3.2)
 - How does this change with annotation source?
 - What annotation sources would you choose and why?
 - Using the KEGG mapper
 - How much of the pathways described by Tiedje et al are found? (Check out Butanoate metabolism and Lysine degradation –KEGG)

What functions did you find?



What functions did you find?



Example 7. Digging deeper

What butyrate synthesis genes are present in class Clostridia in your “best” sample for HMP healthy fecal sample?

Lets look at another way to mine data via the UI!

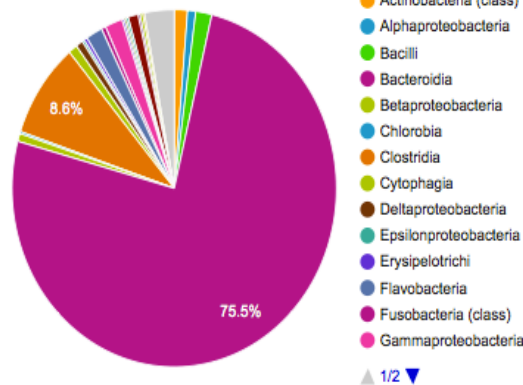
- Go back to Overview page
- What Genera are present in this clade?
 - Limits to how much the workbench can hold! CL-tools/R tools will be better for larger analyses!
- Search table for butyrate-related enzyme functions.
(e.g. 5.4.3.2 in

Select taxon group and move to workbench.

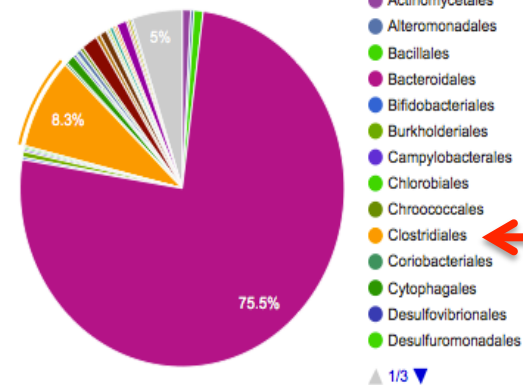
Metagenome 4472164.3

Family *Clostridiaceae*

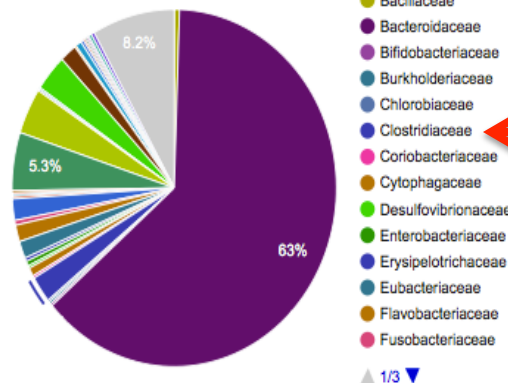
class [Download chart data](#)



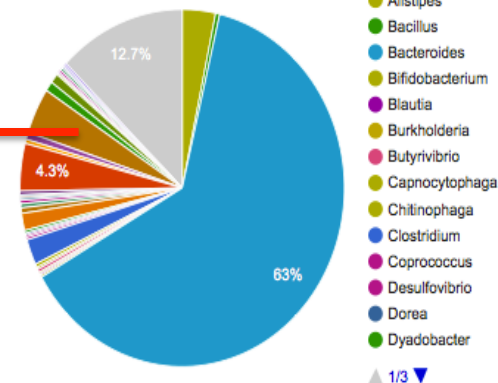
order [Download chart data](#)



family [Download chart data](#)



genus [Download chart data](#)



Moved featured to workbench!

Create Function Table

Metagenome Analysis

1 Data Type

- ORGANISM ABUNDANCE
- Representative Hit Classification
- Best Hit Classification
- Lowest Common Ancestor
- FUNCTIONAL ABUNDANCE
- » Hierarchical Classification
- All Annotations

2 Data Selection

Metagenomes 4472164.3

Annotation Sources

Max. e-Value Cutoff 1e-5

Min. % Identity Cutoff 60 %

Min. Alignment Length Cutoff 15

Workbench ☒ use features from workbench

3 Data Visualization

☐ barchart ☐ tree ☒ table ☐ heatmap ☐ PCoA

KEGG Mapper

Workbench (97108 Features) **Getting Started** **Functional table 1** **Functional table 1**

This data was calculated for metagenome 4472164.3. The data was compared to Subsystems using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases.

Don't forget to check that you want to use workbench data!

Search for function

orkbench (97108 Features) | Getting Started | Functional table 1 | Functional table 1



This data was calculated for metagenome 4472164.3. The data was compared to Subsystems using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases.

Group table by: function change

[download this table](#)

display 15 items per page

available plugins

-  krona graph
-  qiime
- [QIIME report](#)

metagenome	level 1	level 2	level 3	function	abundance	workbench abundance	avg eValue	avg % Ident	avg align len	# hits	workbench # hits	to workbench
4472164	Amino			5.4.3.2	<	<	<	<	<	<	<	
472164.3	Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Lysine degradation	Lysine 2,3-aminomutase (EC 5.4.3.2)	7167	183	-6.92	87.99	29.14	105	4	<input type="checkbox"/>

Example 8. What are community functional differences among 3 HMP body locations?

- Go to matR tutorial
 - Selecting, filtering, normalizing data
 - What are community functional differences among the three body sampling locations?
 - (see tutorial.HMP_subset.9-1-15.R)

Example 9: Exotic operations

- Using the cmd-line tools
 - download all unannotated reads
 - download all dnaK, amoA genes from MANY metagenomes
 - sequence search against all/many data sets