MG-RAST

metagenomics analysis server

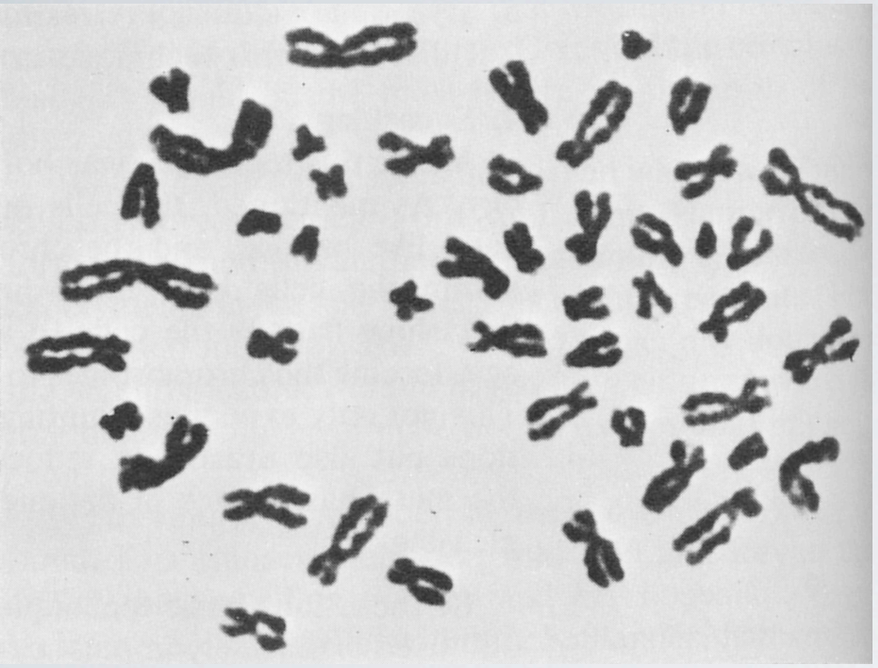
Sequences, MG-RAST, and you

W. Trimble
Argonne National Laboratory

Outline

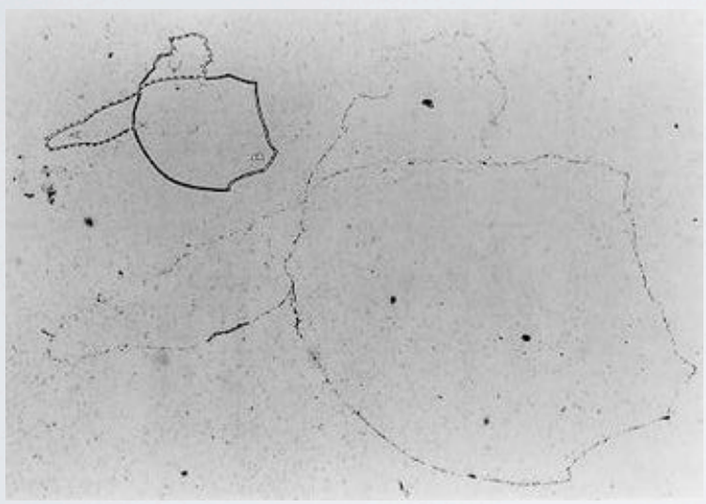
- Sequencing technology -> obligatory data deluge
- Computing technology -> EC2 and Docker
- What MG-RAST does
- How MG-RAST does it
- What MG-RAST does when you query it

How much information?



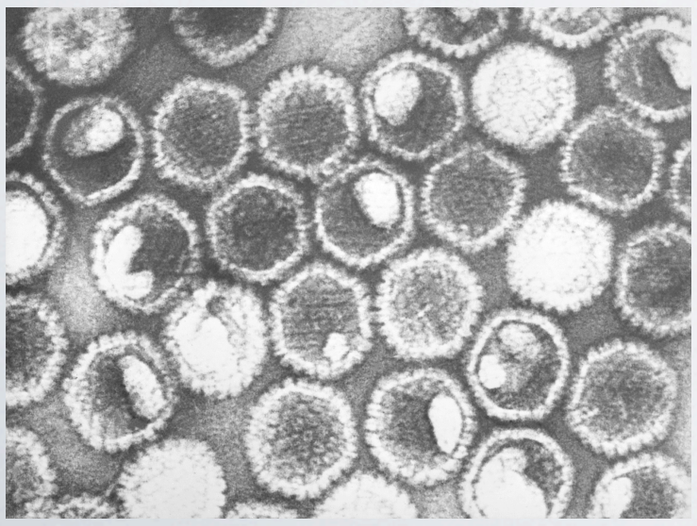
- Animals and plants:
  Two (or more) copies of genome
  10^8 – 10^{10} bp genome
  (Human: 2 copies x $3x10^9$)

BIG



- Microorganisms:
  One copy of a genome
  10^5 --10^7 bp

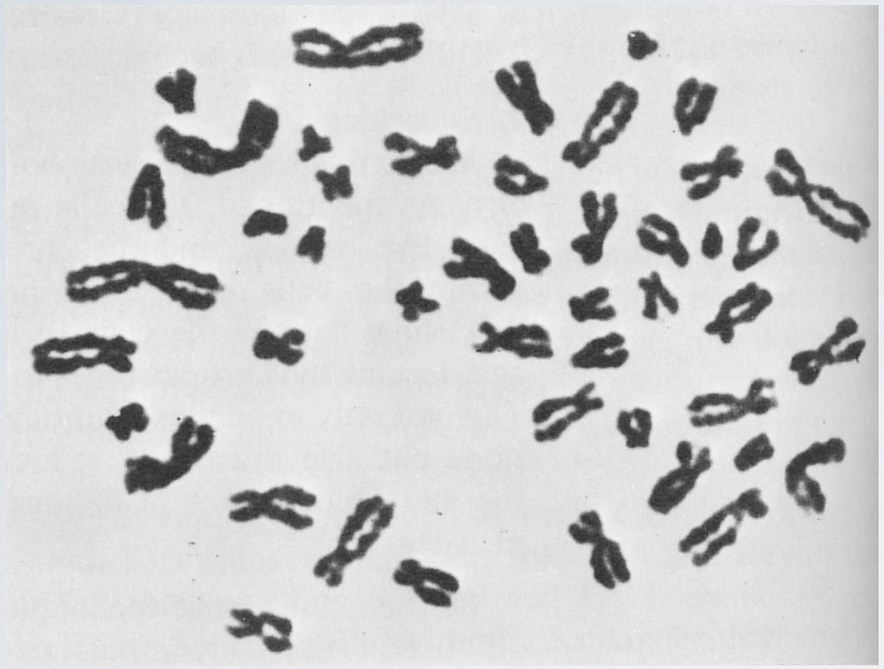
MED



- Viruses:
  One copy of genome
  10^4 --10^6 bp
  highly variable

SMALL

How much information?



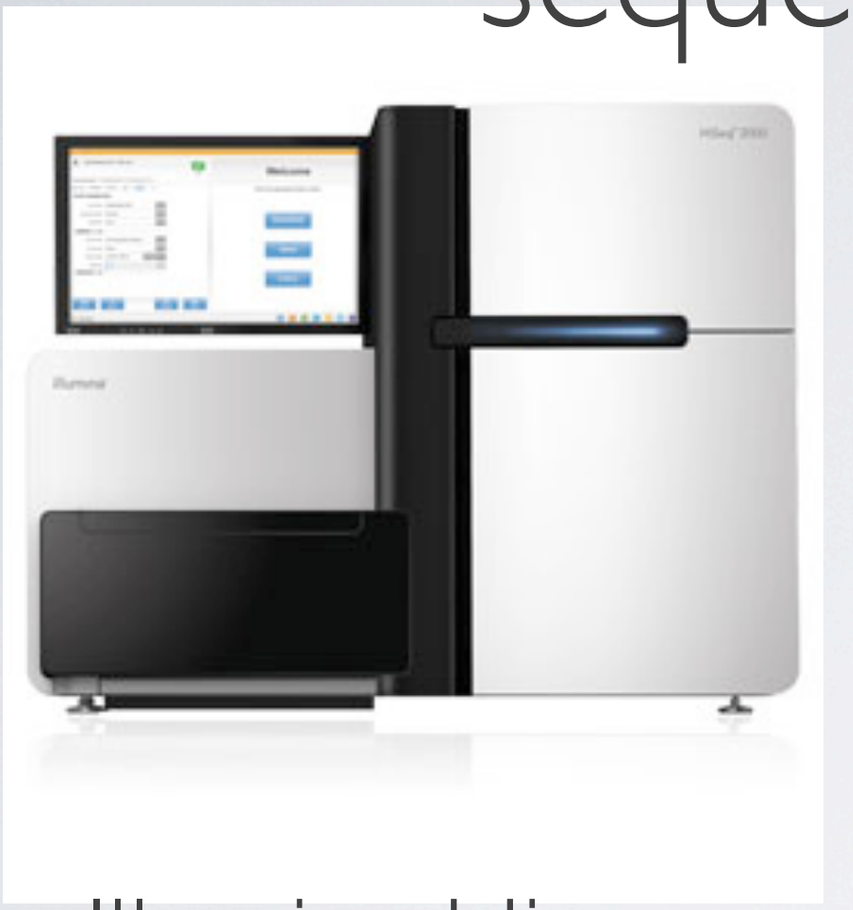
H. sapiens cancersurvivoris

2 copies x $3x10^9$ but we're 99% identical to Craig Venter.

So only a few 10^7 differences from the reference genome will capture most of what can be learned by sequencing

- Environmental samples: How much sequencing did you order? 10^{11} --10^{13} bp?

Market is dominated by synchronized sequencing-by-synthesis



Illumina Hiseq

600 Gbase/run

Illumina Miseq

1 Gbase/run

fast turnaround

for small projects

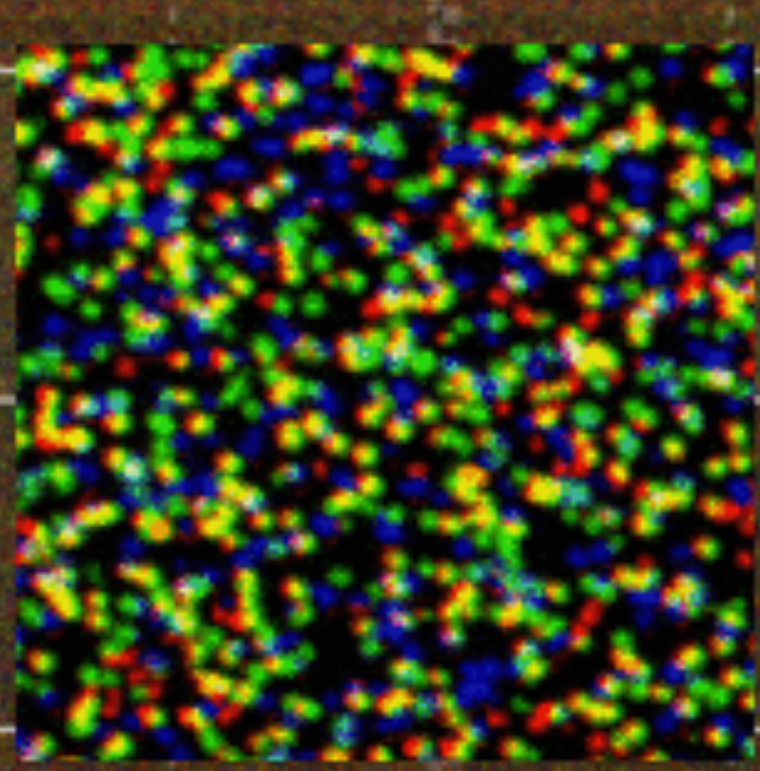
Illumina NextSeq

100 Gbase/run

optimized for

human

That produce
1 billion reads x 200 base apirs
at a time.



Biologists don't want…

Capacity

- Here's a new sequence. What is it?

```
>mysterysequence
CTAAGCACTTGTCTCCTGTTTACTCCCCTGAGCTTGAGGGGTTAACATGAAGGTCATCGATAGCAGGATAATAATACAGTA
```

- Run BLAST, right?

- How many times do you think you can do that?

Computing cost

- I have a sequencing run with 100 million sequences * 200 bp = 20 Gbases.

- 1 Mbase of sequence takes about 1 CPU-year to

  run BLASTX against the 2009 NCBI NR.

- So, running BLASTX to check out every sequence would only take 20,000 CPU-years = 175M hours.

Computing cost

- How much does your computer cost? The depreciation on my laptop is $1/day, so I'm looking at $0.04 / hour.

- Electricity: $0.01 / hour

- You can rent servers from Amazon for $0.10 / hour so my 20Gbase sequncing run will take $17M of compute time.

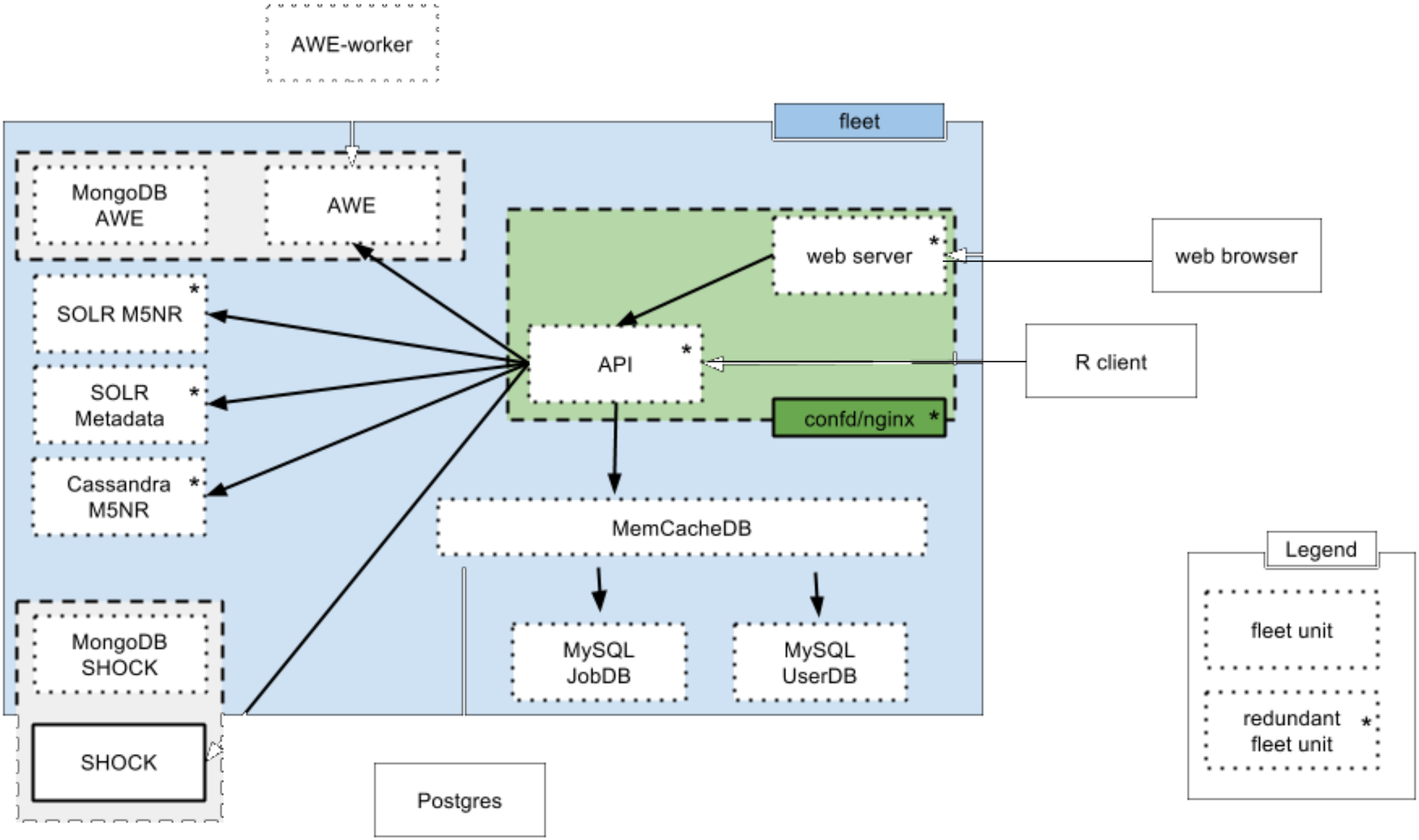
Cloud computing!

- Amazon, Microsoft, and Google rent computers by the hour.
- Pretty simple. You give them a credit card, click "start" and you get a server that you're paying for.
- https://aws.amazon.com/ec2/pricing/on-demand/
- Has anybody ever had a grumpy sysadmin who won't install the newest, greatest data analysis package for you?

amazon web services™

Cloud computing!

- Cloud computing is merely computing by the hour.
- Has anybody ever had a grumpy sysadmin who won't install the newest, greatest data analysis package for you?
- Has anybody ever had so much data that to get it all processed would take (insert time comparison that is way too long here) ?


AWE-worker
fleet
MongoDB
AWE
AWE
web server
*
web browser
SOLR M5NR
*
API
*
R client
SOLR
Metadata
*
confd/nginx
*
Cassandra
M5NR
*
MemCacheDB
MongoDB
SHOCK
MySQL
JobDB
MySQL
UserDB
SHOCK
Postgres
Legend
fleet unit
redundant
fleet unit
*

User Interface

Pipeline

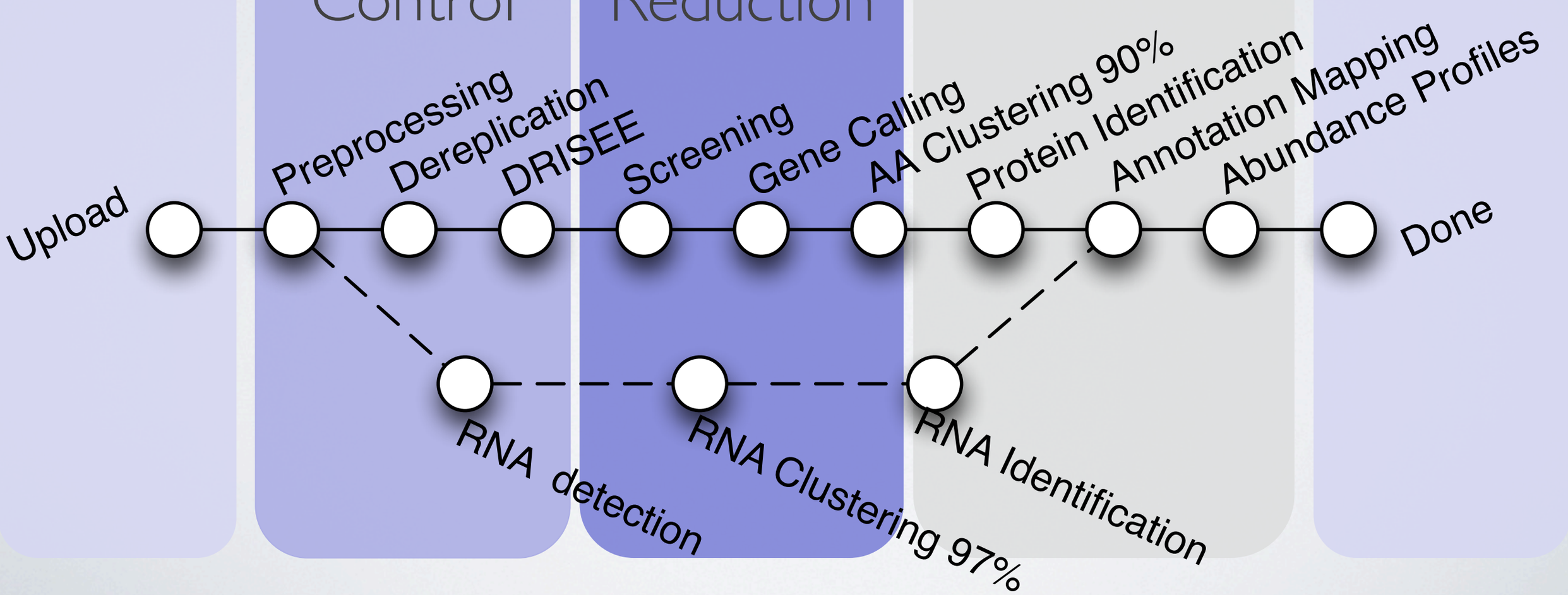
Input

Quality Control

Data Reduction

Analysis

Output



Pipeline

Quality Control

Data Reduction

Analysis

Upload

Preprocessing

Dereplication

DRISEE

Screening

Gene Calling

AA Clustering 90%

Protein Identification

Annotation Mapping

Abundance Profiles

Done

RNA detection

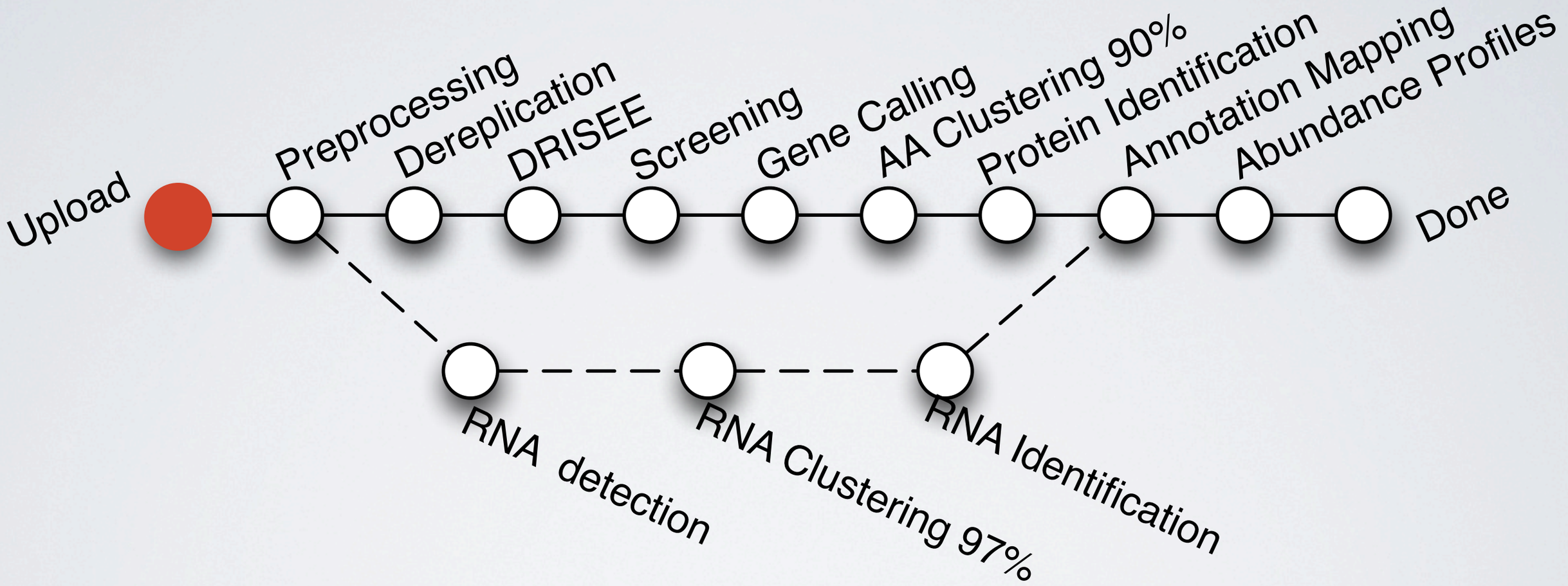
RNA Clustering 97%

RNA Identification

Background

- Modern datasets have 10s of millions of sequences.
- **BLAST is too slow** to run 10 million times against databases of 10s of millions of proteins !!!
- Clustering + faster searching tools (BLAT) allow annotation with reasonable computing resources
- Web interface allows comparisons between annotated datasets.

Pipeline


Upload
Preprocessing
Dereplication
DRISEE
Screening
Gene Calling
AA Clustering 90%
Protein Identification
Annotation Mapping
Abundance Profiles
Done
RNA detection
RNA Clustering 97%
RNA Identification

Background: Sequence formats

FASTA

>sequence_ID | (any other text)
GACTGATCATCTACTAGTCGATGC…
>sequence_ID2
GAGTCAGTCGATCATCGTAGCTAGCTAGCT…

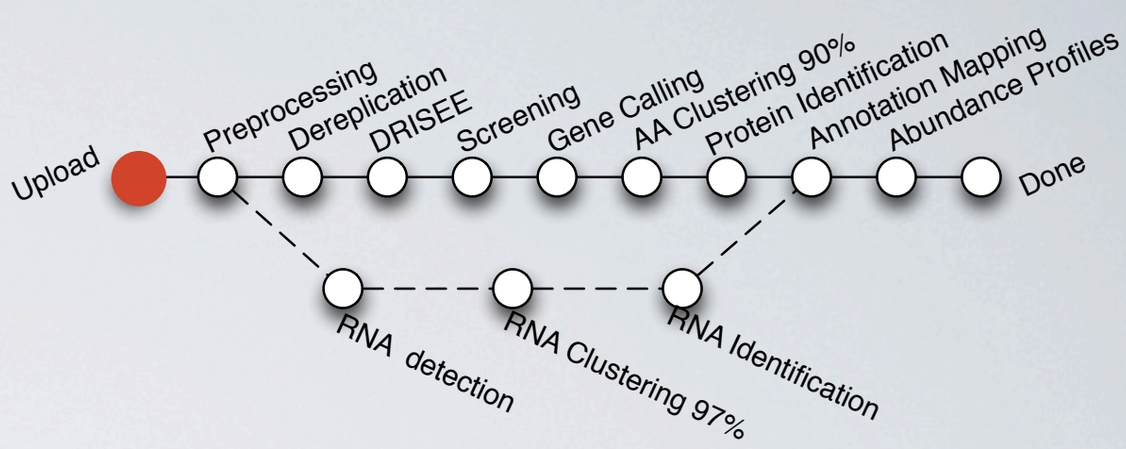
Background

FASTQ

@sequence_ID1 (any other text)
GACTGATCATCTACTAGTCGATGC…
+sequence_ID1
14,12,15,18,56,1,7,34,40,45,43,32,42,…
@sequence_ID2
GAGTCAGTCGATCATCGTAGCTAGCTAGCT…
+sequence_ID2
4,7,34,40,45,43,32,42, 4,11,12,15,18,56,1…
…

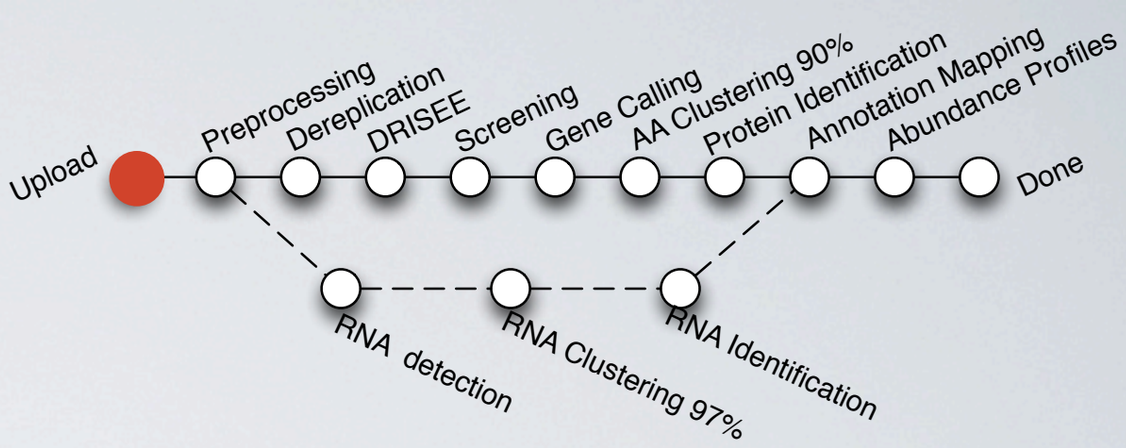
FASTQ includes quality information

Upload



- Pipeline accepts DNA sequences in FASTQ, and FASTA formats.
  - Uploading "raw" sequences w/o any modifications allows QC
  - MG-RAST understands **multiplexed** files
    - Barcode based splitting and extraction is supported
  - Upload checks for appropriate data formats - not always with success
- All sequences in files need **unique identifiers**

Upload



- Multiple-dataset upload
- FASTQ format preferred
- Upload spreadsheets describing samples (facilitating metaanalysis + downstream comparisons)

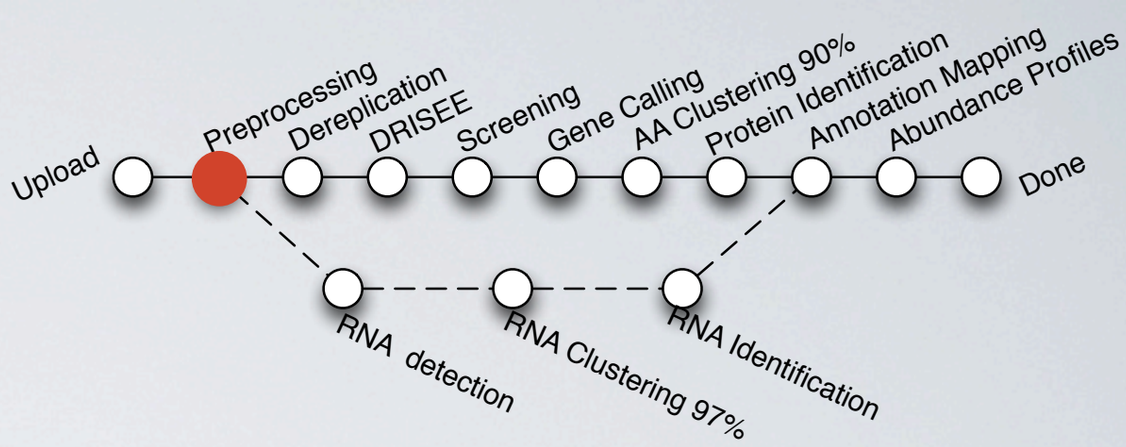
	A	B	C	D	E	F	G
1	sample_nam	sample_id	latitude	longitude	continent	country	location
2	Unique nam	Internal ID o	The geograp	The geograp	The geograp	The geograp	The geogra
3	sample1						
4	sample2						
5	sample3						
6							
7							

GSC MIxS INFO

Investigation Type	**metagenome**
Project Name	**The oral metagenome in health and disease**
Latitude and Longitude	**39.481448, 0.353066**
Country and/or Sea, Location	**Spain Valencia**
Collection Date	**2010-03-01 10:00:00 UTC**
Environment (Biome)	**human-associated habitat**
Environment (Feature)	**human-associated habitat**
Environment (Material)	**human-associated habitat**
Environmental Package	**human-oral**
Sequencing Method	**454**

More Metadata

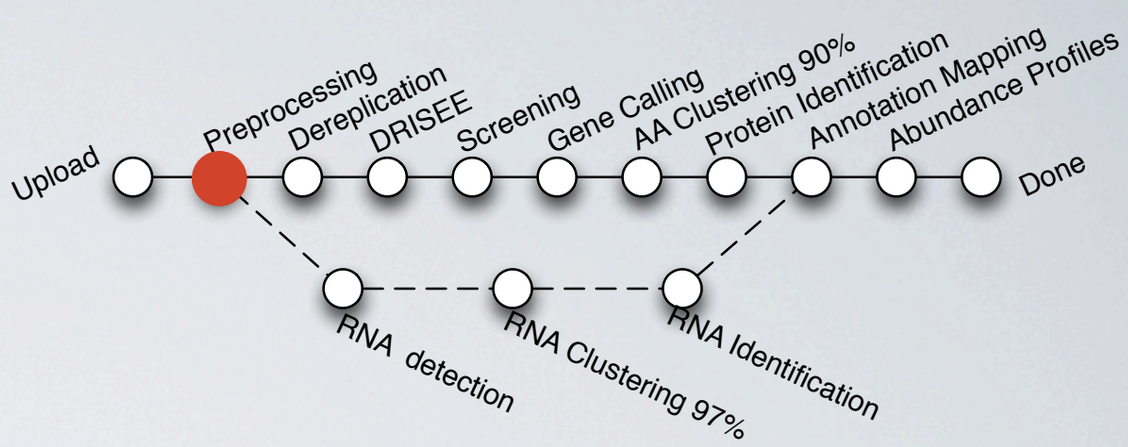
Preprocessing



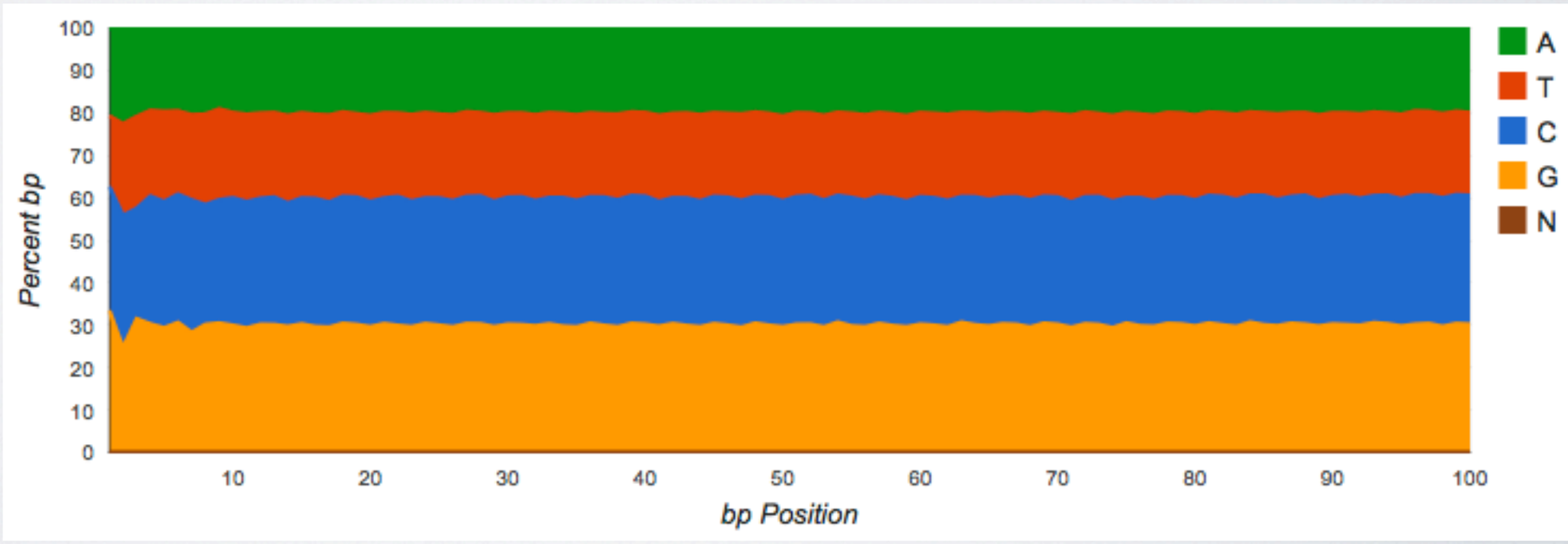
- Optional step to discard low-quality reads, turned on by default
- For **FASTA** files (no qualities available),
  - discard reads with too many **ambiguous basecalls** (Ns)

  and
  - Discard reads with **extreme lengths** (>= 2 StdDev)
- Based on (Huse et al. PMID: 17659080)

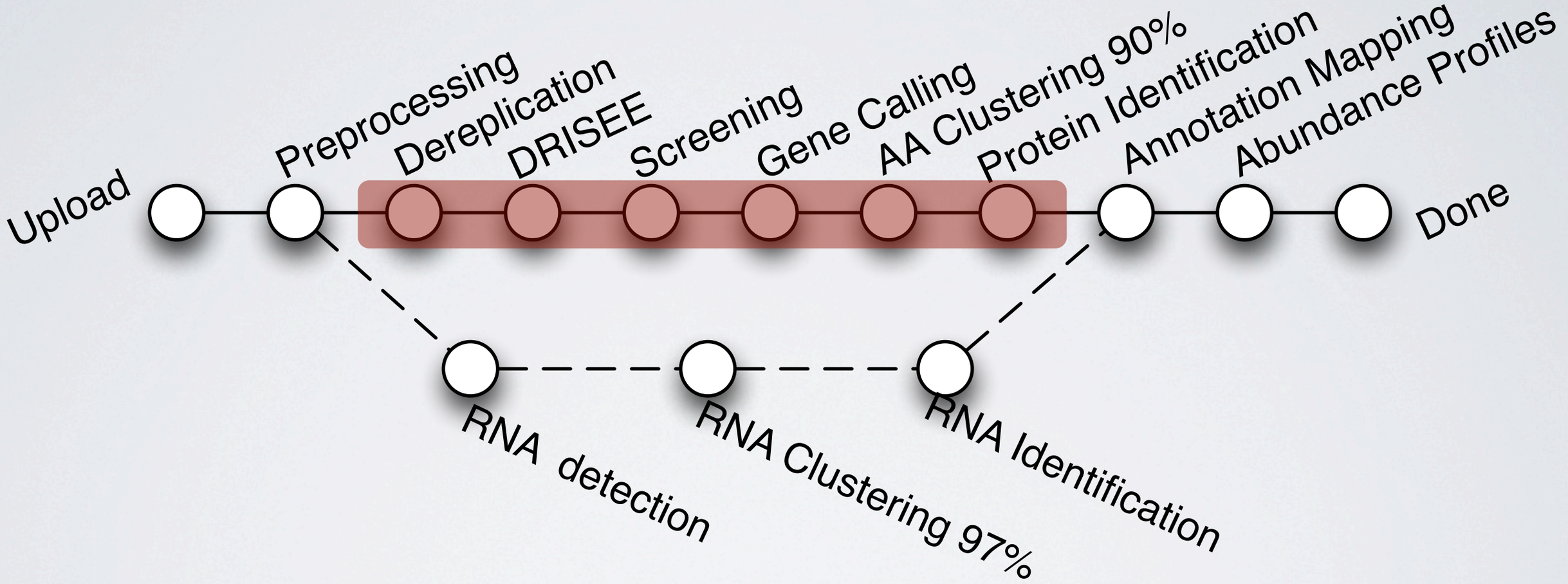
Preprocessing



- If quality info present: **perform quality-based read trimming** (DynamicTrim Cox et al. PMID: 20875133)

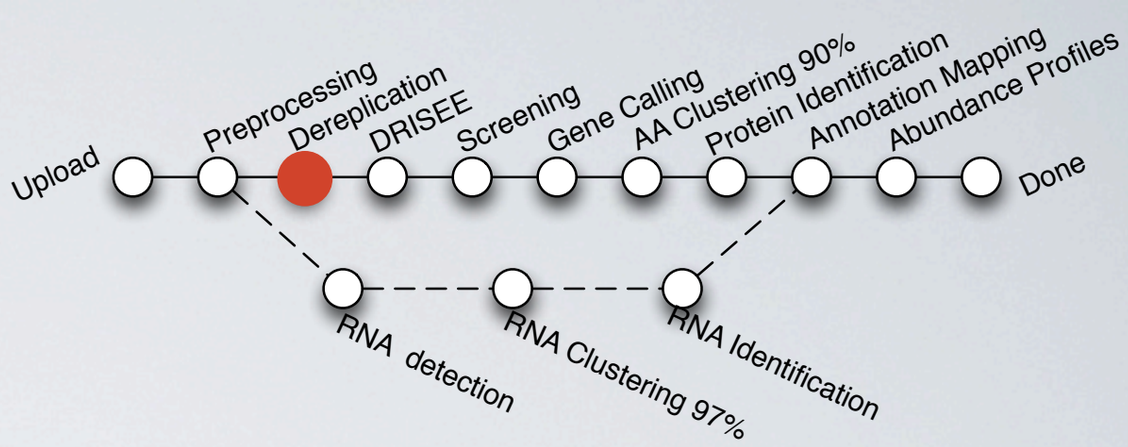


Shotgun reads



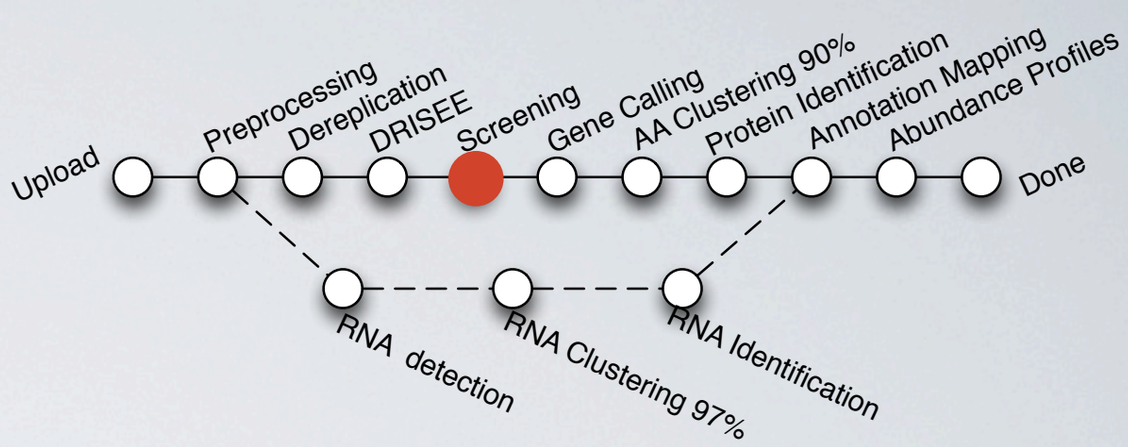
whole genome fragments

Dereplication



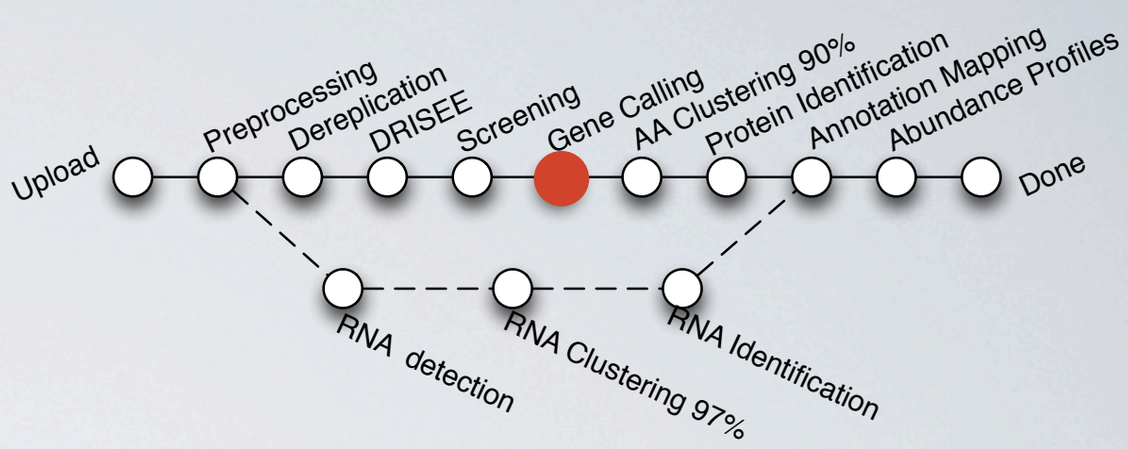
- Optional step to remove artifactually-duplicated sequences, called technical duplicates caused by PCR (see Gomez-Alvarez PMID: 19587772)
  - reads with first 50-bp identical are grouped together
  - all but the longest sequence are discarded
  - NOT APPROPRIATE FOR AMPLICON DATA

Screening



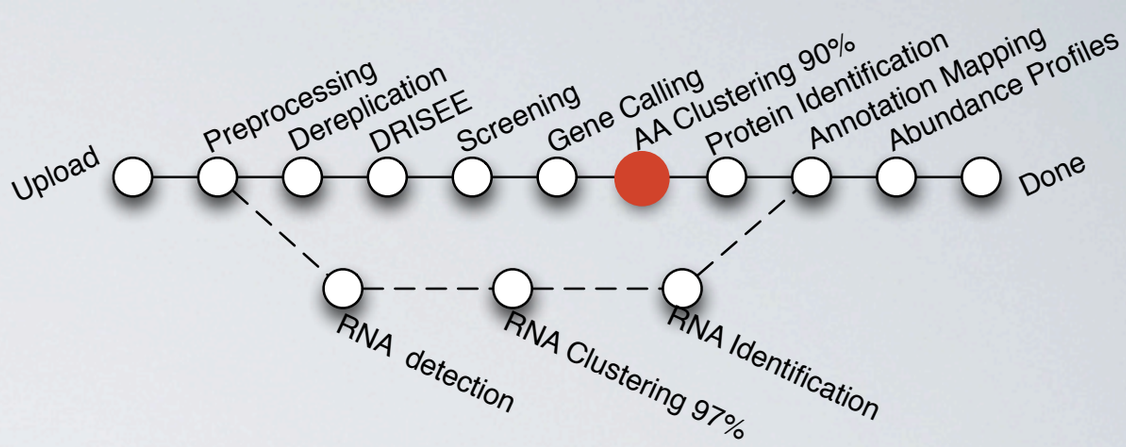
- Optional step to **remove sequences from model organisms**
- Reference genomes selected by user (*H. sapiens*, *S. cerevisiae*, *E. coli*, others)
- Sequences identified using DNA level matching (**`bowtie`**)
  - Maximal 2 mismatches in the first 20 basepairs
- Matching sequences are removed and don't proceed to annotation.
- **Note:** *H. sapiens* screen is on by default

Gene Calling



- Predicts coding regions on the remaining reads using FragGeneScan (Rho et al PMID: 20805240)
  - translates DNA reads into AA sequence fragments
    - Uses codon information from already sequenced genomes
    - Provides error tolerance (frameshifts, in/dels, etc.)
- using predicted coding sequences as new input for following steps

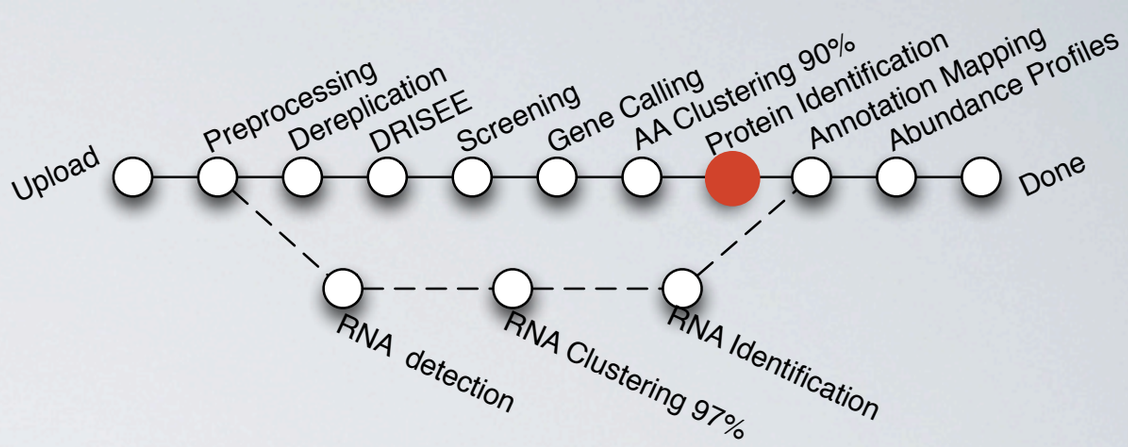
Clustering



Group amino acid sequences

- cluster by 90% AA sequence similarity
- select cluster representative sequence **and singlets** for further processing
- keep cluster abundances

Similarity

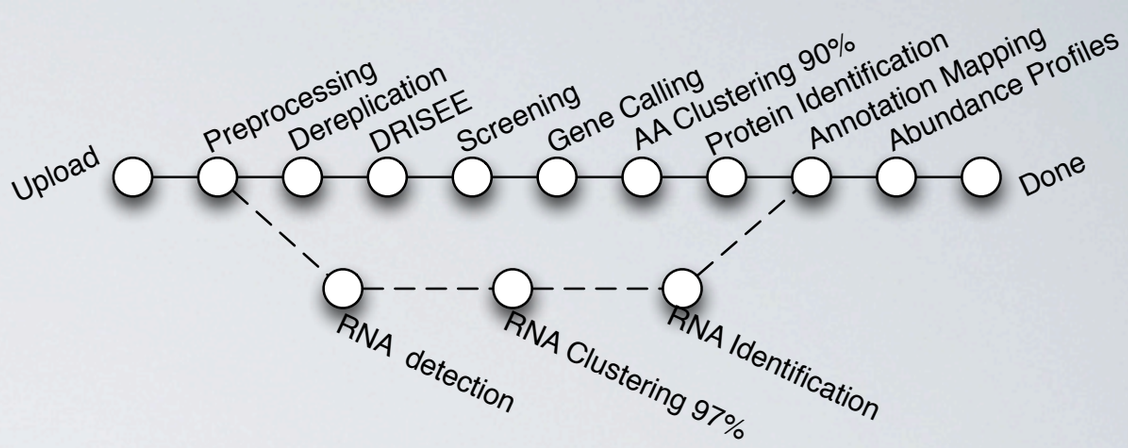


- Similarities are computed using BLAT(Kent PMID: 11932250) against a **M5NR**

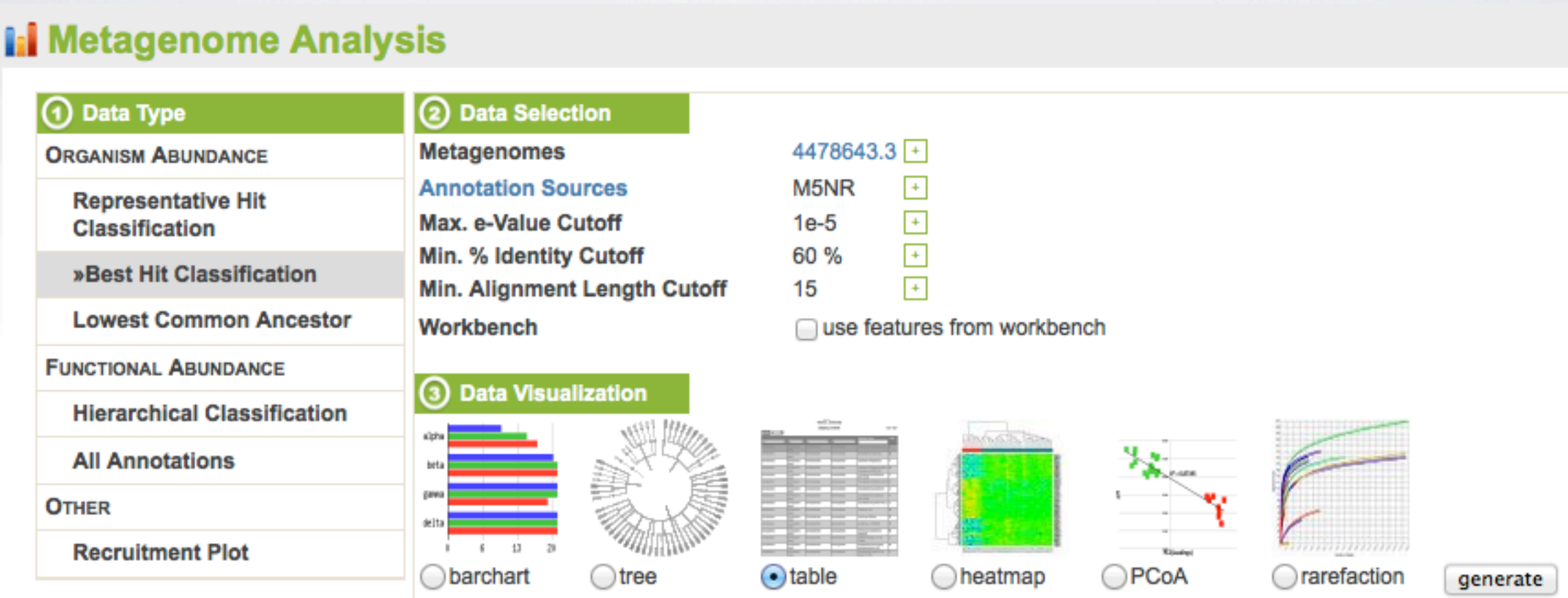
- Most expensive step because protein database and input dataset are large

  - Note: slightly less sensitive than BLAST but computationally a lot cheaper

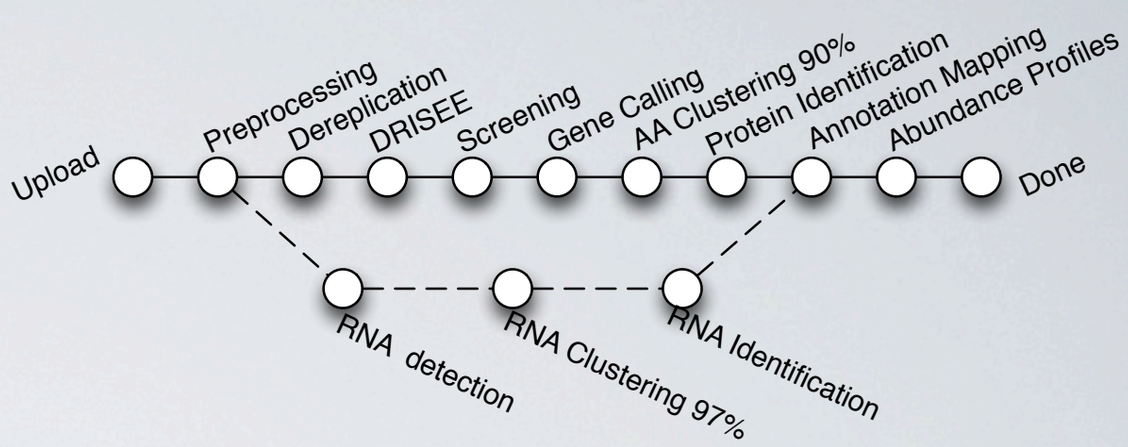
Similarity II



- We store the best 100 hits per input sequence
- By default we use the **best hit**
- LCA (Huson PMID CITE) integrates the best 10 hits



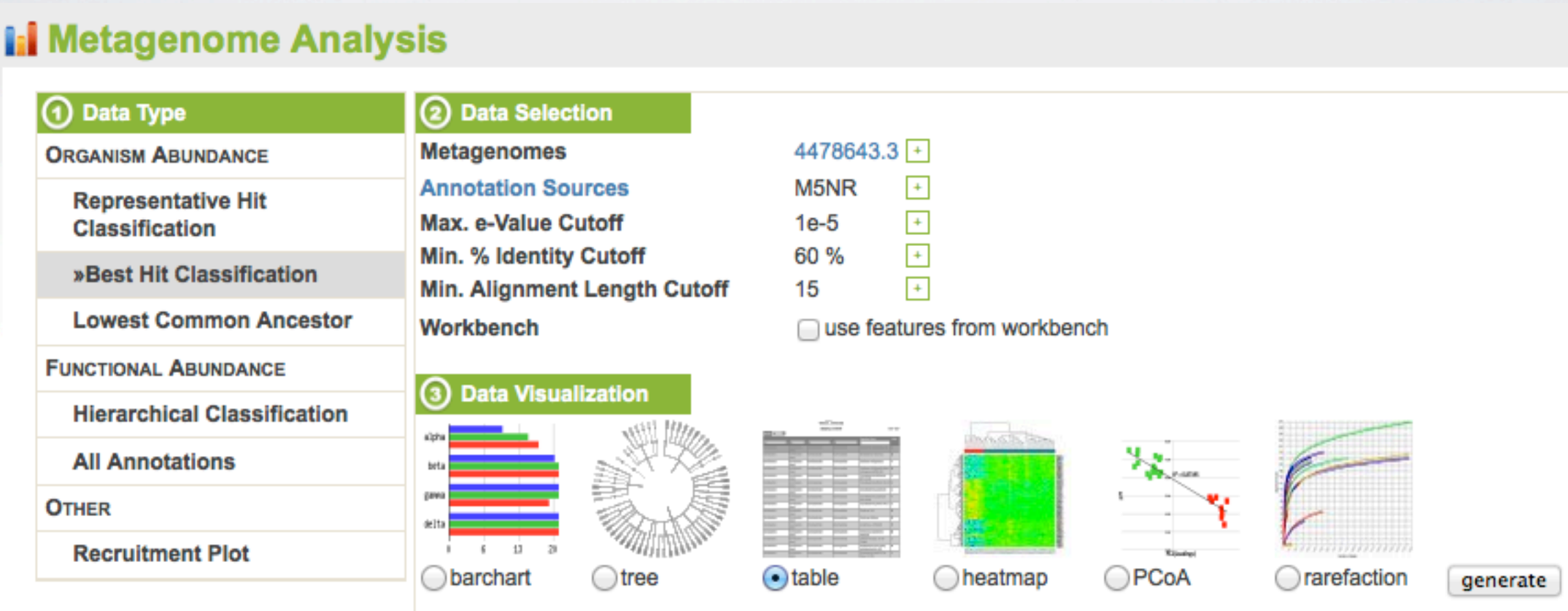
Similarity III



- Web interface summarizes annotations on demand.

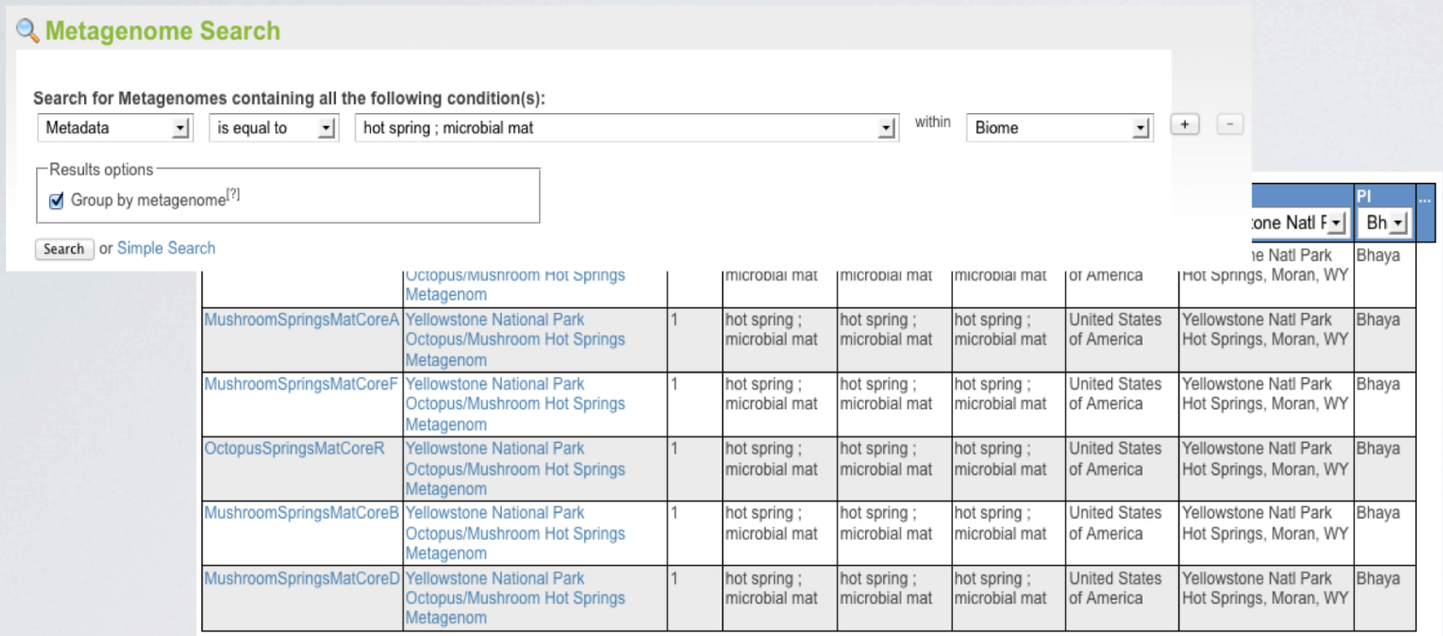
  Can choose multiple datasets

  Can choose annotation database

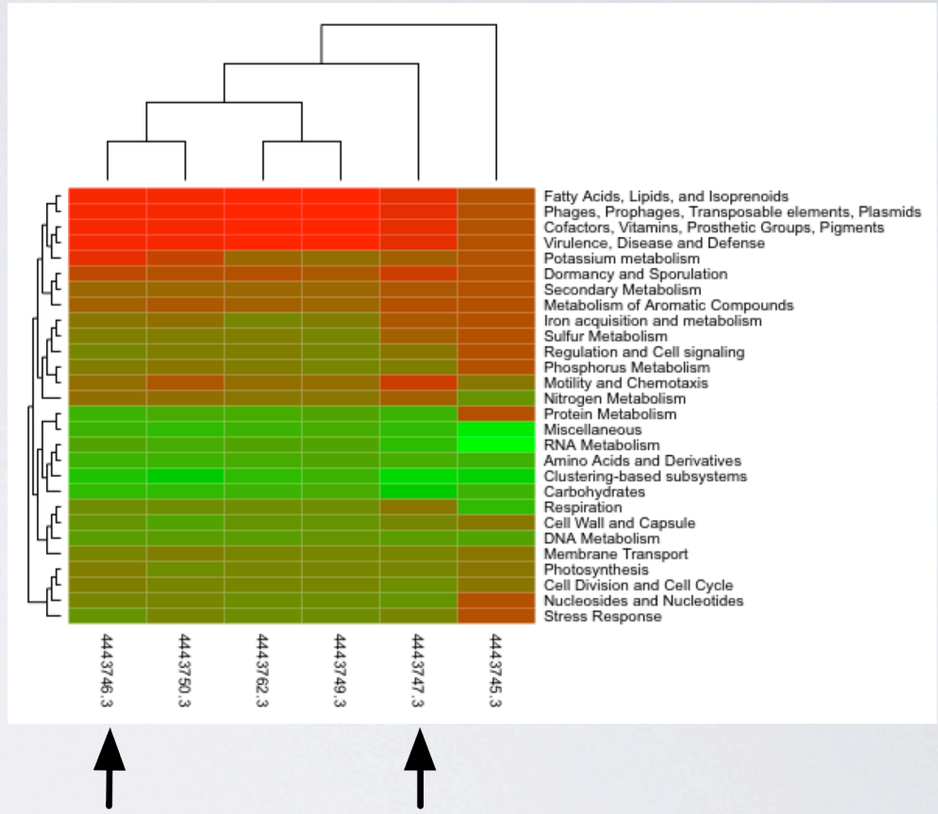


Web interface visualizations

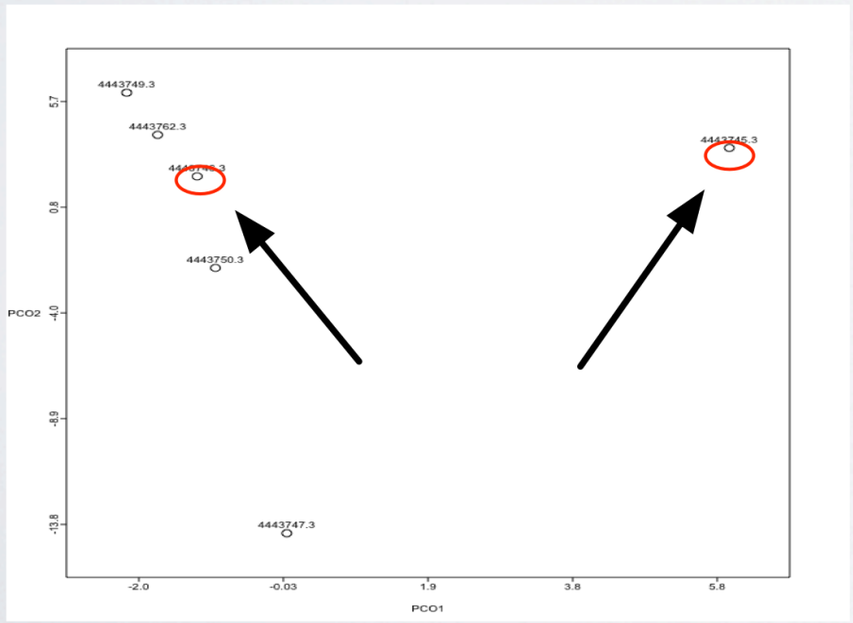
a)



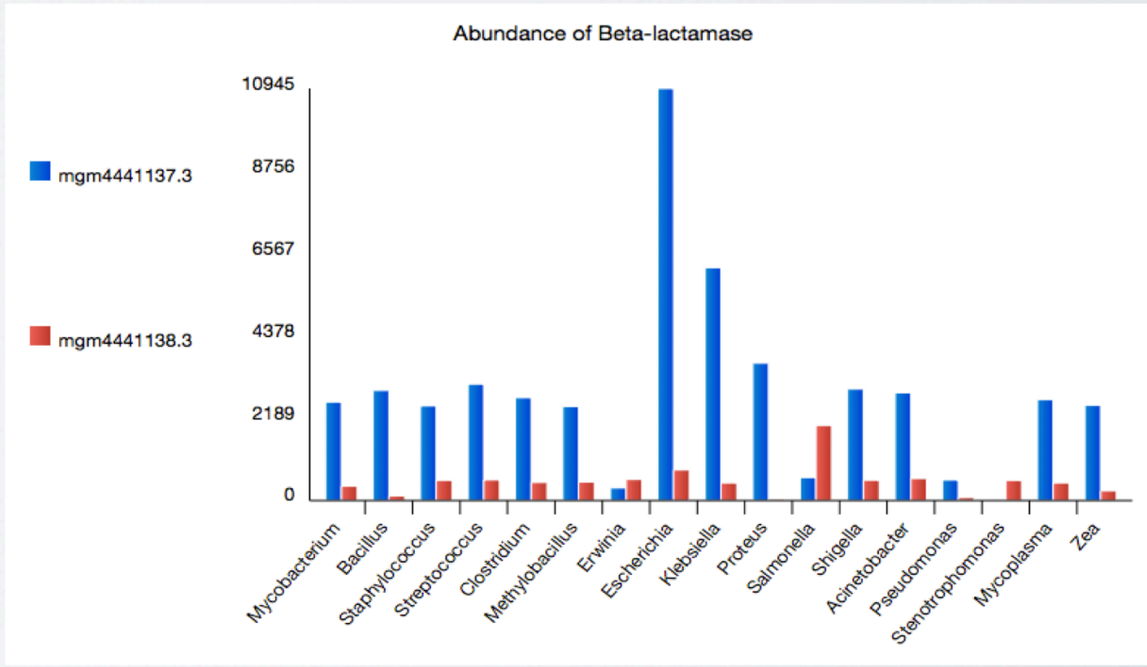
b)



c)

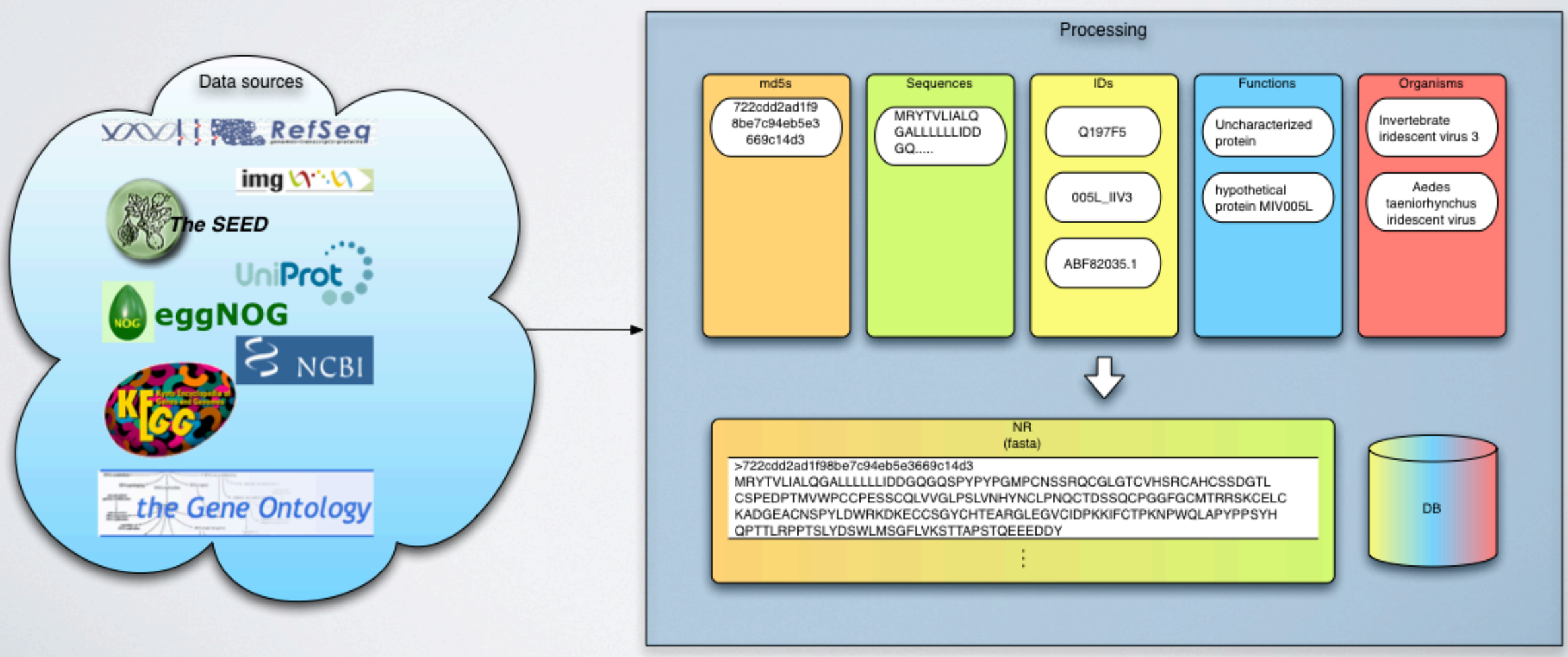


d)

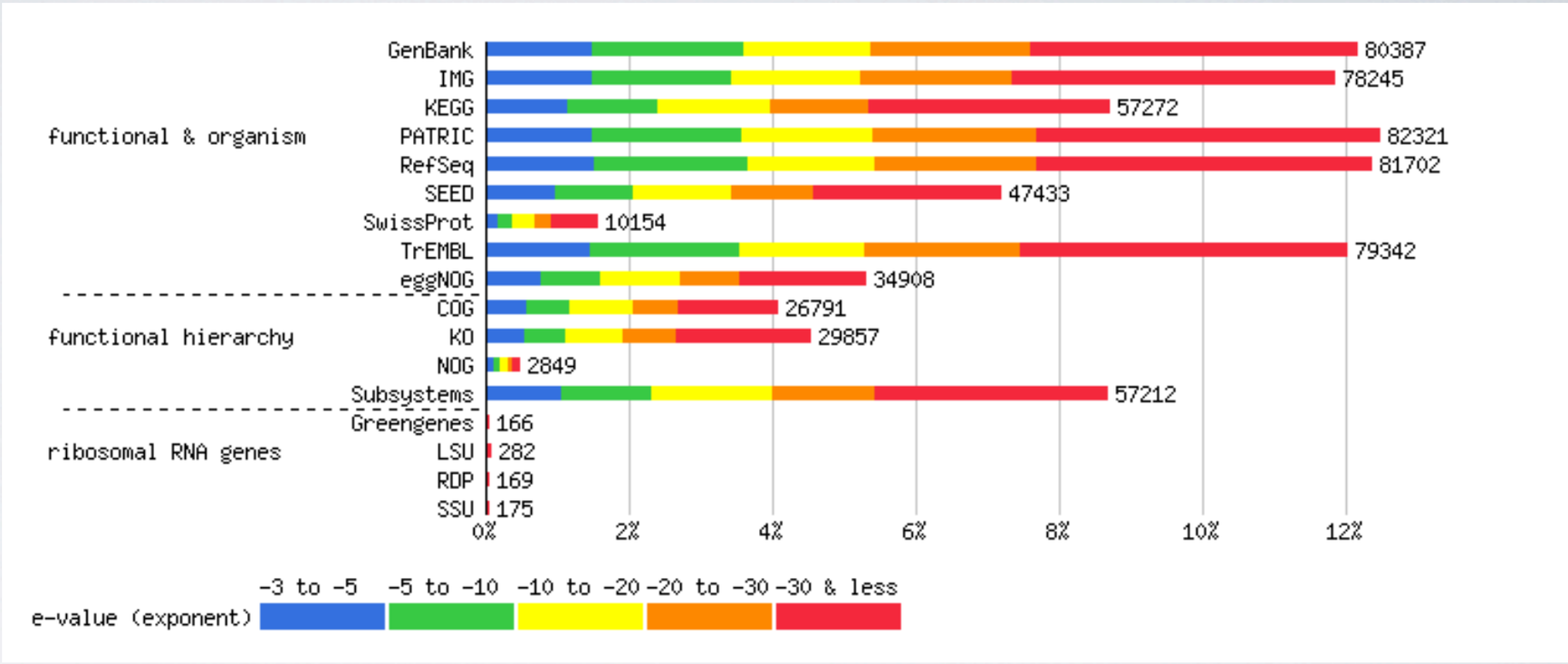


Non redundant Database

8 billion amino acid protein database :
union of Refseq, GO, NCBI, SEED.
Database choice made AFTER similarity search.



Functional hierarchies

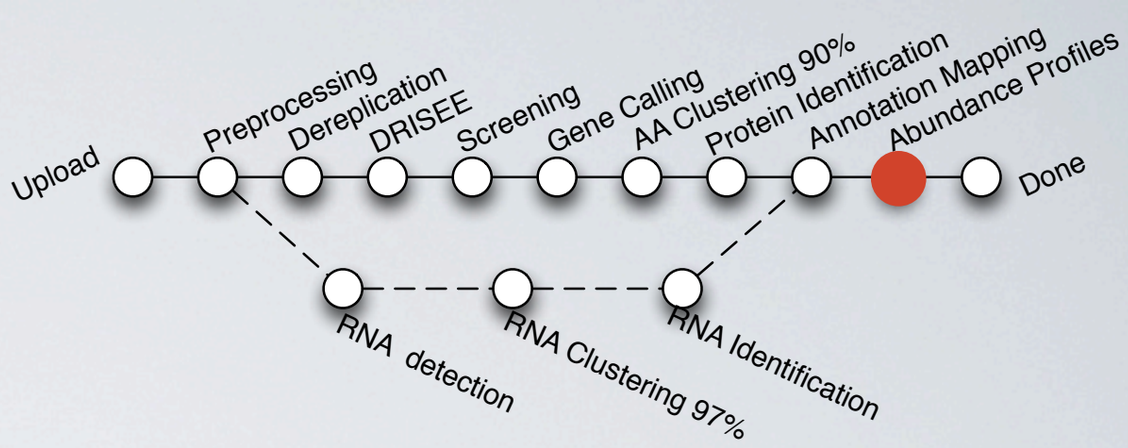


- Functional hierarchies allow comparison of multiple metagenomes
- Derived from similarity search results
- Fewer hits in functional hierarchies

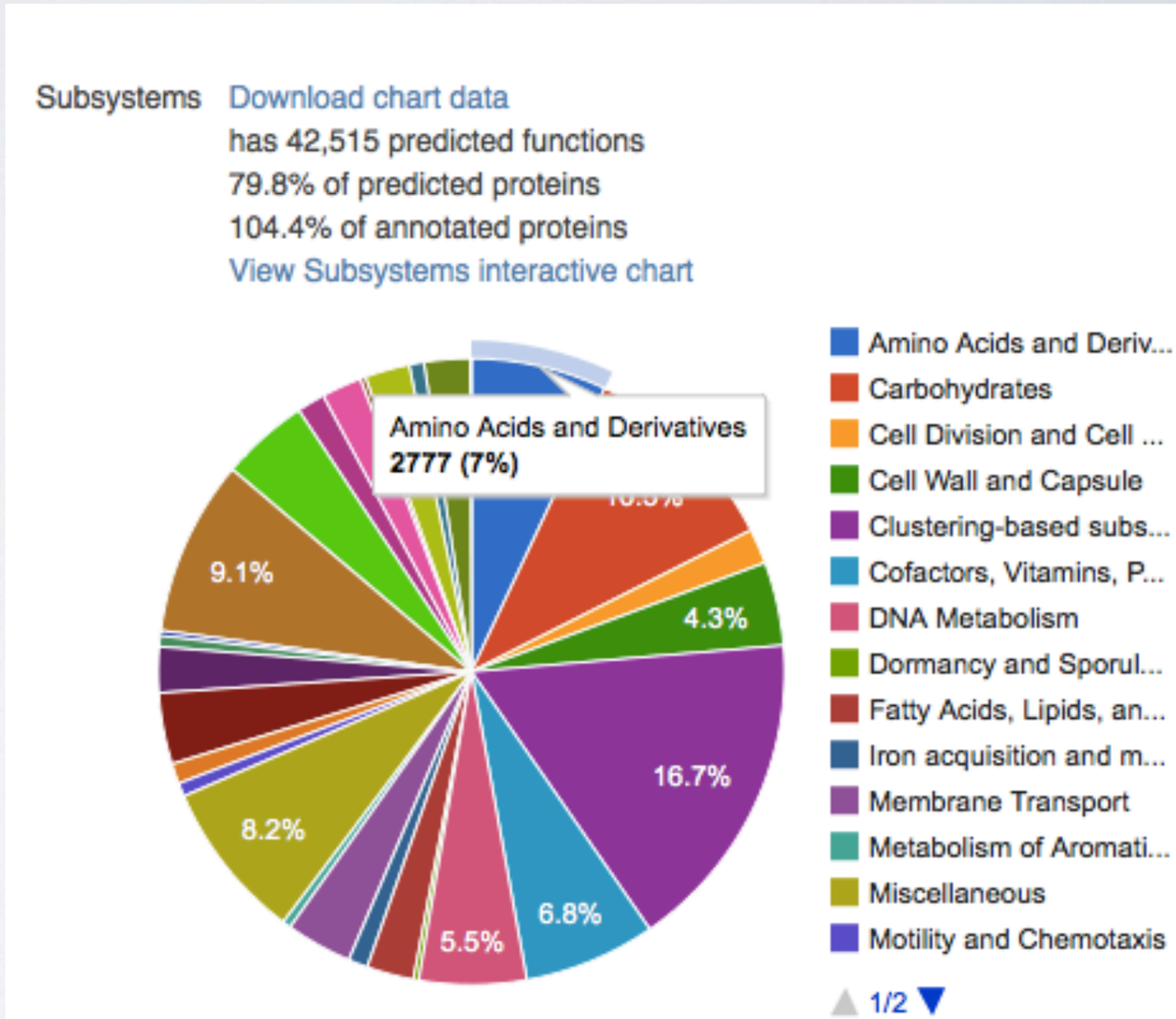
How can I get annotations out?

- On website – through browser (ape limitations)
- Data files from download page (limited by ape's patience)
- https://api.mg-rast.org programming interface for auotmated access to data
- MG-RAST-Tools python tools to talk to the API (full)
- R package matR to talk to the API (read-only)

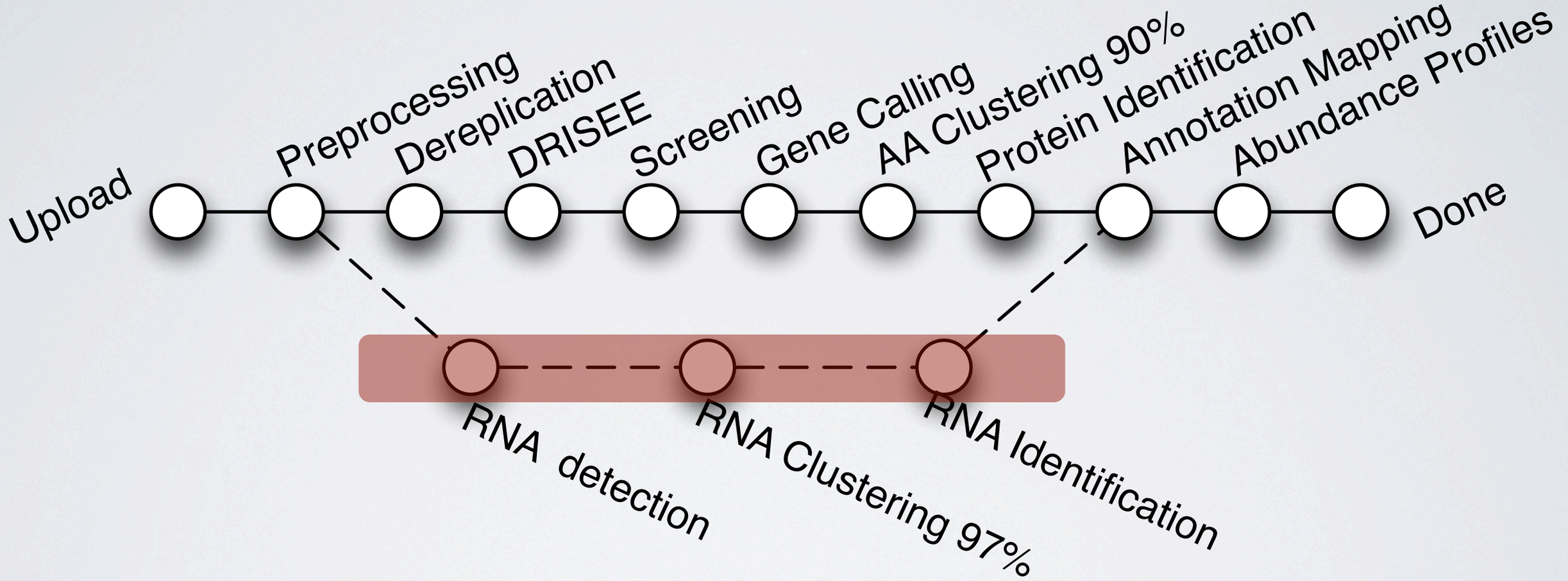
Abundance Profile



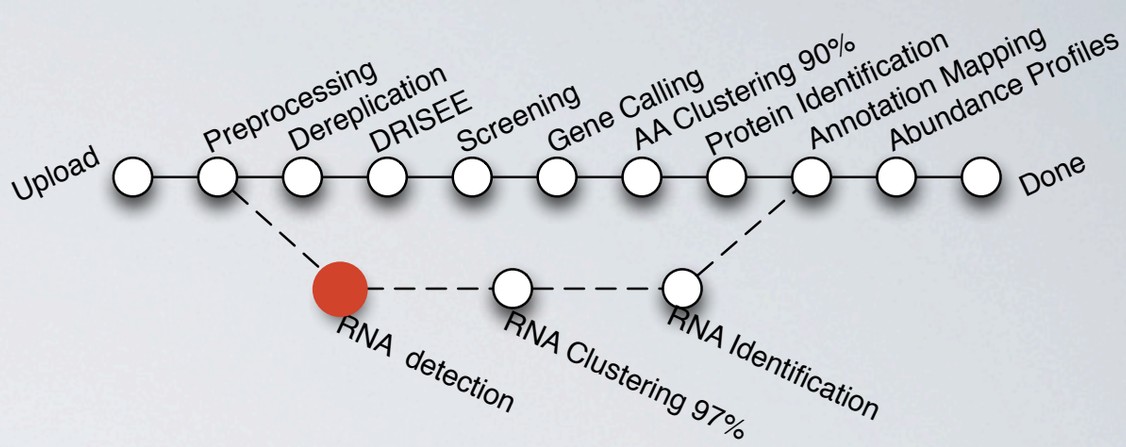
- Functional classifications
  - Using COG, NOG, KO and SEED Subsystems
- Taxonomic classifications
  - based on NCBI taxonomy , Silva , Greengenes and RDP
- LCA
  - compute LCA per fragment
  - NCBI taxonomy
  - group LCA



RNA

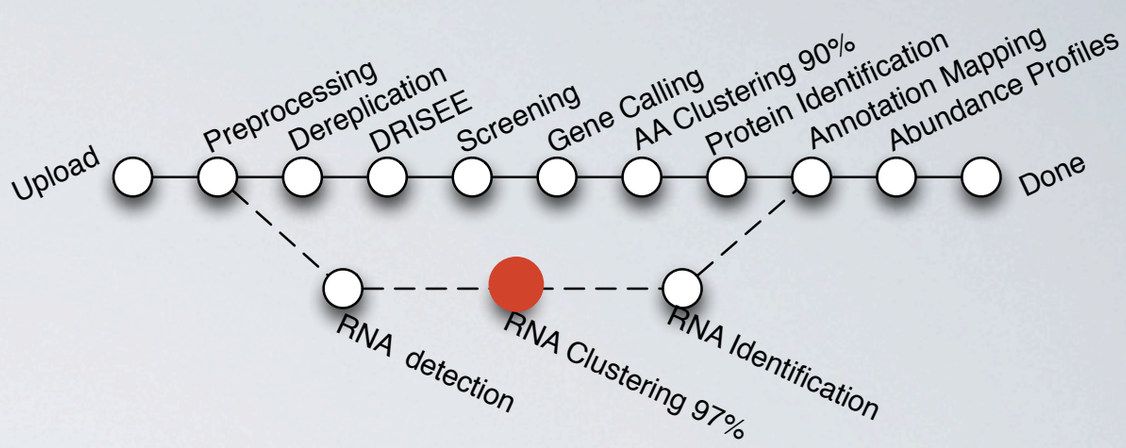

Upload
Preprocessing
Dereplication
DRISEE
Screening
Gene Calling
AA Clustering 90%
Protein Identification
Annotation Mapping
Abundance Profiles
Done
RNA detection
RNA Clustering 97%
RNA Identification

RNA Detection



- Extract RNA sequences from shotgun pool
- Using reduced RNA database
- Fragments with hits are processed
- Applicable for WGS data sets

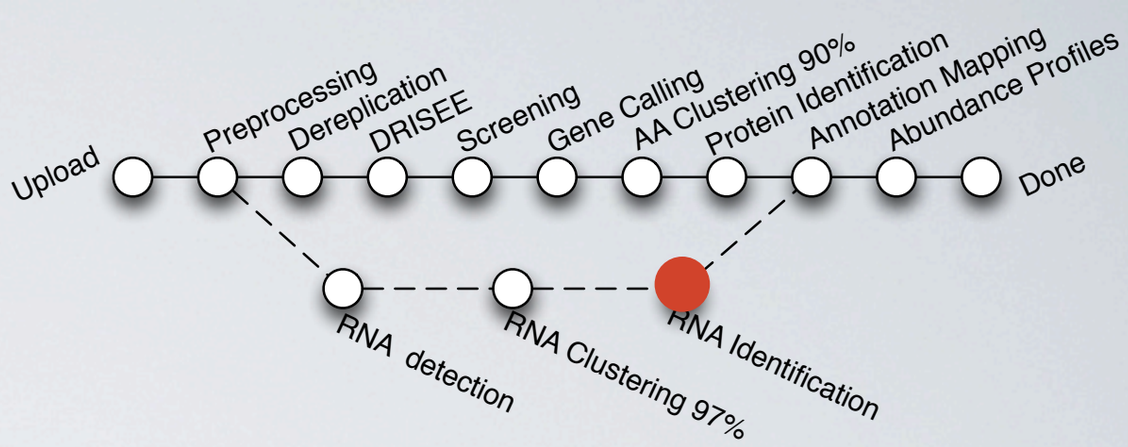
RNA Clustering



Group DNA sequences

- cluster by 97% sequence similarity
- select cluster representative sequence and singlets for further processing
- keep cluster abundances

Similarity



- Identify similar sequences in a reference database
- Similarities are computed using blat (Kent PMID: 11932250)
- Input
  - Cluster representatives
  - M5 RNA non redundant database
- Output
  - Similarity file

- e identity alignment length number of mismatches number of gap openings query start query end hit starhit end e-value bit score semicolon separated list of annotations

- mgm4447943.3|GF8803K01A6AYB|KO 00013cfb233c3250470387f8978f8e16 96.43 28 1 13 40 40 67 8.8e-08 55.0 [dihydroxy-acid dehydratase [EC:4.2.1.9]]

- mgm4447943.3|GF8803K01CA4XS|KO 00041978678ba63ab6057840be689d42 83.93 56 9 69 124 1 56 2.8e-19 93.0 [fused signal recognition particle receptor]

- mgm4447943.3|GF8803K01A4XTJ|KO 0007c0e8723384b66f266a60ea8a0219 86.67 15 2 17 31 10 24 1.6e+02 24.0 [lactoylglutathione lyase [EC:4.4.1.5]]

- mgm4447943.3|GF8803K01BLNT3|KO 0007c0e8723384b66f266a60ea8a0219 86.24 109 15 1 109 18 126 2.2e-53 206.0 [lactoylglutathione lyase [EC:4.4.1.5]]

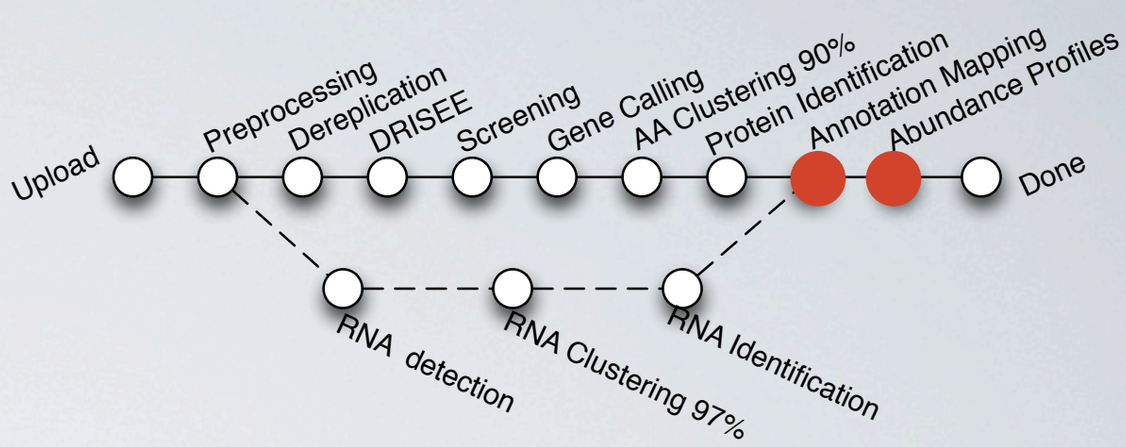
- mgm4447943.3|GF8803K01CHTGX|KO 0007c0e8723384b66f266a60ea8a0219 84.27 89 14 23 111 33 121 2.6e-41 166.0 [lactoylglutathione lyase [EC:4.4.1.5]]

- mgm4447943.3|GF8803K01CW3DH|KO 0007c0e8723384b66f266a60ea8a0219 98.81 84 1 8 91 15 98 8.9e-44 174.0 [lactoylglutathione lyase [EC:4.4.1.5]]

- mgm4447943.3|GF8803K01D893O|KO 000b6fd4f8a1e46f0441a7ca0e60a17d 89.71 68 7 1 68 303 370 1.5e-26 117.0 [O-acetylhomoserine (thiol)-lyase [EC:2.5.1.49]]

- mgm4447943.3|GF8803K01CX9J1|KO 000d9afd2dde083d470eb9c8ea33b080 80.69 145 28 1 145 2 146 1.2e-60 231.0 [L-lactate dehydrogenase (cytochrome) [EC:1.1.2.3]]

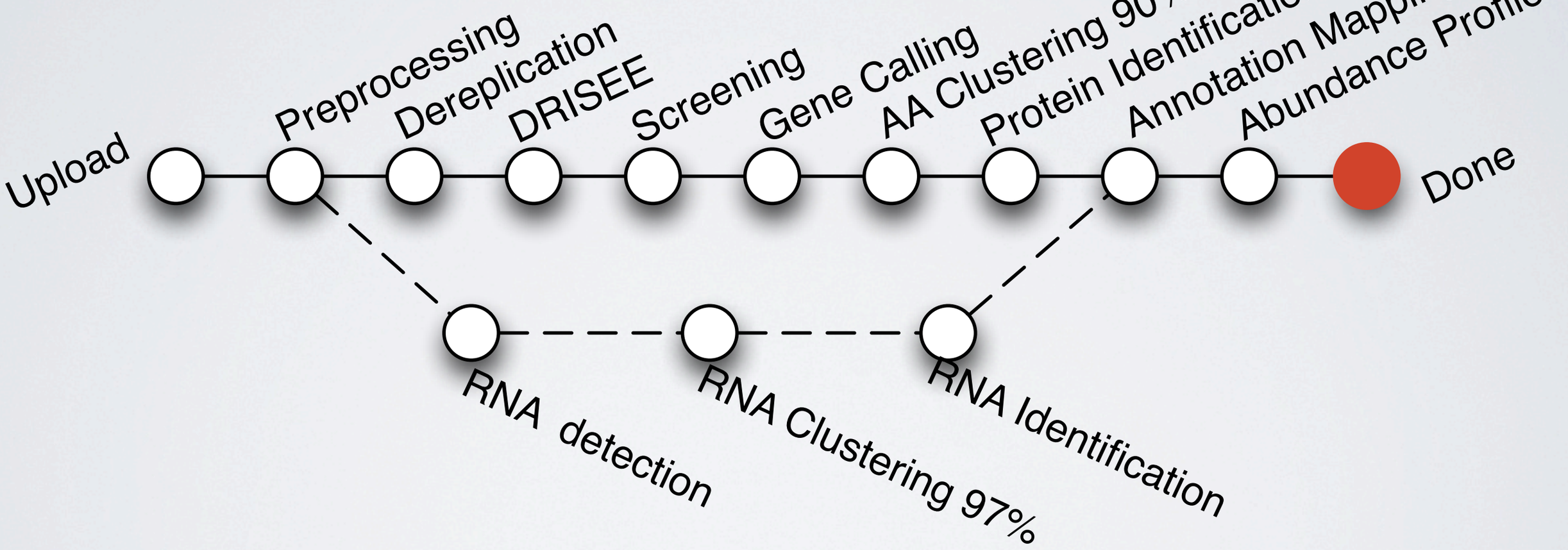
Annotation Mapping and Abundance Profiles



Same process as described before:

- Lookup organism for md5s
- Compute summary counts

Standard Pipeline

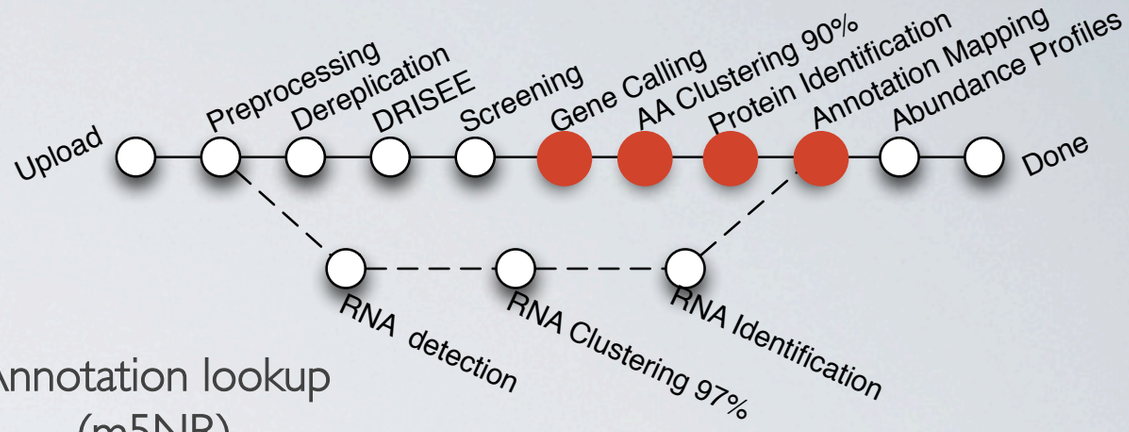


What does MG-RAST do when I query it?

https://api.mg-rast.org/matrix/function?

id=mgm4447943.3&
id=mgm4447192.3&
id=mgm4447102.3&
group_level=level3&
source=Subsystems&
identity=80&
evalue=5

Under the hood



Gene calling (FGS)

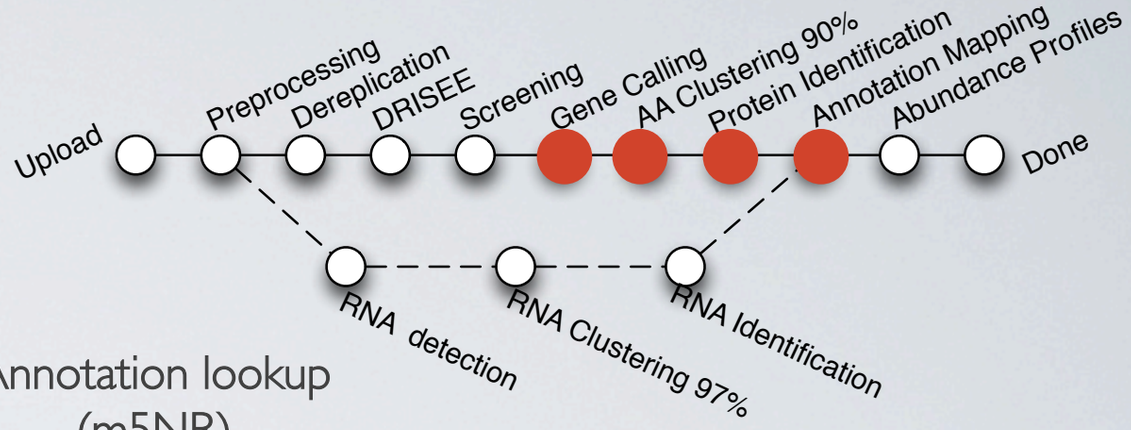
Protein clustering (90%)

Protein similarity search (blat)

Annotation lookup (m5NR)

(reads)	(features)	(clusters)	(md5 ids)	(database identification)
read1	read1_gene1	cluster1 (2)	prot43634	RefSeq: DNA topoisomerase III L. drancourtii
			prot17372	Trembl: DNA topoisomerase III H. elongata
			prot93882	SEED: DNA topoisomerase III (EC 5.99.1.2)
read2	read2_gene1			
read3	read3_gene1	read3_gene1	prot21846	RefSeq: Hypothetical protein, H. aurantiacus
			prot49918	Trembl: Putative uncharacterized protein, H. aurantiacus
read4	read4_gene1			
	read4_gene2	read4_gene2	prot01873	RefSeq: beta-ketoacyl synthase, H aurantiacus
			prot27256	Trembl: beta-ketoacyl synthase, H aurantiacus
			prot96100	SEED: Malonyl CoA-acyl transacylase (EC 2.3.1.39)
read5	read5_gene1			
read6	read6_gene1	cluster2 (3)	prot23848	RefSeq: Ribonucleotide reductase, C. tuberculostearicum
read7	(no predicted orfs)			
read8	read8_gene1	cluster3 (2)	prot24851	RefSeq: LuxR transcriptional regulator, B. pseudomallei
			prot028379	Trembl: transcriptional regulator, XRE family, B. pseudomallei
read9	read9_gene1			

Under the hood



Gene calling (FGS)

Protein clustering (90%)

Protein similarity search (blat)

Annotation lookup (m5NR)

read1 — read1_gene1 — cluster1 (2)
- prot43634 — DNA topoisomerase III L. drancourtii
- prot17372 — DNA topoisomerase III H. elongata
- prot93882 — DNA topoisomerase III (EC 5.99.1.2)

read2 — read2_gene1

read3 — read3_gene1 — read3_gene1
- prot21846 — Hypothetical protein, H. aurantiacus
- prot49918 — Putative uncharacterized protein, H. aurantiacus

read4 — read4_gene1

read4_gene2 — read4_gene2
- prot01873 — beta-ketoacyl synthase, H aurantiacus
- prot27256 — beta-ketoacyl synthase, H aurantiacus
- prot96100 — Malonyl CoA-acyl transacylase (EC 2.3.1.39)

read5 — read5_gene1

read6 — read6_gene1 — cluster2 (3)
- prot23848 — Ribonucleotide reductase, C. tuberculostearicum

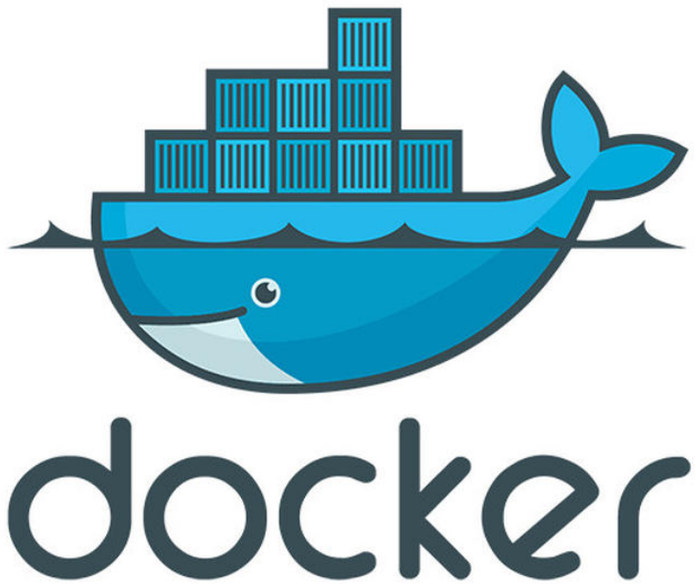
read7 — (no predicted orfs)

read8 — read8_gene1 — cluster3 (2)
- prot24851 — LuxR transcriptional regulator, B. pseudomallei
- prot028379 — transcriptional regulator, XRE family, B. pseudomallei

read9 — read9_gene1

(reads) (features) (clusters) (md5 ids)

(database identification)
- RefSeq
- Trembl
- SEED



Containers!

- Much, if not most of the difficulty in scientific computing is getting the software to work.

- We could install the software, and all the software that it depends on, and that's what lots of bioinformatics classes do.

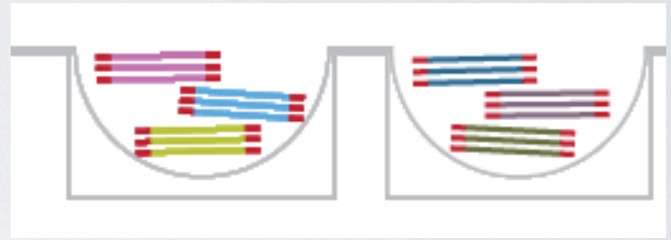
- We could get a "container" that has the operating system, the programs, and the dependencies all installed.

- What can containers do? (What can computers do?)

The sequencing technology is changing again!

Pac Bio systems

Oxford nanopore



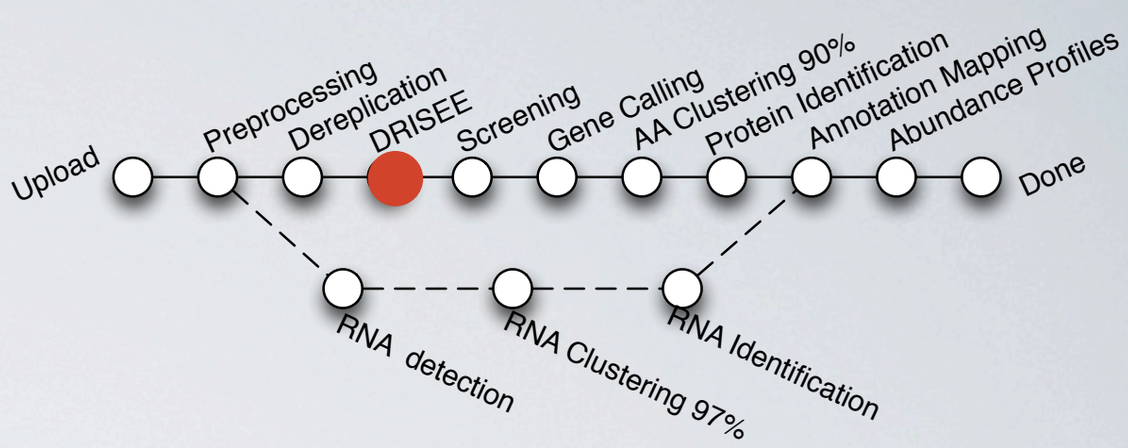
Dilution-amplification

Very poor (12% error rate)
very long (5-50kbases) reads
Computationally expensive to align

Moleculo, 10X genomics

END

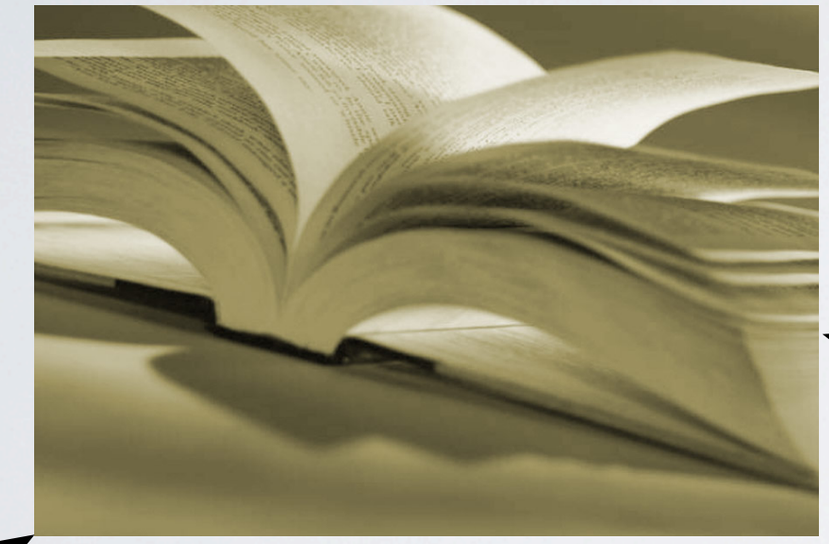
DRISEE



- Estimates sequence error based on technical replicates

- Unexpected values indicate poor agreement with shotgun assumptions—some datasets merit technical scrutiny.

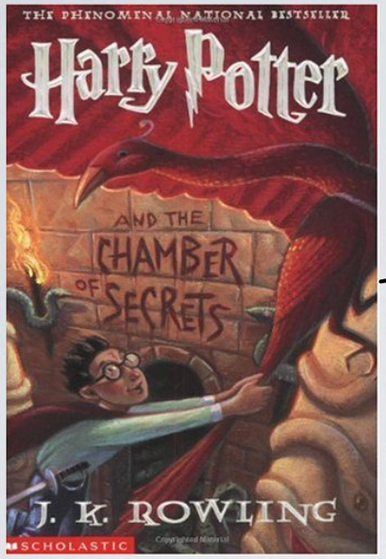
MD5 check sum describes content



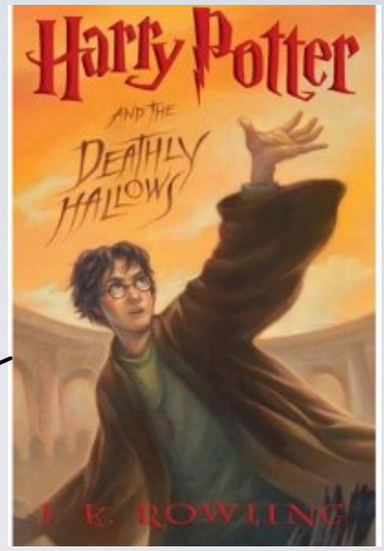
Cover

Content

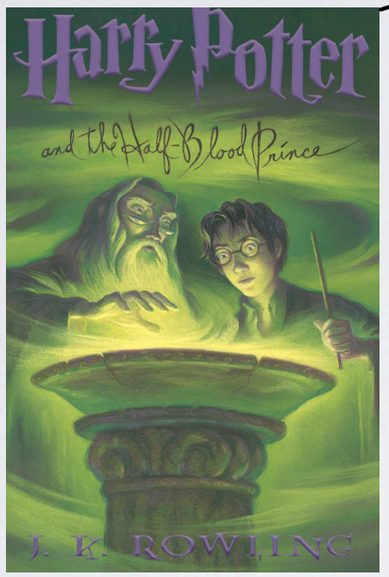
>gblABF82035.1 hypothetical protein MIV005L (Aedes taeniorhynchus iridescent virus)
MRYTVLIALQGALLLLLLIDDGQGQSPYPYPGMPCNSSRQCGLGTCVHSRCAHCSSDGTL
CSPEDPTMVWPCCPESSCQLVVGLPSLVNHYNCLPNQCTDSSQCPGGFGCMTRRSKCELC
KADGEACNSPYLDWRKDKECCSGYCHTEARGLEGVCIDPKKIFCTPKNPWQLAPYPPSYH
QPTTLRPPTSLYDSWLMSGFLVKSTTAPSTQEEEDDY



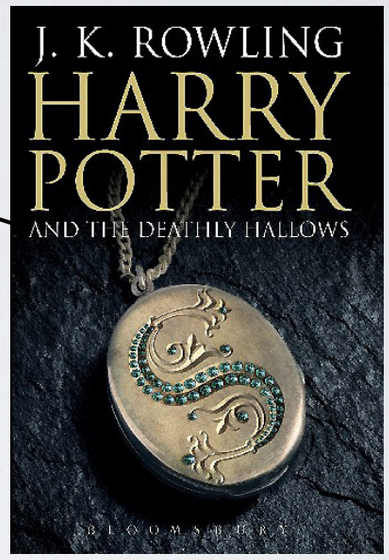
891cdd2ad1f98be7c94eb5e3669c16d5



921cdd2ad1f98be7c94eb5e3669c23e6



722cdd2ad1f98be7c94eb5e3669c14d3



522ccc2ab1f11be8c94eb9f3669c18d1