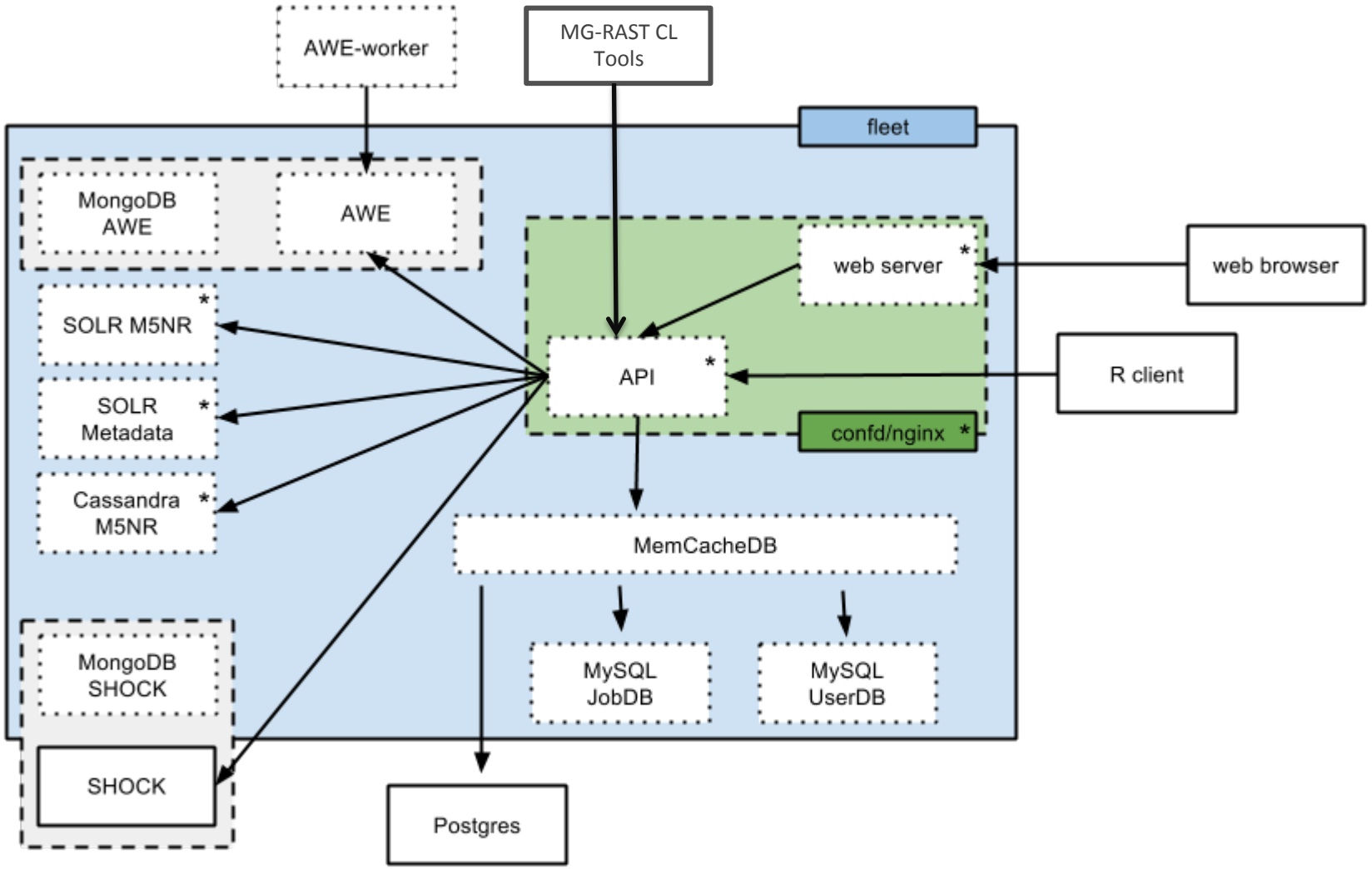
Folker Meyer

Argonne National Laboratory

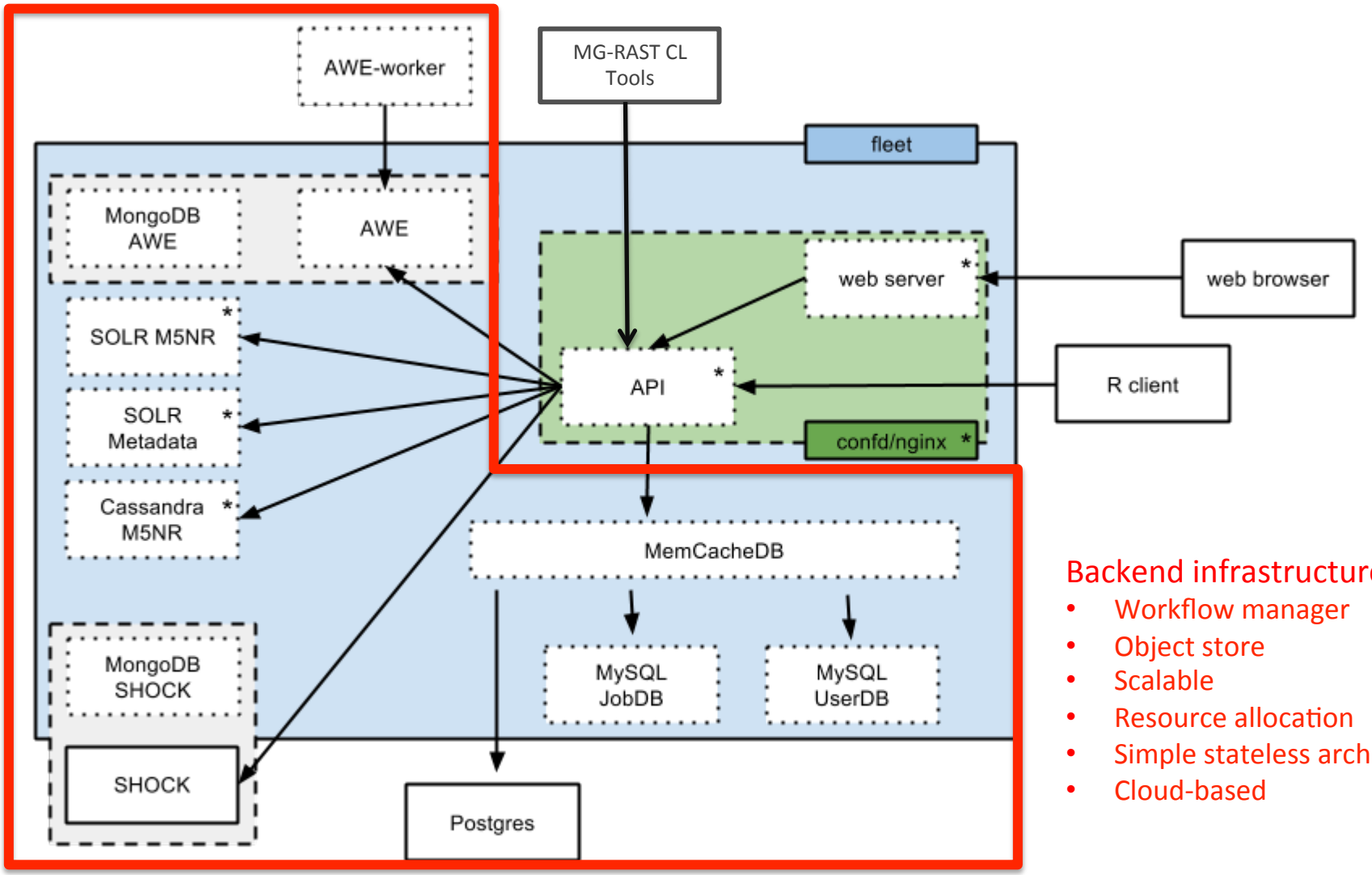
MG-RAST is

- 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

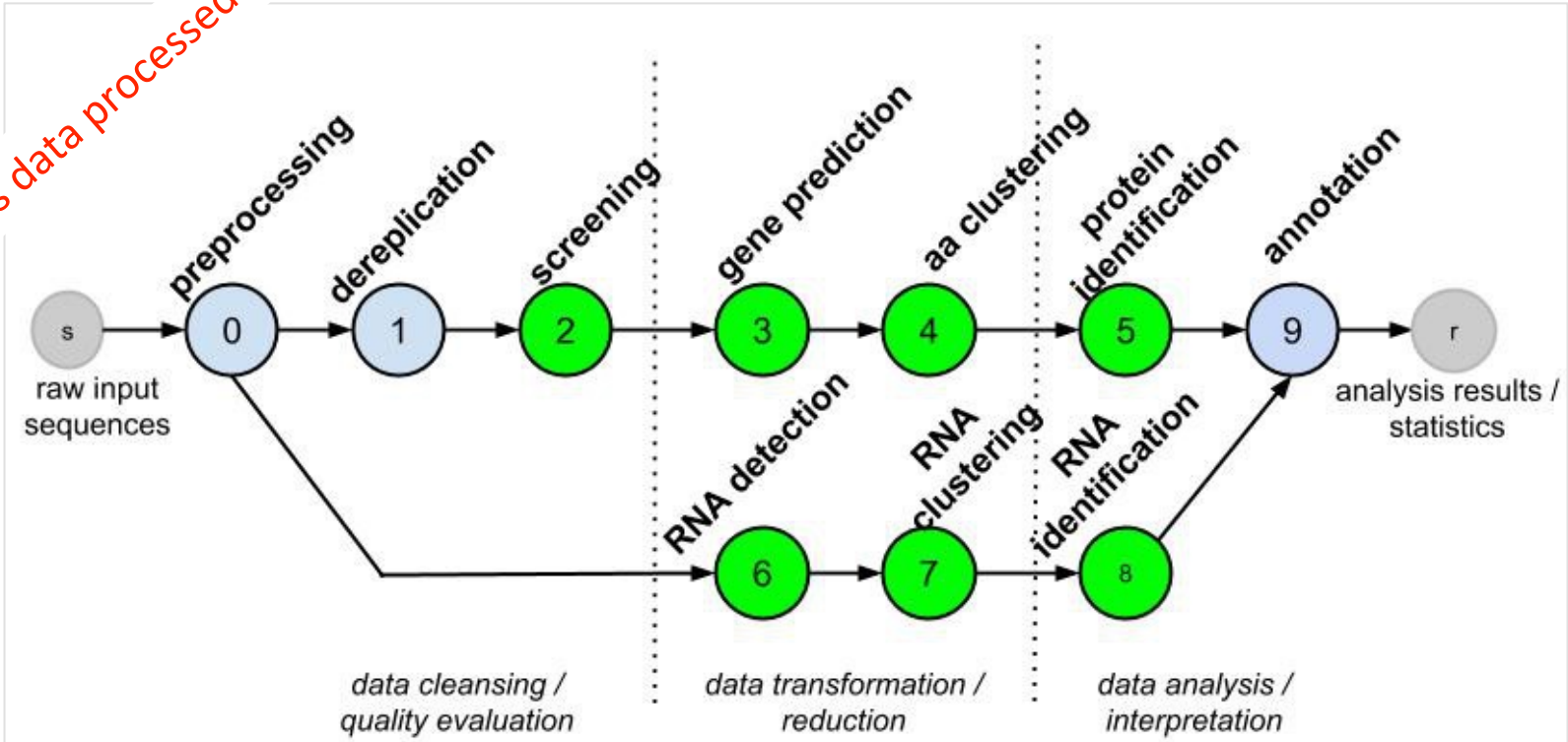


Backend infrastructure

- Workflow manager
- Object store
- Scalable
- Resource allocation
- Simple stateless architecture
- Cloud-based

How is data processed and annotated?

The Big Picture

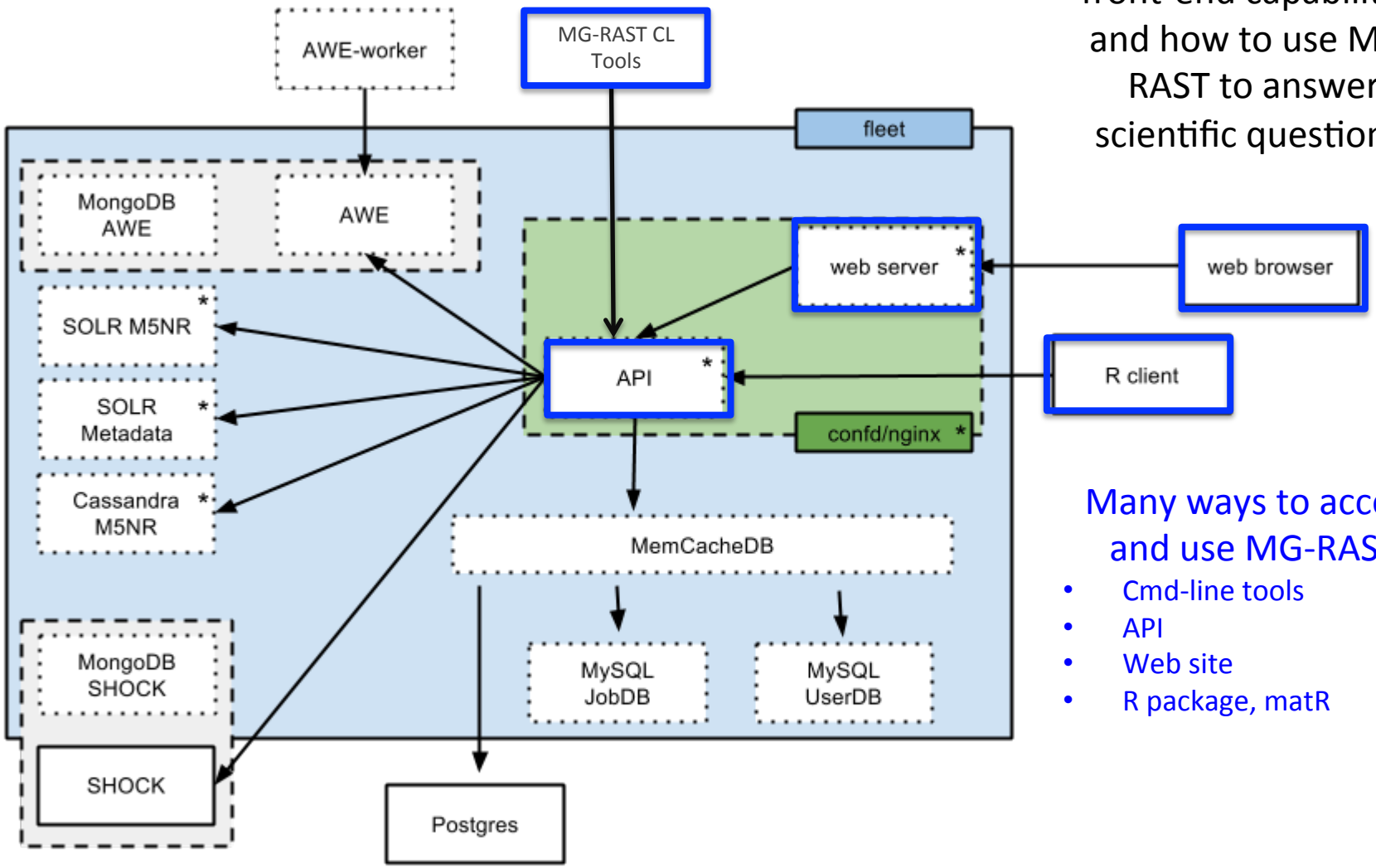


Data submitted to MG-RAST goes through our analysis pipeline.

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.

Many ways to access and use MG-RAST

- Cmd-line tools
- API
- Web site
- R package, matR

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, KEGG, Phantome
GO, NCBI, UniProt
JGI, SEED, 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	E0R0K8	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 exsiting 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

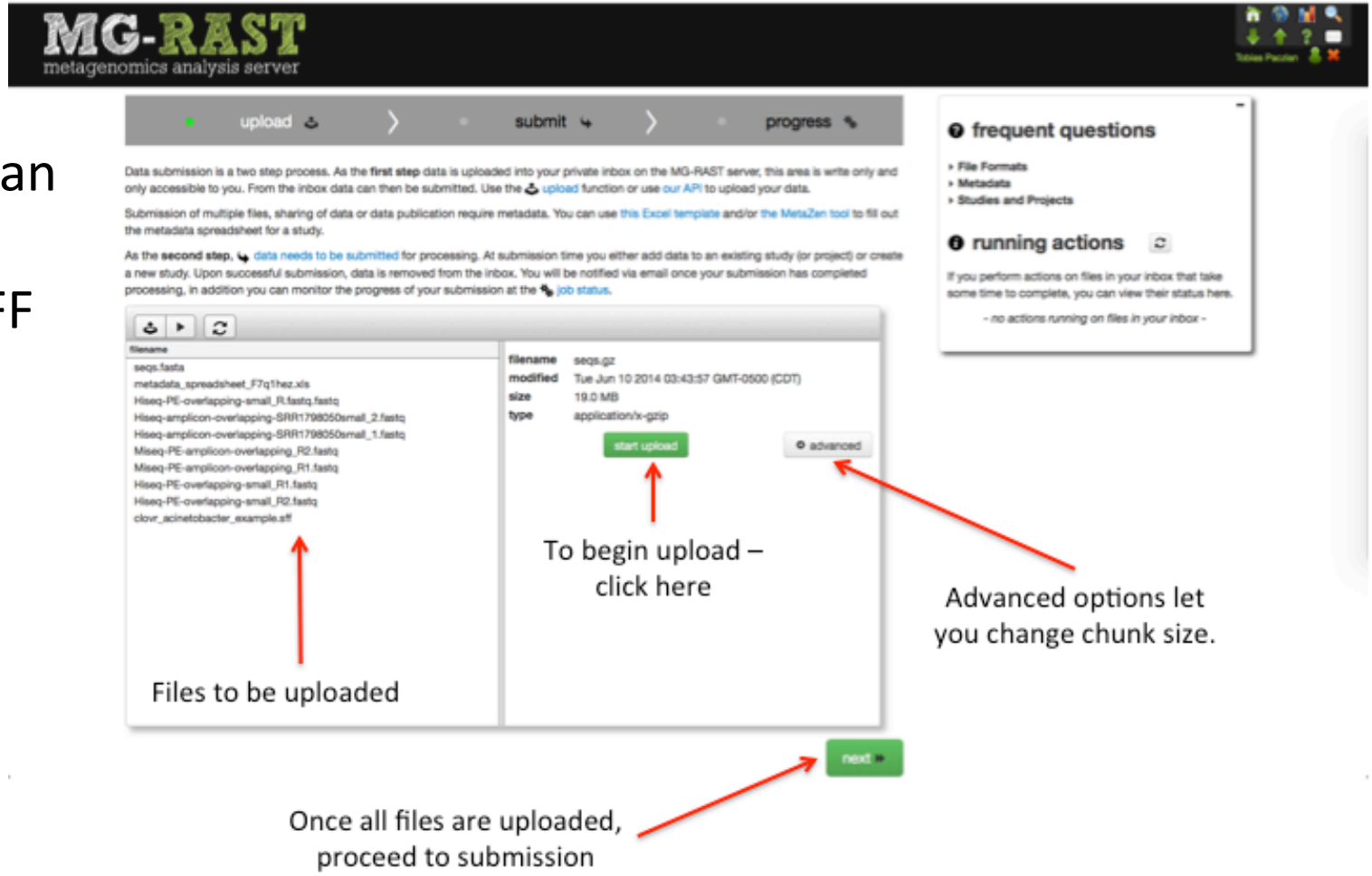
- Typically, users will come in with their own data and/or looking to analyze public data.
  - *You can use the data you just downloaded to try out the upload if you don't have data of your own.*
- In order to analyze your data – follow some easy steps to upload and submit data.
- Time needed to process 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

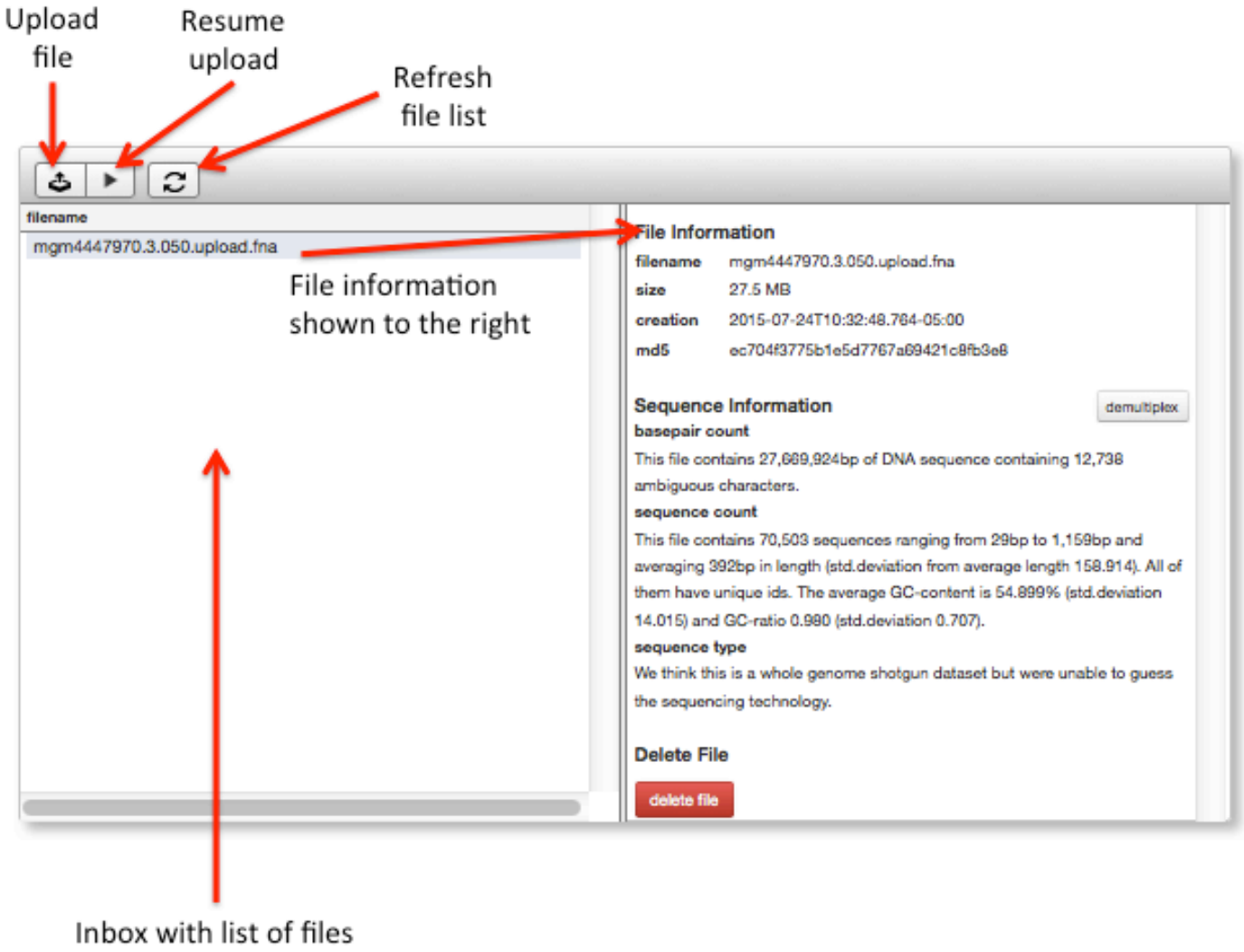
- Green “up” arrow takes you to upload

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

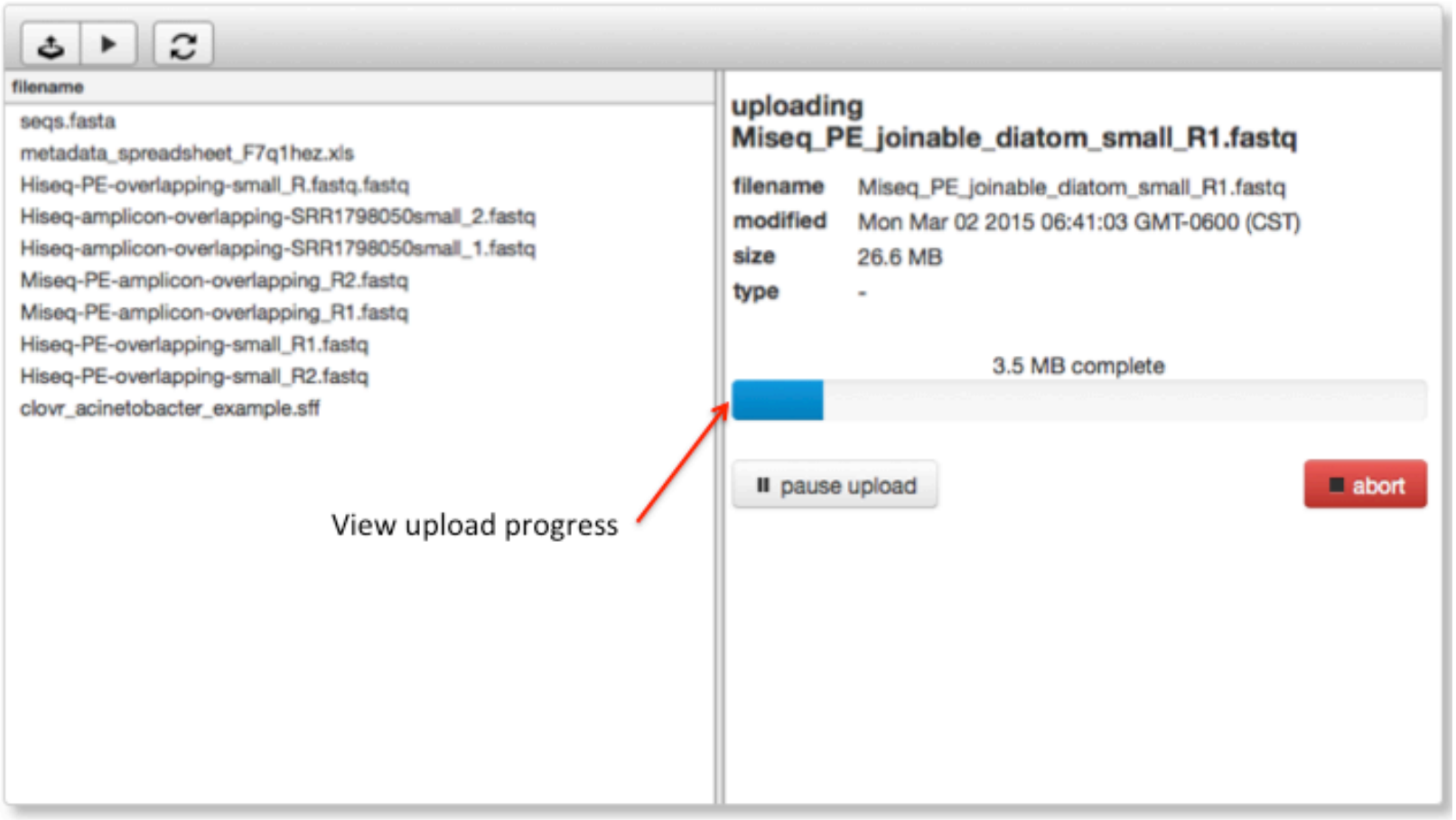
- Can be compressed files.



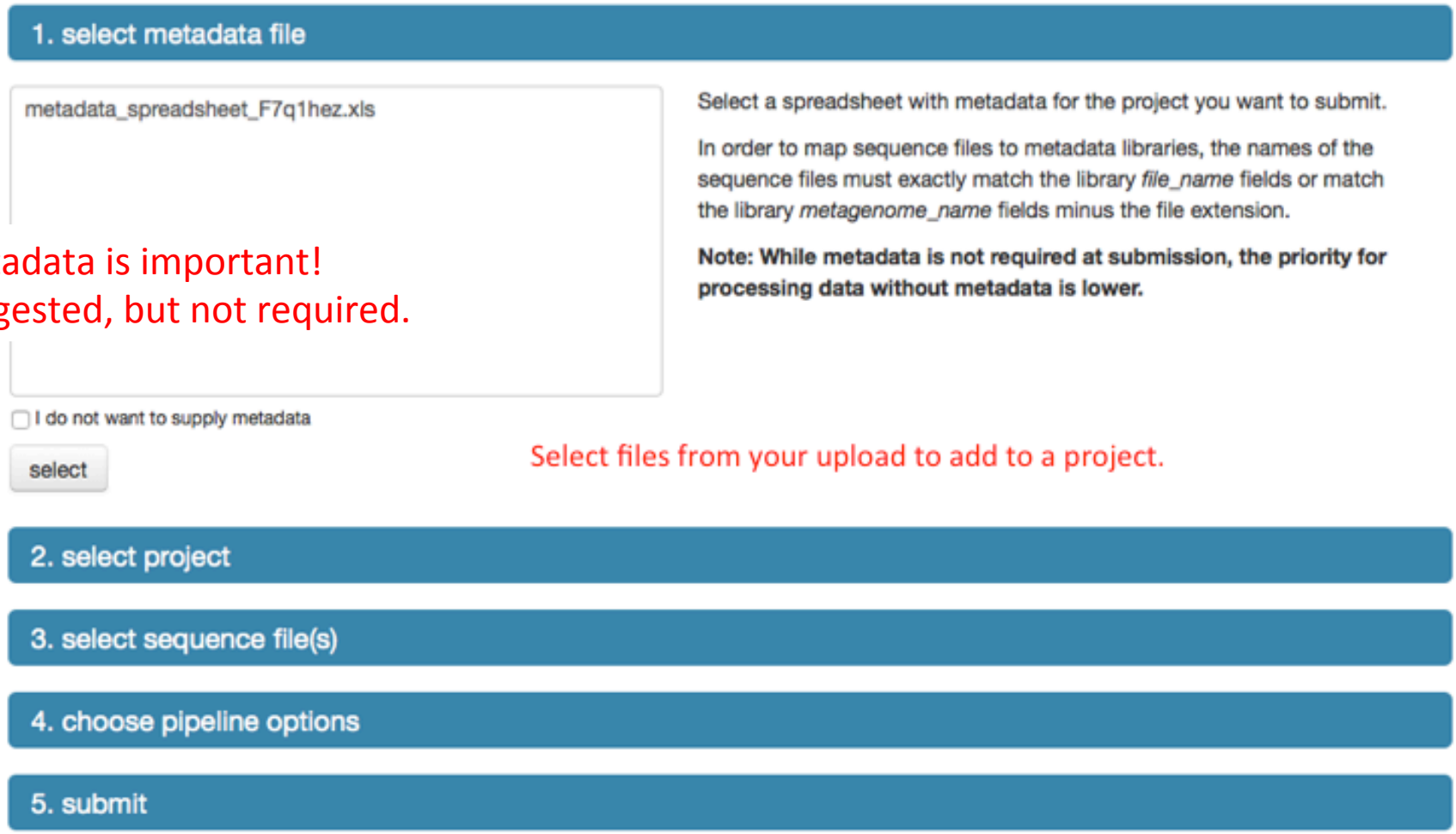
Elements of the file browser



Upload progress



Now to submit your data

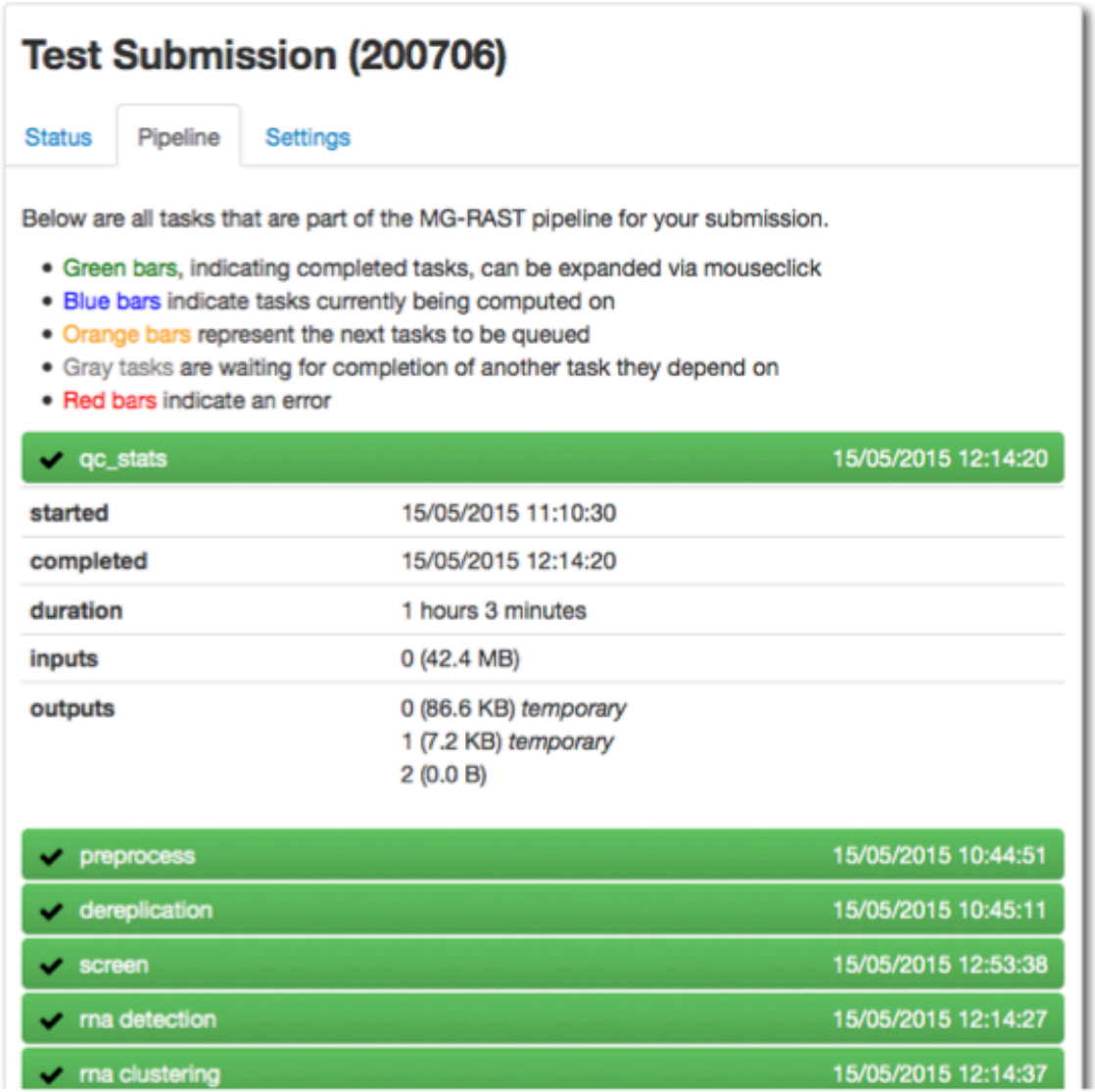


Metadata is important!
Suggested, but not required.

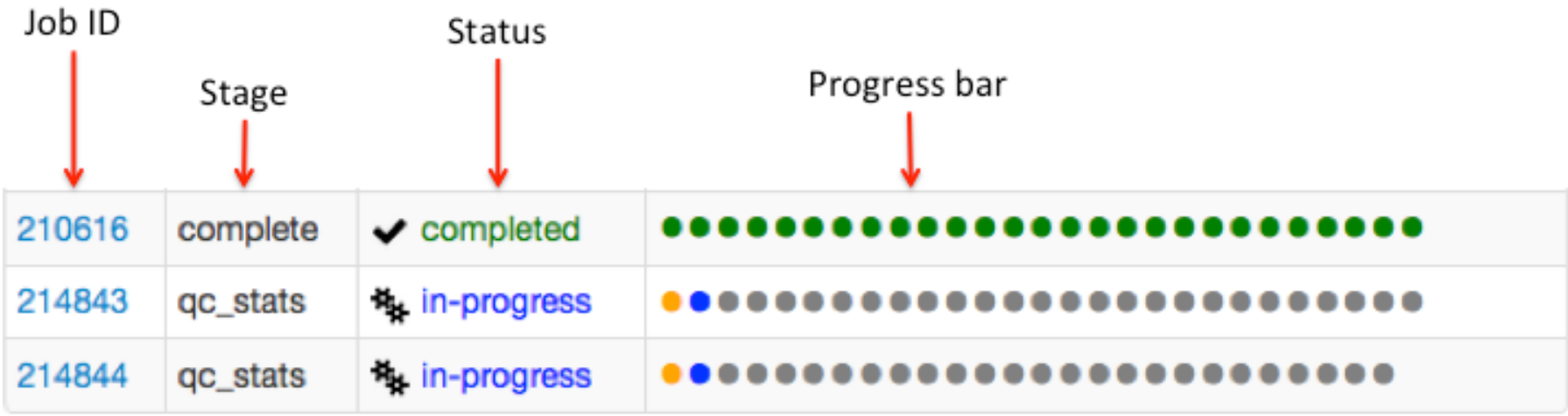
Select files from your upload to add to a project.

Submission is done …

when the completed sections turn green.



How far along has my data progressed?



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

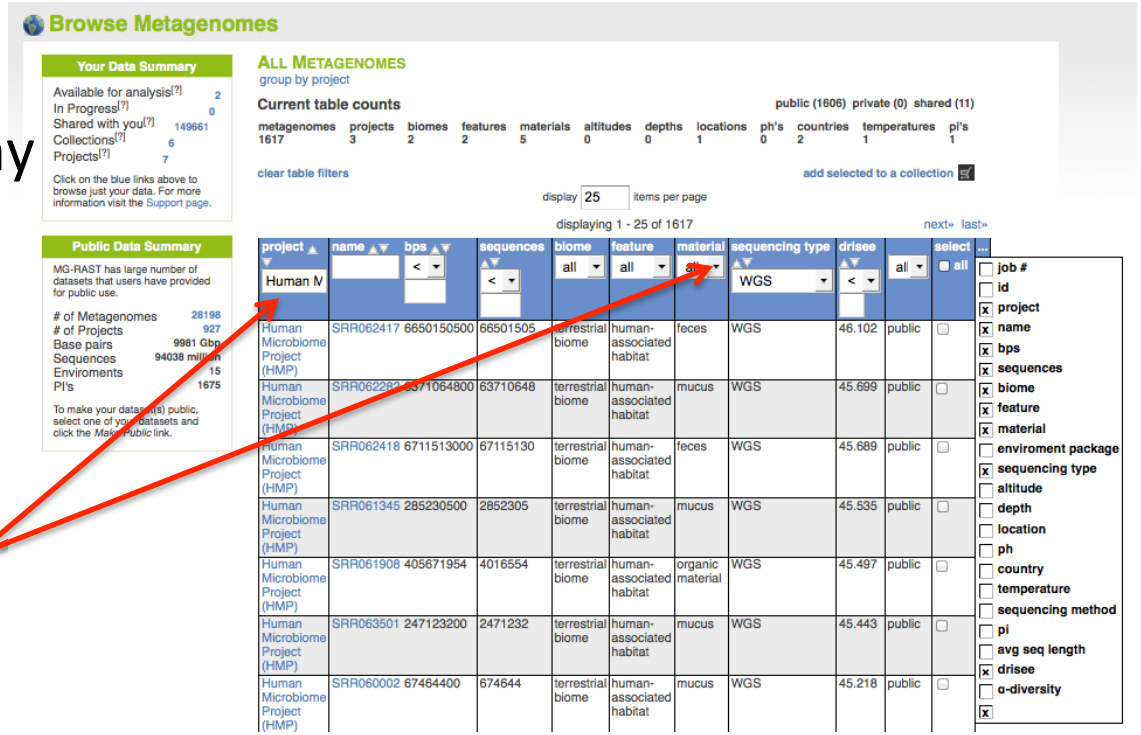
- Feces samples from HMP (Human Microbiome Project) consortium.
- Study citation:
  - Vital M, Howe AC, Tiedje JM. ***Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data***. MBio. 2014 Apr 22;5(2):e00889. PubMed PMID: 24757212
- They screened for butyrate synthesis pathways 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.

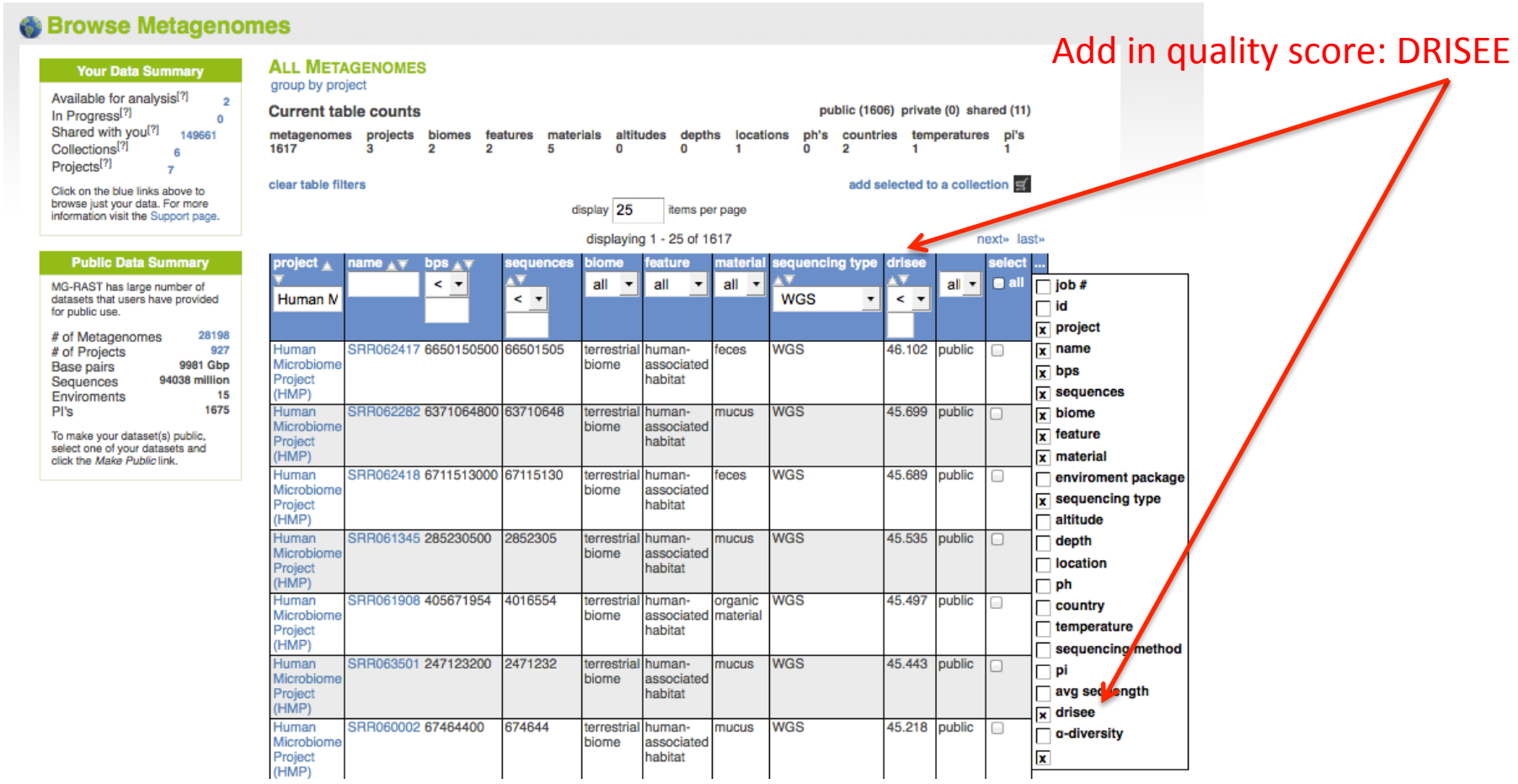
- Go to browse page and filter table by:
  - Project: Human Microbiome Project
  - Material: Feces
  - Disease state: Healthy

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.



Example 4: Data Quality - Examine Samples

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

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

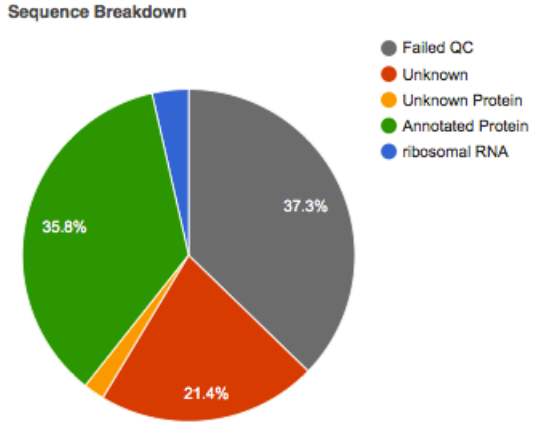
Note: A DRISEE score of zero means the metagenome did not meet standards for calculation.

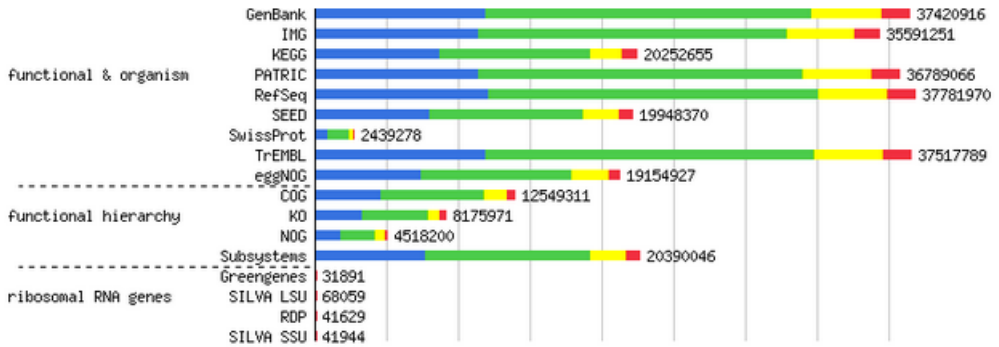
Clicking on the metagenome IDs takes you to the Overview page with a summary of the annotation and analysis.

Metagenome overviews

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

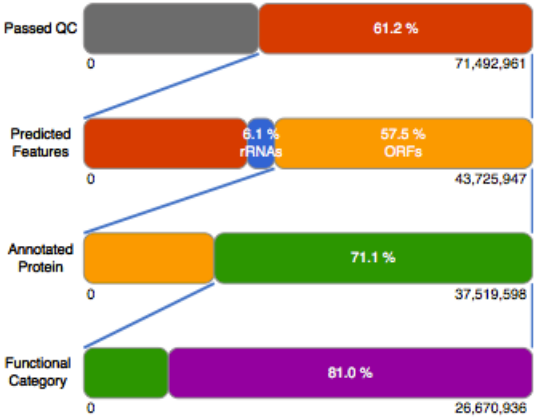
Example Sections of Overview



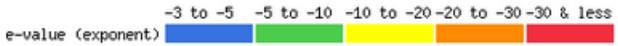


Analysis Flowchart

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.

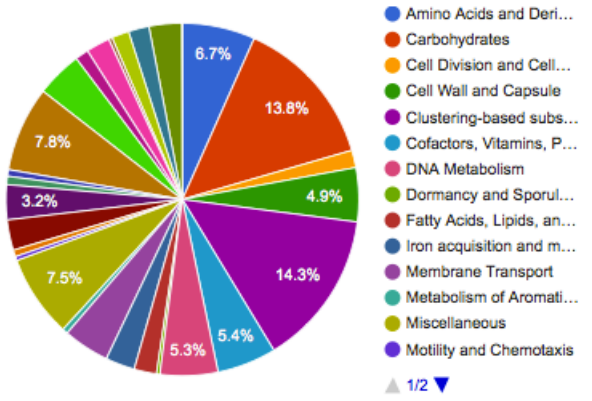


Analysis Statistics



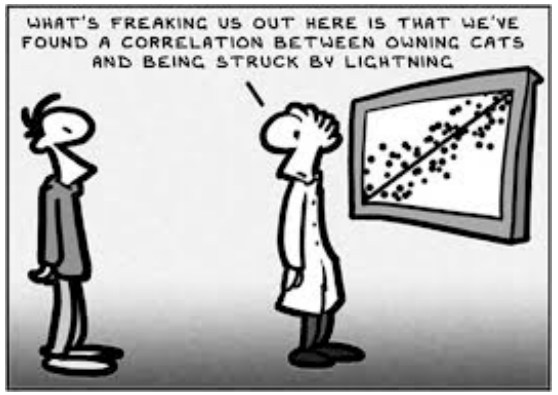
Upload: bp Count	**7,149,296,100 bp**
Upload: Sequences Count	**71,492,961**
Upload: Mean Sequence Length	**100 ± 0 bp**
Upload: Mean GC percent	**42 ± 9 %**
Artificial Duplicate Reads: Sequence Count	**23,940,065**
Post QC: bp Count	**4,131,507,015 bp**
Post QC: Sequences Count	**43,725,947**
Post QC: 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.



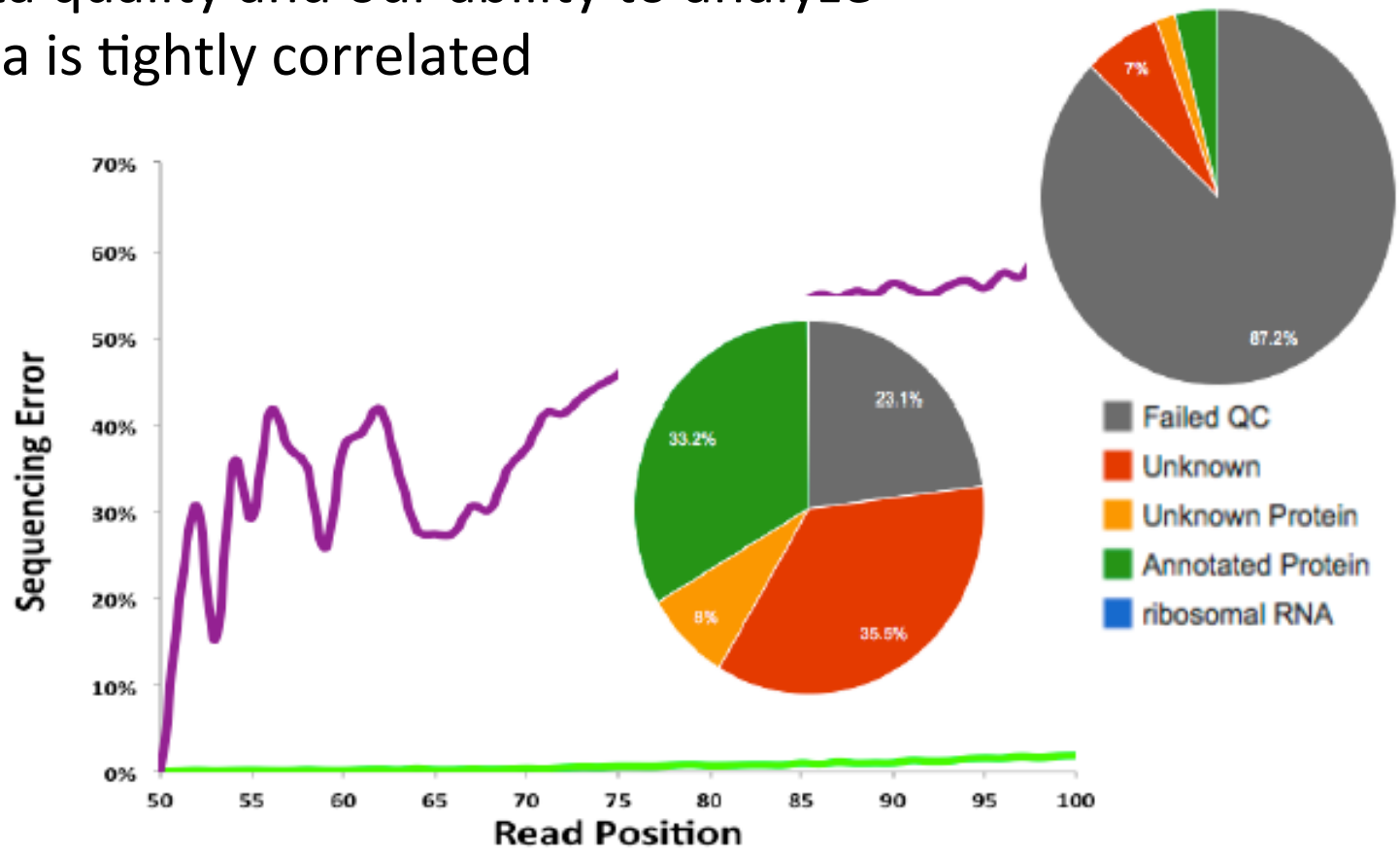
Why is it so important to know?

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

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

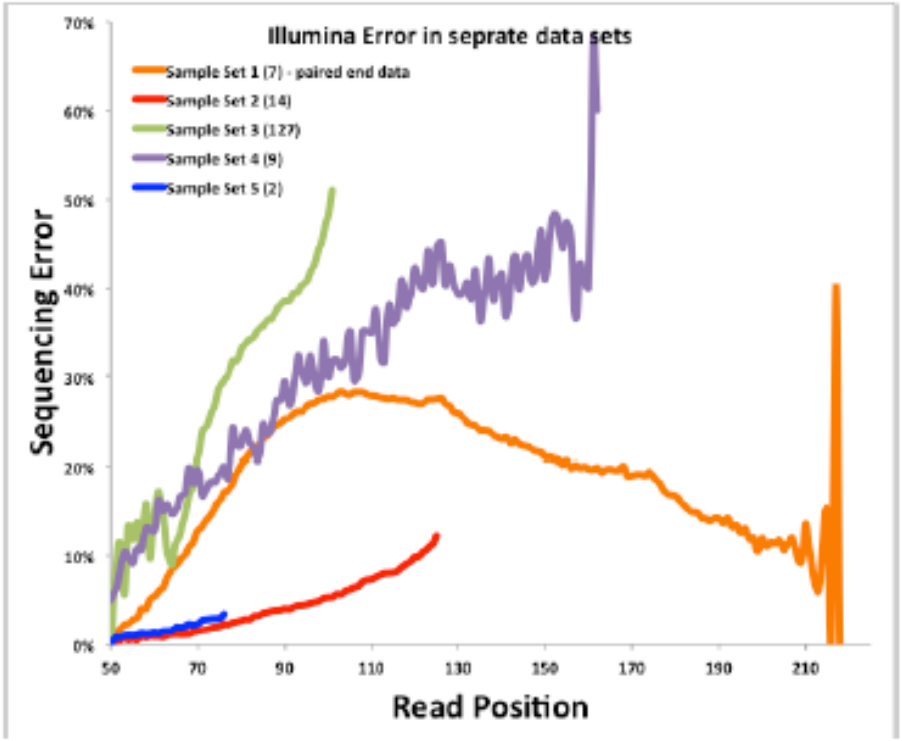
Data Quality Challenge

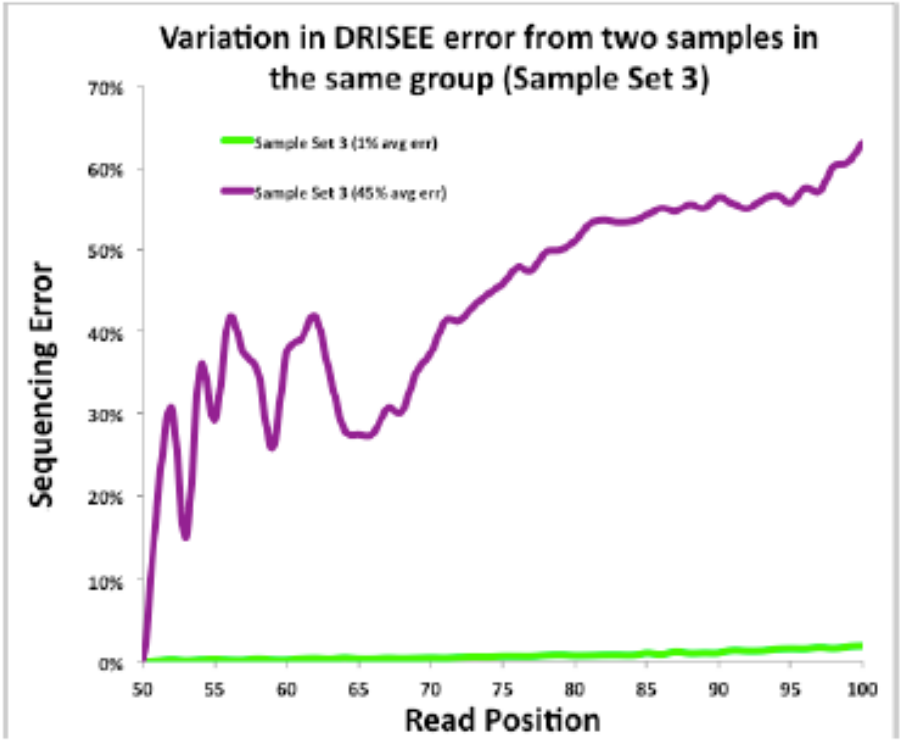
Data quality and our ability to analyze data is tightly correlated



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

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



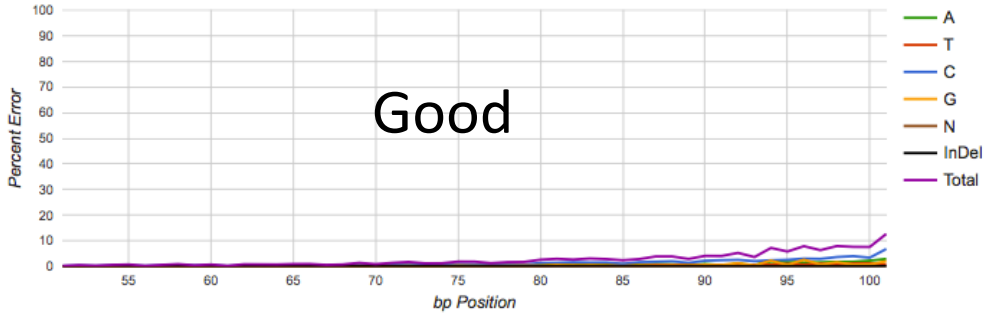


Extreme cases of quality

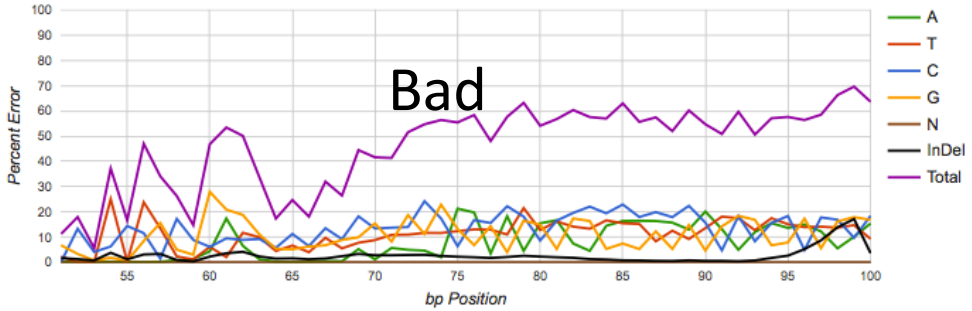
DRISEE [?] HIDE

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

Total DRISEE Error = 2.700 %

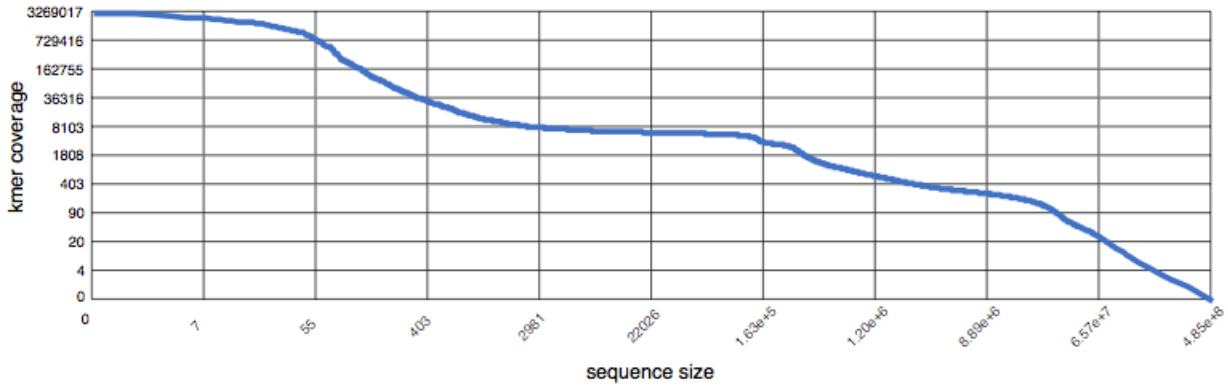


Total DRISEE Error = 46.102 %



Kmer profiles

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

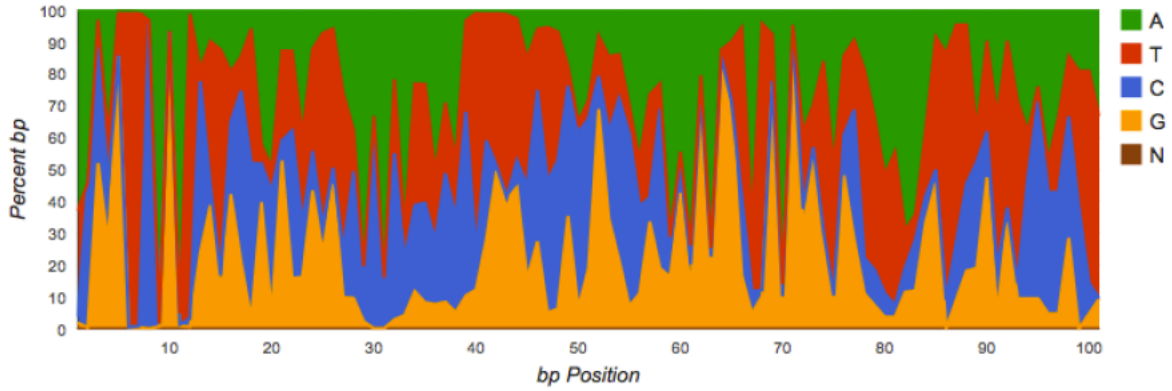


Answers the question “What is the coverage of the *n*th most-abundant kmer?”.

Nucleotide histogram

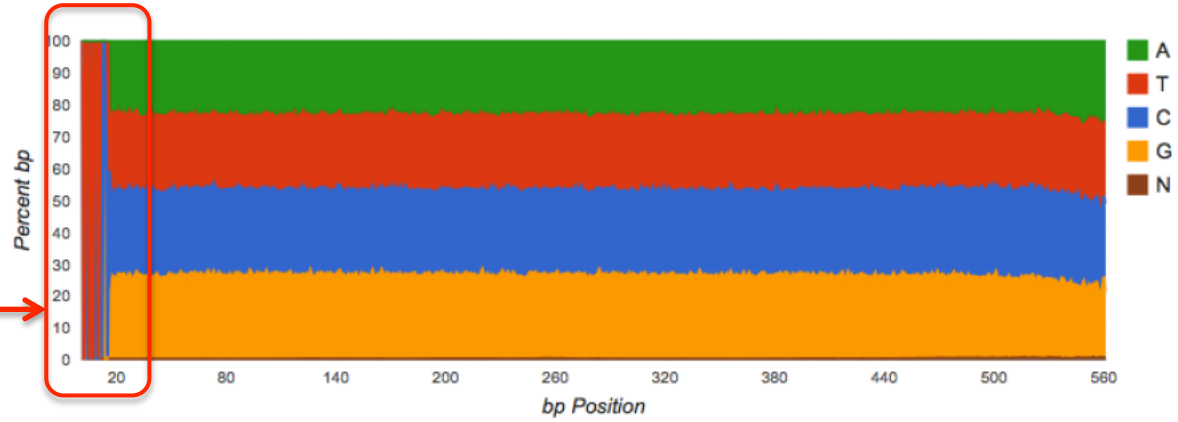
- Amplicon datasets should show biased distributions of bases at each position.

Reflects conservation and variability in the recovered sequences:



- WGS datasets should have roughly equal proportions of basecalls.

An example of untrimmed barcodes.

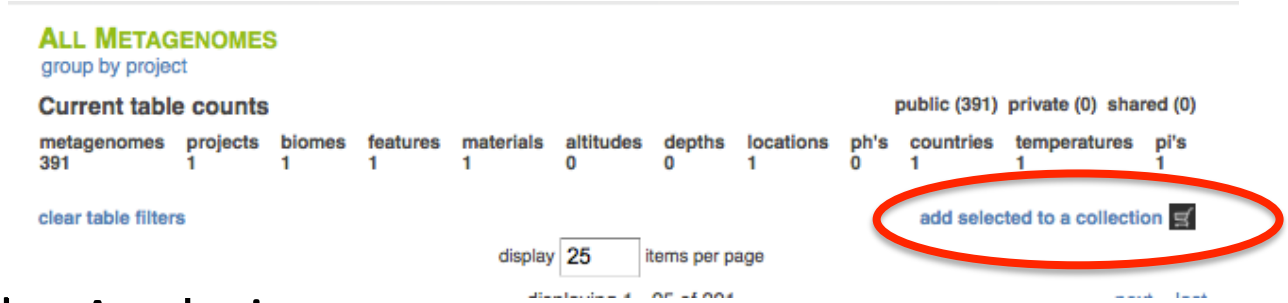


What do the samples look like?

- 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?

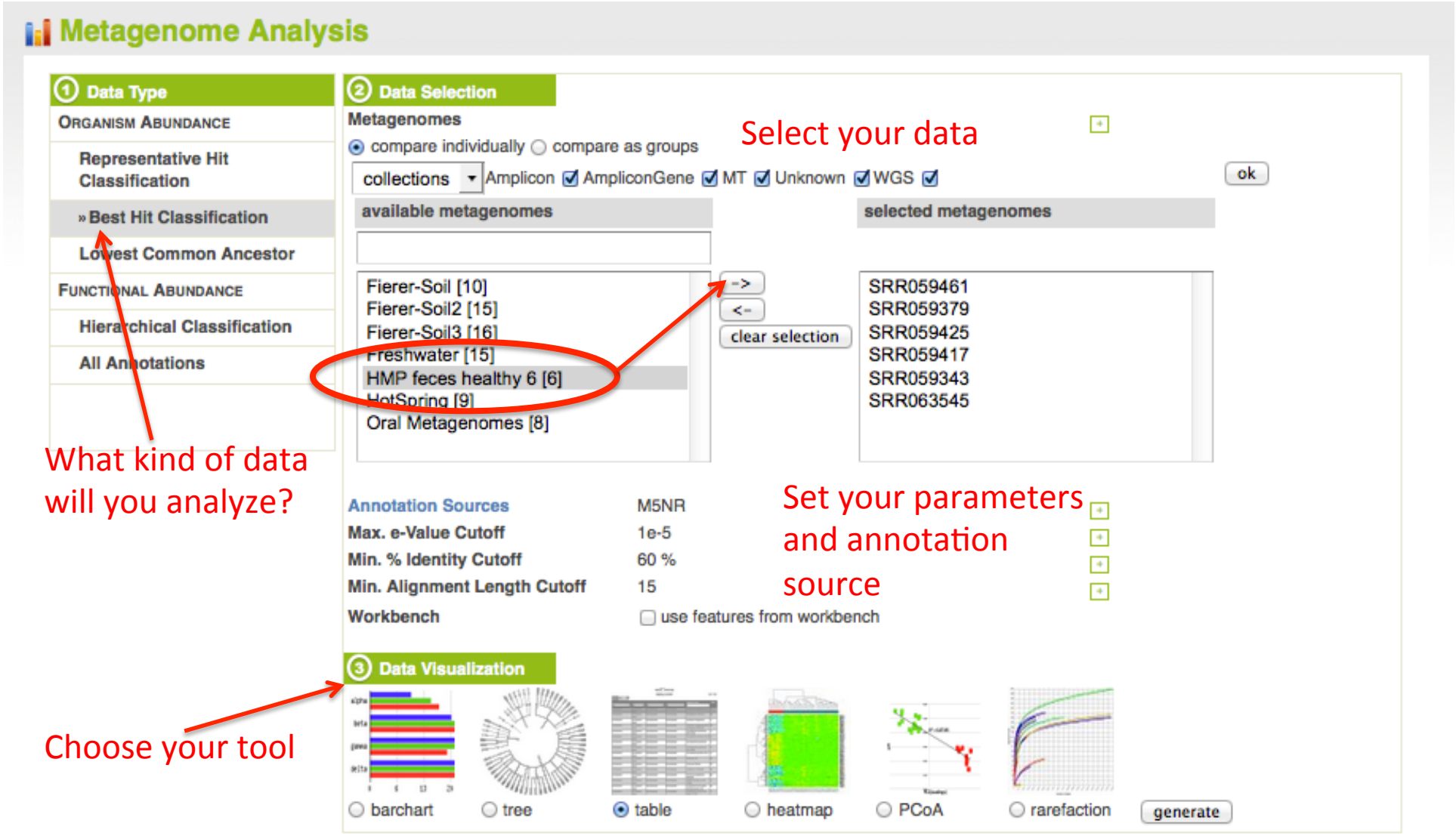
Example 5. How taxonomically diverse are your samples?

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

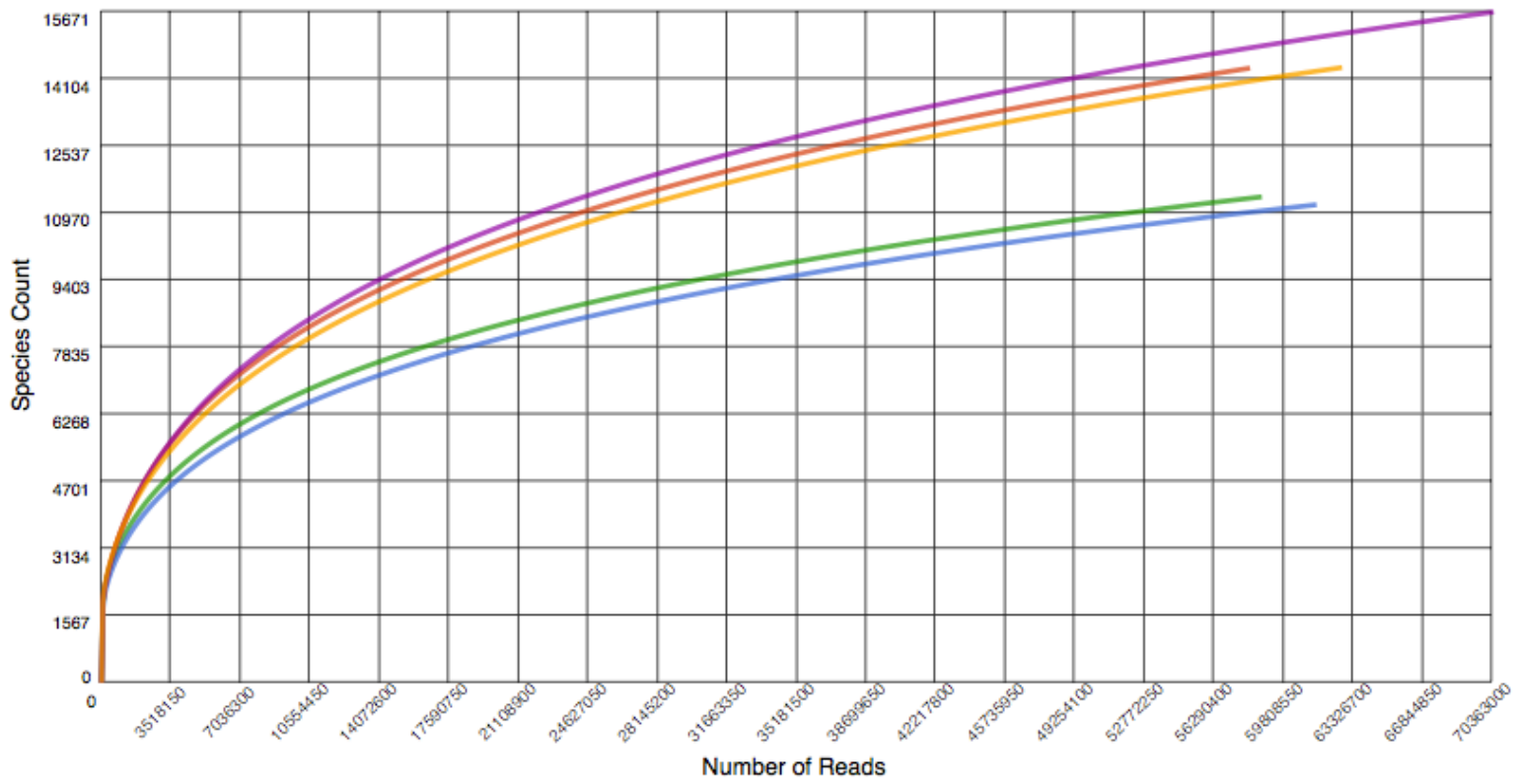


- Go to the Analysis page
  - Create rarefraction curve for your collection
  - Create heatmap
  - Create table
    - Select on order Clostridiaceae
    - Subselect and add to workbench
    - Oops – if you can't use MG-RAST web UI with that many features - move over to cmd-line!

Array of analyses

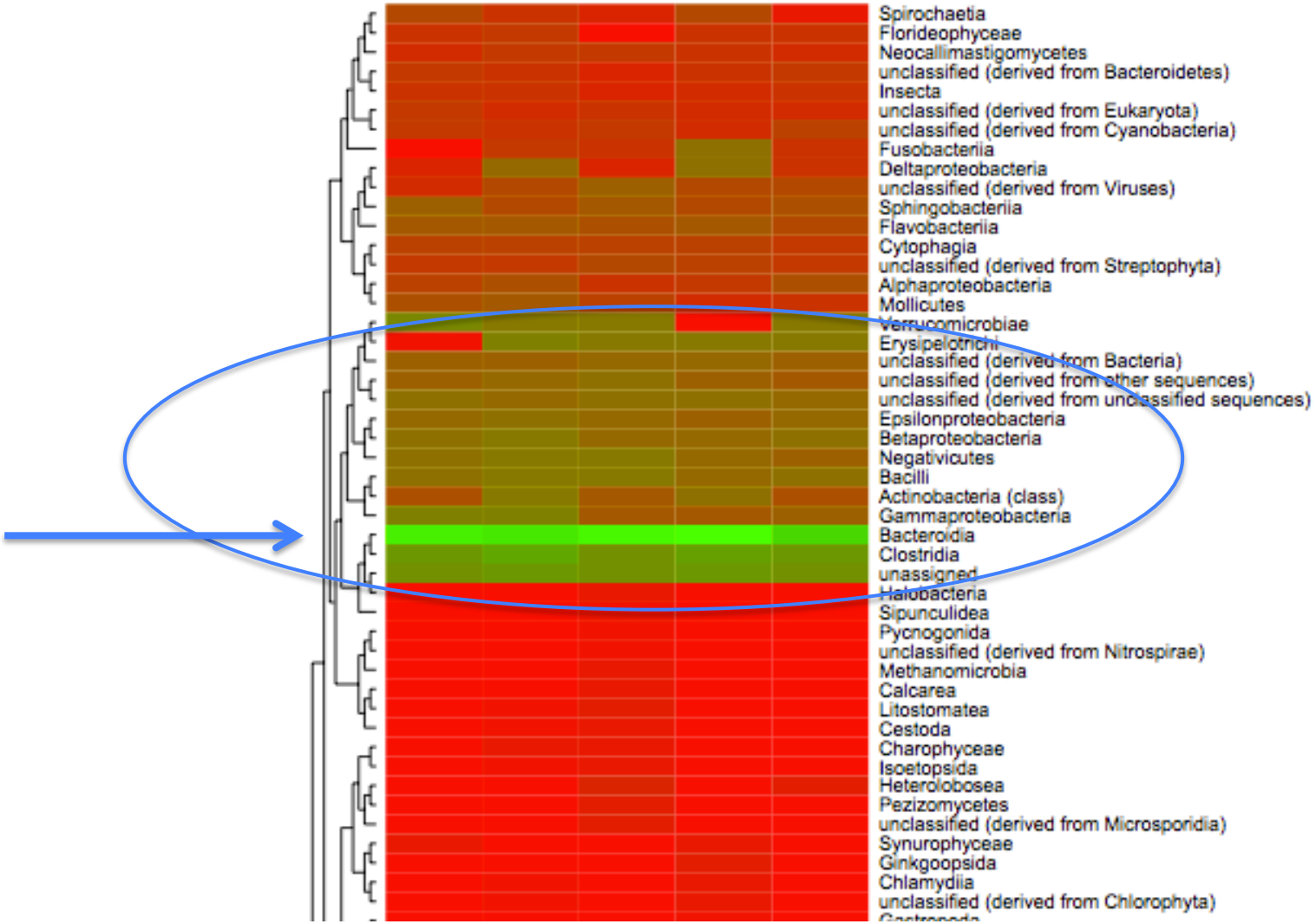


Sample diversity



rarefaction curve	metagenome	alpha diversity
	4472223.3	79.30
	4472165.3	80.61
	4472199.3	133.28
	4472191.3	161.97
	4472129.3	184.78

What taxa are in common?



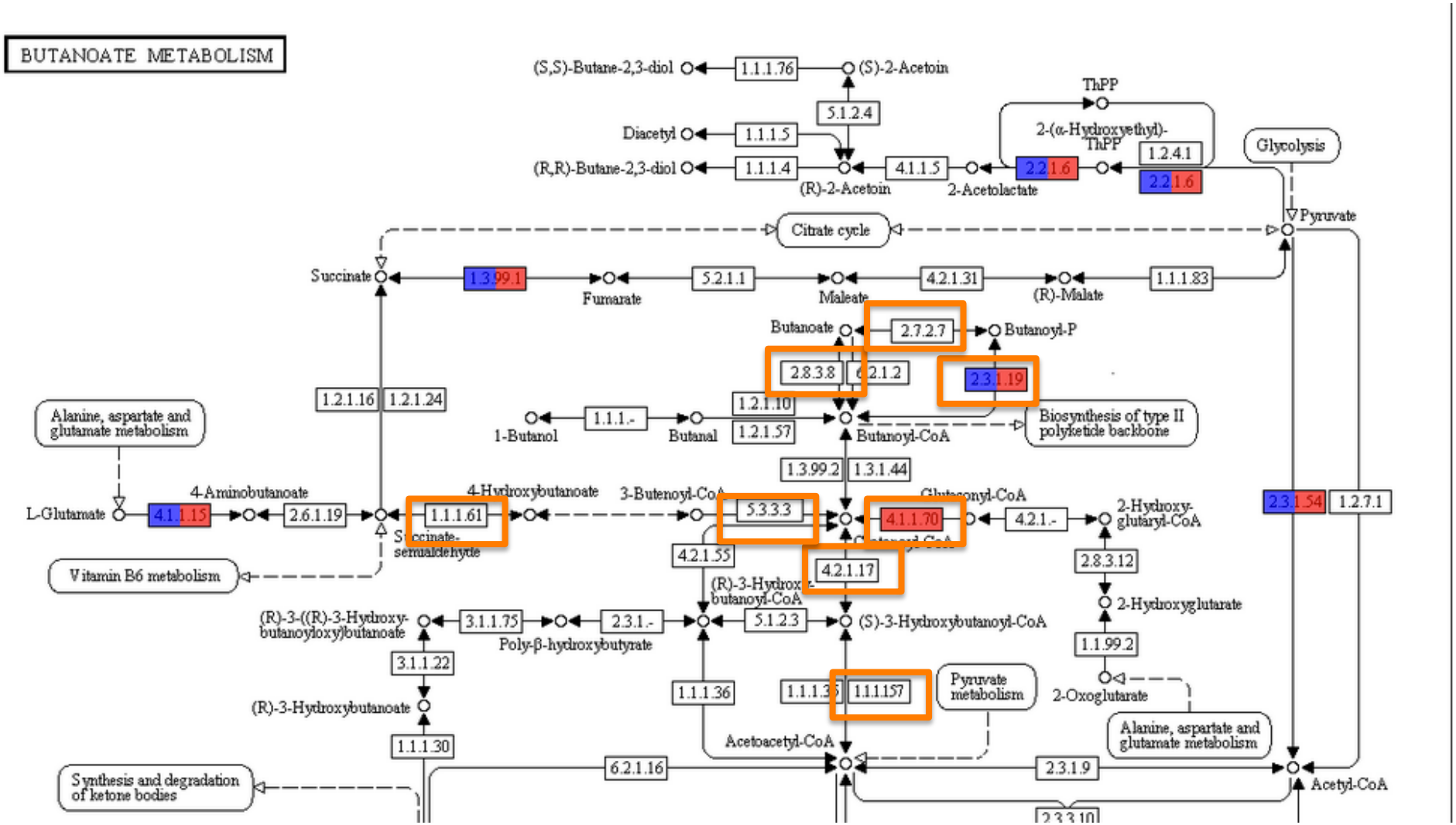
Are there major differences in the samples?

- 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?

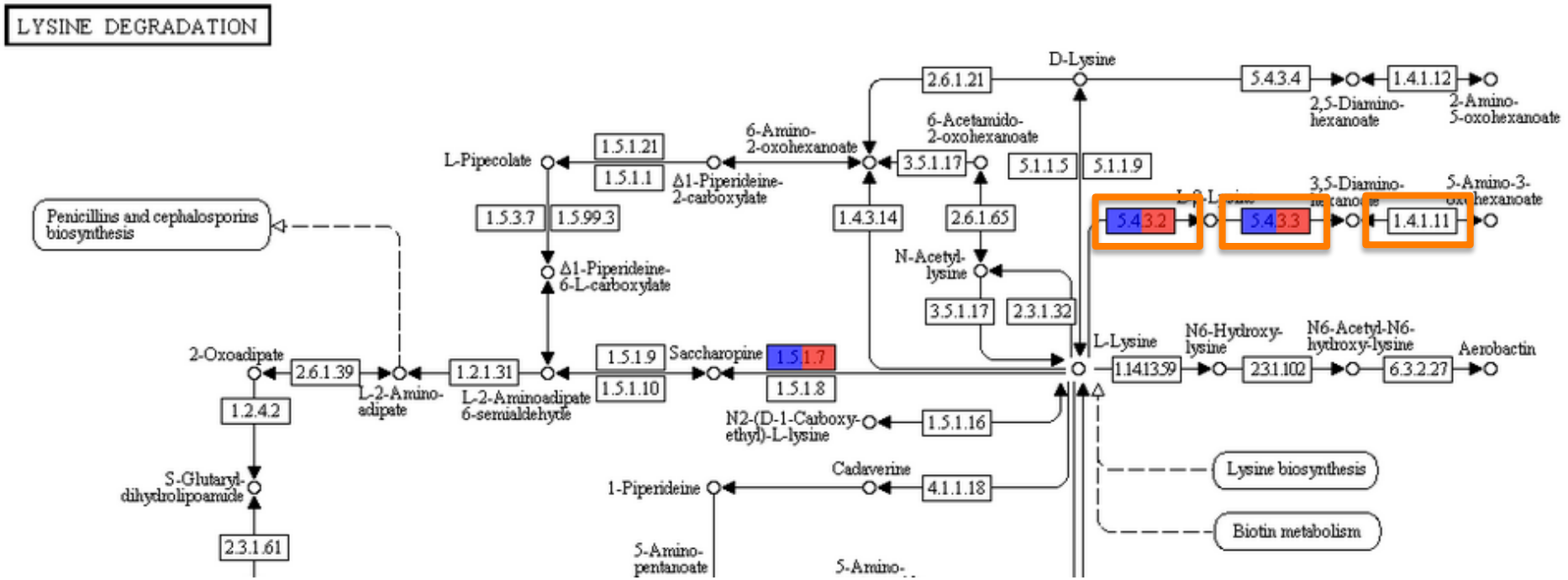
Example 6. Functional diversity

- 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 etal are found? (Check out Butanoate metabolism and Lysine degradation –KEGG)

What functions did you find?



What functions did you find?



Example 7. Digging deeper

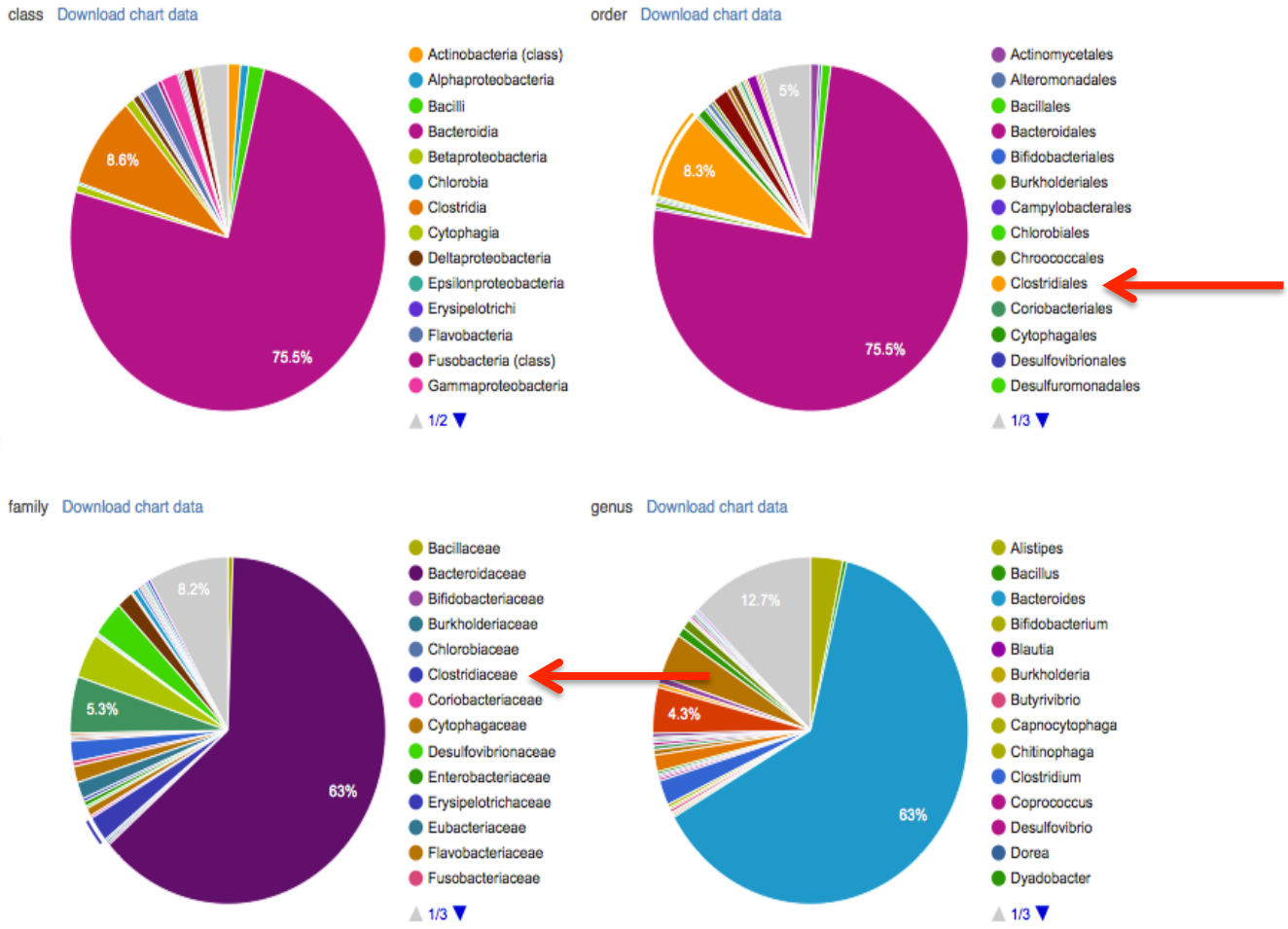
What butyrate synthesis genes are present in class Clostridia in your “best” sample for HMP healthy fecal sample?

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

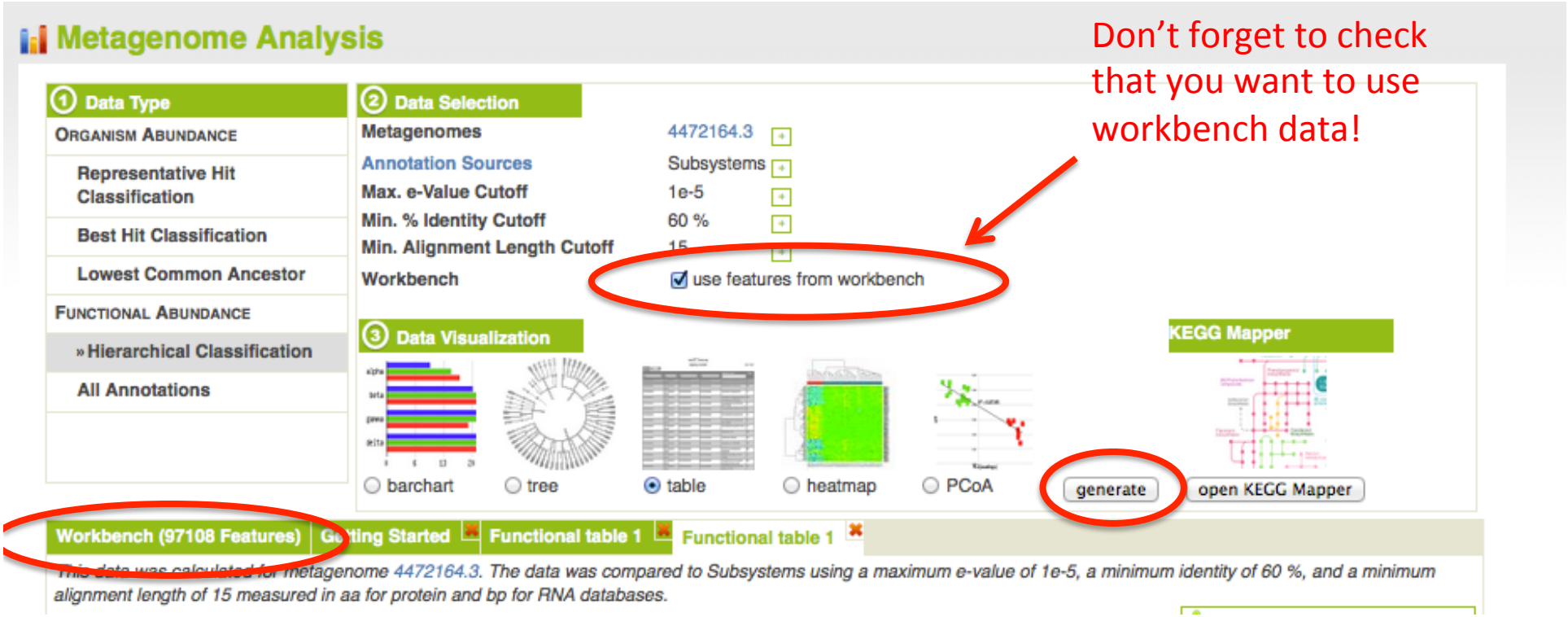
Select taxon group and move to workbench.

Metagenome 4472164.3
Family *Clostridiaceae*

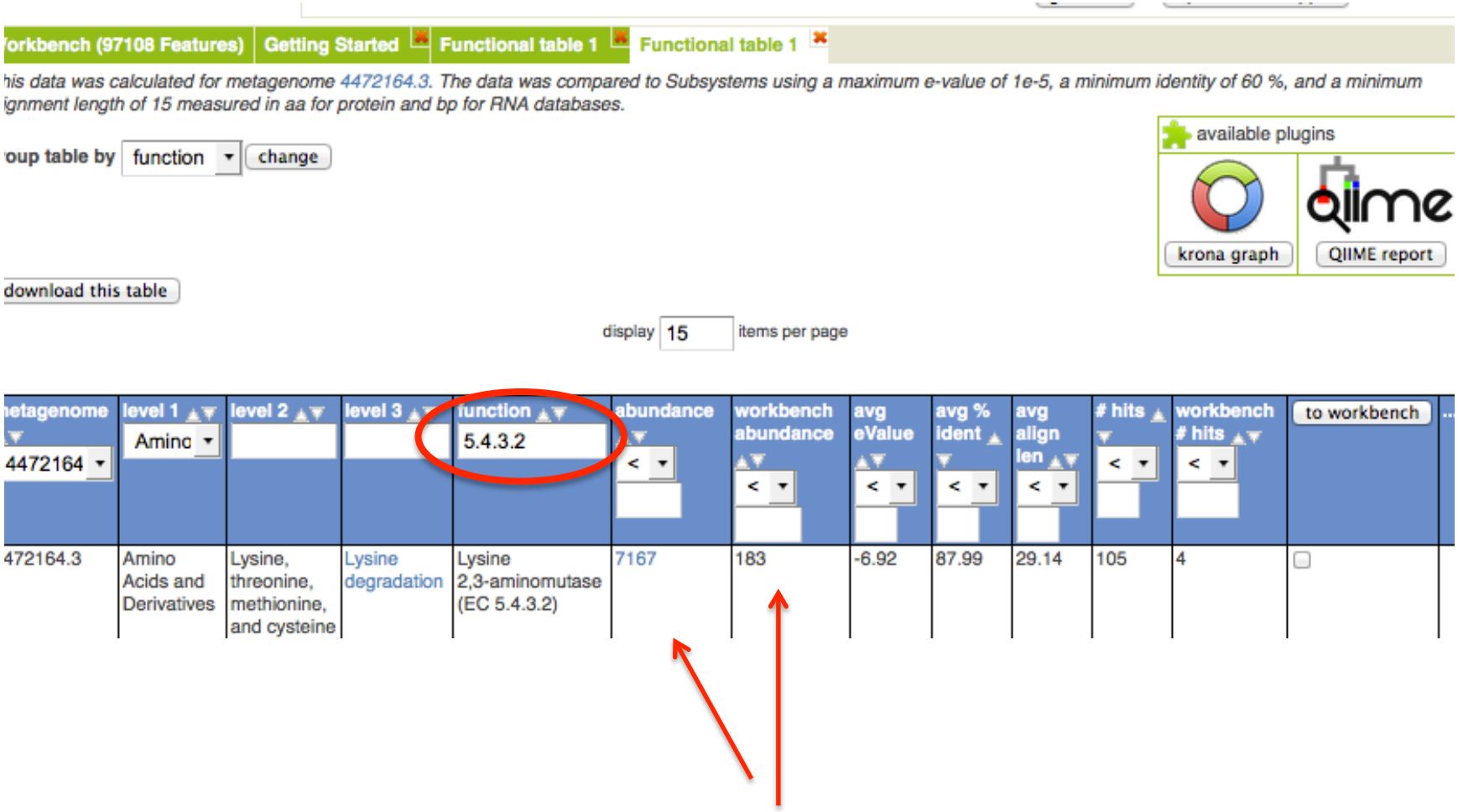


Moved featured to workbench!

Create Function Table



Search for function



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