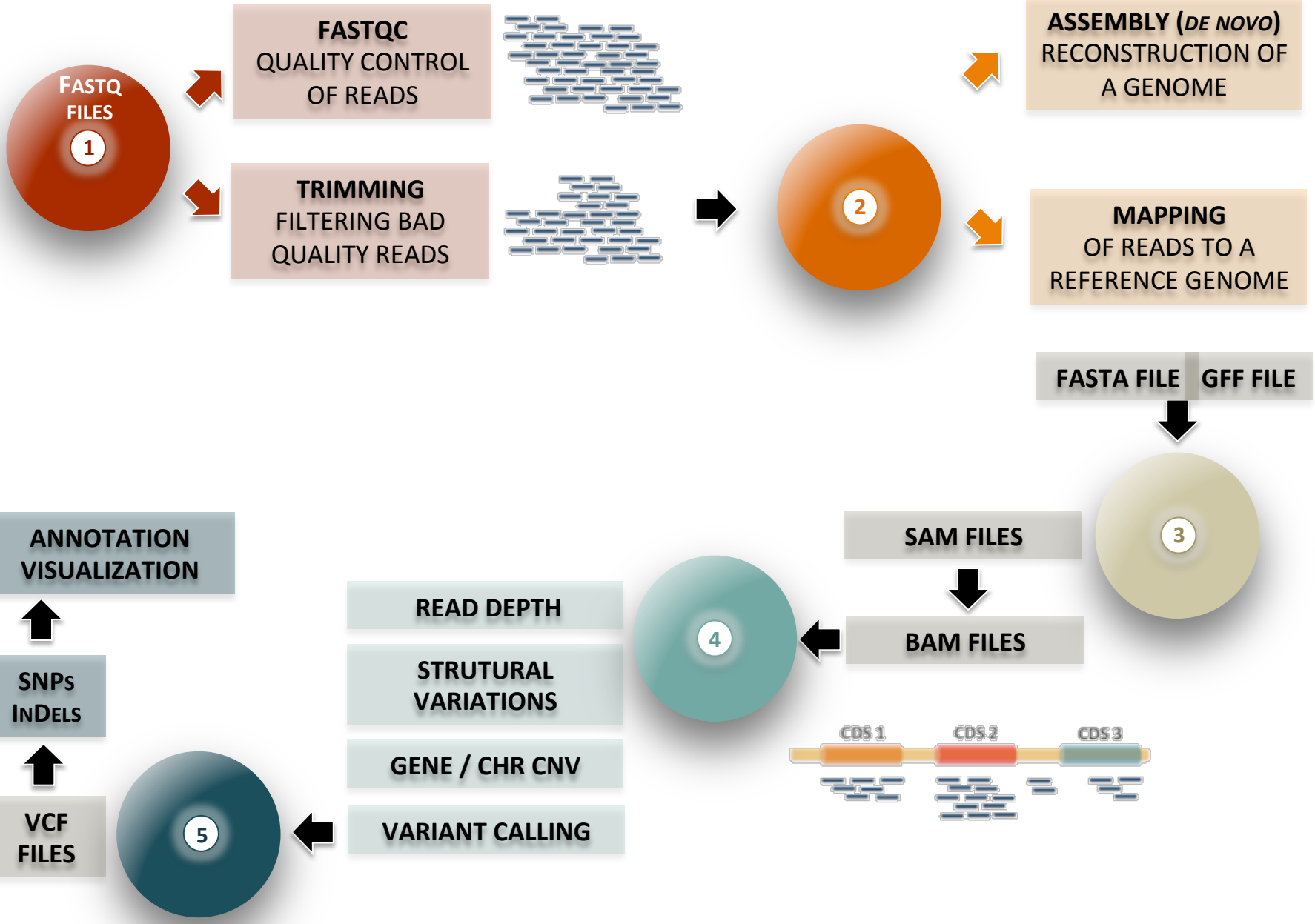
H3ABioNet

Pan African Bioinformatics Network for H3Africa

Machine Learning and Metagenome Analysis

Chris Fields's slides
presented by Amel Ghouila

Overview of analysis workflow



Overview of metagenome analysis

- What is metagenomics?
  - The study of the collective genomic material from environmental samples, for example
    - **Environment** : soil, water
    - **Medical** : fecal, skin, kidney stone
    - **Industrial** : bioreactors, fermenters, enrichments
    - Pretty much anything


Resource

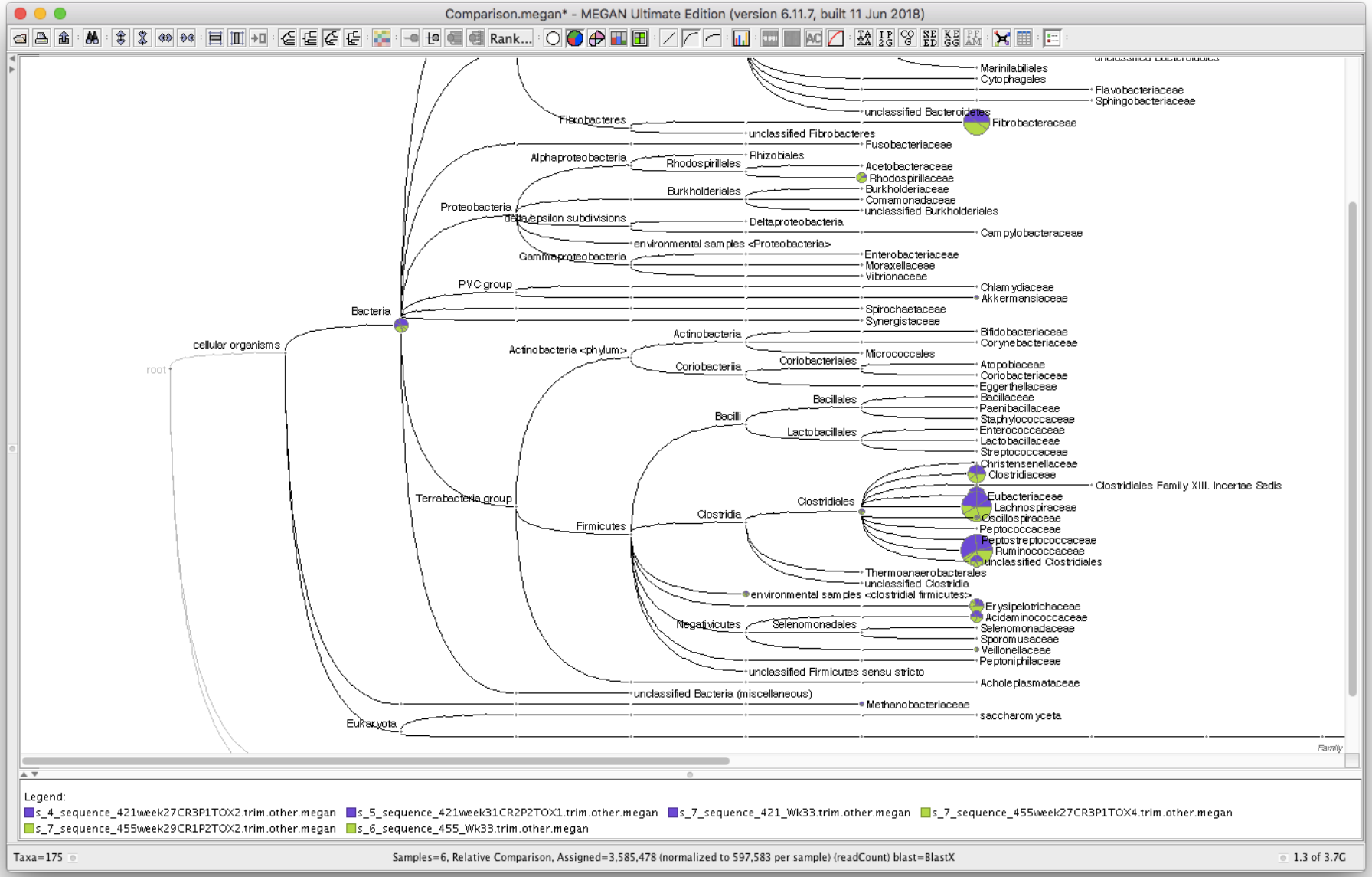
Windshield splatter analysis with the Galaxy metagenomic pipeline

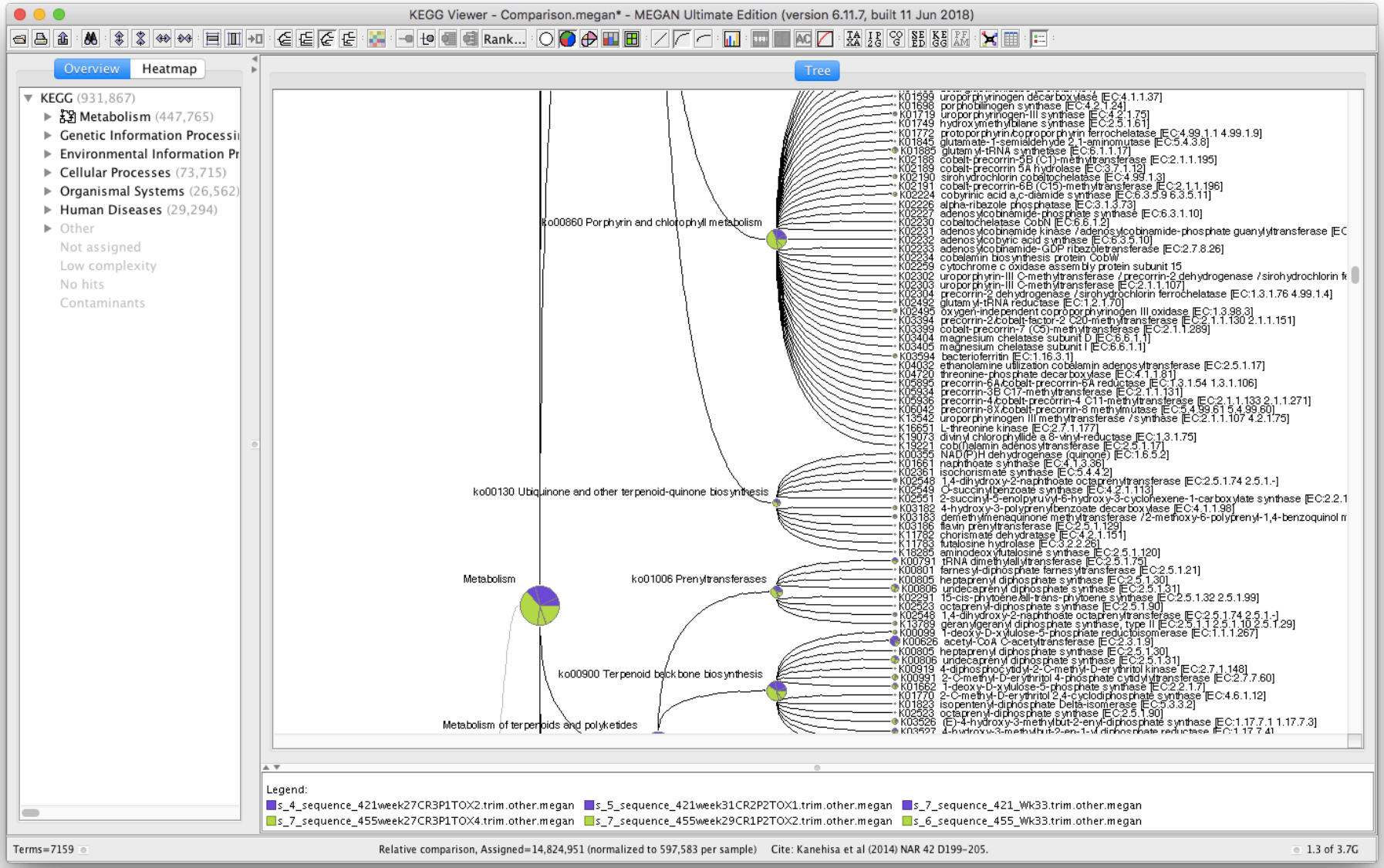
Overview of metagenome analysis

- Why?
  - Characterize a sample that may be of “biological interest”, but…
  - The vast majority of microorganisms cannot be cultured
  - Methods used to culture from environmental samples miss these
- **Solution**: isolate DNA from samples, sequence it, then break down what is there.
  - Yes, it’s as difficult as it sounds

Overview of metagenome analysis

- **Solution**: isolate DNA from samples, sequence it, then break down what is there.
  - **Taxonomic** – what is present?
  - **Functional** – what can be done metabolically (e.g. metabolic potential)?
    - Note, this cannot be done with 16s directly


Comparison.megan* - MEGAN Ultimate Edition (version 6.11.7, built 11 Jun 2018)
Rank...
root
cellular organisms
Bacteria
Eukaryota
unclassified Bacteria (miscellaneous)
Terrabacteria group
Firmicutes
Bacilli
Bacillales
Lactobacillales
Clostridia
Clostridiales
Negativicutes
Selenomonadales
Actinobacteria <phylum>
Actinobacteria
Coriobacteriia
Coriobacteriales
Micrococcales
PVC group
Proteobacteria
Alphaproteobacteria
Rhodospirillales
Rhizobiales
Burkholderiales
delta/epsilon subdivisions
Deltaproteobacteria
environmental samples <Proteobacteria>
Gammaproteobacteria
Enterobacteriaceae
Moraxellaceae
Vibrionaceae
Fibrobacteres
unclassified Fibrobacteres
Fusobacteriaceae
Acetobacteraceae
Rhodospirillaceae
Burkholderiaceae
Comamonadaceae
unclassified Burkholderiales
Campylobacteraceae
Chlamydiaceae
Akkermansiaceae
Spirochaetaceae
Synergistaceae
Bifidobacteriaceae
Corynebacteriaceae
Atopobiaceae
Coriobacteriaceae
Eggerthellaceae
Bacillaceae
Paenibacillaceae
Staphylococcaceae
Enterococcaceae
Lactobacillaceae
Streptococcaceae
Christensenellaceae
Clostridiaceae
Clostridiales Family XIII. Incertae Sedis
Eubacteriaceae
Lachnospiraceae
Oscillospiraceae
Peptococcaceae
Peptostreptococcaceae
Ruminococcaceae
unclassified Clostridiales
Thermoanaerobacterales
unclassified Clostridia
environmental samples <clostridial firmicutes>
Erysipelotrichaceae
Acidaminococcaceae
Selenomonadaceae
Sporomusaceae
Veillonellaceae
Peptoniphilaceae
unclassified Firmicutes sensu stricto
Acholeplasmataceae
Methanobacteriaceae
saccharomyceta
Marinilabiliales
Cytophagales
Flavobacteriaceae
Sphingobacteriaceae
unclassified Bacteroidetes
Fibrobacteraceae
Family
Legend:
s_4_sequence_421week27CR3P1TOX2.trim.other.megan
s_5_sequence_421week31CR2P2TOX1.trim.other.megan
s_7_sequence_421_Wk33.trim.other.megan
s_7_sequence_455week27CR3P1TOX4.trim.other.megan
s_7_sequence_455week29CR1P2TOX2.trim.other.megan
s_6_sequence_455_Wk33.trim.other.megan
Taxa=175
Samples=6, Relative Comparison, Assigned=3,585,478 (normalized to 597,583 per sample) (readCount) blast=BlastX
1.3 of 3.7G


KEGG Viewer - Comparison.megan* - MEGAN Ultimate Edition (version 6.11.7, built 11 Jun 2018)
Overview
Heatmap
Tree
KEGG (931,867)
Metabolism (447,765)
Genetic Information Processi
Environmental Information Pr
Cellular Processes (73,715)
Organismal Systems (26,562)
Human Diseases (29,294)
Other
Not assigned
Low complexity
No hits
Contaminants
ko00860 Porphyrin and chlorophyll metabolism
ko00130 Ubiquinone and other terpenoid-quinone biosynthesis
Metabolism
ko01006 Prenyltransferases
ko00900 Terpenoid backbone biosynthesis
Metabolism of terpenoids and polyketides
Legend:
s_4_sequence_421week27CR3P1TOX2.trim.other.megan
s_5_sequence_421week31CR2P2TOX1.trim.other.megan
s_7_sequence_421_Wk33.trim.other.megan
s_7_sequence_455week27CR3P1TOX4.trim.other.megan
s_7_sequence_455week29CR1P2TOX2.trim.other.megan
s_6_sequence_455_Wk33.trim.other.megan
Terms=7159
Relative comparison, Assigned=14,824,951 (normalized to 597,583 per sample)
Cite: Kanehisa et al (2014) NAR 42 D199-205.
1.3 of 3.7G

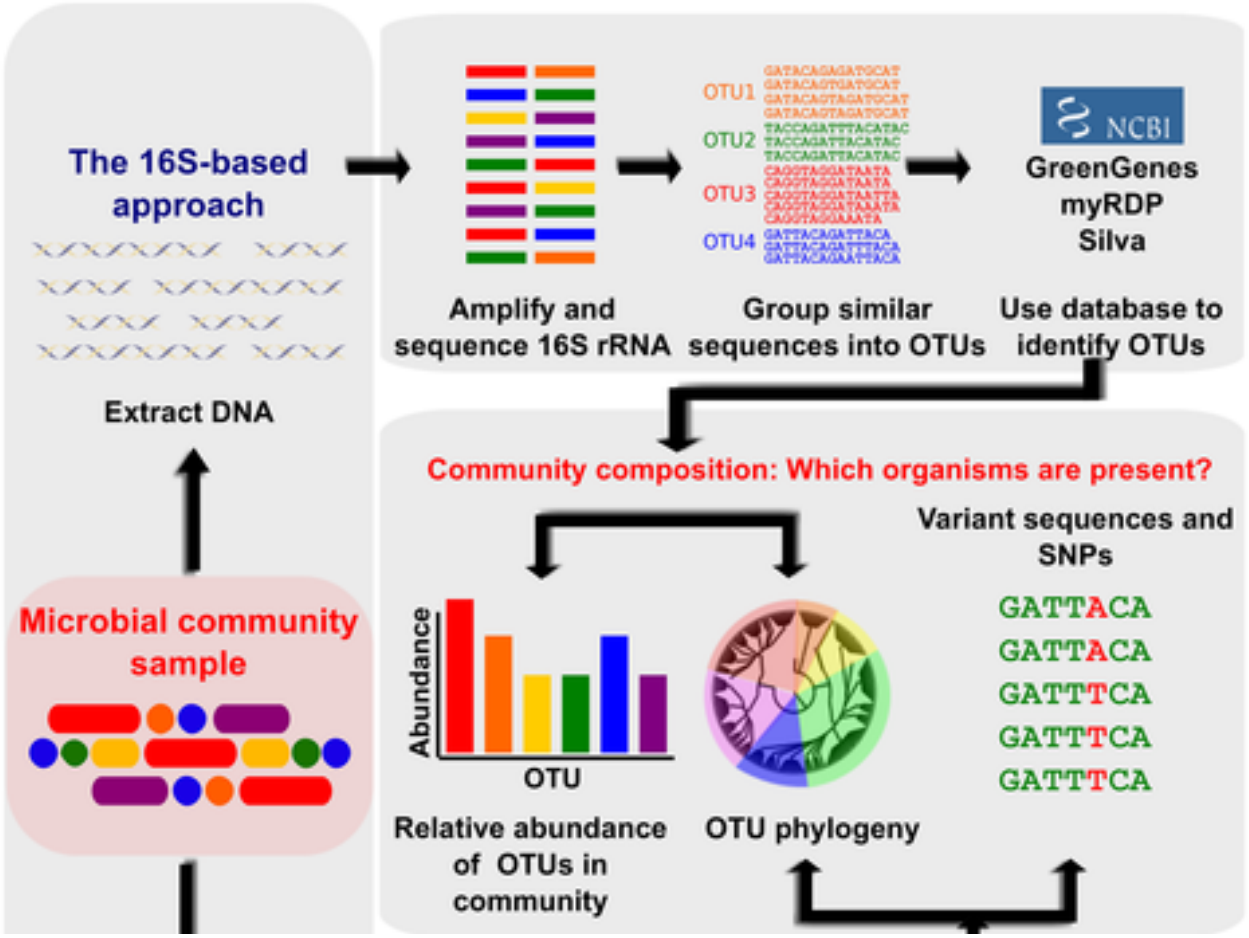
Overview of metagenome analysis

- Note: depending on the question, may be complementary (and similarly difficult) data
  - **Metatranscriptome** – what is being expressed in environmental samples (RNA)
  - **Metabolome** – metabolites produced
  - **Proteome** – proteins present in sample

Overview of metagenome analysis

- **Two general approaches**
  - Targeted sequencing (e.g. 16s variable regions)
  - Shotgun (whole) metagenome sequencing

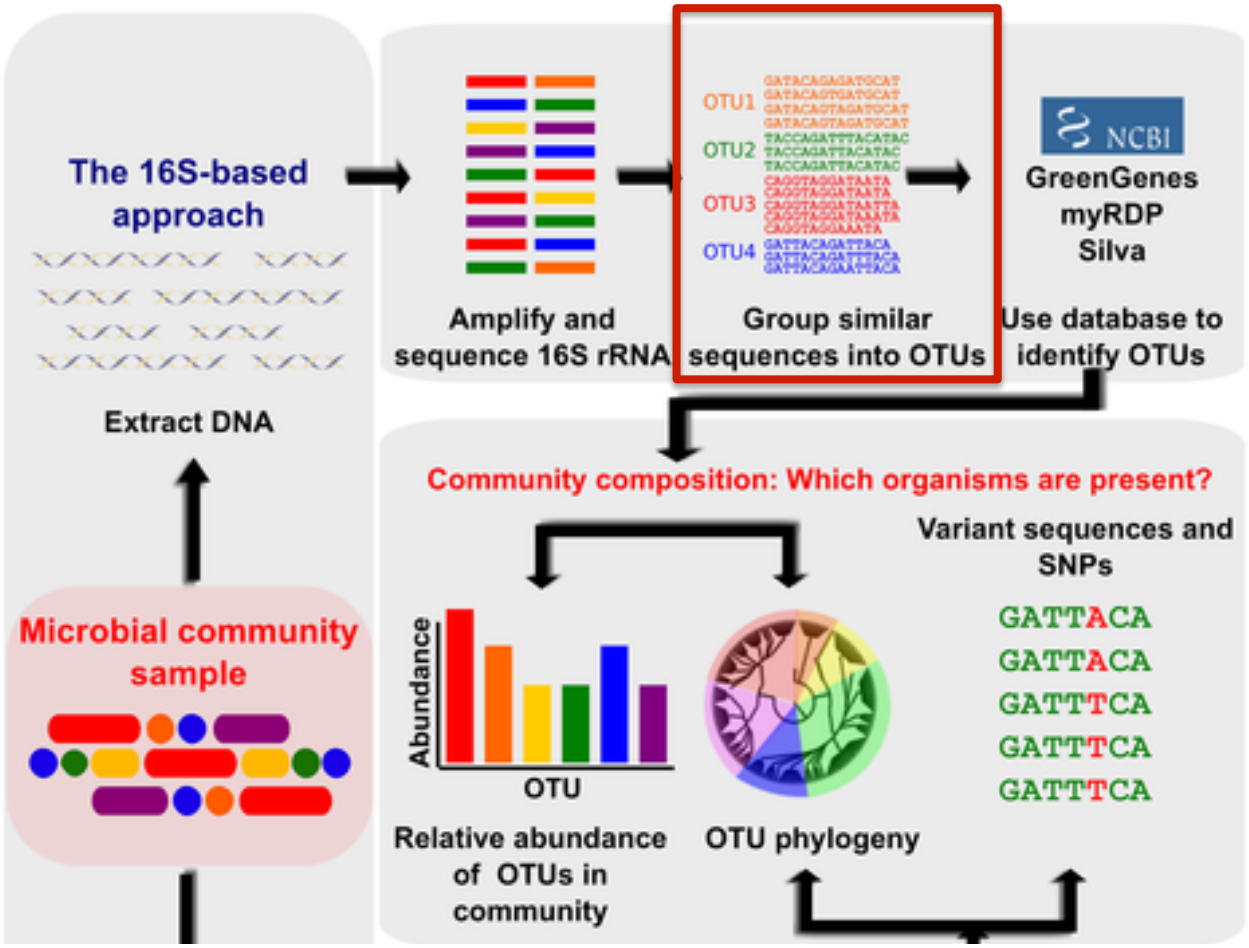
Targeted analysis



Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLOS Computational Biology 8(12): e1002808.

OTU: Operational Taxonomic Unit (cluster of similar sequence variants) used to categorize bacteria

Targeted analysis

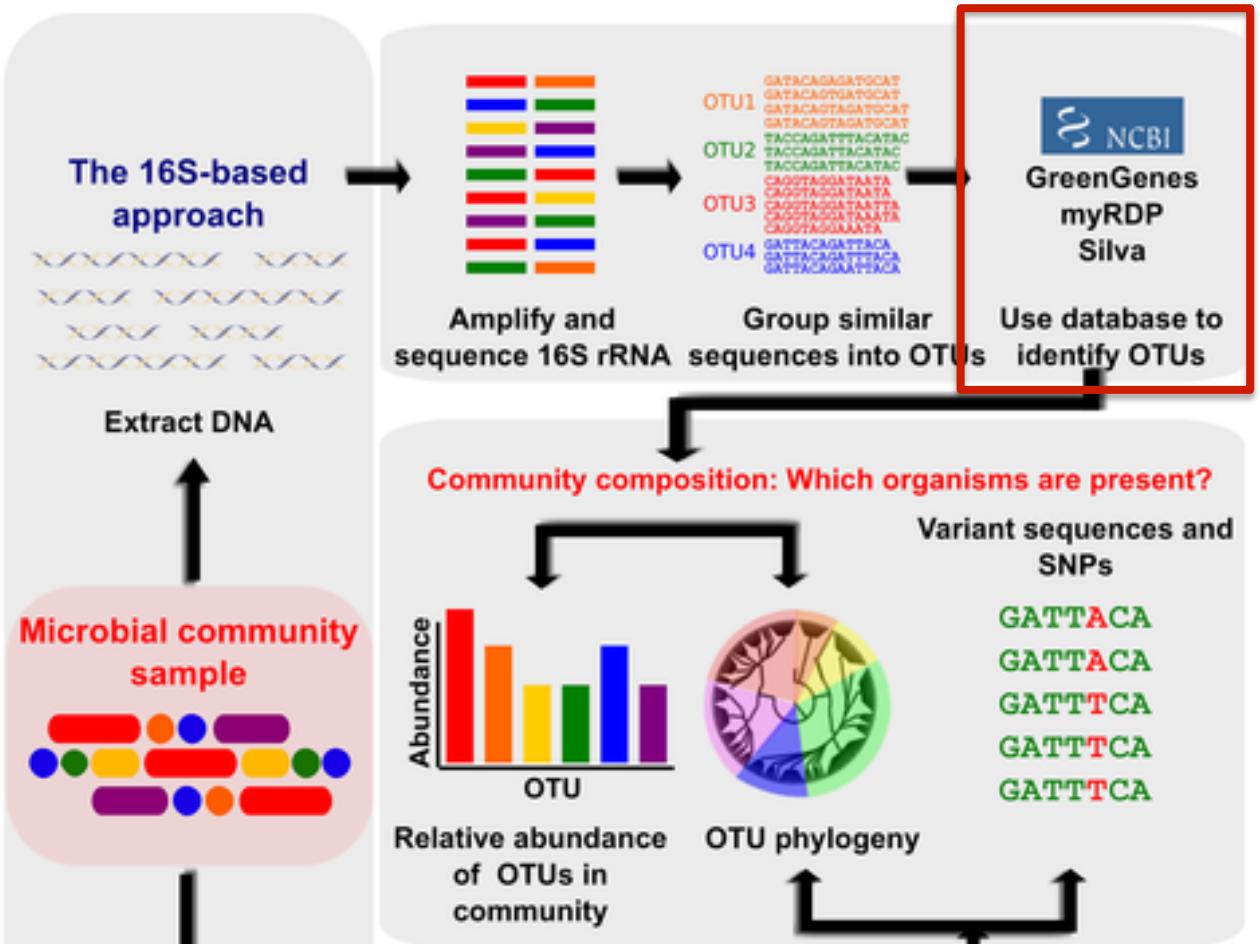


Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLOS Computational Biology 8(12): e1002808.

k-NN
Hierarchical clustering
Bayesian clustering
Greedy heuristic clustering

Tools
Mothur
USEARCH/UCLUST/UPARSE
CD-HIT

Targeted analysis



Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLOS Computational Biology 8(12): e1002808.

Linear model
Random forest

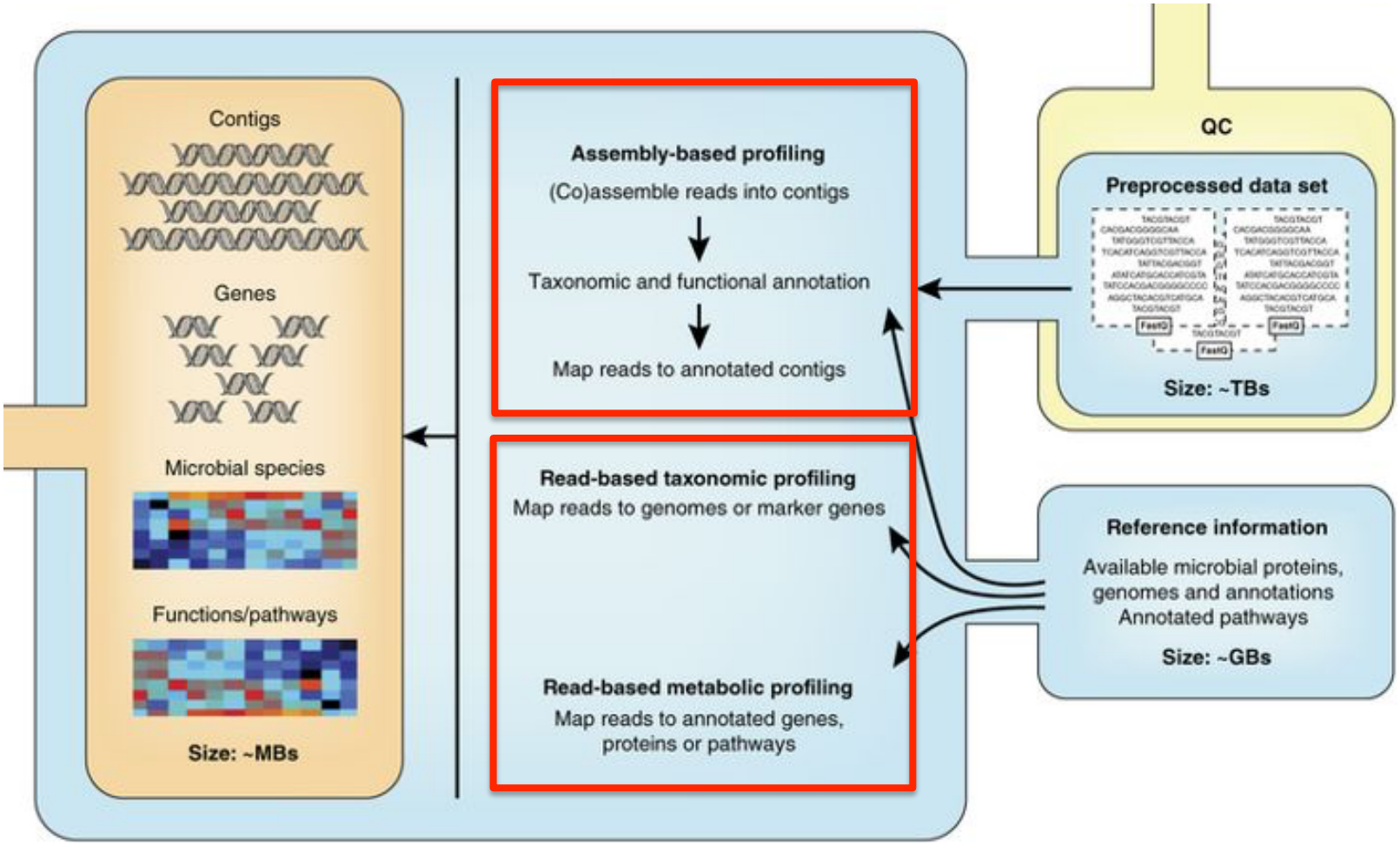
Tools

RDP Classifier
16s Classifier
PhyloSift
PhyloPithia

Shotgun metagenome analysis

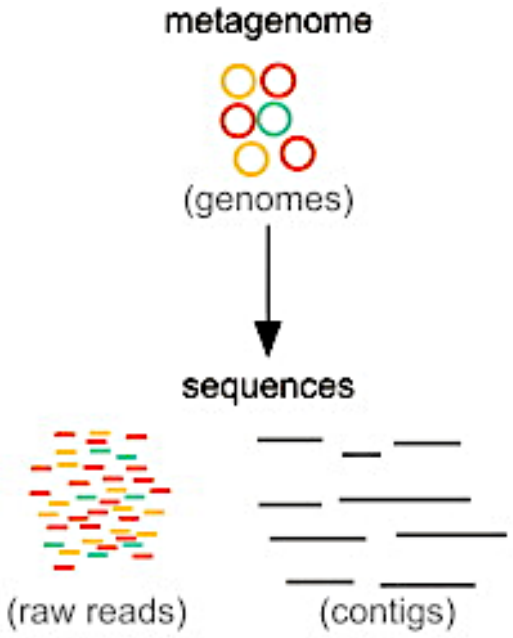
- Full sequencing of the genomic content of an environmental sample.
- Two general methods in analysis:
  - **Assembly-based:** assemble the sequences, then classify the contigs from the assembly into 'bins', followed by gene prediction, annotation, and some form of quantifying and normalizing data for comparison across samples
  - **Read-based:** analyse the unassembled reads directly against a database of interest, then assign taxonomy and function when possible

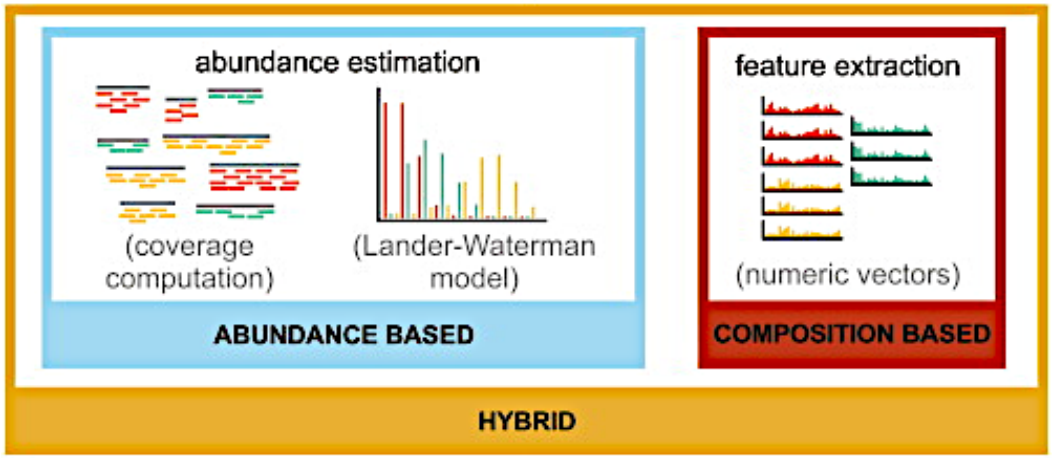
Shotgun metagenome analysis



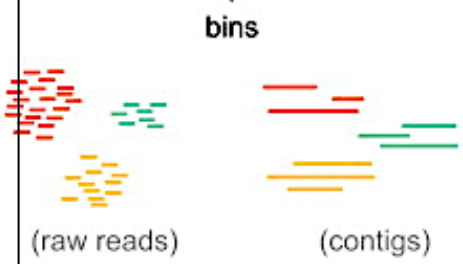
Quince, C *et al.* Shotgun metagenomics, from sampling to analysis, (2017) Nature Biotechnology (35):833–844

Metagenome analysis - Binning





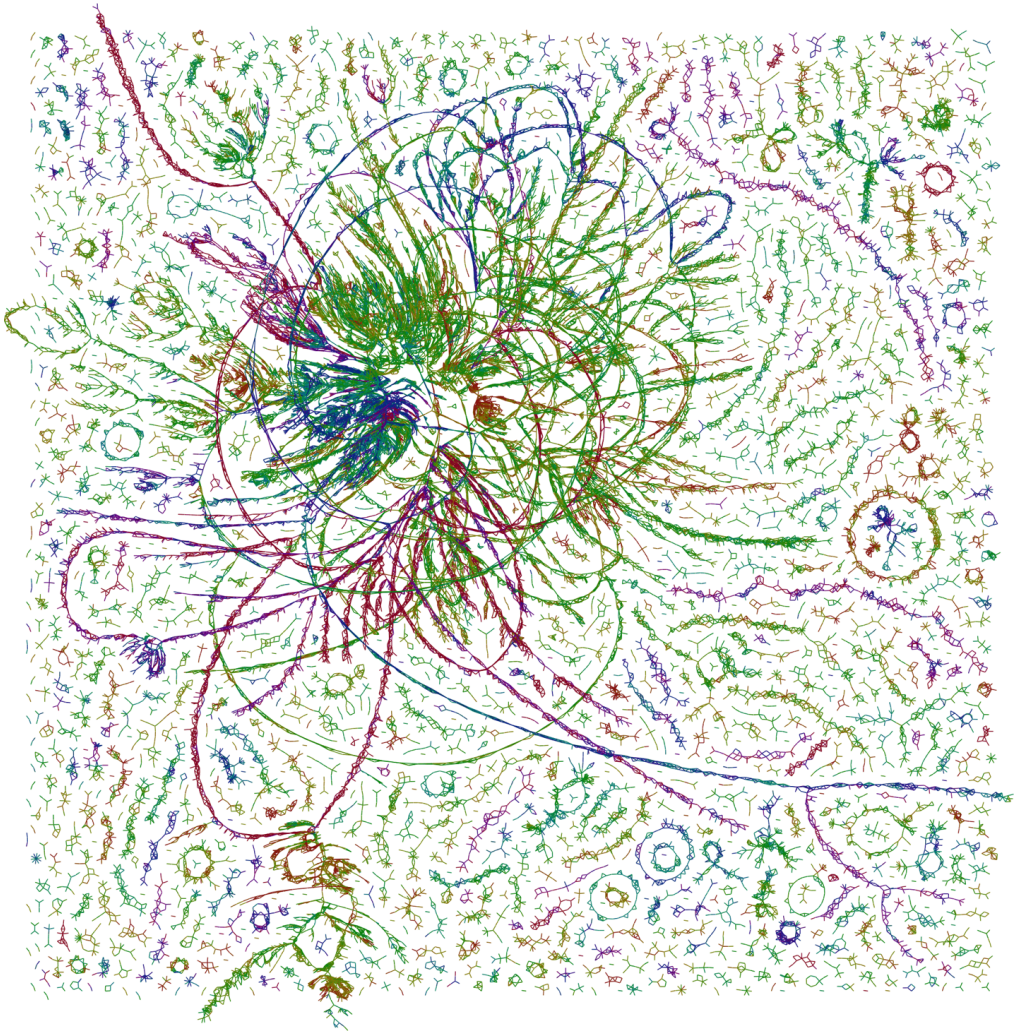
ML Model
Linear regression
Int. Markov Model
PCA
SVD
Lots of Clustering!
k-means
k-medioids
Gaussian mixture model
Greedy heuristic
Bayesian clustering
Spectral clustering



Tools
CONCOCT
MetaBAT
MaxBin

Sedlar, K *et* al, Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Computational and Structural Biotechnology Journal 15:48-55. 2017

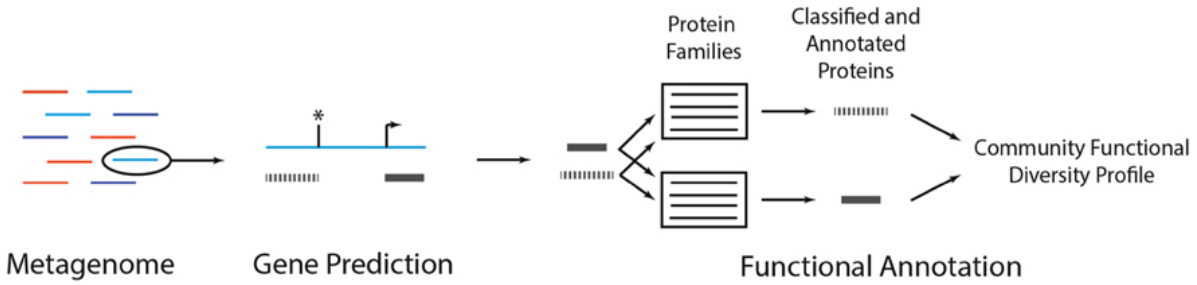
Shotgun metagenome analysis



http://armbrustlab.ocean.washington.edu/seastar

Shotgun metagenome analysis

- Let's say you have a metagenome assembly
- Now you have to annotate it to get functional information



ML Model
HMM
Neural network
Int. Markov models

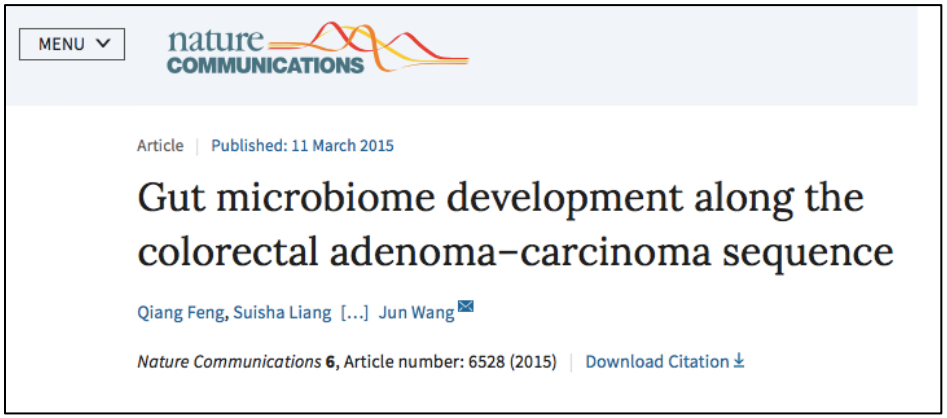
Tools
MetaProdigal
MetaGeneMark
FragGeneScan

Sharpton, T. An introduction to the analysis of shotgun metagenomic data. Front. Plant Sci., 16 June 2014

What next?

- At the end, you normally end up with quantitative information related to:
  - Taxonomic counts
  - Feature counts (genes, protein families)
- These can go into standard downstream packages for analysis (phyloseq, MEGAN, etc)
  - Normally involves performing some form of ordination (PCoA, MDS, etc)

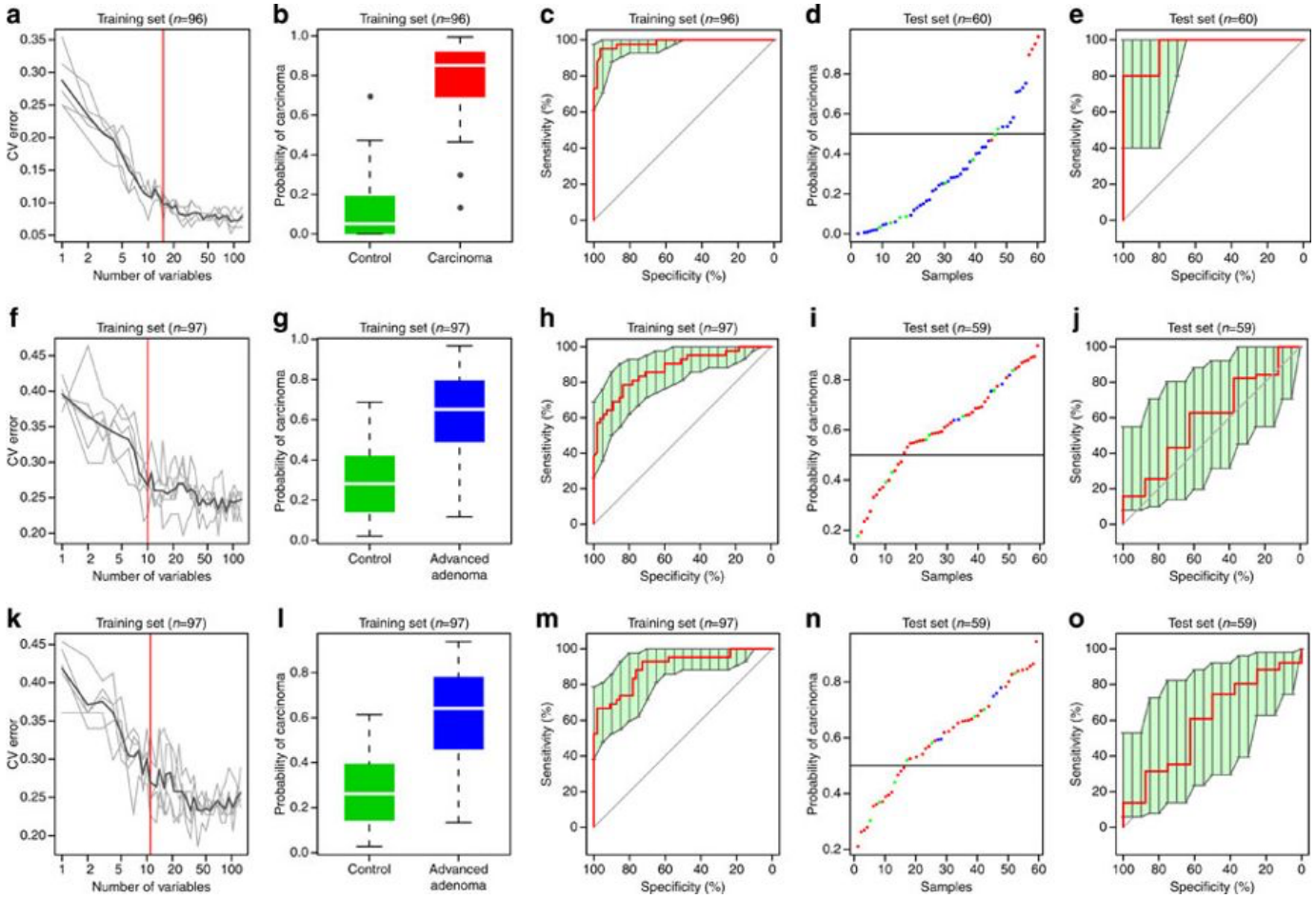
ML used for classification



MLG-based classification of adenoma or carcinoma

To illustrate diagnostic value of the faecal microbiome for colorectal cancer, we constructed a random forest classifier that could detect carcinoma samples. Five repeats of 10-fold cross-validation (that is, 50 tests) in the training set consisted of 55 controls and 41 carcinoma samples led to the optimal selection of 15 MLG markers that performed nicely on the training set (Fig. 5a–c, Supplementary Data 1 and 5). The

Figure 5 : Gut MLGs classify colorectal carcinoma and adenoma samples from healthy controls.



Nice literature overview

https://arxiv.org/pdf/1510.06621.pdf

Machine learning for metagenomics: methods and tools

Hayssam Soueidan, Macha Nikolski

(Submitted on 22 Oct 2015 (v1), last revised 8 Mar 2016 (this version, v2))

Owing to the complexity and variability of metagenomic studies, modern machine learning approaches have seen increased usage to answer a variety of question encompassing the full range of metagenomic NGS data analysis. We review here the contribution of machine learning techniques for the field of metagenomics, by presenting known successful approaches in a unified framework. This review focuses on five important metagenomic problems: OTU-clustering, binning, taxonomic profiling and assignment, comparative metagenomics and gene prediction. For each of these problems, we identify the most prominent methods, summarize the machine learning approaches used and put them into perspective of similar methods. We conclude our review looking further ahead at the challenge posed by the analysis of interactions within microbial communities and different environments, in a field one could call "integrative metagenomics".

Subjects: **Genomics (q-bio.GN)**
Cite as: **arXiv:1510.06621 [q-bio.GN]**
(or **arXiv:1510.06621v2 [q-bio.GN]** for this version)

ML – Overview

	Representations	Learning	Applications example	Tools example
Classification/ Regression	**Instances**	*k*-NN	binning, OTU clustering	DOTUR
		Support Vector Machines	gene prediction, Comparative MG	MetaDistance
	Linear models	linear regression	binning	Tetra
		logistic regression	gene prediction, Comparative MG	MetaGene, MetaPhyl
		large scale linear model	taxonomic assignment	Vowpal Wabbit
	Decision trees	Random Forest	taxonomic assignment	16S Classifier
	Neural networks	neural networks	gene prediction	Orphelia
	Network	Hidden Markov Model	gene prediction	FragGeneScan
		Interpolated Markov Model	binning, gene prediction	Glimmer-MG, SCIMM
Dimension reduction	**Linear combinations of features**	PCA	binning	CONCOCT
		SVD	binning	LSA
Clustering	**Means**	*k*-means	binning	SCIMM, MetaCluster
	Medoids	*k*-medoids	binning	MultiBin
	Dendrogram	hierarchical clustering	OTU clustering	ESPRIT, MOTUR
	Mixtures / Soft partitions / Likelihoods	Gaussian mixture models	binning	CONCOCT
		Bayesian clustering / stochastic search clustering	OTU clustering, binning	CROP,BACDNAS, BEBaC, LikelyBin
		spectral clustering	binning	CompostBin
	NA	greedy heuristic clustering	binning, OTU clustering	DNAClust, USEARCH

ML – OTU Clustering

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
Dotur	16S, rpoB	multiple seq. alignment	hierachical clustering	github.com/mothur/DOTUR	2005	2005	none
Mothur	16S	alignment against a reference DB	hierachical clustering	github.com/mothur/Mothur	2009	2015	none
ESPRIT	16S	*k*-mer prefiltering, pairwise alignment	hierachical clustering	upon request	2009	unknown	Dotur
Usearch / Uclust	16S	pairwise alignment	greedy clustering	drive5.com/usearch	2010	unknown	CD-HIT
GramCluster	16S	suffix tree to access sequences, grammar-based distance, Lempel-Ziv sequence representation, grammar-based comparisons	greedy clustering	bioinfo.unl.edu/ gramcluster.php	2010	2010	CD-HIT, Uclust
ESPRIT-Tree	16S	*k*-mer prefiltering, probabilistic sequences, pairwise alignment	pseudo-hierarchical clustering	web access	2011	unknown	Uclust, CD- HIT, ESPRIT
DNAclust	16S	suffix trees, *k*-mer prefiltering, length sorting, pairwise alignment	greedy clustering	dnaclust.sourceforge.net	2011	2013	Uclust, CD-HIT
CROP	16S	Gaussian mixture model to describe the data, MCMC, pairwise alignment	Baeysian clustering	github.com/tingchenlab/CROP	2011	2014	ESPRIT, mothur
CD-HIT	16S	*k*-mer prefiltering, length sorting, pairwise sequence alignment	greedy clustering	cd-hit.org	2012	2015	Uclust
BEBaC	16S	*k*-means, pre-clustering based on k-mers, multiple seq. alignment	stochastic search clustering	available upon request	2012	unknown	ESPRIT- Tree, Uclust, CROP
Uparse	16S, ITS	maximum parsimony model, abundance sorting, pairwise alignment	greedy clustering	drive5.com/uparse	2013	unknown	QIIME, mothur
BACDNAS	16S	sequences modeled as Markov chains (no alignment!)	Bayesian Clustering with the Dirichlet process prior	www.helsinki.fi/bsg/software/ BACDNAS/	2013	2013	ESPRIT-Tree, BEBaC, CROP
Swarm	16S	*k*-mer prefiltering, pairwise alignment	single-linkage agglomerative clustering	github.com/torognes/swarm	2014	2015	CD-HIT, DNAclust, Usearch
OTUCLUST	16S-10, ITS-10, 16S-R	abundance sorting, pairwise alignment	greedy clustering	github.com/ compmetagen/ micca/wiki	2015	2015	Uparse, QIIME

ML - Binning

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
TETRA	only 16s	4-mers	linear regression	www.megx.net/tetra	2004	unknown	none
CompostBin	bacterial genomes	6-mers + alignment-based weighting scheme	PCA+spectral clustering	bobcat.genomecenter.ucdavis.edu/souravc/compostbin/	2008	2012	none
LikelyBin	bacterial genomes	*k*-mers	partitions / stochastic search (MCMC)	ecotheory.biology.gatech.edu/downloads/likelybin	2009	unknown	none
SCIMM	only 16s	preclustering	Interpolated Markov Models + *k*-means	www.cbcb.umd.edu/software/scimm	2010	2012	PHYSCIMM
AbundanceBin	bacterial genomes	*k*-mer abundances	stochastic search, expectation-maximization	omics.informatics.indiana.edu/AbundanceBin	2011	2013	MetaCluster
MetaCluster	bacterial, whole metagenome (viral included)	*k*-mer frequencies + Spearman footrule distance	*k*-means	i.cs.hku.hk/~alse/MetaCluster	2012	2014	AbundanceBin, Toss
MultiBin	bacterial genomes	pairwise alignment	similarity graph + *k*-medoids	none	2012	NA	AbundanceBin
Toss	bacterial genomes	unicity of *k*-mers	similarity graphs + MCL	www.cs.ucr.edu/~tanaseio/toss.htm	2012	unknown	CompostBin
MultiMetaGenome	bacterial genomes	scaffold coverage, 4-mers frequencies, GC content, ORF, marker proteins, taxonomic assignement	linear regression + local PCA	github.com/MadsAlbertsen/multi-metagenome	2013	2014	ESOM
CONCOCT	bacterial genomes	*k*-mers and coverage	Gaussian mixture models	github.com/BinPro/CONCOCT	2014	2015	MetaWatt, CompostBin, SCIMM, LikelyBin
MaxBin	bacterial genomes	4-mers, coverage, marker genes analysis	stochastic search, expectation maximization	sourceforge.net/projects/maxbin/	2014	2015	ESOM
LSA	bacterial genomes, some remarks on phage analysis	locally sensitive hashing of *k*-mers,	SVD + *k*-means	github.com/brian-cleary/LatentStrainAnalysis	2015	2015	CONCOCT, GroopM
MetaBAT	bacterial genomes	4-mers, coverage, alignment, pre-assembled contigs	custom *k*-medoids	bitbucket.org/berkeleylab/metabat	2015	2015	CONCOCT, Canopy, MaxBin, GroopM
DNAClust	AMD dataset	*k*-mers + filtering	greedy clustering	dnaclust.sourceforge.net	2011	2013	Uclust, CD-HIT

ML – Taxonomic Classification

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
RDP classifier	16S, LSU	8-mers, with specific priors and genus-specific conditional probabilities	naive Bayes / *k*-NN	rdp.cme.msu.edu	2007	2015	none
NBC	16S, ITS, LSU, viruses, fungal	*k*-mers	naive Bayes	nbc.ece.drexel.edu	2008	2012	BLAST
Phymm	whole metagenome	variable-length *k*-mers	Interpolated Markov Models	www.cbcb.umd.edu/software/phymm/	2009	2012	PhyloPythia, Carma
pplacer	16S	none	likelihood based phylogenetics	matsen.fredhutch.org/pplacer/	2010	2015	RAxML
TAC-ELM	whole metagenome	GC-content, 3–4 mers	extreme learning machines, neural networks	cs.gmu.edu/~mlbio/TAC-ELM	2010	unknown	TAC-ELM, Phymm, BLAST, PhymmBL, PhyloPythia
FCP	bacterial genomes	10-mers + Laplace Smoothing	custom Naive Bayes	kiwi.cs.dal.ca/Software/FCP	2011	2015	PhyloPythiaS, TACOA, BLAST, LCA
Taxy	whole metagenome	*k*-mer composition of the sample	mixture modeling	gobics.de/peter/taxy	2011	unknown	Carma, Treephyler, Phymm, Galaxy
PhyloPithia / PhyloPithiaS	bacterial, whole metagenome	4, 5 and 6-mer frequencies	structural SVM	web access	2012	unknown	NBC, best BLASTN
Quikr*	16S	*k*-mer frequency	dimensionality reduction (compressed sensing), convex optimization	sourceforge.net/projects/quikr	2013	2014	RDP
WGSQuikr*	whole metagenome	k-mer frequency	dimensionality reduction (compressed sensing), convex optimization	sourceforge.net/projects/wgsquikr	2014	2014	RDP
SEK*	16S	k-mer composition of the sample	Kernel density estimator and mixture density models	github.com/dkoslicki/SEK	2014	2015	Quikr, Taxy, BeBAC
ARK*	16S	k-mer frequency	k-means then SEK	github.com/dkoslicki/ARK	2015	2015	Quikr, SEK, RDP
AKE	bacterial genomes	*k*-mers (length normalized, importance weight and over/under representation)	H2SOM classifier (neural network)	web access	2014	unknown	NBC, PhyloPythiaS, WebCarma
PhyloSift	16S	alignment against HMMs of gene families	Bayesian model	github.com/gjospin/PhyloSift	2014	2014	QIIME
16S Classifier	16S	2–6 normalized *k*-mers frequencies	Random Forest	metabiosys.iiserb.ac.in/16Sclassifier/application.php	2015	2015	RDP, BLAST
Vowpal Wabbit	whole metagenome	4–12 mers	large scale linear models, feature hashing	unavailable	2015	NA	BWA-MEM, NBC
CSSS	viral, bacterial	*k*-mer based distances + alignment	1-NN	collaborators.oicr.on.ca/vferretti/borozan_csss/csss.html	2015	2015	Phymm, NBC, PAIPhy, Kraken, PAUDA

ML – Gene Prediction

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
MetaGene	bacterial	GC%, ORF lengths, distance from left start codons, distance between neighboring ORFs	Hidden Markov Model	omics.informatics.indiana.edu/FragGeneScan	2006	2015	Glimmer, MetaGene
MetaGeneMark	bacteria, archaea	3, 4, 5 and 6-mer and nucleotide frequencies	three-periodic Markov chain model	web access	2010	unknown	GeneMarkS, MetaGeneAnnotator, MetaGene
FragGeneScan	bacteria, whole metagenome	codon usage bias, sequencing error models and start/stop codon patterns	Hidden Markov Model	omics.informatics.indiana.edu/FragGeneScan	2010	2015	BLASTX, MetaGene
Orphelia	bacterial	codon, dicodon usage, orf length, tis, gc content	neural network	orphelia.gobics.de	2008	unknown	MetaGene
MetaGeneAnnotator	bacteria, archaea, prophage	as MetaGene + models for prophage genes and RBS	logistic regression	metagene.cb.k.u-tokyo.ac.jp	2008	unknown	GeneMarkS, Glimmer, MetaGene
MetaProdigal	bacterial	start site information, translation table, hexamer statistics, RBS motifs and upstream base composition	log-likelihood function on subsets of training data	github.com/hyattpd/Prodigal	2012	2015	MetaGeneAnnotator, MetaGeneMark
Glimmer-MG	whole metagenome	gene length, start/stop codon presence, TIS, rbs, start codon usage	Interpolated Markov Models	www.cbcb.umd.edu/software/glimmer-mg	2012	2014	MetaGeneAnnotator, MetaGeneMark, FragGeneScan