Lessons learned from operating a big metagenomics resource

Folker Meyer, PhD

Argonne National Laboratory
and
University of Chicago

Introduction

Sequence quality

Assembly

Gene prediction

Functional annotation

Current MG-RAST work

Metagenomics

Environmental clone libraries (“functional metagenomics”)

- Sanger sequencing of BAC clones with env. DNA

→ low throughput but supports in vitro screens

➔ **Discovery of novel functions**

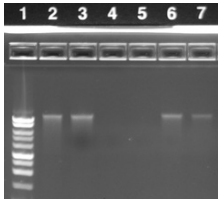
Amplicon studies (single gene studies, 16s rDNA)

- Sequencing of PCR amplified ribosomal genes

→ sequence quality limited (→ rare biosphere debate)

→ often can’t distinguish between individual strains

➔ **Ecology**

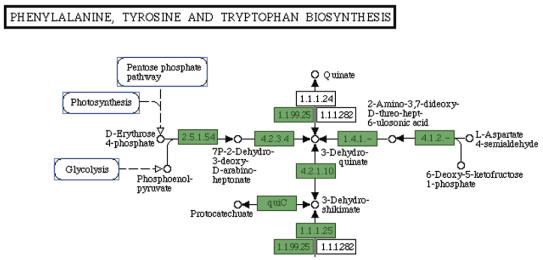


Shotgun metagenomics

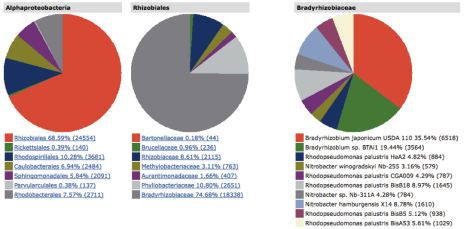
- “random shotgun DNA sequencing applied directly to environmental samples”

➔ **Consensus sequences & Ecology**

What are they doing?



Who are they?



Why I am talking about this?

MG-RAST
metagenomics analysis server

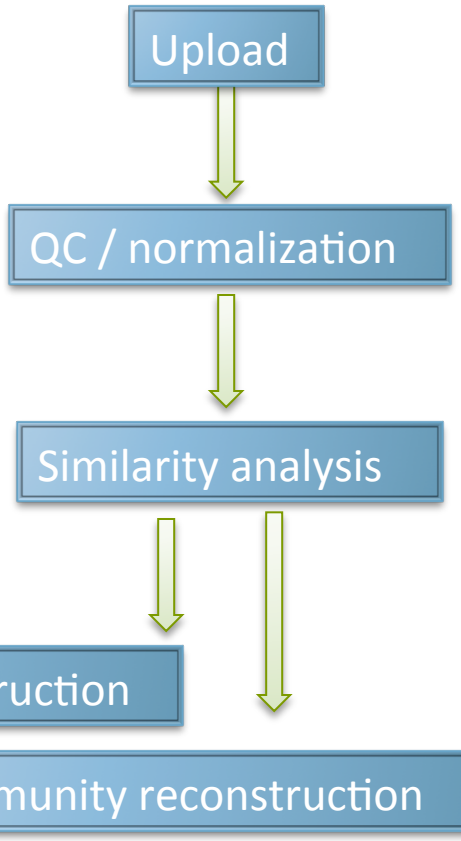
# of metagenomes	111,797
# base pairs	44.18 Tbp
# of sequences	380.7 billion
# of public metagenomes	16,791

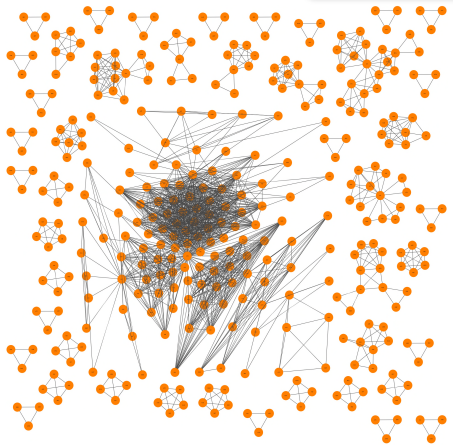
13,000 users
>100,000 data set
>~1.5 Terabases per month

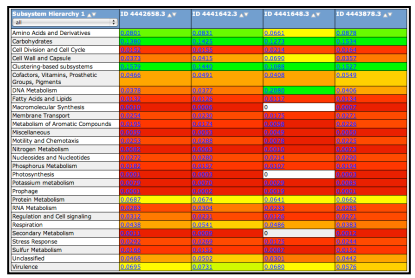
Meyer et al., BMC Bioinformatics, 2008

Brief history of MG-RAST

- December 2007 (v1)
  - 100+ groups and ~250 data submitters
  - 100+ data sets, **~10+ GBp** total size
- October 2009 (v2)
  - Pre-publication sharing available
  - ~1500 data submitters, ~300 public data sets
  - 6000+ data sets
  - **200+ GBp**
- March 2011 (release v3)
  - 2500+ data submitters
  - ~2000 public data sets
  - 25,000 data sets total
- March 2012 (v 3.1.2)
  - **12 Terabasepairs** (10^12 bp)
- May 2012 (3.20)
  - **13.8 TBp** (10^12bp)
  - 128 billion sequences
- December 2013 (3.3.7.3)
  - **39 TBp**
  - 340 billion sequences
  - 100,000 jobs / 14,000 public







simplified

Introduction

Sequence quality

Assembly

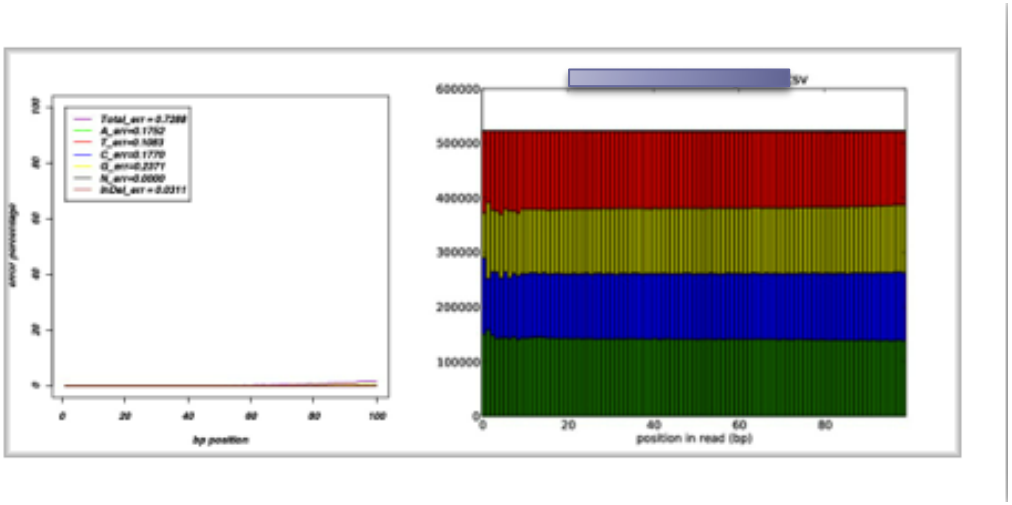
Gene prediction

Functional annotation

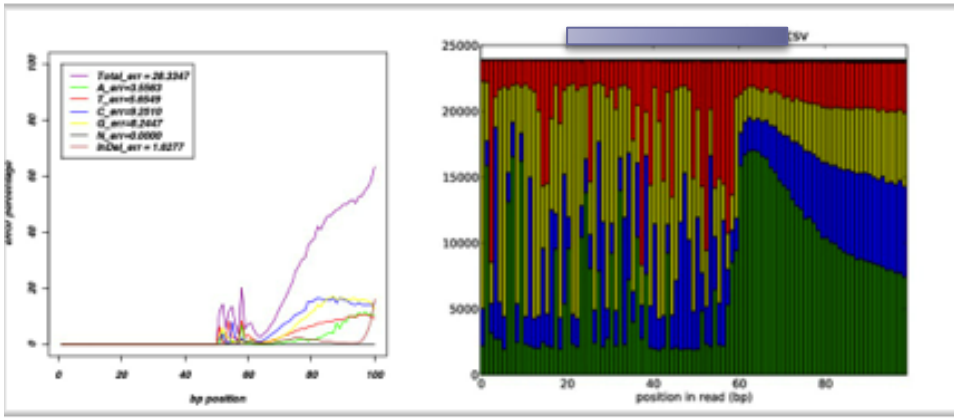
Functional annotation

Current MG-RAST work

Not all sequence data created equal



DRISEE: Keegan, et al, PLoS Comp. Bio, 2012



Every platform has unique issues

- Not all samples from one platform are the same
- Massive operator variation

- Most users don't understand sequence quality
- Majority of users lacks tools to gain insight into" sequence space"
- Sequence providers work hard to obscure issues

- FAQ: Shouldn't my student do the QC locally?
  - → sure he will do a much better job than us, we have only seen 10,000 similar data sets

Introduction

Sequence quality

Assembly

Gene prediction

Functional annotation

Current MG-RAST work

Two provoking examples:

- **Example 1:**
- **high diversity host microbiome sample**
  - One physical sample
  - two libraries, two sequencing runs

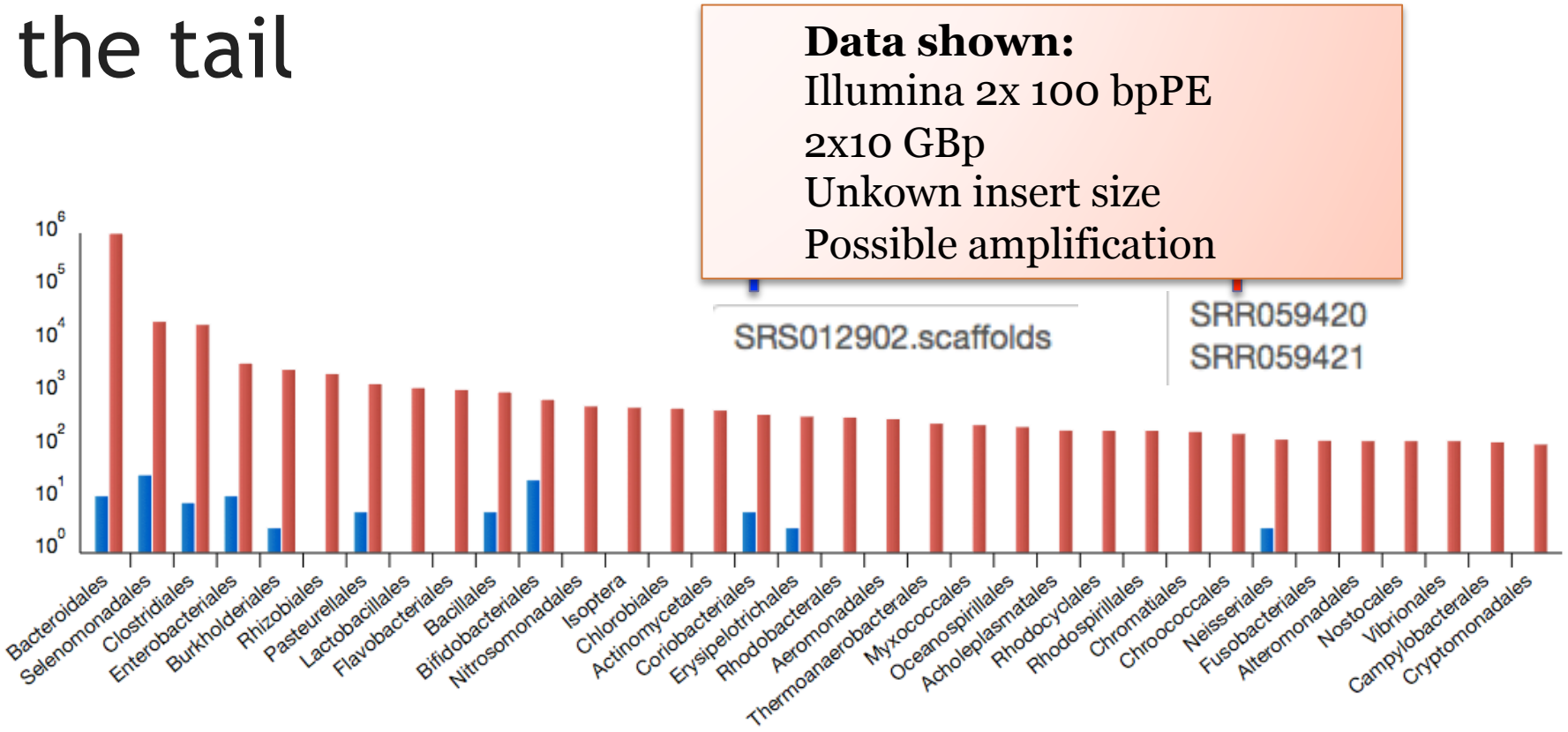
  → unexpected results

- **Example 2:**
- **Medium diversity environmental sample**
  - One physical sample
  - One library, one sequencing run

  → Very unexpected results

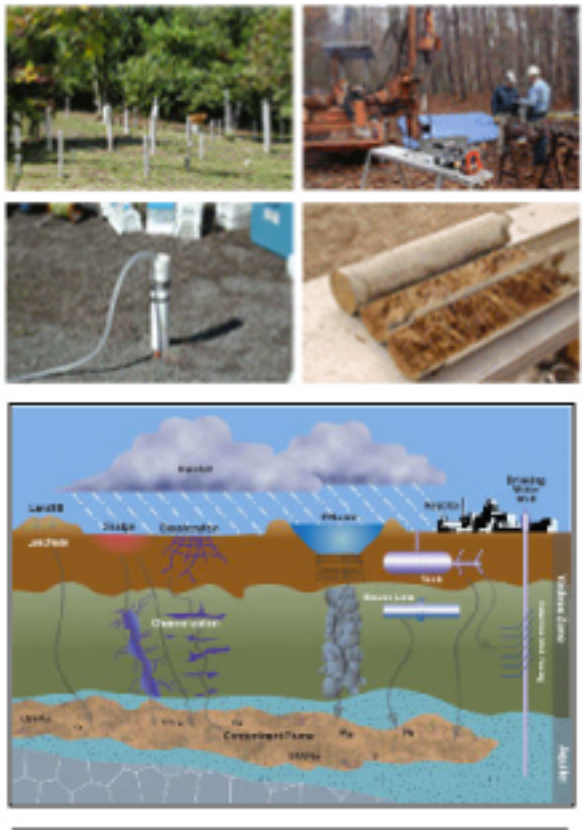
Example 1: Human microbiome

We are missing more than the rare taxa at the tail



I spent less time finding a sample with this problem then plotting the graph.

Example 2: A subsurface community



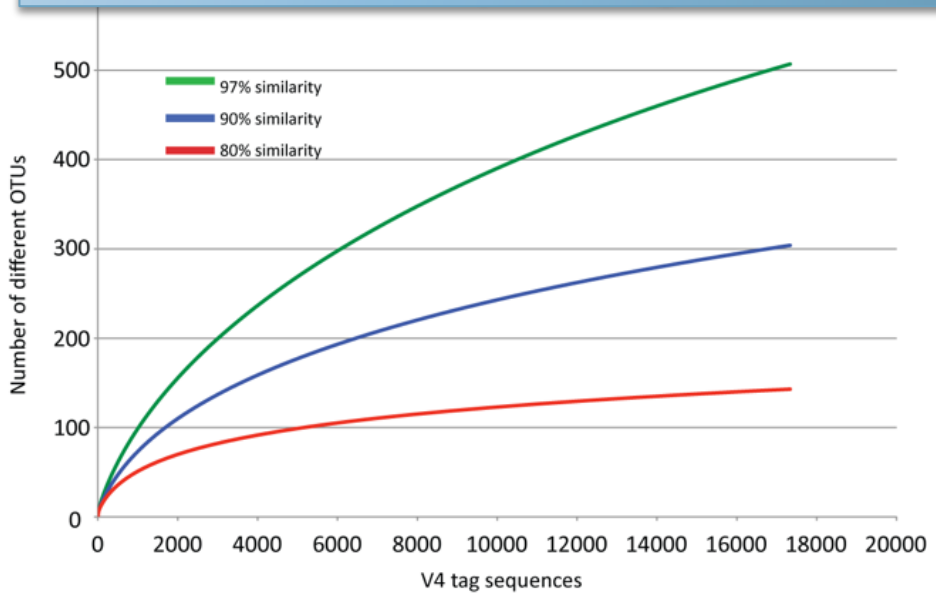
Images from DOE UMTRA web site

Data shown:
Illumina 2x 75bpPE
9.1GBp
440bp insert size
no amplification

- Sample from DOE cold war site (uranium contaminated)
- Research plan:
  - use unassembled Illumina data
  - Determine alpha diversity and species
  - Determine protein complement
  - Later attempt assembly
    - Use mate-pairs to verify

How complex is the community?

16s Rarefaction analysis using RDP classifier

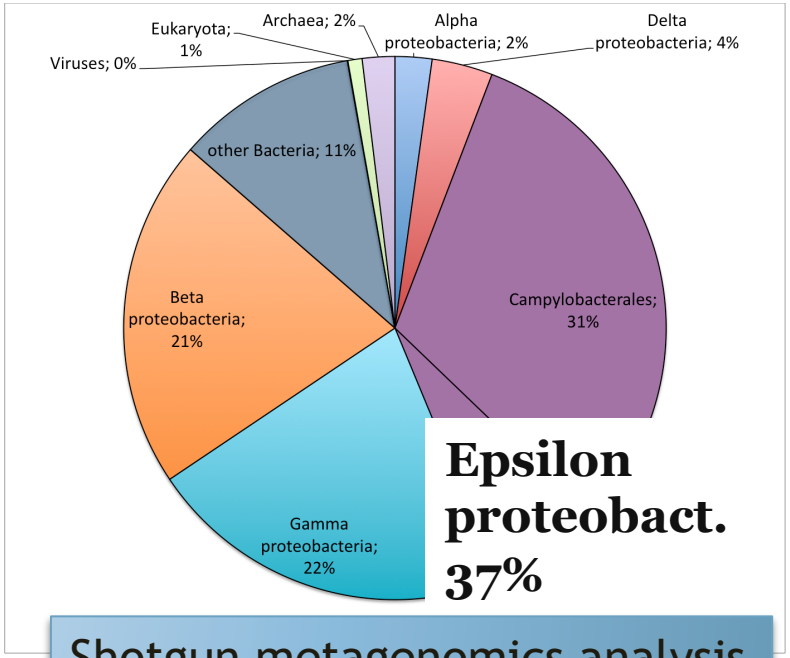


- Based on 17,334 V4 tag sequences
- 500+ "species" inferred at 97% sequence similarity to define the OTU
- **Proteobacteria (86.2%)**
  - **genus Sulfuricurvum (46.9%)**

➔ Assumption: easy assembly

RDP: Cole, et al, NAR, 2007

Shotgun metagenomics analysis



- Shotgun metagenome data confirm
- **~37% epsilon proteobacteria**
  - no Sulfuricurvum genome sequence

➔ Assumption: easy assembly

MG-RAST: Meyer, et al, BMC Bioinformatics. 2008

- Sulfuricurvum reads == ~38% of all reads
- Approx. 3.45GBp are *Sulfuricurvum* reads ➔ 45 million reads
- Assuming 3MBp genome size

Assembly should be easy ….

Expected:

```
Length of genome           G = 3000000 bp
Number of sequences        N = 45000000
Average length of sequences L = 75 bp
Minimum overlap            T = 30 bp

Redundancy of coverage                    = 1125.0
Expected number of contigs                =  1.0
The expected number of seqs in a contig   = 45000000.0
The expected length of each contig        = 3000000
Percentage of coverage of genome          = 100%
```

➔ **1100X coverage of genome**

Observed:

Resulting contigs:

244,782 Contigs
67Mbp
N50: 227bp

Long contigs >= 1000

5,967 Contigs

No long contigs from Epsilon proteobacteria

Lander and Waterman: Genomics, 1988

Zerbino et al, Genome Res. 2008

Sequence quality the source of the problem?

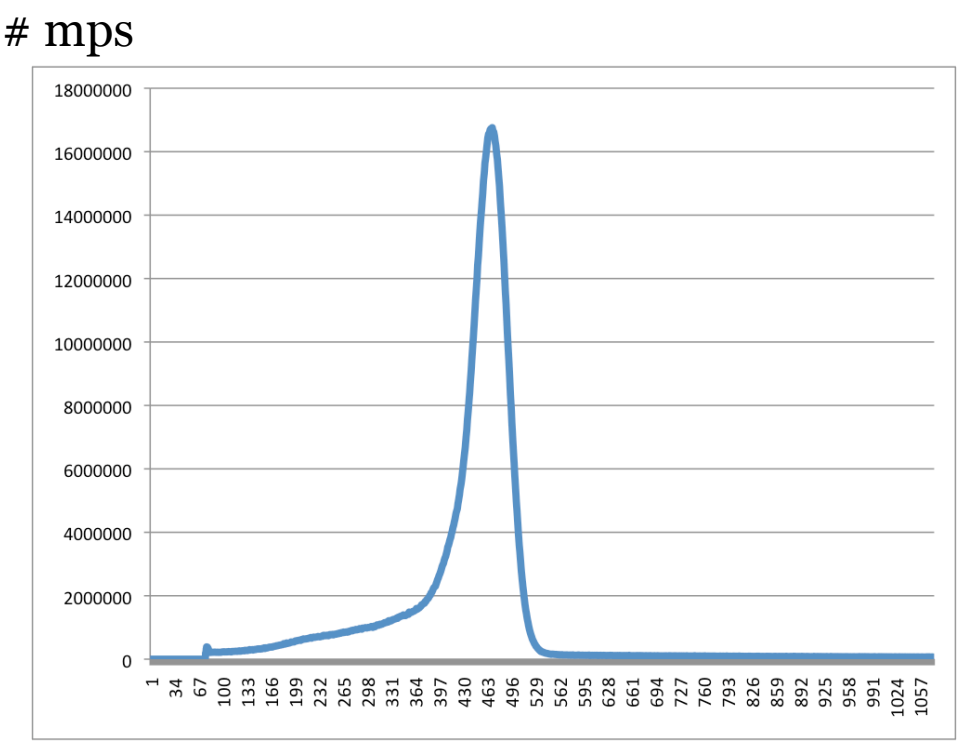
Two arguments against that:

a) Very good DRISEE score
INSERT SCORE HERE

b) Mate pairs map nicely to contigs
- mate percentage: 86.015%
- average distance: 446
  - As designed
- Using exact matches (300-1000 window)

➔ **NO! Look for other reasons**

Distribution of mate-pair "insert lengths"



Using exact bowtie matches

Langmead et al, Genome Biol. 2009

Assembly tool(s) or parameters?

Two options:

1) Velvet parameters are incorrect

→ we varied the k-mer length parameter

2) Velvet does not work properly?

→Will other assemblers produce similar result?

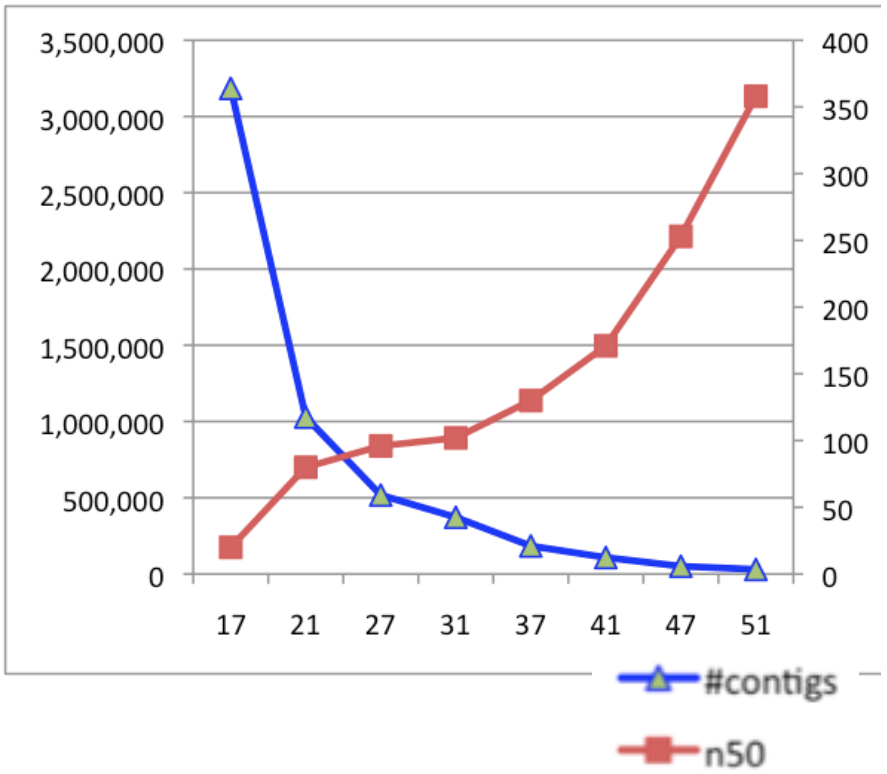
→We tried: SOAP de-novo, Newbler, Abyss, AMOS, AllPaths

- We learned many valuable lessons
- Requires significant computational resources
- Most tools can't handle 45 million reads

➔ similar results with other assemblers

Mihai Pop in 2010 @ HMP in St. Louis, MO
"metagenomes can't be assembled"
and
"all assemblers are equally bad"

Effects of changing k-mer lengths on contig number and sizes



Short k-mers == many contigs but short
Long k-mers == few long contigs

Too much coverage?

- **Observations:**
  - Contigs yield is very low
  - varying parameters yields different contigs
  - Genome yield is low for all cases
- What if…
  - Coverage was not too low, but too high?

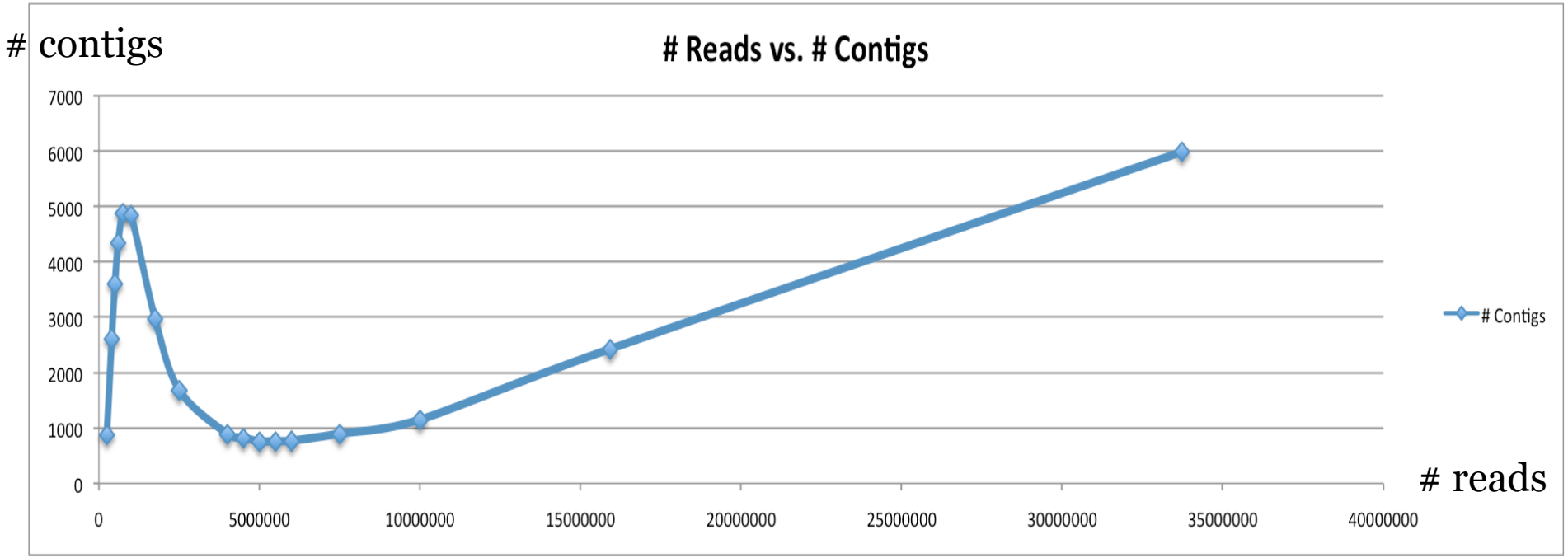
- **Hypothesis**:

→ Strain variation is confusing assembly tools

- IFF high coverage confuses assemblers by exposing strain variation, lowering coverage might help.
  - **Test strategy**: Subsampling

In addition: Integrating different parameter sets will overcome parameter set artifacts

The effects of subsampling



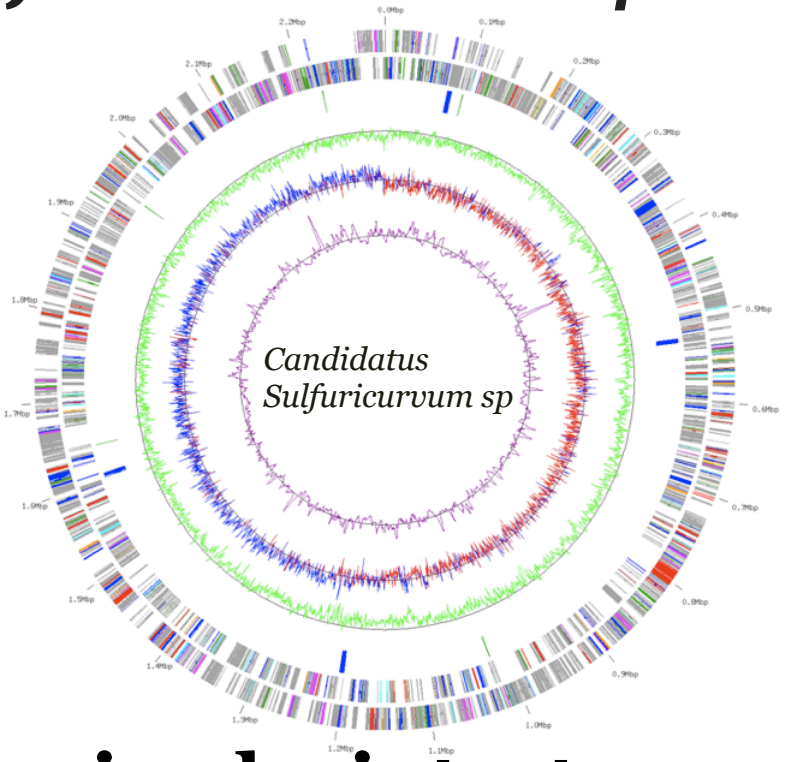
Plotting number of contigs vs. number of reads
→ More reads with variation confuse velvet (and the other tools)

→ Existing assembly tools have been built for clonal situations

→ Even modern tools like metaVelvet, meta-IDBA, metaSOAP will succumb to too much strain variation

Result: *Candidatus Sulfuricurvum sp.*

- new genus
- Epsilon proteo bacteria
- *Campylobacteriales*
- 2.3 MBp genome
- 1 contig
- 43% GC
- High quality



- **Verified by many overlapping by intact mate-pairs**

→ likely to not be chimeric

Genome annotation with RAST: Aziz et al, BMC Genomics, 2008
Genome published: Handley et al, AEM, 2014 (in press)

State of the art
(yes I am being provocative)

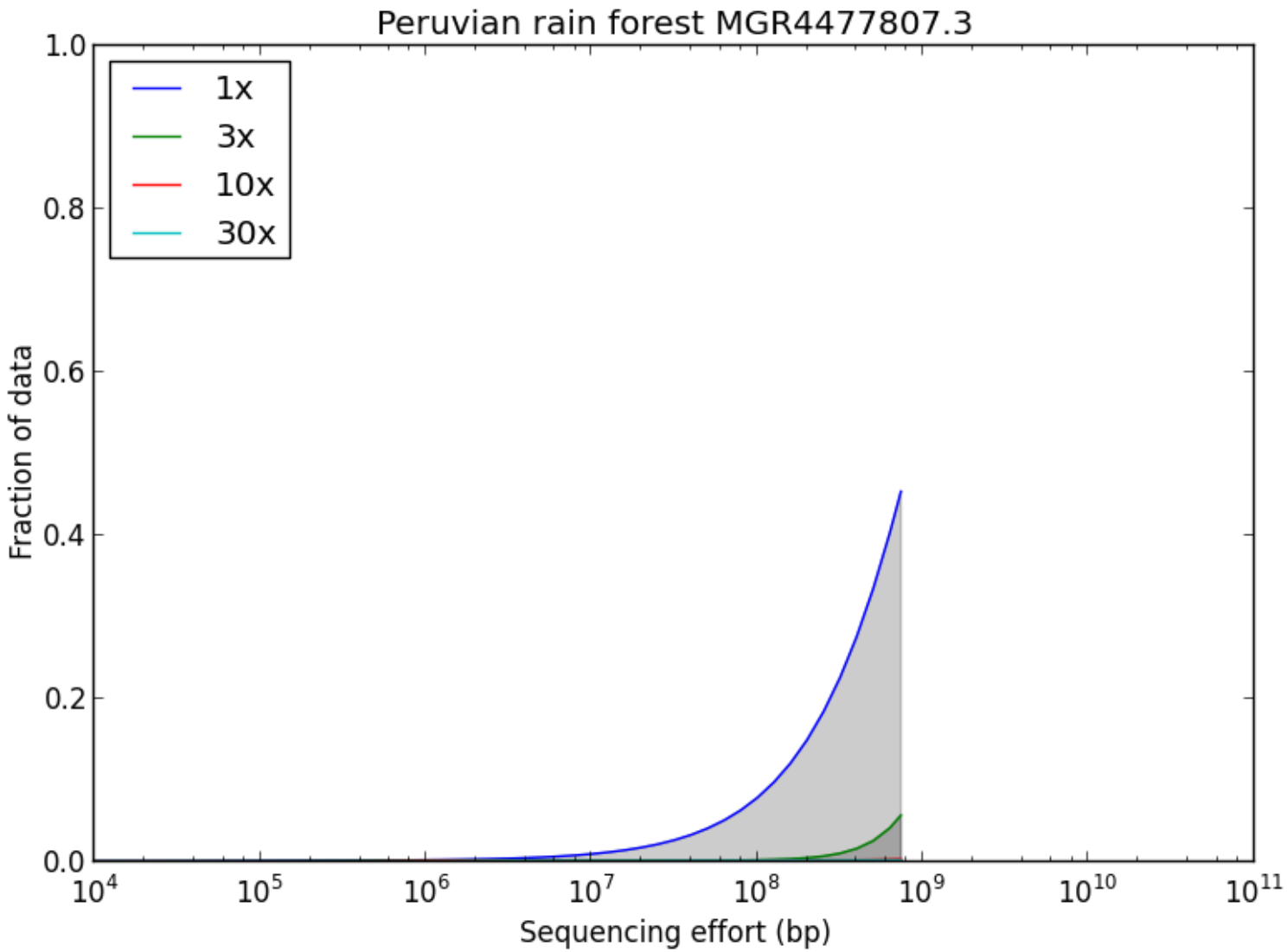
- Apply a single algorithm
  - Assembly, clustering, etc.
- Assume optimal results
- Proceed
- Data sets are sizable, trying to reduce them to manageable size quickly

Alternative:
Using nonpareil-K (npK) to understand coverage strata **first**

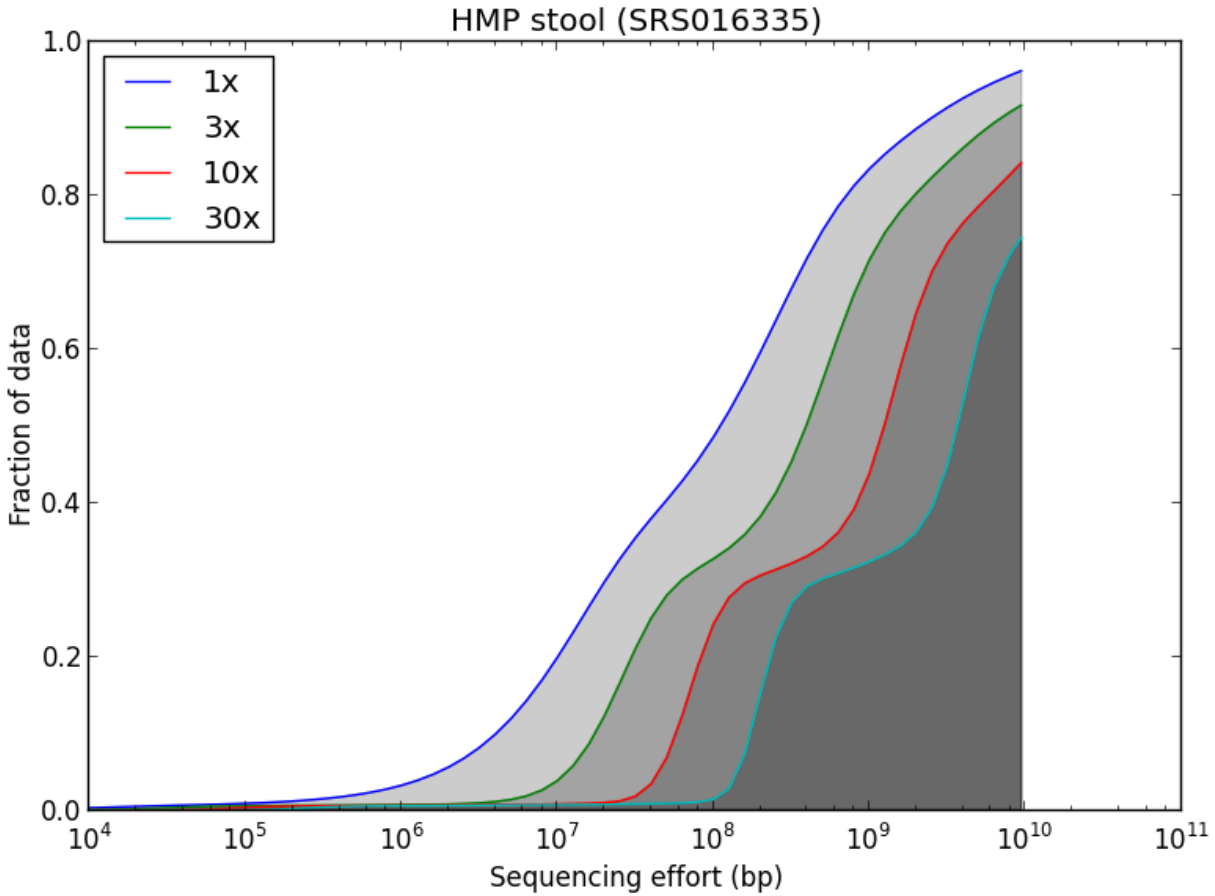
- npK evolved from
  - Nonpareil (Rodriguez & Konstantinidis, Bioinformatics, 2013)
  - and Kmer-spectrum-analyzer (Williams et al, BMC Genomics 2014)

- npK: does not use sub-sampling, instead uses Kmer index
- Allows rapid classification “strata”
  - Next slides explain the term ☺

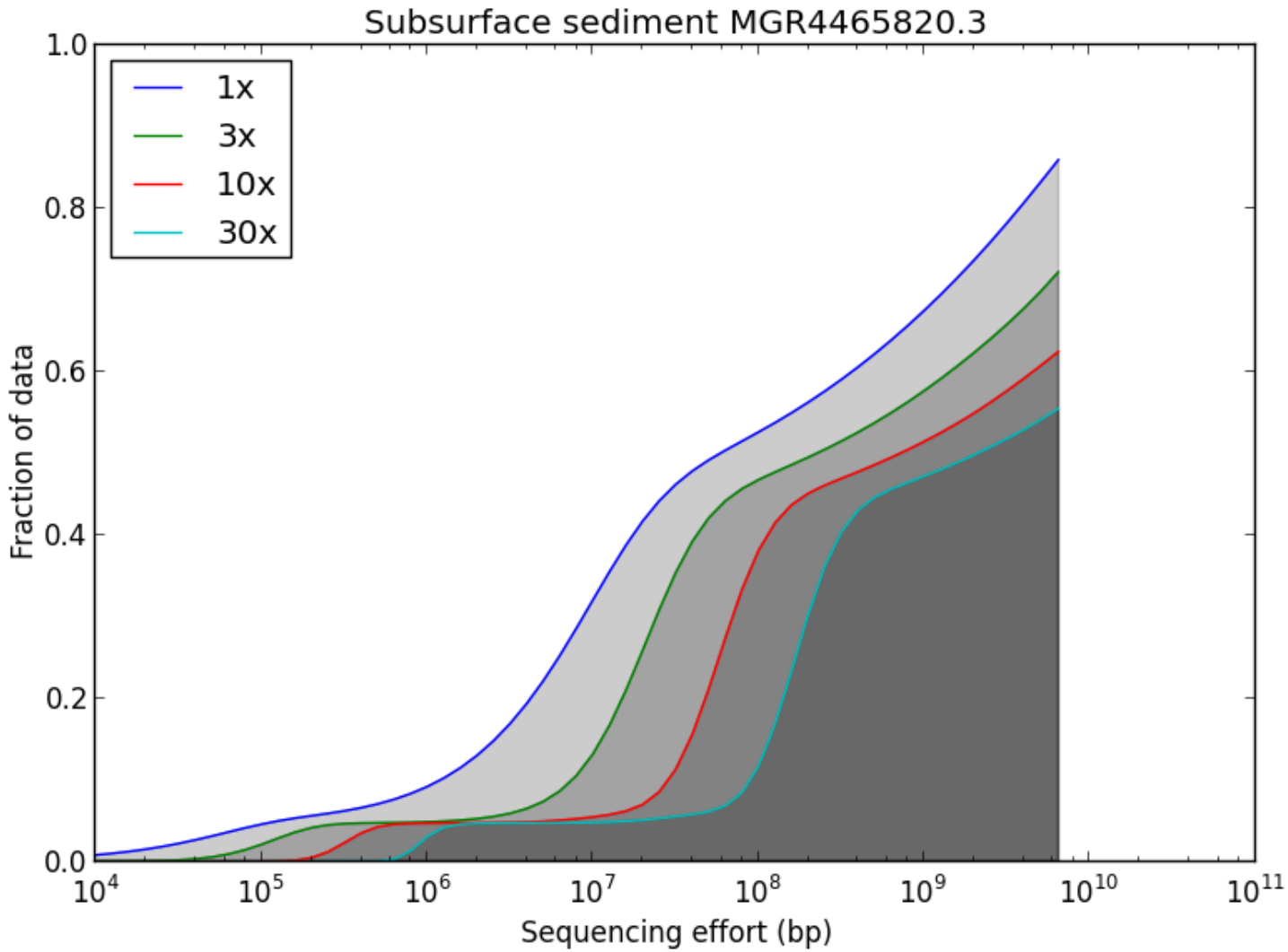
A soil sample



A microbiome sample



A subsurface sample



Introduction

Sequence quality

Assembly

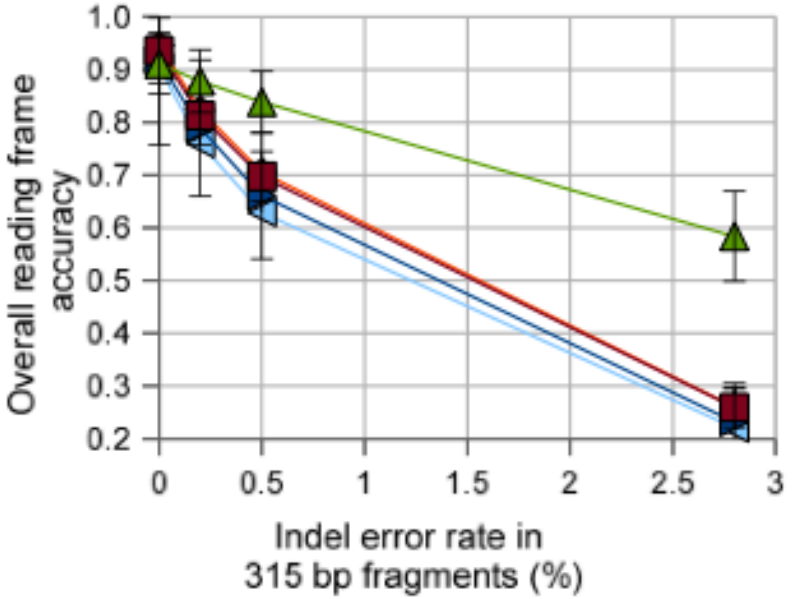
Gene prediction

Functional annotation

Current MG-RAST work

Not all gene callers created equal

- Question:
  - What happens if I vary the tool chain?
- **Existing approaches** rely on:
  - Compare results of different studies (ie multiple pipelines)
- Here we study 5 different popular gene finding tools for metagenome on **simulated data**
- **Effects are dramatic**
  - Accuracy goes drops dramatically with moderate error
- Comparison of data requires identical tool chain



From: Trimble et al, BMC Bioinformatics, 2012

Introduction

Sequence quality

Assembly

Gene prediction

Functional annotation

Current MG-RAST work

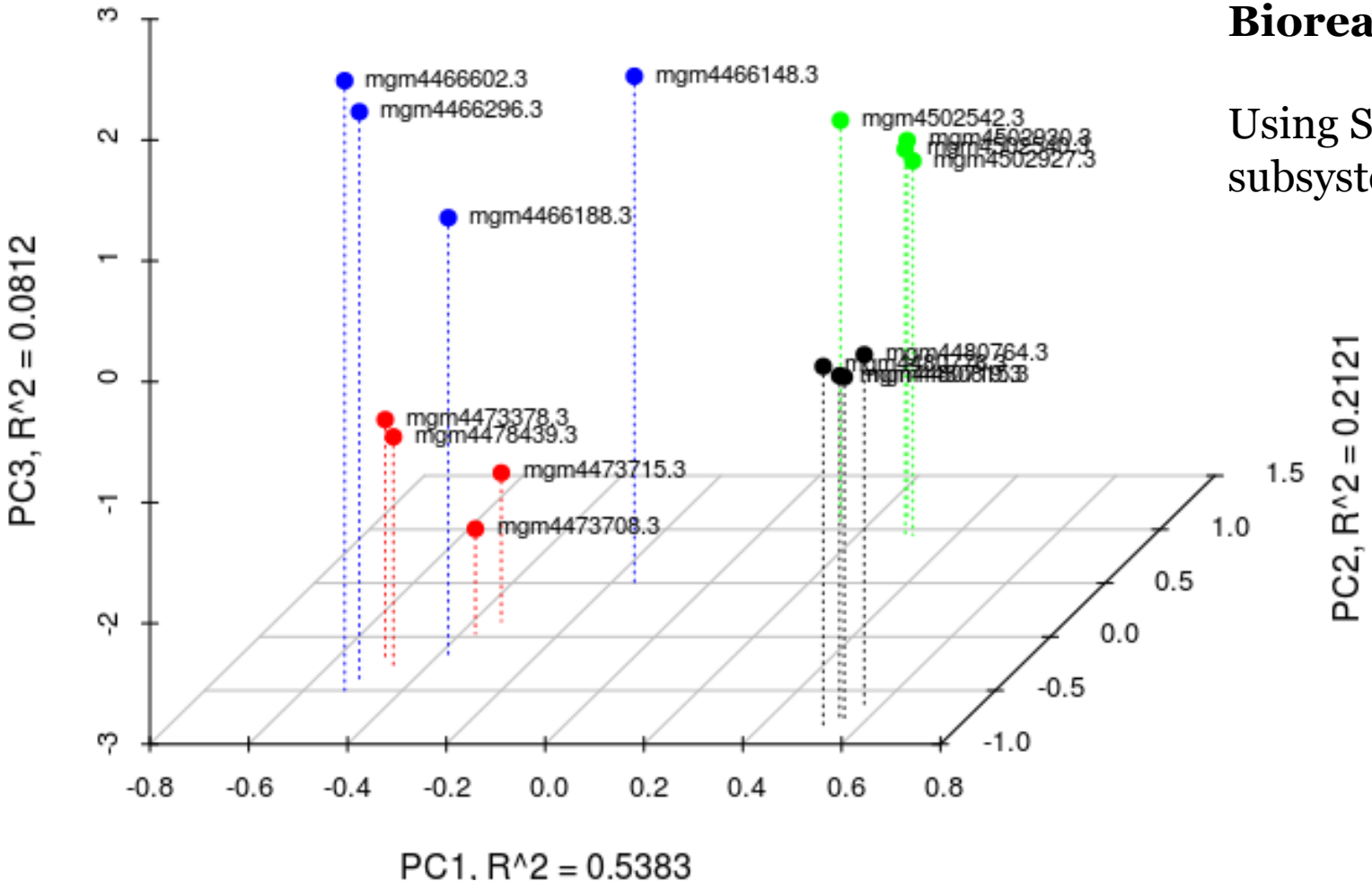
Experiment: Does metagenomics work?

- Simple test, pick 4 different sets of samples

- Red: Human microbiome, different body parts (sizes from 1-16GBp each)
- Blue: Feces from pregnant women in Finland (~12GBp each)
- Green: Soil in Illinois (~12GBp each)
- **Black**: Bioreactor (Beer-to-Caproate) 4 points in a time series (3GBp each)

PCoA - four projects

Shotgun metagenomics works → Yeah!



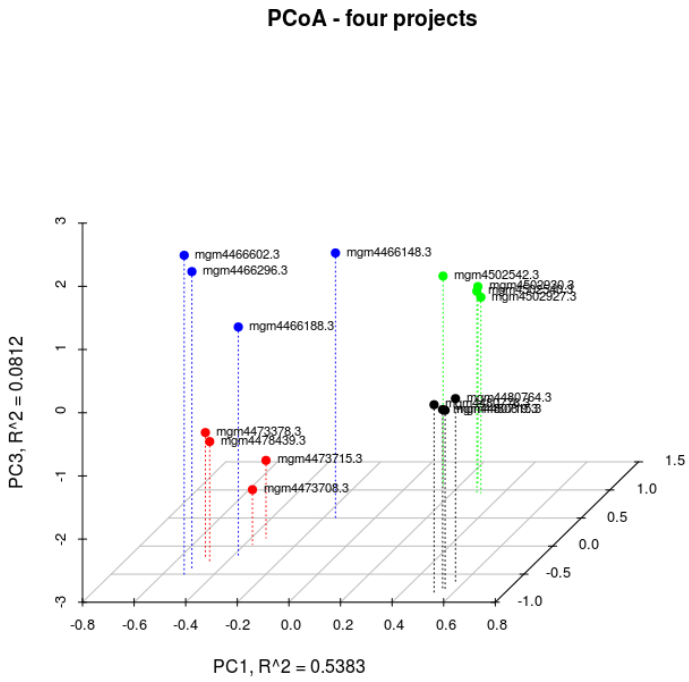
Human microbiome
Feces
Soil
Bioreactor

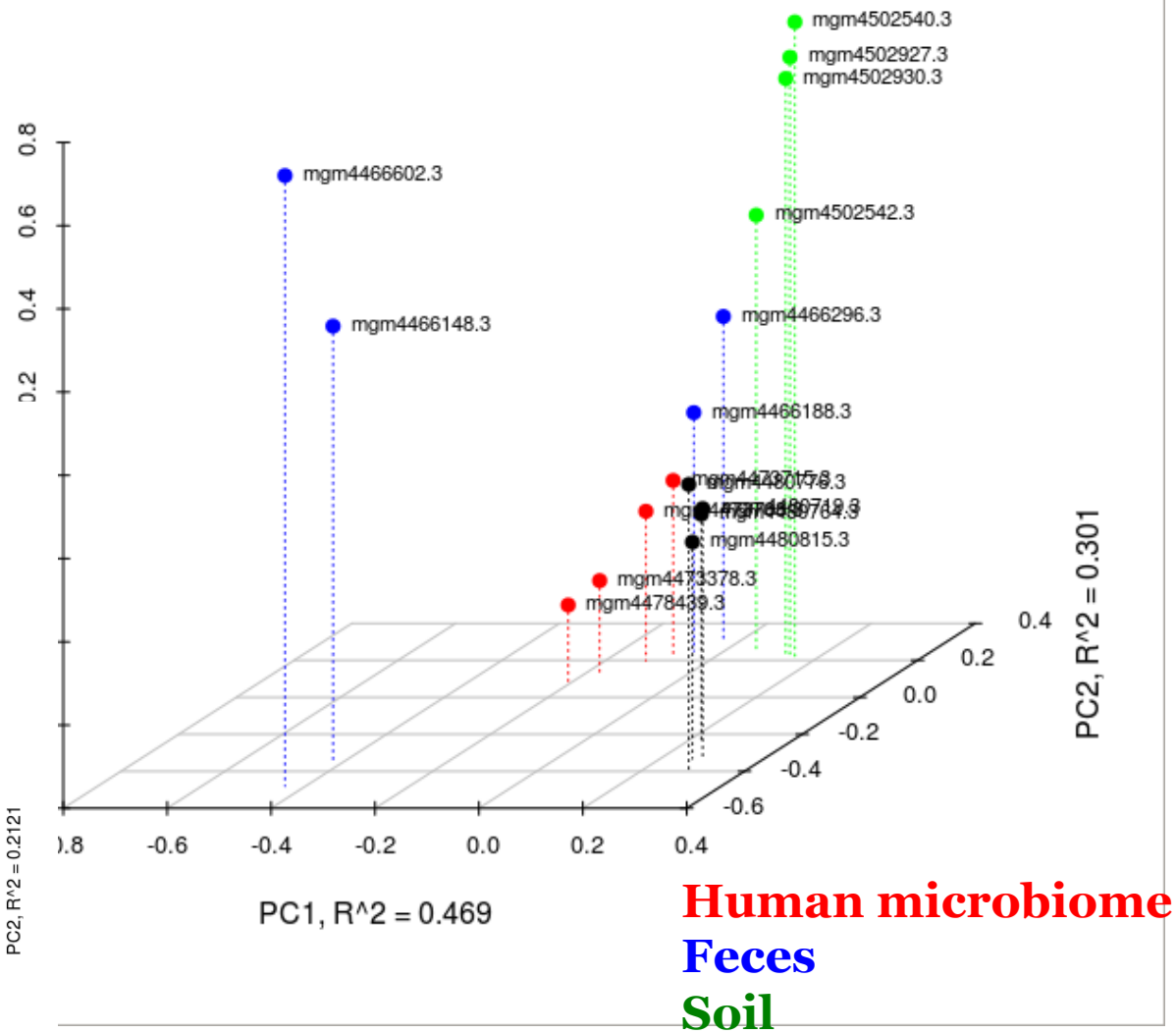
Using SEED subsystems Level 3

Or does it?

Human microbiome
Feces
Soil
Bioreactor

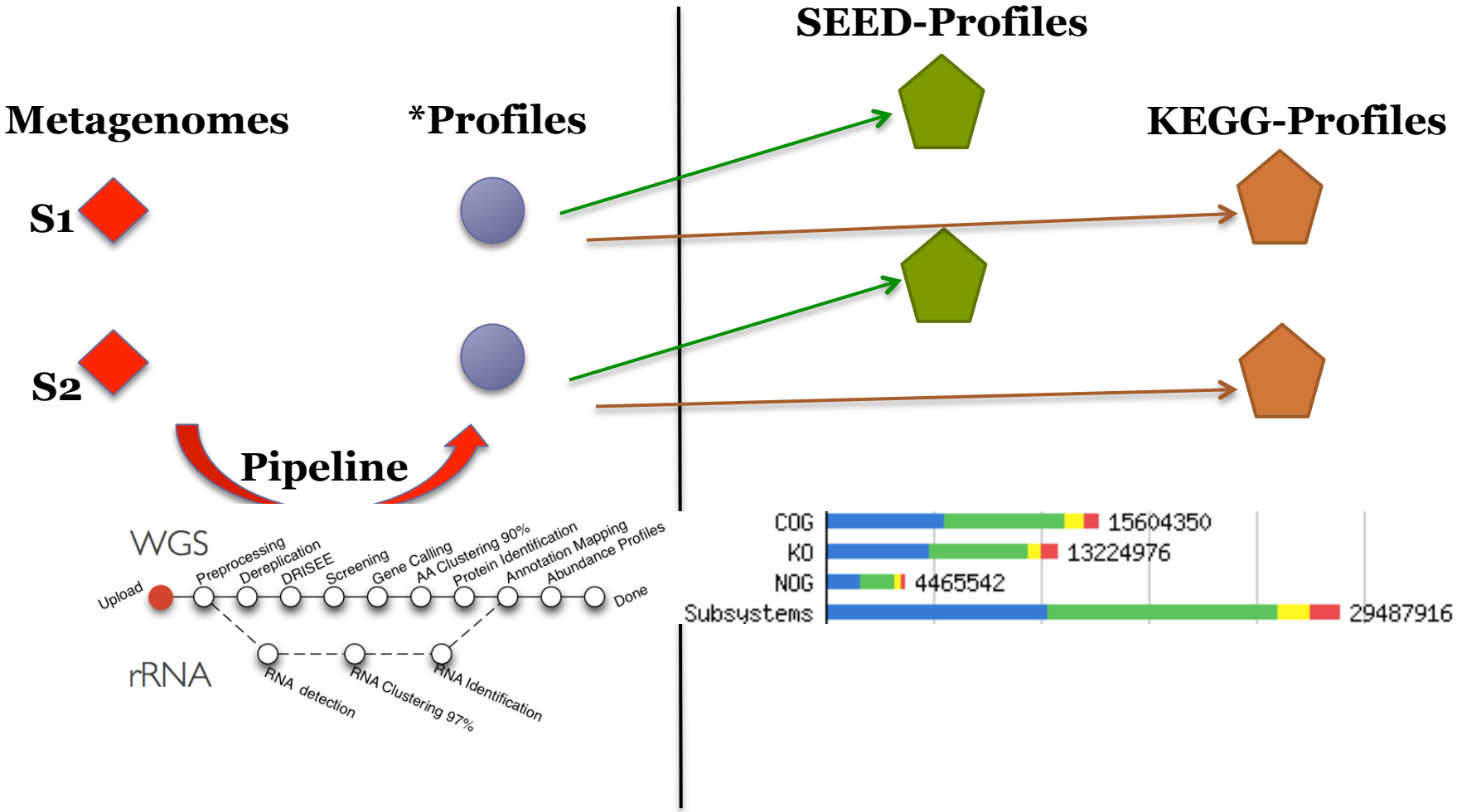
KEGG instead of SEED





Human microbiome
Feces
Soil
Bioreactor

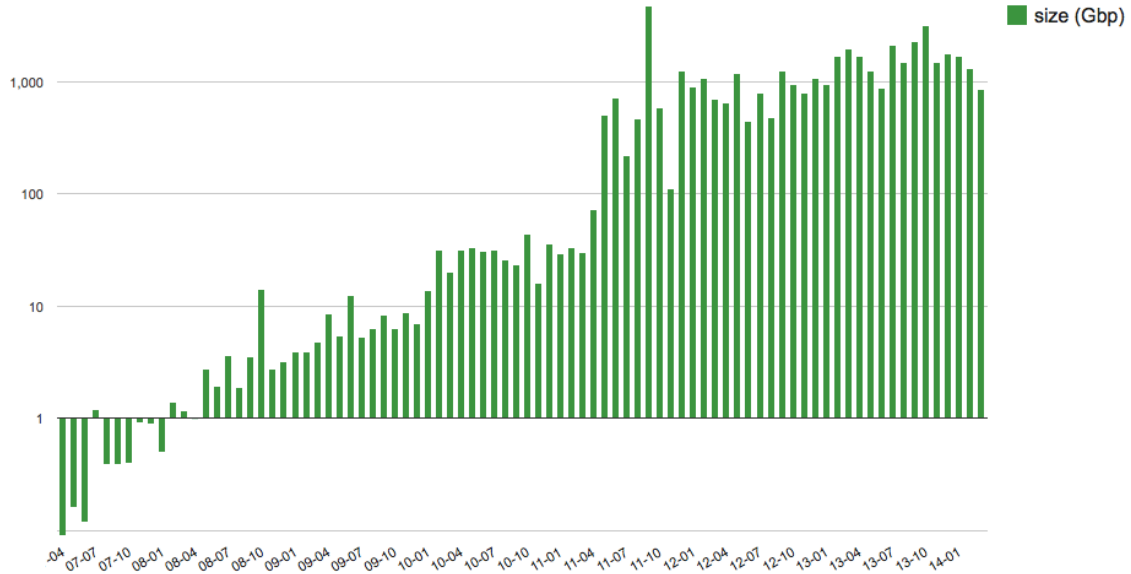
What has happened?



Introduction
Sequence quality
Assembly
Gene prediction
Functional annotation
Current MG-RAST work

Today

MG-RAST
metagenomics analysis server

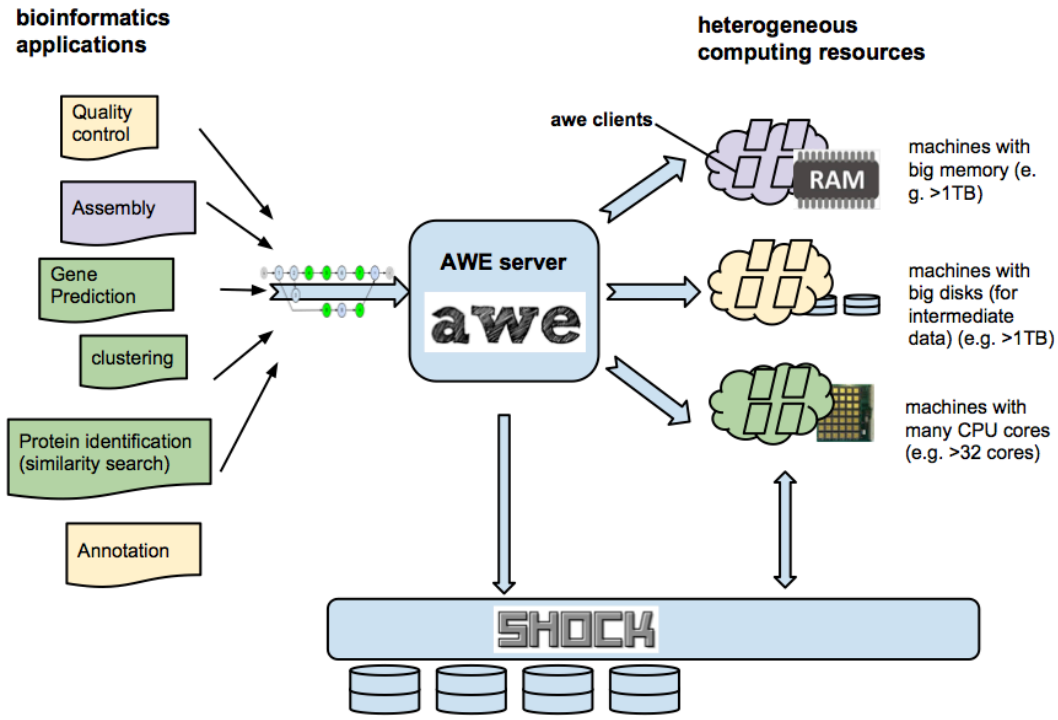


KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

From a capability problem to a capacity problem

built engineering tools accelerate platform



Common analysis tasks

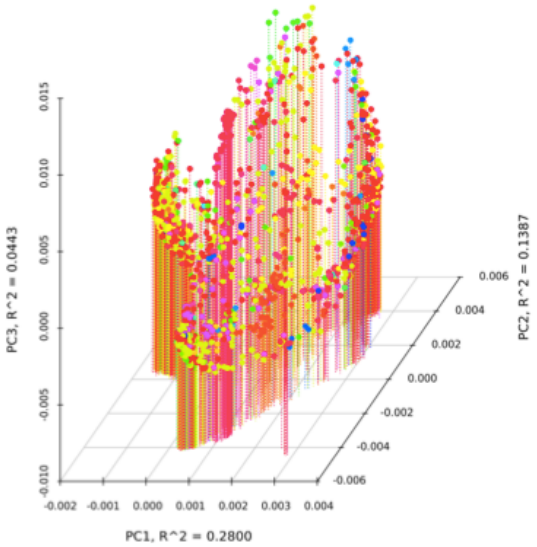
KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

- Assign taxonomy and function at scale
  - User sets parameters for annotation transfer at analysis time
- analyze a single sample
  - Taxonomy, rarefaction, function (if WGS)
  - K-mer profile, functional categories
  - QC tools
- Compare many samples wrt function and taxonomy
  - Extract subsets e.g. functions for set of species or species for a set of functions
  - Use subsets for comparison
- Compute normalizations, heatmaps, PCoA, tables

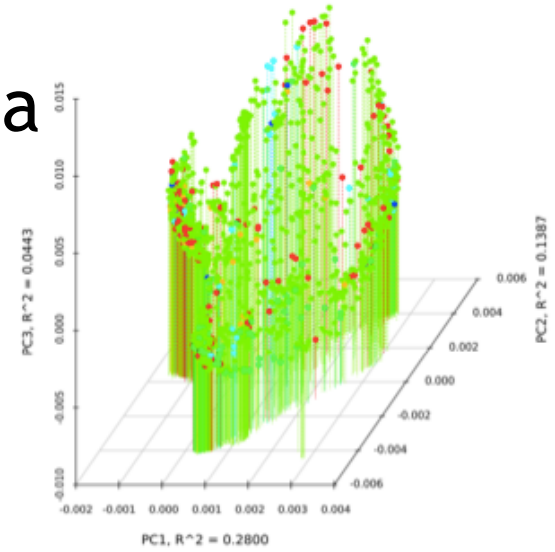
KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Example: **large scale comparison**

- Analysis of **2075 soil WGS data sets**
- clustered by normalized subsystem abundance



- Painted with metadata



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Example: **scaling**

- Use of computational appliances for metagenomic analysis
  - Appliances are self contained virtual machines (VMs)
  - Throughput **over 100GBp/day**, can scale much higher
  - Using KBase and non-KBase computers

- Have **third party users contribute computing** from their institution to the analysis of their data
  - e.g. UOregon (Jessica Green's group), OSU, Hudson Alpha, University of Oklahoma

- Provide a unifying API
- Move to Kmer based analysis whenever possible
- Re-Build in Kbase using novel technology
- Apply for big data analytics

Unifying API

- REST and JSON-RPC API
- Supports multiple levels of access
- R, Perl, Python, Java, …

- Re-building Web-UI using API currently

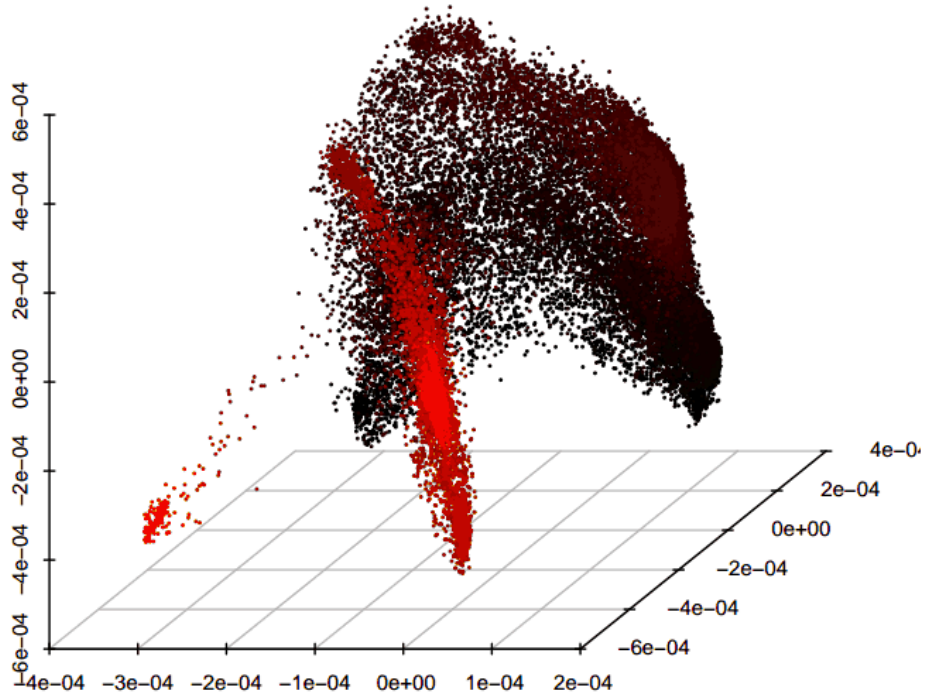
- Opening up formerly closed application is liberating

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Big data analytics in Bio:: compare many environmental samples

28k shotgun metagenomes
Normalized Subsystem level 3
Red=foreground

Using matR
R language client



Comparing metabolic models of a microbial community using different classes of sequence data and computational methods.

- Metabolic models can help understand interactions and dependencies of community members
- Identification of cultivation conditions for “unculturable” members enables use of additional research techniques, e.g. wet lab
- Deeper understanding of microbial communities may allow to predict the effect of modifications in culture conditions
- Possible applications in industry and medicine (e.g. optimization of the production of biofuels and biopharmaceuticals)

- Different various methods produce differing results

➔ **VERY MUCH WORK IN PROGRESS** ←

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

From reads to taxa / functions

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

- Functional genes
  - From assembled contigs (e.g. velvet, meta-velvet, KiKi, IDBA-UD, ….)
  - From assembled contigs and singletons
  - From clustering
- 16s ribosomal genes
  - From assembled contigs (see above)
  - From assembled contigs and singletons
  - From clustering
  - From EMIRGE (Miller et al, Genome Biology, 2011)

From taxa / reads to models

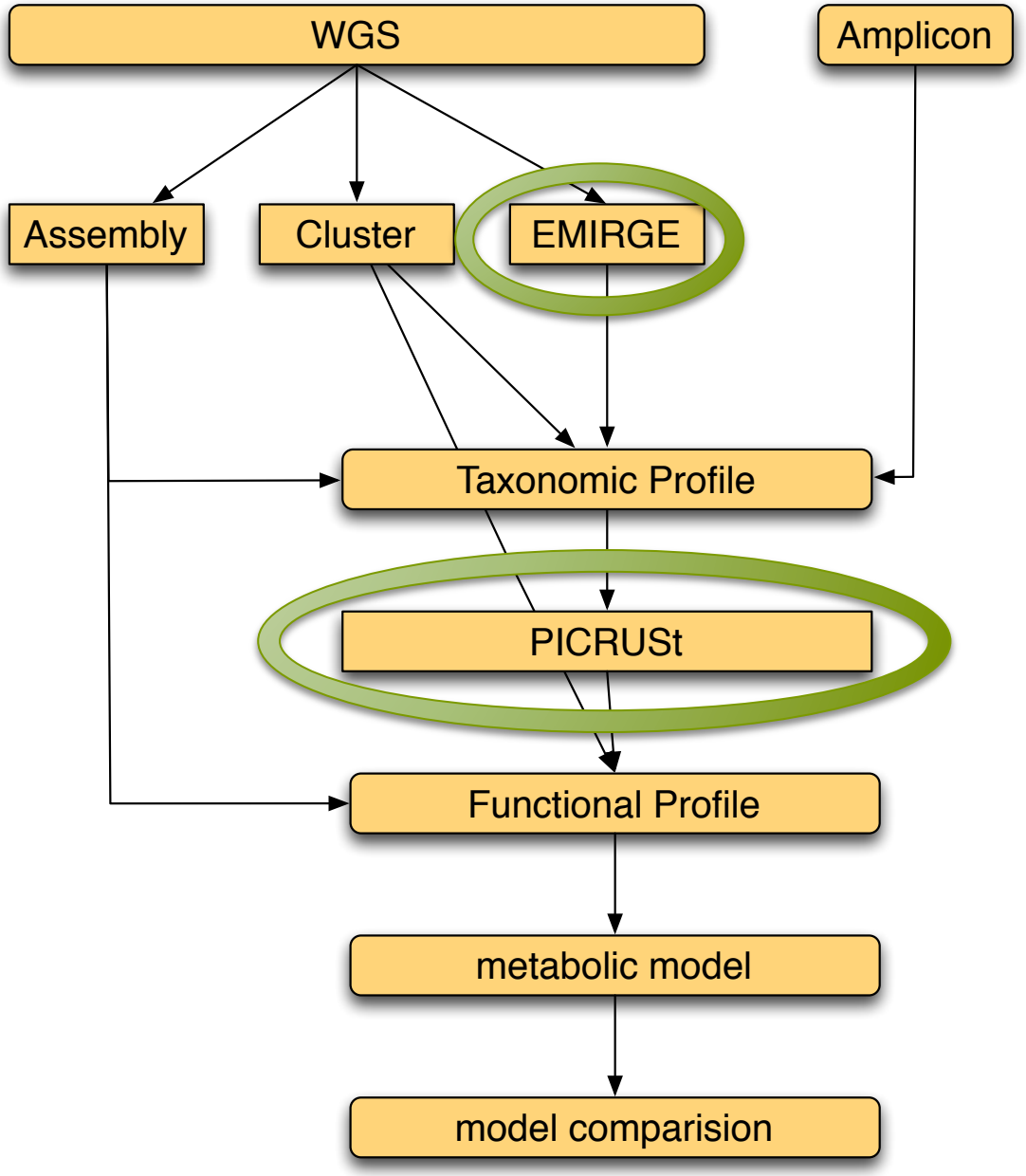
- Functions derived from observed genes
  - using Kbase Communities pipeline (based on Meyer *et al*, BMC Bioinf. 2008)

- Functions inferred via 16s proxy
  - using PICRUSt (Langille, *et al* Nature Biotechnology, 2013)

➔Next ModelSEED to create model and fill gaps
  - (Henry, *et al* Nature Biotechnology, 2010)

For all steps: lengthy, expensive computational tasks

- **Input**: environmental shotgun sequences
- **Output**: several models of community metabolism



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Taxonomic reconstructions on the same data

treatment	GBp	sequences	# 16S	# OTUs-97%	ID
16S amplicon (Meyer et al, BMC Bioinformatics 2008)	0.005	20,913	3,650	2269	mgm4506684.3
Clustering (Meyer et al, BMC Bioinformatics 2008)	4.763	47,165,484	630,700	15999	mgm4509402.3
EMIRGE (Miller et al, Genome Biology, 2011)	4.763	47,165,484	20	4	mgm4509402.3
KiKi (unpublished)	0.604	3,400,711	429,985	1021	mgm4512893.3
Khmer (Pell, et al PNAS 2012)	1.208	3,652,356	14	n/A	mgm4510006.3

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Creates mixed bag metabolic model

View Metabolic Model Details - Output 14:50:58, 2/10/2014

mgm4506694.3.otu.picrust.kegg2ss.anno.raw.001

1917 reactions, 1556 compounds, 1 gapfill runs

Reactions | Compounds | Compartment | Biomass | Gapfill | Gapgen

Show 10 entries Search all:

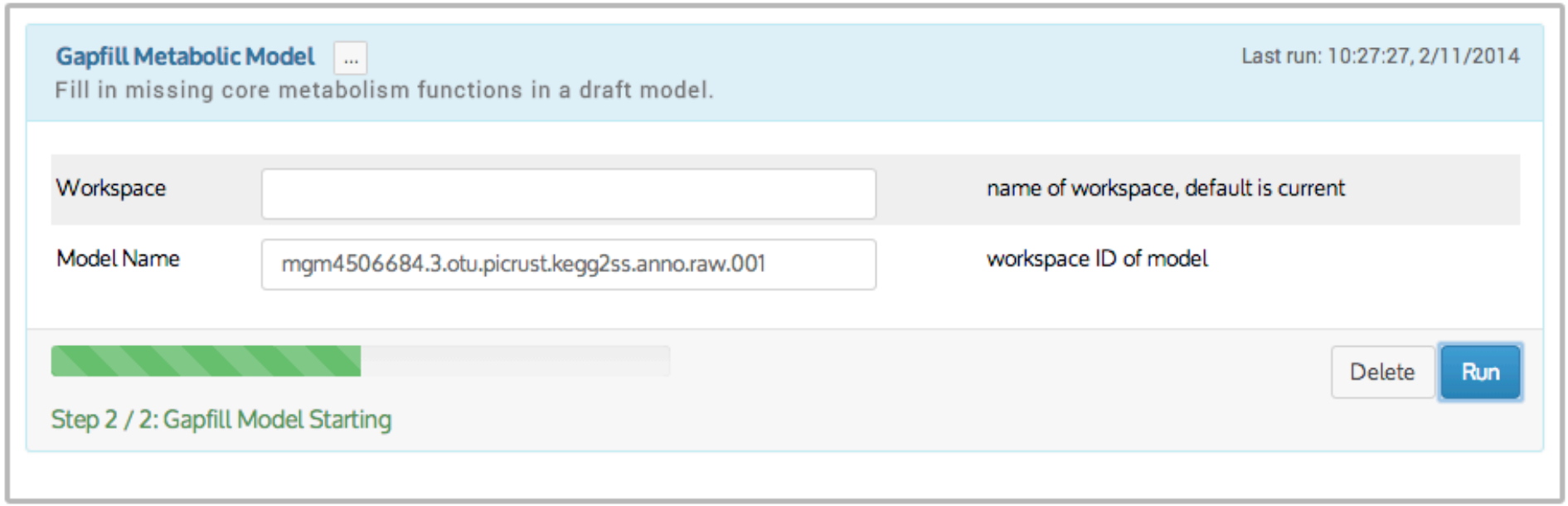
Id (compartment)	Name	Equation	Genome Features Mapped to this Reaction	
rxn00001 (c0)	Pyrophosphate phosphohydrolase	H2O[c0] + PPi[c0] => (2)H+[c0] + (2)Phosphate[c0]	kb|mganno.167.tail.0.g.0.peg.6652 kb|mganno.167.tail.0.g.0.peg.4886 kb|mganno.167.tail.0.g.0.peg.6194	rxn00001 (c0)
rxn00002 (c0)	Urea-1-carboxylate amidohydrolase	(3)H+[c0] + Allophanate[c0] + H2O[c0] => (2)NH3[c0] + (2)CO2[c0]	kb|mganno.167.tail.0.g.0.peg.603 kb|mganno.167.tail.0.g.0.peg.177 kb|mganno.167.tail.0.g.0.peg.5928	rxn00002 (c0)
rxn00006 (c0)	hydrogen-peroxide:hydrogen-peroxide oxidoreductase	(2)H2O2[c0] => (2)H2O[c0] + O2[c0]	kb|mganno.167.tail.0.g.0.peg.4256	rxn00006 (c0)
rxn00007 (c0)	alpha,alpha-Trehalose glucohydrolase	H2O[c0] + TRHL[c0] <=> (2)D-Glucose[c0]	kb|mganno.167.tail.0.g.0.peg.348 kb|mganno.167.tail.0.g.0.peg.3887	rxn00007 (c0)
rxn00010 (c0)	Glyoxylate carboxy-lyase (dimerizing)	H+[c0] + (2)Glyoxalate[c0] => Tartronate semialdehyde[c0] + CO2[c0]	kb|mganno.167.tail.0.g.0.peg.5640	rxn00010 (c0)

Showing 1 to 5 of 1,917 entries First Previous 1 2 3 4 5 Next Last

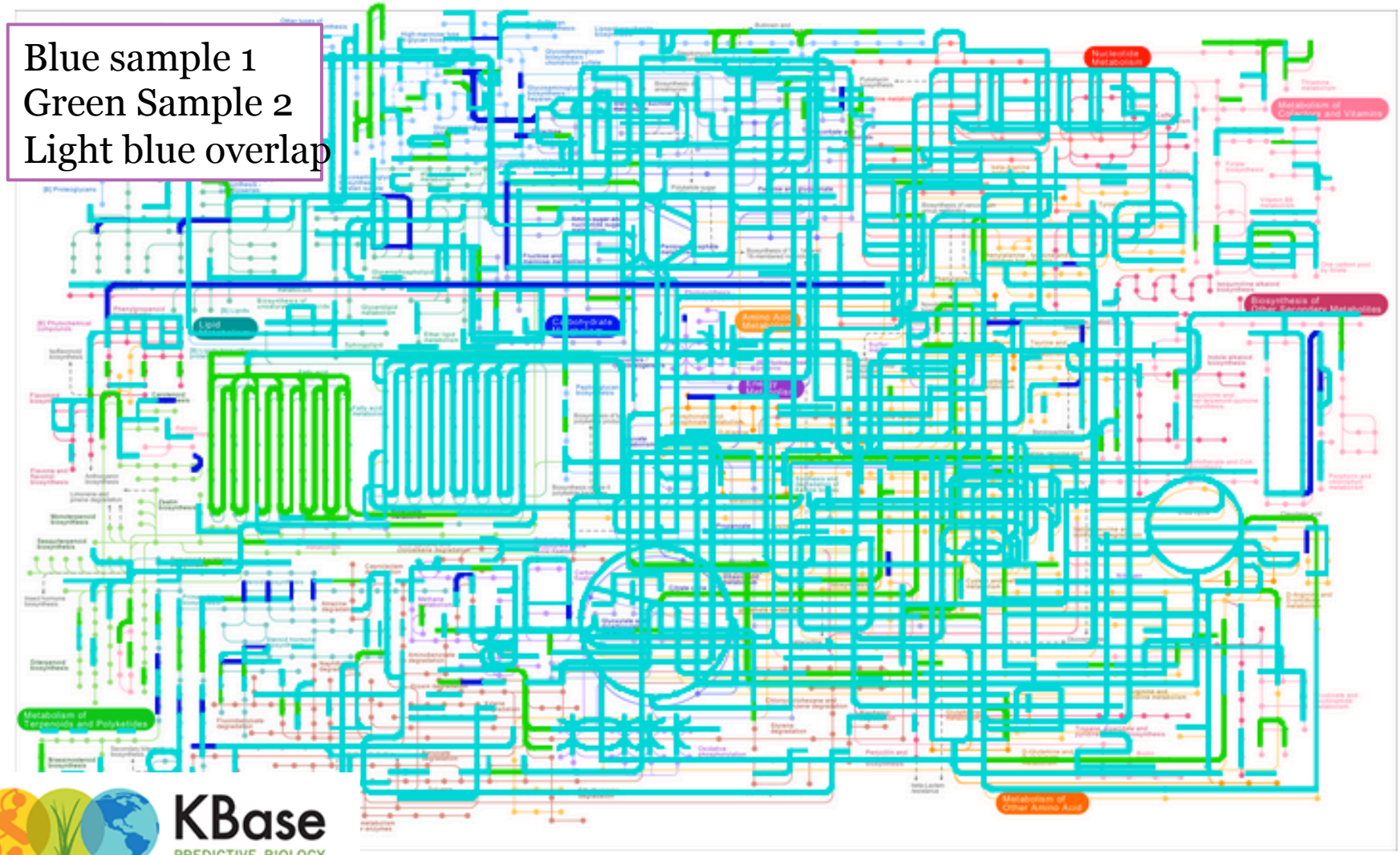
KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KBase has UIs for:
EMIRGE, PICRUSt, ModelSEED, …

e.g.
Gapfilling the models

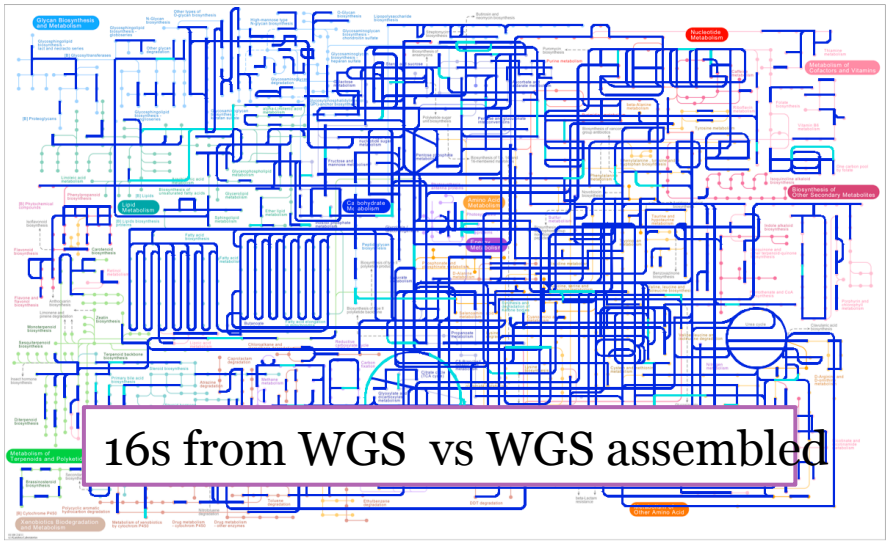


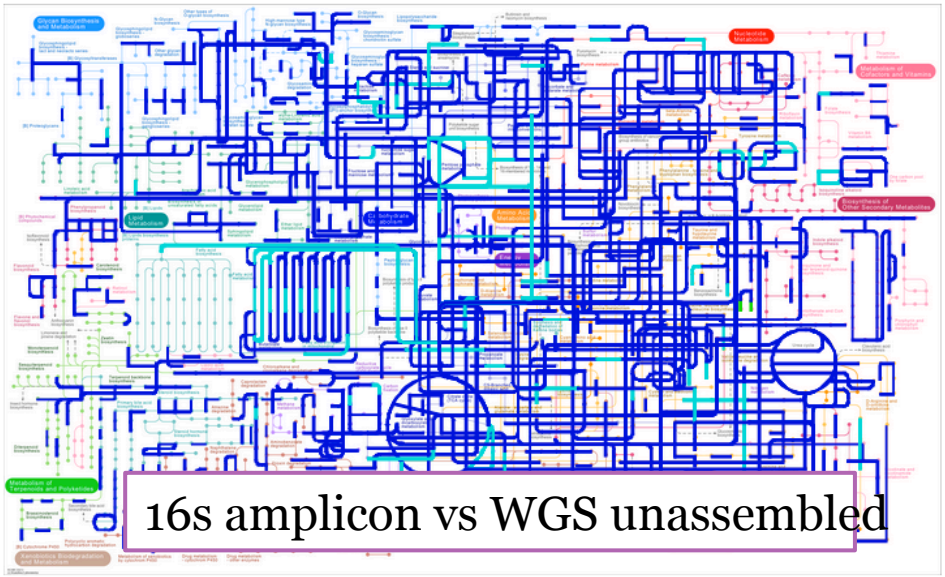
Visual Comparison two different samples



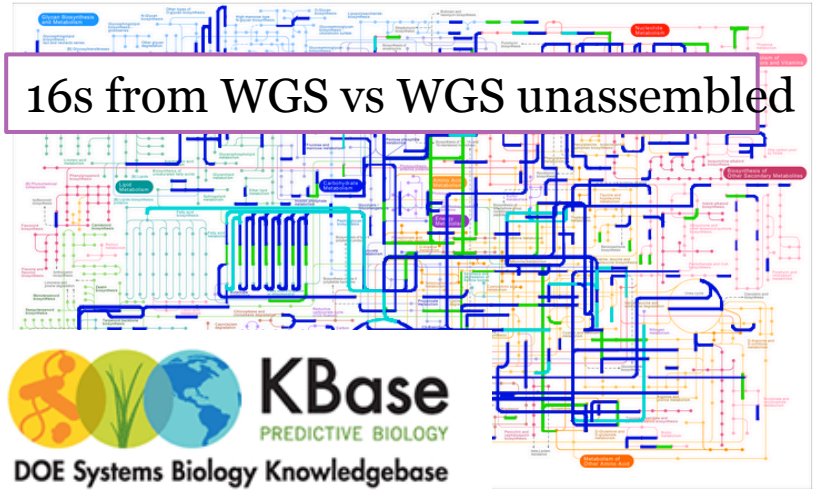
KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

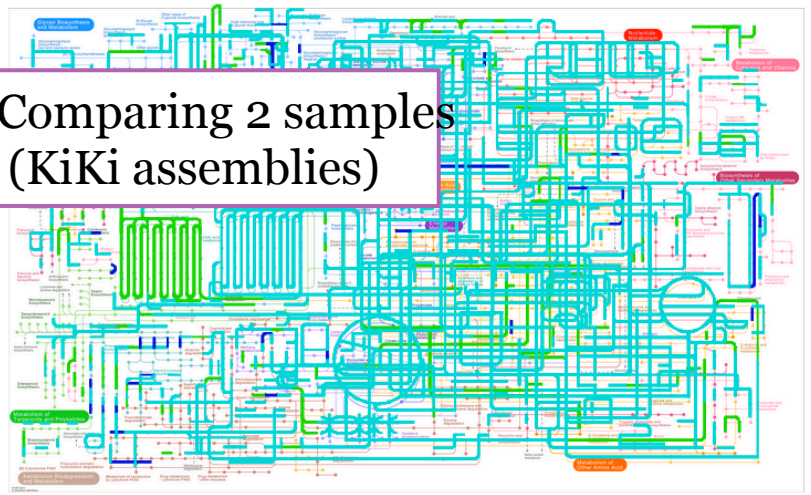
Comparing multiple technologies





Data from Steve Allison, Princeton,





Acknowledgements

Team:

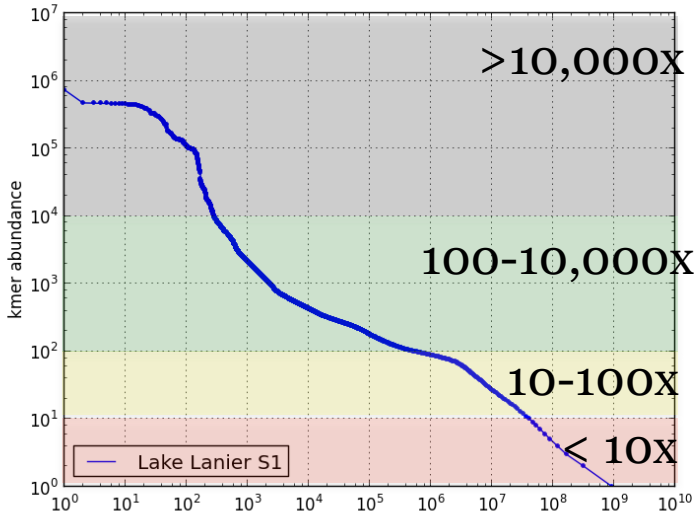
- Jared Bischof
- Dan Braithewaite
- Adina Chuang-Howe
- Narayan Desai
- Mark d' Souza
- Katya Drybinski
- Elizabeth M. Glass
- **Wolfgang Gerlach**
- Travis Harrison
- **Kevin Keegan**
- Tobias Paczian
- Hunter Matthews
- **Wei Tang**
- **Will Trimble**
- **Andreas Wilke**
- Jared Wilkening

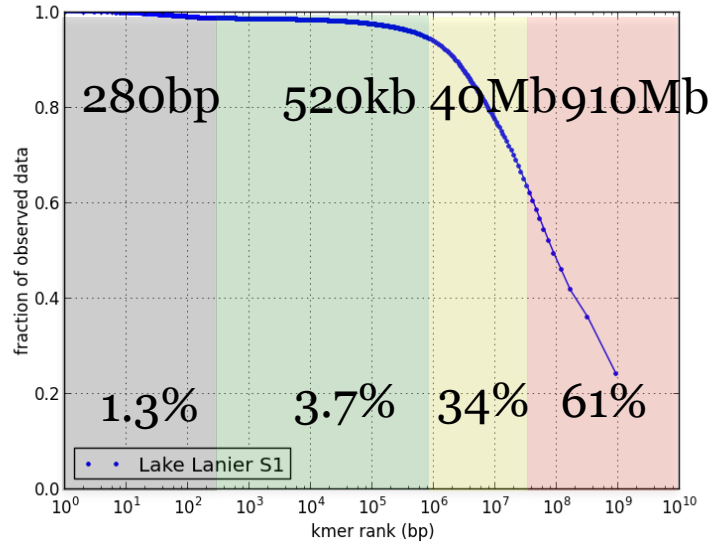
Collaborators:

- D. Antonopoulos (Argonne)
- A. Arkin (Berkeley)
- E. Chang (UChicago)
- Dawn Field (Oxford)
- F.-O. Glöckner (MPI Bremen)
- Jack Gilbert (Argonne)
- Jeff Grethe (CalIT2, CAMERA)
- Sarah Hunter / Guy Cochrane (EBI)
- Ken Kemner (Argonne)
- Rob Knight (Colorado)
- Nikos Kyrpides (DOE JGI)
- J. Tiedje (MSU)
- Owen White (UMaryland, HMP DACC)

Thank you

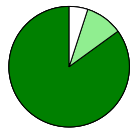
Stratifying kmers – heterogeneous population of underlying reads



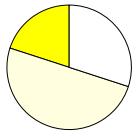


Example: Lake Lanier S-1

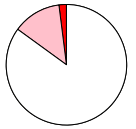
expect *but can't prove* mostly complete



expect mostly fragmented / incomplete



expect assembly is hopeless



Average completeness must be some sort of weighted average of the completeness of the components.

I can deliver estimates of approximate total genome size(s) in each coverage band.

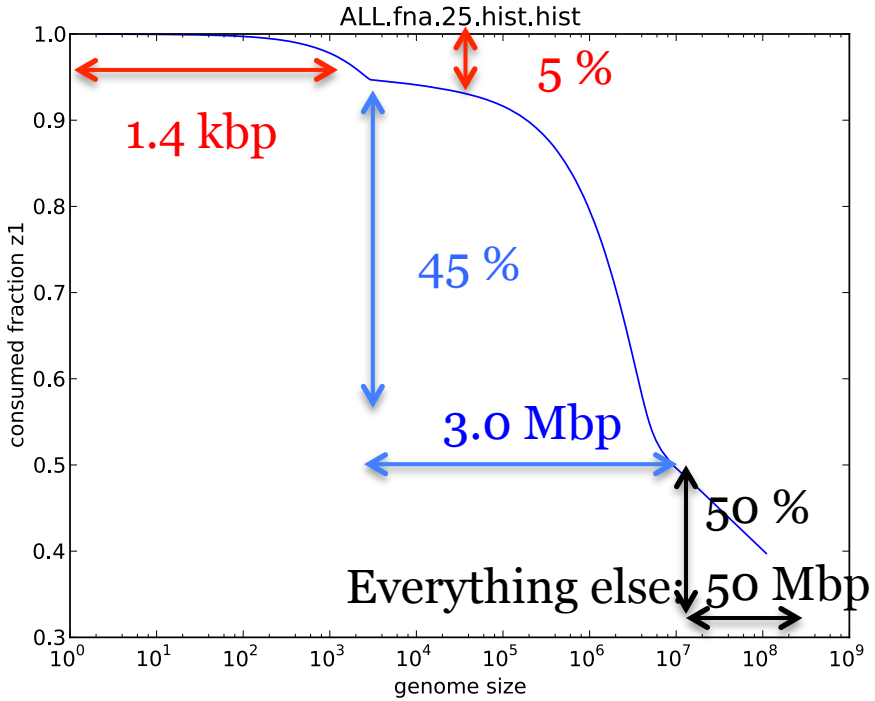
Decadal coverage bands (1-9, 10-99, 100-999) look go

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Predicting replicons before assembly

- Using **K-mer spectra** to predict (pan-) genome size
  - K-mer= unique word, easily computed
- In addition to alpha diversity
  - 300 OTU data set
- Using k-mer size 25

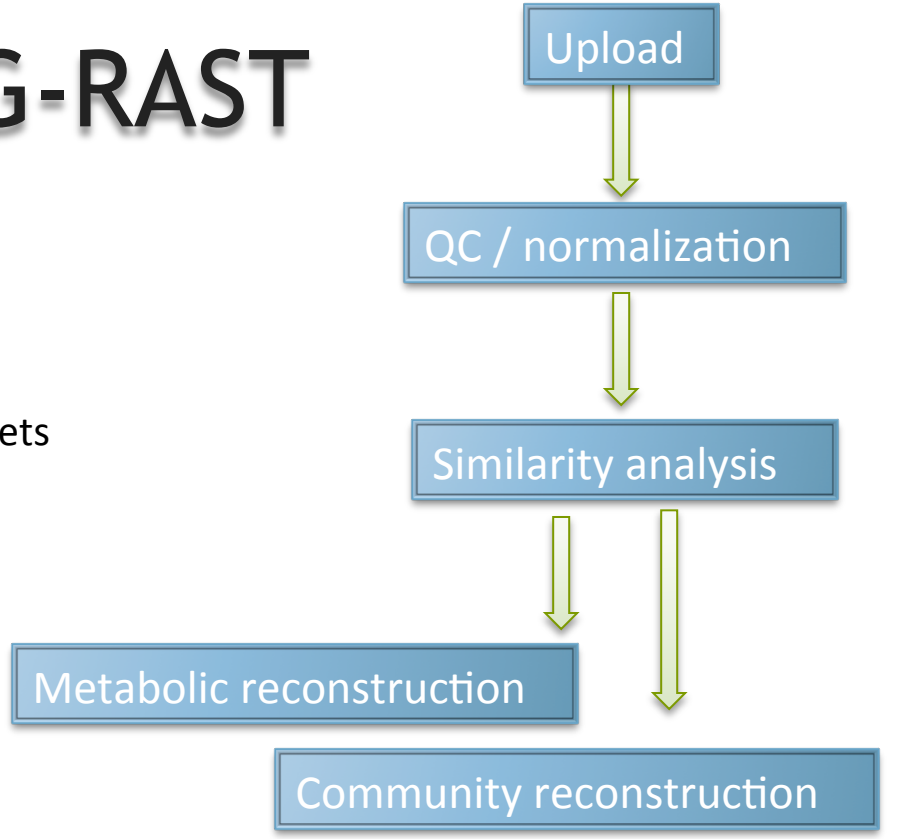
- **Red** and **blue** replicons were missing in assembly
  - Allows adjustment of parameters

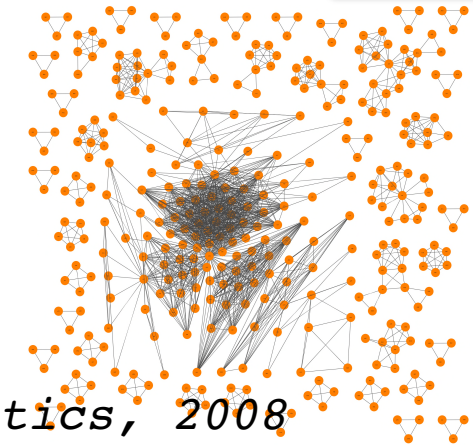


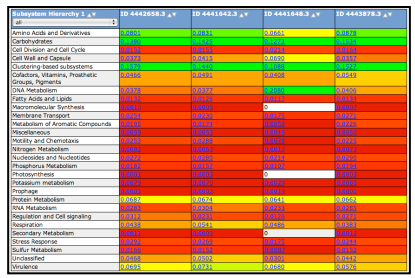
From: Williams et al, BMC Genomics, 2013

Brief history of MG-RAST

- December 2007 (v1)
  - 100+ groups and ~250 data submitters
  - 100+ data sets, ~10+ GBp total size
- October 2009 (v2)
  - **Pre-publication sharing available**
  - ~1500 data submitters, ~300 public data sets
  - 6000+ data sets
  - 200+ GBp total data set size
  - About ~30 GBp/month throughput
- March 2011 (release v3)
  - 2500+ data submitters
  - ~2000 public data sets
  - 25,000 data sets total
  - Throughput:
    - 47GBp in 24h
    - 3000 submissions in 24h
- March 2012 (v 3.1.2)
  - 12 Terabasepairs (10^12 bp)
- May 2012 (3.20)
  - 13.8 TBp (10^12bp)
  - 128 billion sequences







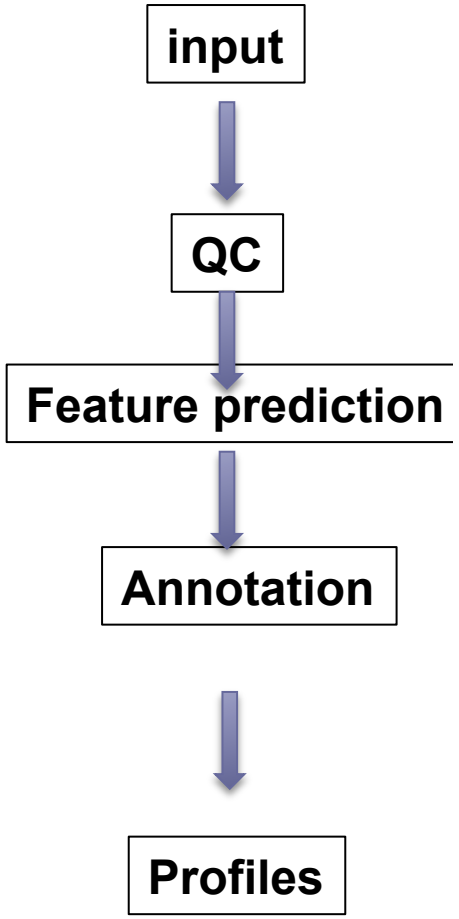
Meyer et al., BMC Bioinformatics, 2008

simplified

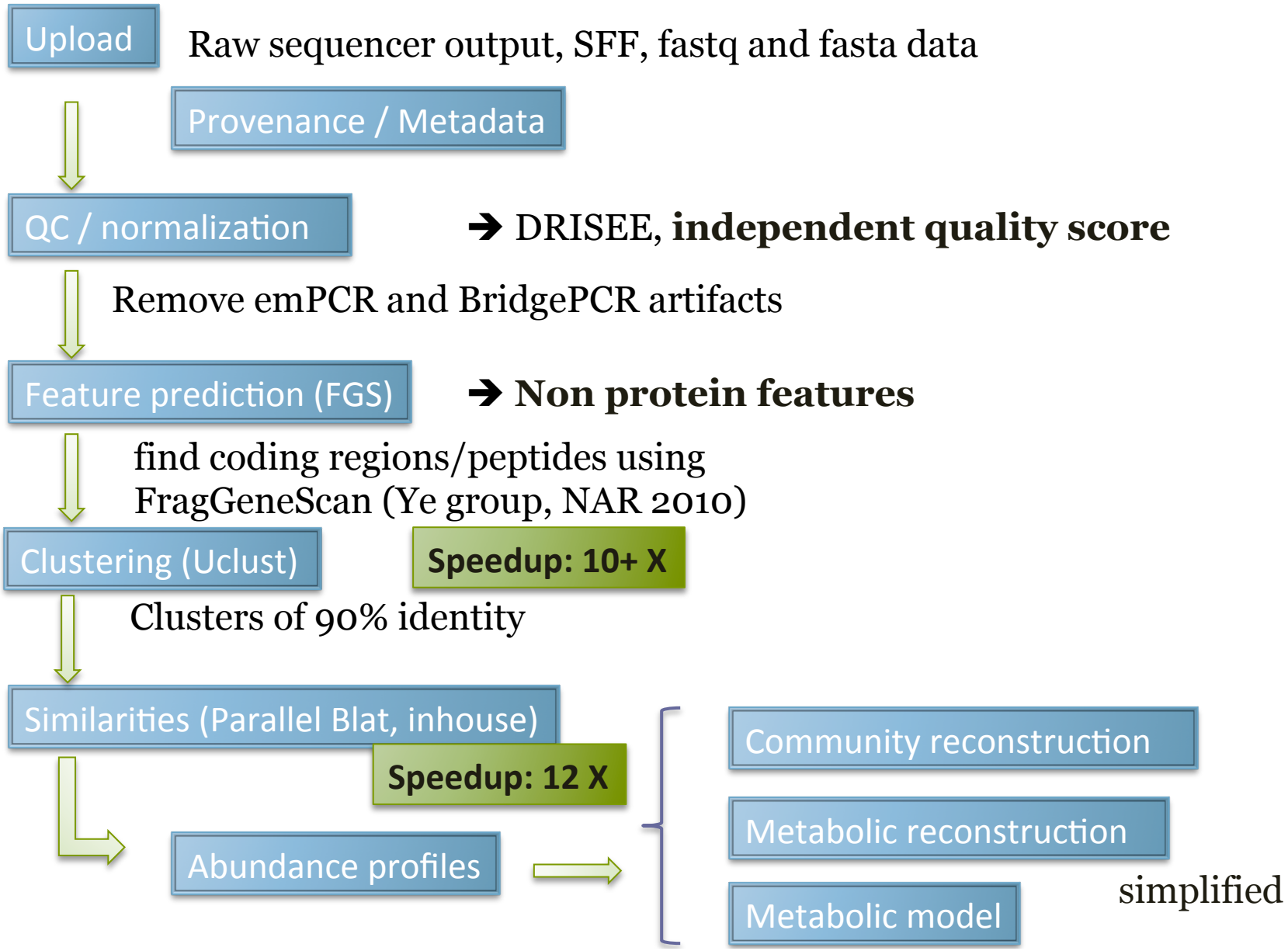
MG-RAST in a slide

- 80,000 data sets, 10,000+ public
  - 27TBp analyzed, 240 billion sequences
- Normalized data from many groups
- Intensive QC (innovative QC methods)
- Mapping against known annotations
  - Using many sources (SEED, KEGG, COG, NOGs, …)
- Allow comparison and meta-analysis
  - Require metadata → partnership with GSC
- Creates abundance profile
  - Counting occurrence of taxa, phyla, …

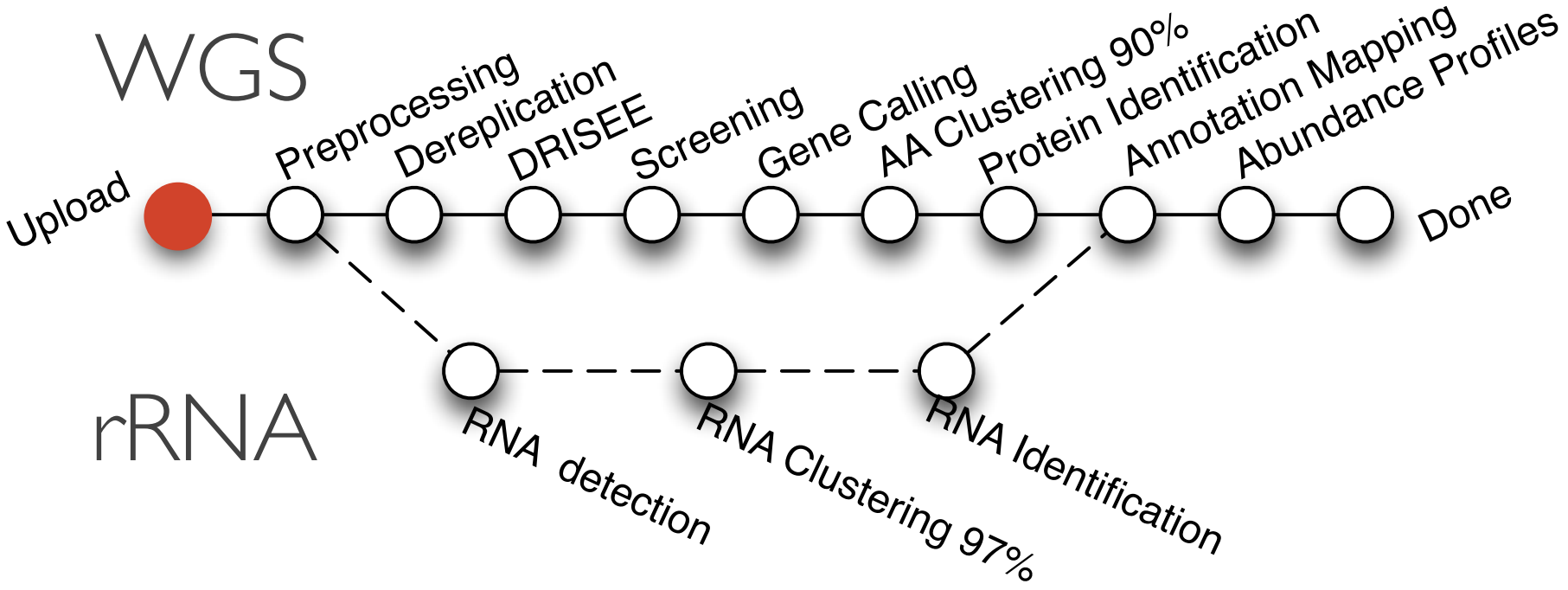
  and
  - Count functions (e.g. subsystems, KEGG pathways, etc.)



Improved Algorithm



Pipeline Overview



MG-RAST is an automated pipeline for annotating individual NGS-sequencing reads, transforming raw sequences into *meaningful* output

Access to MG-RAST

- via web-UI **(http://metagenomics.anl.gov)**
- ftp/http downloads

- Web services API [BETA]
  - http://api.metagenomics.anl.gov
  - C, C++, Java, Perl, Python, Ruby, Javascript, …
  - RESTful, JSON objects
- R client (matR) [BETA]

Pipeline changes?

- MGA vs FGS?
  - MGA misses 70+%
- BLAST vs BLAT?
  - BLAST is more sensitive
- What is the effect of a different protein database?
  - E.g. TrEMBL
- What is the effect of using protein motifs?
  - E.g. interpro?
- **Assembly is the biggest unknown…**

Lessons learned from GOS

- Old style:
  - "Lets sequence as much as we can afford"
  - "Metagenomics is like genomics"

- Today:
  - Often 16s amplicon study first
  - replicates (biological and technical)
    - "design for statistics"
    - "replicate or lie" (Jim Prosser)

  - metadata
    - Genomics Standards Consortium provides tools

  - Provide good QC
    - Identify signal vs. noise ratio
    - Throw away bad data when needed (!)

  - Identify appropriate analysis workflow
  - Perform assembly?

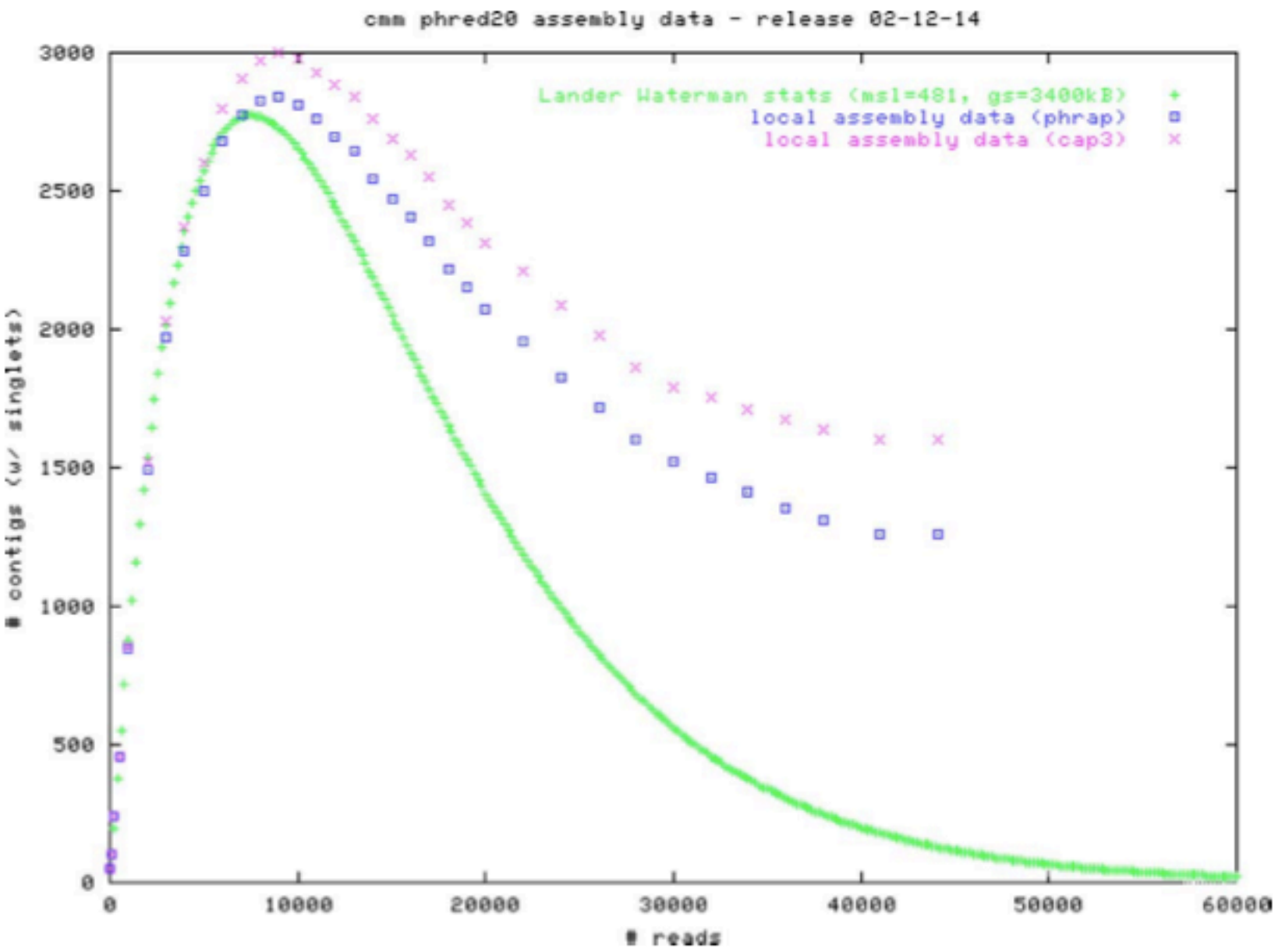
➔ **Design for statistics**

➔ **Metadata (r)evolution**

➔ **Data hygiene**

➔ **Tool chain matters**

From: Knight, et al, Nature Biotechnology, 2012



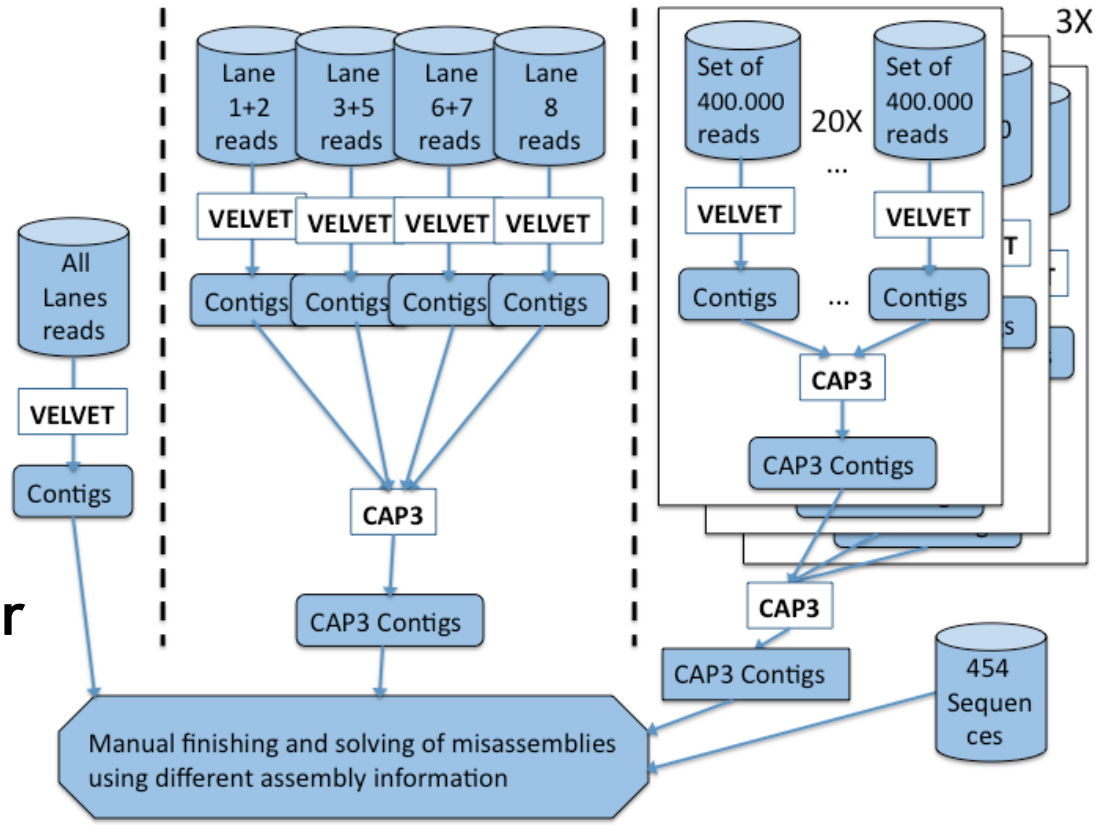
Kaiser et al, J. Biotech, 2003

Our subsampling pipeline

- Build multiple sets of reads
  - All
  - 2 lanes at a time
  - Sets of 400k reads
- Run velvet on each set
  - K-mer= 51
  - K-mer=17
- Initially used cap3 to integrate between sets

- replaced cap3 with **newbler**
  - **added 50% to contig yield!**

- Class of small repeats 100-200bp requires longer reads (using 454)
  - Manually integrated

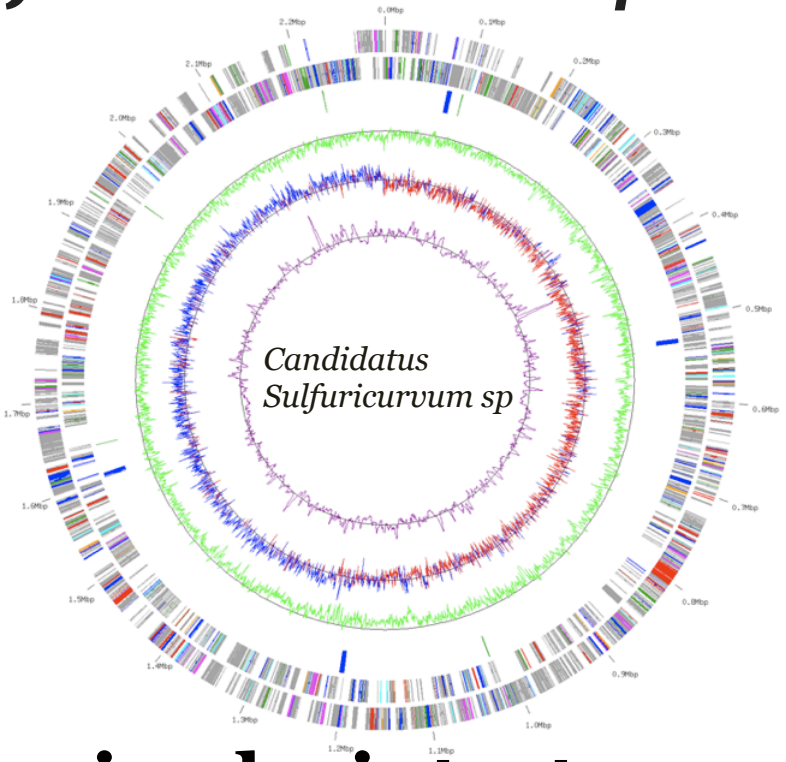


Newbler: J. Knight, unpublished

Cap3: Huang et al, Genome Res. 2003

Result: *Candidatus Sulfuricurvum sp.*

- new genus
- Epsilon proteo bacteria
- *Campylobacteriales*
- 2.3 MBp genome
- 1 contig
- 43% GC
- High quality



- **Verified by many overlapping by intact mate-pairs**

→ likely to not be chimeric

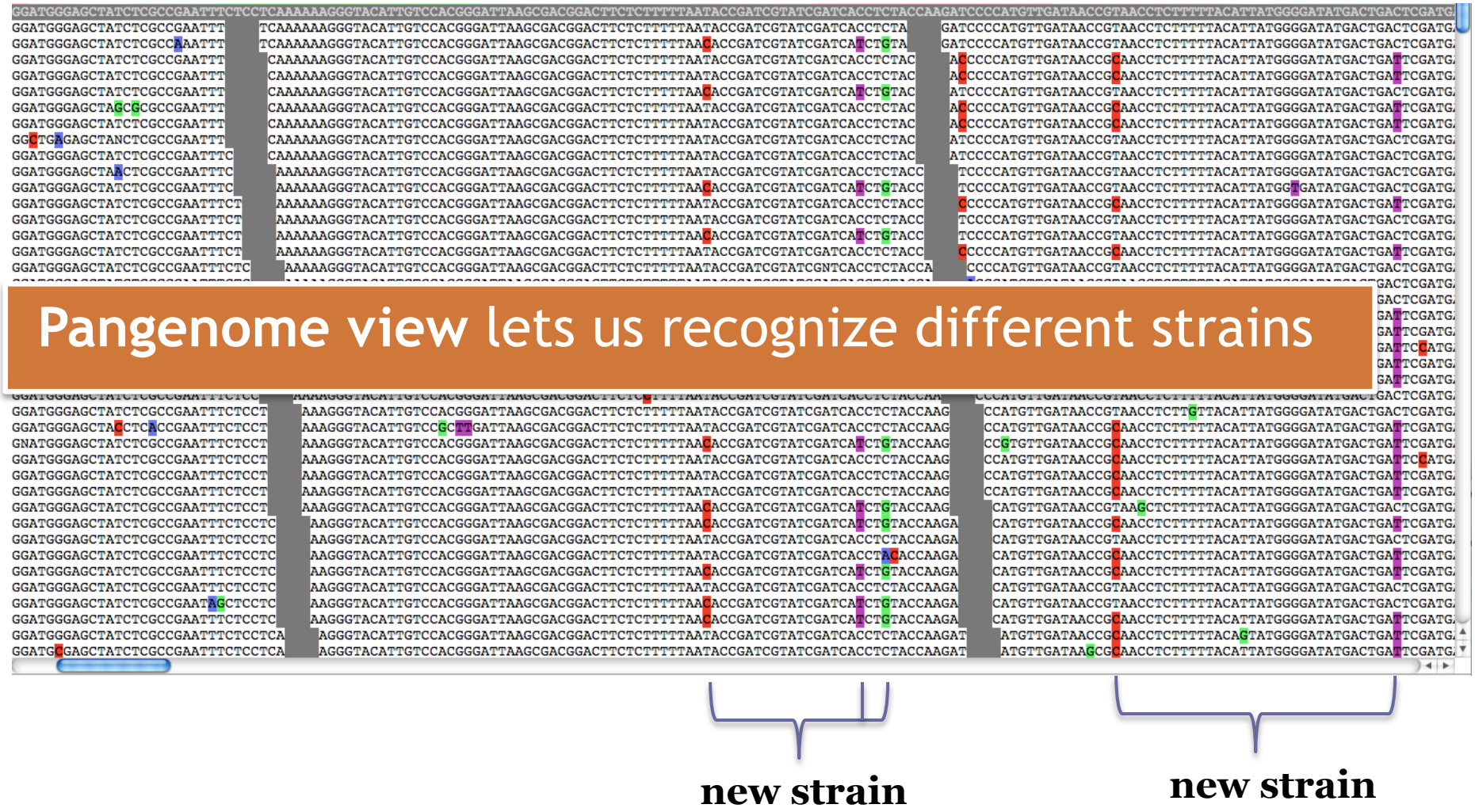
Genome annotation with RAST: Aziz et al, BMC Genomics, 2008

What does having 45 million reads give us?

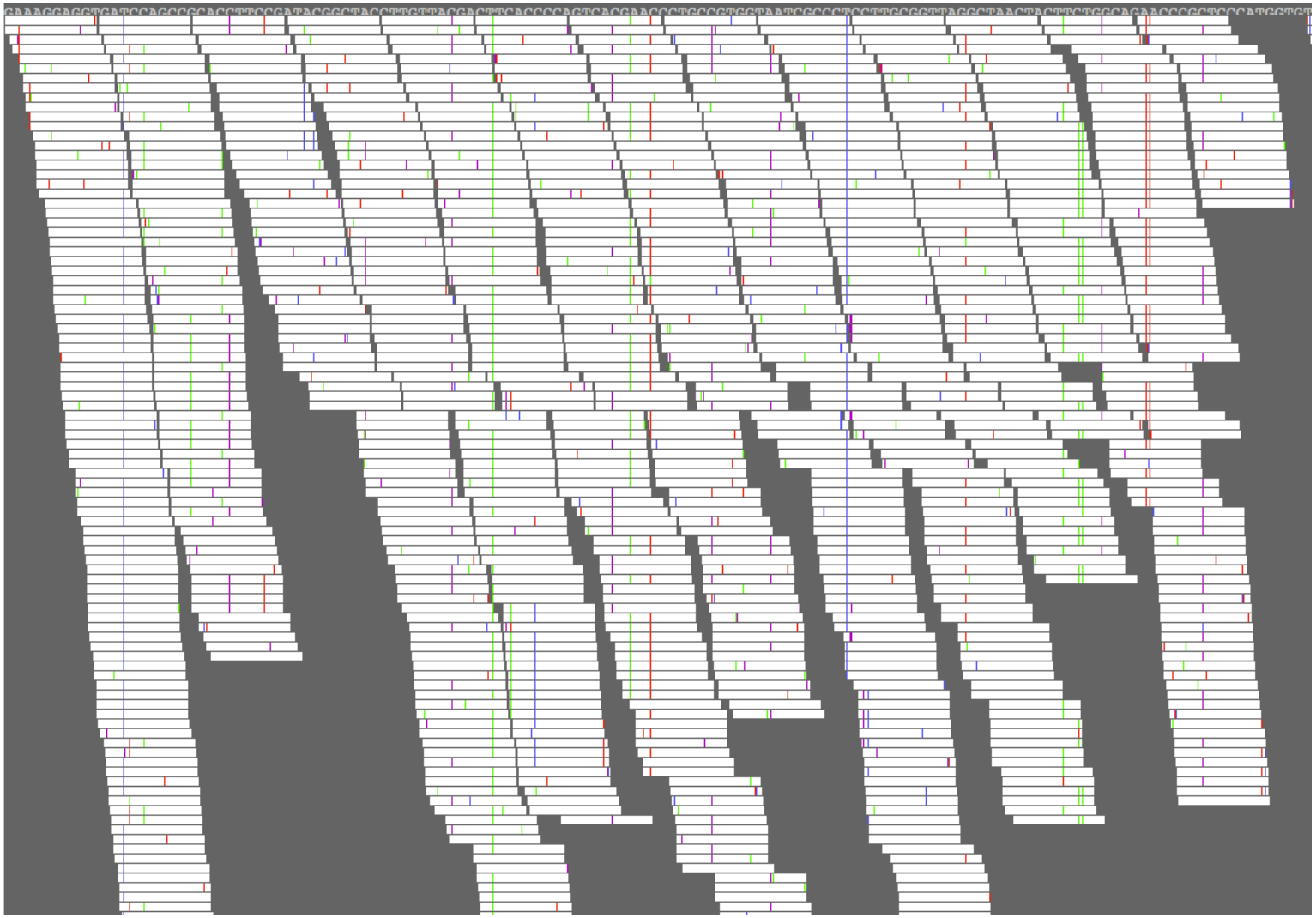
- GOS assumed nature was *pseudo* clonal
- In GOS assembly reduced diversity to consensus sequences

- Now…

we can study strain variation ➔ Evolution



There are many strains in the data

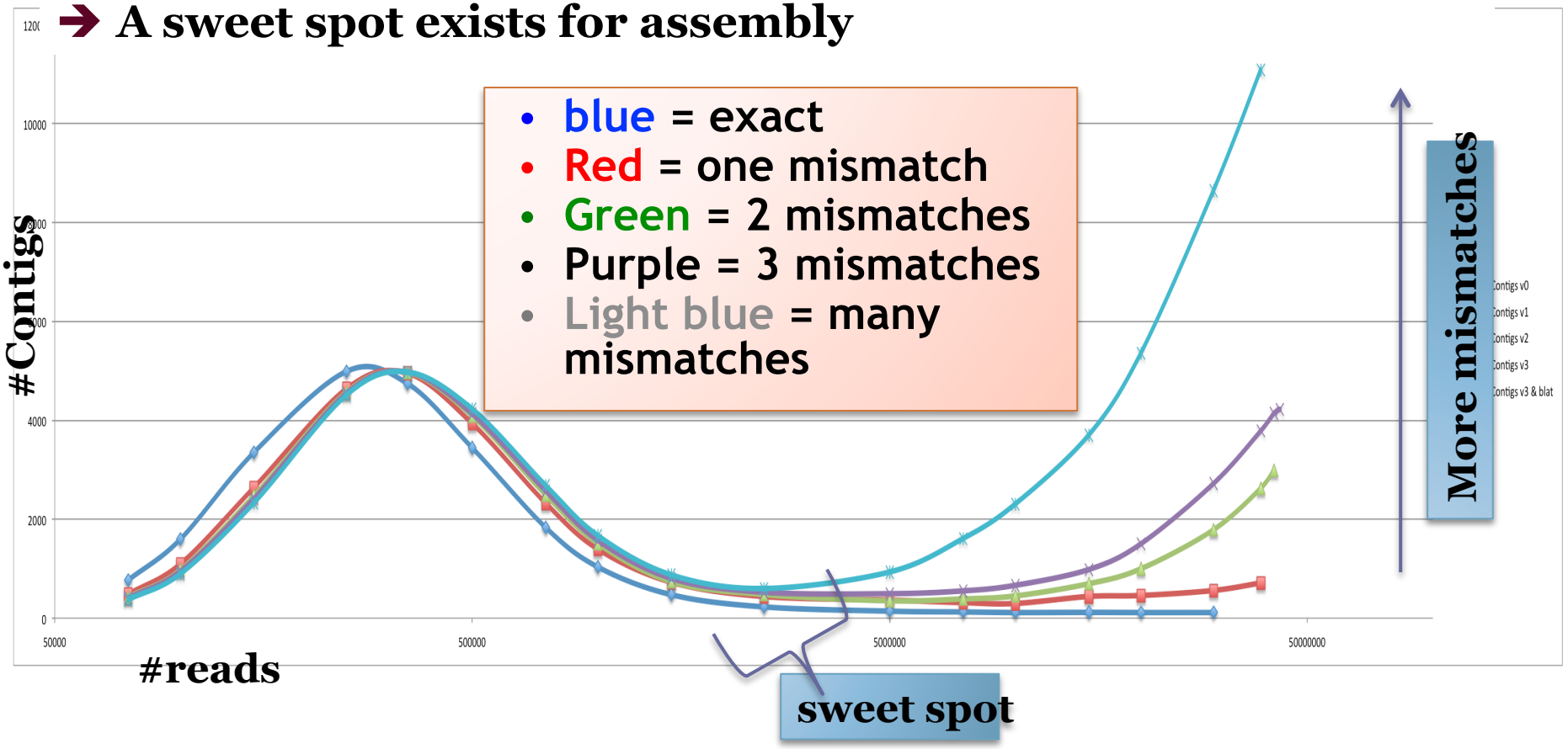


Effect of strain variation on assembly

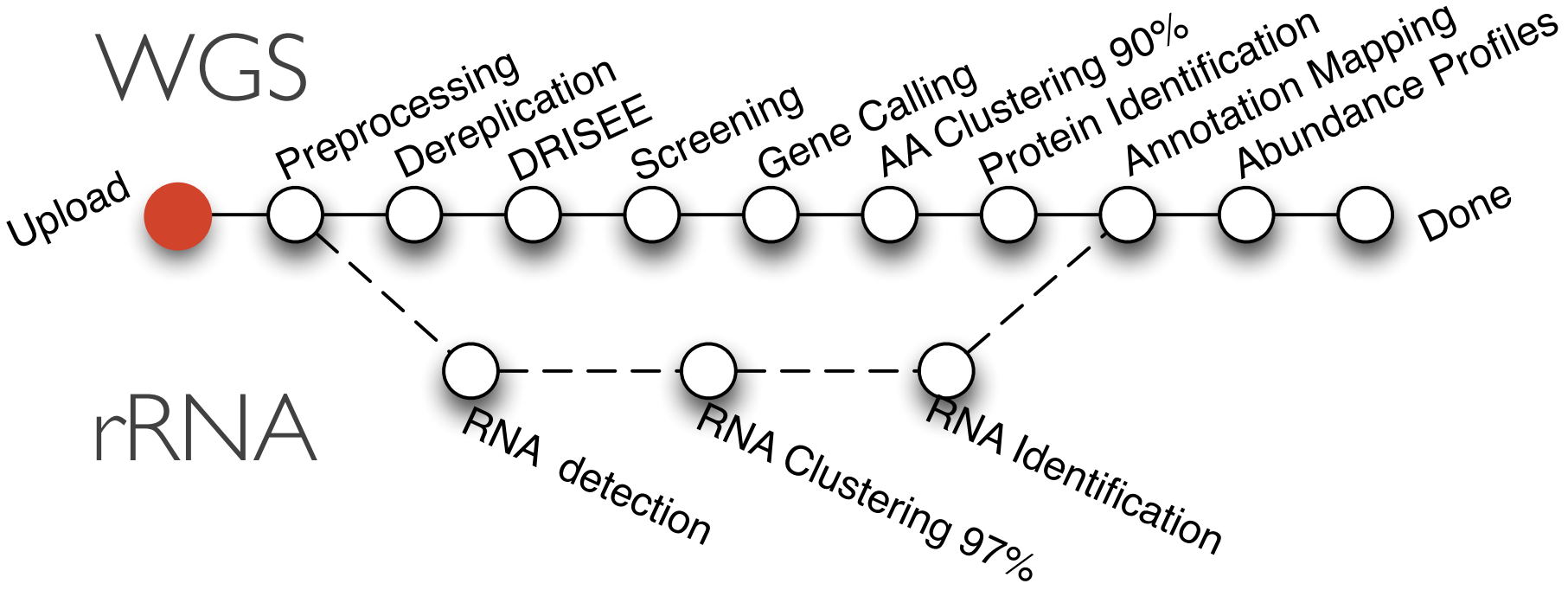
- we created several artificial subsets using the genome as reference
  - allowing more and more variation (repeated with boot strapping)

➔ **Confirms strain variation breaks assembly tools**

➔ **A sweet spot exists for assembly**



Pipeline Overview



MG-RAST is an automated pipeline for annotating individual NGS-sequencing reads, transforming raw sequences into *meaningful* output