

## THE NEED FOR REPRODUCIBLE BIOINFORMATICS



# Towards solving the metagenomics reproducibility crisis with CWL and RO



**FOLKER MEYER**

[folker@anl.gov](mailto:folker@anl.gov)

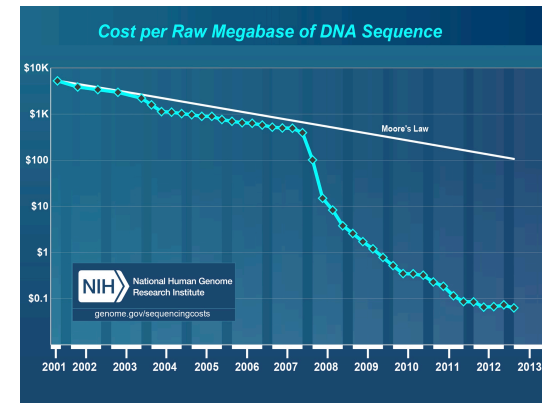
Argonne National Laboratory  
and  
University of Chicago

Amsterdam, April, 2018



# Microbiome informatics

- Example of big data analytics in bio-medical science
  - **Used to be unique stunt, is routine today**
- Analyzing and contextualizing DNA sequence data
  - “environmental DNA” (aka Metagenomics)
- We study complex microbiomes (“getting hands dirty”)
  - SOIL and Human GI tract
- Building tools and integrations

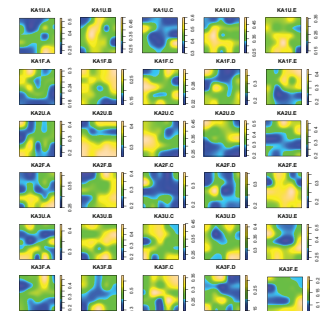


Source: genome.gov

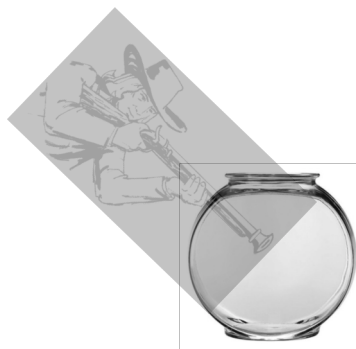
MG-RAST [<https://mg-rast.org>]



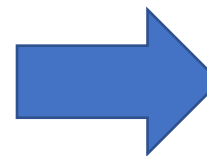
Fermi Bioenergy plot



hundreds of samples



Genome hunt



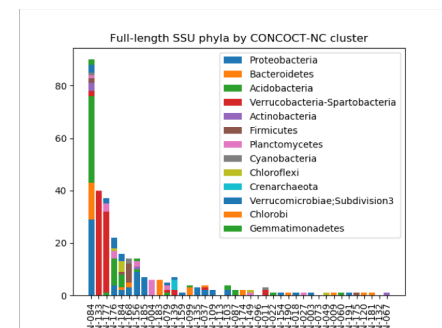
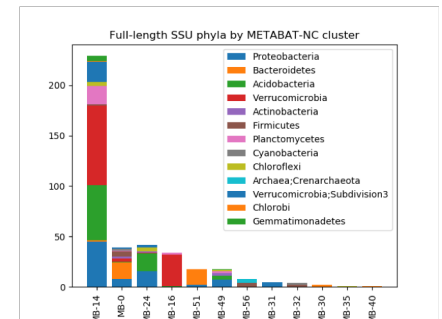
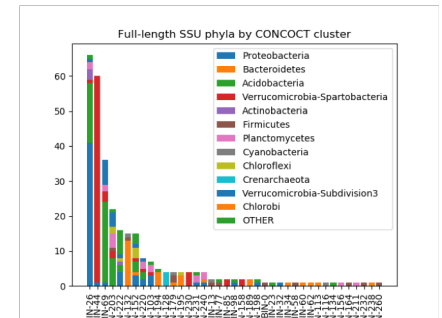
MongoDB  
MySQL  
Cassandra  
ES  
MemCacheDB  
SOLR  
..

# Microbiome puzzle challenge

- Current controversy: can we or can't we reconstruct **"complete Species genome"** (3-5 MB) from many short reads (150 byte)?
- Several "species binning tools" exist ("state of the art")
- Same data, different algorithms
- Using real data not synthetic as benchmark
- **Supposedly clean results are impure at Phylum level**
- Computational provenance is never reported

➔ **Reproducible bioinformatics is missing** ⬅

Kingdom → **Phylum** → Class → Family → Order → Genus → **Species**

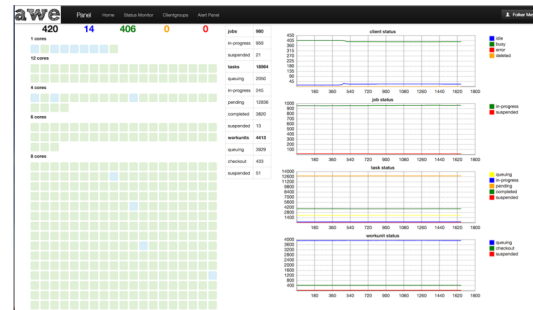
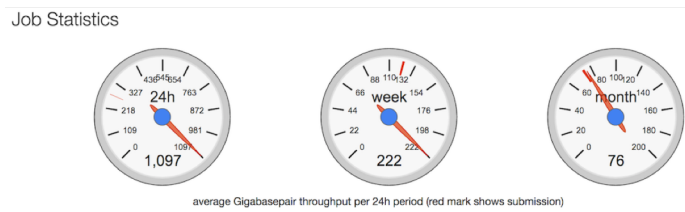
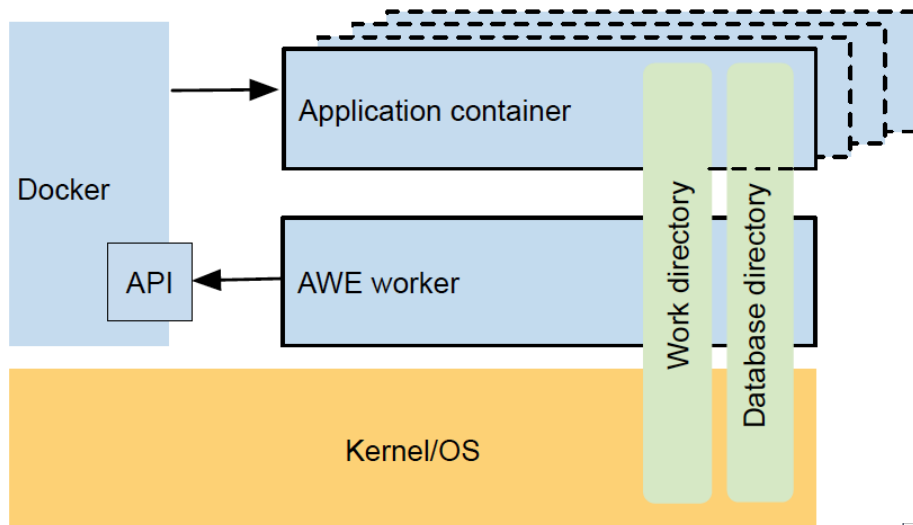


# Distributed containerized workflows

Skyport -- a scalable platform for distributed data centric reproducible computing

Skyport [Gerlach et al, IEEE Data-Intensive Computing in the Clouds, 2014]

- Uses **object store** instead of shared filesystem (→ Shock)
- Resource manager distributes work (→ AWE)
  - Multiple backends (“cloud”, bare-metal, legacy)
  - AWE2 supports → CWL
- Provide provenance info (→ <http://researchobject.org>)
- Training class available



# Current provenance in MG-RAST

## AWE v1

- Example mgm4441680.3

- <http://www.mg-rast.org/mgmain.html?mgpage=download&metagenome=mgm4441680.3>
- [Environment](#)
  - Specific github release/commit
- [workflow document](#)

```
{
  "info": {
    "pipeline": "mgrast-prod-3.0",
    "name": "[% job_id %]",
    "project": "[% project_name %]",
    "user": "[% user %]",
    "xref": "[% xref %]",
    "clientgroups": "[% clientgroups %]",
    "priority": 1,
    "userattr": {
      "id": "[% mg_id %]",
      "job_id": "[% job_id %]",
      "name": "[% mg_name %]",
      "created": "[% job_date %]",
      "status": "private",
      "owner": "[% user %]",
      "sequence_type": "[% seq_type %]",
      "bp_count": "[% bp_count %]",
      "project_id": "[% project_id %]",
      "project_name": "[% project_name %]",
      "type": "metagenome",
      "pipeline_version": "3.0"
    }
  },
  "tasks": [
    {
      "cmd": {
        "name": "pipeline_qc",
        "args": "-j [% job_id %] -s @[% inputfile %] -n raw -p 4 -k 6,16",
        "description": "qc_stats"
      },
      "dependsOn": [],
      "inputs": {
        "[% inputfile %]": {
          "host": "[% shock_url %]",
          "node": "[% shock_node %]"
        }
      },
      "outputs": {
        "[% job_id %].075.assembly.abundance": {
          "host": "[% shock_url %]",

```

Processing Steps

Data are available from each step in the MG-RAST pipeline. Each section below corresponds to a step in the processing pipeline. Each of these sections includes a description of the input, output, and procedures implemented by the indicated step. Buttons to download data processed by the step and details statistics (click on "show stats" to make collapsed table visible).

0. Upload

This is the original submitted sequence file. This is a sequence file in either fastq or fastq format. It may have been edited to change all end-of-line characters into UNIX format.

mgm4441680.3.000.upload.fna

filesize 18.2 MB

MD5 68271081731254c4e4b6b30c00f6b6

1. Initial sequence statistics

Compute statistics for the sequence, determine coverage information and preserve it for later stages. The script executed at this step is available [here](#). It uses the following software:

DRIVEE [download](#) [citation](#)

drivee -r -t -f format -f <input>

Jellyfish [download](#) [citation](#)

jellyfish count -C -m <dlis> -c 12 -s 16 <input>

mgm4441680.3.075.assembly.abundance (temporary)

filesize NaN B

mgm4441680.3.075.upload\_stats (temporary)

filesize NaN B

mgm4441680.3.075.qc\_stats (temporary)

filesize NaN B

2. Adapter Trimming

Detection and removal of adapter sequences using a bit-mapped k-difference matching algorithm. The script executed at this step is available [here](#).

3. Denoising and normalization

Depending on the options chosen, the preprocessing step filters sequences based on length, number of ambiguous bases and quality values if available. The FASTX formatted file 100.preprocess.filtered.fna contains the sequences which were accepted and will be passed on to the next stage of the analysis pipeline. The FASTX formatted file 100.preprocess.removed.fna contains the sequences which were rejected and will not be passed on to the next stage of the analysis pipeline. The script executed at this step is available [here](#). It uses the following software:

DynamicTrim [download](#) [citation](#)

dynamictrim.pl <infile> -k <min\_qual> -n <max\_len>

mgm4441680.3.100.preprocess.filtered.fna

filesize 802.7 KB

MD5 -

mgm4441680.3.100.preprocess.removed.fna

filesize 18.3 MB

MD5 -

4. Removal of sequencing artifacts

PCR artifacts require removal: sequences are artificially duplicated during the preparation for sequencing (see <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2007242/>) for metagenomes and metatranscriptomes, the technique cannot be used for amplicon reads. The optional deduplication step removes redundant "technical replicator" sequences from the metagenomic sample. Technical replicates are identified by finding reads with identical first 50 base pairs. One copy of each 50-base-pair identical bin is retained. The FASTX formatted file 100.deduplication.filtered.fna contains the sequences which were retained and will be passed on to the next stage of the analysis pipeline. The FASTX

mgm4441680.3.100.deduplication.filtered.fna

filesize 18.0 MB

MD5 -

## General Information

Your fna dataset of 18.2 Mbp was submitted to version 3.0 of the MG-RAST pipeline at 2008-11-05 07:39:03 with priority 1.

You chose the following pipeline options for this submission:

assembled	no
dereplication	yes
screening	H. sapiens, NCBI v36
publication	never
length filtering	yes
length filter deviation multiplier	2.0
ambiguous base filtering	yes
maximum ambiguous basepairs	5

The computational environment and workflow can be downloaded below:

- [environment](#)
- [workflow document](#)

# From MG-RAST Pipeline Products to Research Object

- AWE v2 – CWL workflow
  - Could use CWLprov ?
- API call for RO json manifest
  - [http://api-dev.mg-rast.org/researchobject/manifest/\\${metagenomeID}](http://api-dev.mg-rast.org/researchobject/manifest/${metagenomeID})
- Command line script to create RO like directory
  - ***mg-export-research-object.py --metagenome mgm4441680.3***

> ls

data manifest-sha1.txt metadata snapshot tagmanifest-sha1.txt workflow

```
{
  aggregates: [
    {
      bundledAs: {
        folder: "/data/",
        filename: "mgm4441680.3.050.upload.fna"
      },
      mediatype: "text/plain; charset=UTF-8",
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=050.1"
    },
    {
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=100.1",
      mediatype: "text/plain; charset=UTF-8",
      bundledAs: {
        folder: "/data/",
        filename: "mgm4441680.3.100.preprocess.passed.fna"
      }
    },
    {
      mediatype: "text/plain; charset=UTF-8",
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=100.2",
      bundledAs: {
        filename: "mgm4441680.3.100.preprocess.removed.fna",
        folder: "/data/"
      }
    },
    {
      bundledAs: {
        filename: "mgm4441680.3.150.dereplication.passed.fna",
        folder: "/data/"
      },
      mediatype: "text/plain; charset=UTF-8",
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=150.1"
    },
    {
      bundledAs: {
        folder: "/data/",
        filename: "mgm4441680.3.150.dereplication.removed.fna"
      },
      mediatype: "text/plain; charset=UTF-8",
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=150.2"
    },
    {
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=299.1",
      bundledAs: {
        filename: "mgm4441680.3.299.screen.passed.fna",
        folder: "/data/"
      },
      mediatype: "text/plain; charset=UTF-8",
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=350.1"
    },
    {
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=425.1",
      bundledAs: {
        filename: "mgm4441680.3.425.genecalling.coding.faa",
        folder: "/data/"
      },
      mediatype: "text/plain; charset=UTF-8",
      uri: "http://api-dev.mg-rast.org/download/mgm4441680.3?file=425.1"
    }
  ]
}
```

Making request https://api.mg-rast.org//researchobject/manifest/mgm4441680.3  
Cloning into './H0Nabate3w'...  
remote: Enumerating objects: 47, done.  
remote: Compressing objects: 100% (32/32), done.  
remote: Total 47 (delta 14), pack-reused 9138  
Unpacking objects: 100% (47/47), done.  
Downloading mgm4441680.3.050.upload.fna ... Done  
Making request https://api.mg-rast.org//download/mgm4441680.3?file=100.1  
Downloading mgm4441680.3.150.dereplication.removed.fna ... Done  
Making request https://api.mg-rast.org//download/mgm4441680.3?file=100.2  
Downloading mgm4441680.3.150.dereplication.passed.fna ... Done  
Making request https://api.mg-rast.org//download/mgm4441680.3?file=150.1  
Downloading mgm4441680.3.150.dereplication.removed.fna ... Done  
Making request https://api.mg-rast.org//download/mgm4441680.3?file=299.1  
Downloading mgm4441680.3.299.screen.passed.fna ... Done  
Making request https://api.mg-rast.org//download/mgm4441680.3?file=350.1  
Downloading mgm4441680.3.350.genecalling.coding.faa ... Done  
Making request https://api.mg-rast.org//download/mgm4441680.3?file=425.1

# Open Issues / Gaps (in our understanding?)

- End user tooling to view/edit/parse RO's
  - Our domain has a reproducibility crisis today
- How do domain specific parts creep in

# The MG-RAST team and friends

## ANL/UChicago

- **Wolfgang Gerlach**
- Travis Harrison
- Will Trimble
- **Andreas Wilke**
- Sarah Owens
- Stephanie Greenwald
- Dion Antonopoulos

## Purdue collaborators:

- Ananth Grama
- Saurabh Bagchi
- Somali Chaterji

## EBI collaborators:

- **Rob Finn**
- **Guy Cochrane**
- Alex Mitchell

## GSC

Pelin Yilmaz, MPI Bremen

## UAMS, Little Rock

Dave Ussery, UAMS  
Intawat Nookaew, UAMS  
Tip Wongsurawat, UAMS

## MG-TAP

Martin Hartmann  
Michael Crusoe







## ***mg-export-research-object.py --metagenome mgm4441680.3***

```
# mg-export-research-object.py --metagenome mgm4441680.3
Making request https://api.mg-rast.org//metagenome/mgm4441680.3
Making request https://api.mg-rast.org//researchobject/manifest/mgm4441680.3
Cloning into './H0Nabate3w'...
remote: Enumerating objects: 47, done.
remote: Counting objects: 100% (47/47), done.
remote: Compressing objects: 100% (32/32), done.
remote: Total 9185 (delta 21), reused 29 (delta 14), pack-reused 9138
Receiving objects: 100% (9185/9185), 8.35 MiB | 4.15 MiB/s, done.
Resolving deltas: 100% (5874/5874), done.
Making request https://api.mg-rast.org//download/mgm4441680.3?file=050.1
Downloading mgm4441680.3.050.upload.fna ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=100.1
Downloading mgm4441680.3.100.preprocess.passed.fna ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=100.2
Downloading mgm4441680.3.100.preprocess.removed.fna ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=150.1
Downloading mgm4441680.3.150.dereplication.passed.fna ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=150.2
Downloading mgm4441680.3.150.dereplication.removed.fna ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=299.1
Downloading mgm4441680.3.299.screen.passed.fna ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=350.1
Downloading mgm4441680.3.350.genecalling.coding.faa ... Done
Making request https://api.mg-rast.org//download/mgm4441680.3?file=425.1
```