Metagenomics 2

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

Learning Outcome:

- Explain fundamental questions raised by the study of microbial communities that can be addressed through metagenomics.

- Able to select appropriate bioinformatics tools and process of determining the taxonomic and functional composition of metagenomic samples.

Recap: What is Metagenomics?

Microbial community genomics

- study of complete microbial communities directly in their natural environments, including
  - skin, mouth, and gut samples to understand human related diseases,
  - soil samples to study plant soil-microbe interactions, and
  - marine water samples
- metagenomics does not require isolation and culturing of individual microbes: "culture independent"
- focus on microbial communities: a sample can contain more than 10,000 species

What is a microbiome?

- The complete set of microbes that live in a particular microbial ecosystem: within and on the human body (human microbiome), specific plants (grapevine microbiome), or marine areas
- Includes also the entire collection of microbial genes provided by these microbial communities

DID YOU KNOW?

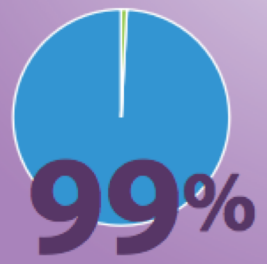
Our bodies contain 37 trillion human cells - but 100 trillion bacterial cells! That's **100,000,000,000,000 microbes!**

This may sound scary, but not all microbes make us sick. In fact, the vast majority of microbes living in and on our bodies play a role in many fundamental life processes:

- digestion
- development
- protection from pathogens

2.5 LBS = WEIGHT
of the microbiome

3 PINTS = VOLUME
of the microbiome



Microbes contribute an extra 2,000,000 genes to each person. Compare this to the 20,000 genes in the human genome.

Learning more about your microbiome:

This brochure is just an introduction to the human microbiome, and there are many more things to learn and stories to tell - you can find the unabridged version of this document at:

American Academy of Microbiology:
http://bit.ly/HumanMicrobiome

The best information on the human microbiome comes from the experts studying it. If you would like to learn more about the human microbiome, you can turn to the source below:

NIH Human Microbiome Project
https://commonfund.nih.gov/hmp

The American Academy of Microbiology is the honorific branch of the American Society for Microbiology, a non-profit scientific society with almost 40,000 members. Fellows of the AAM have been elected by their peers in recognition of their outstanding contributions to the field of microbiology. Through its colloquium program, the AAM draws up on the expertise of these fellows to address critical issues in microbiology.

The Human Microbiome FAQ report is based on the deliberations of 13 expert scientists who gathered for a day to develop science-based answers to questions the public might have about the human microbiome. The report was reviewed by all participants, and every effort was made to ensure that the information is accurate and complete. However, the report is not intended to advocate any particular position or actions, nor to replace the advice of an individual's health care provider.

The AAM does not offer medical advice and encourages consumers to be educated and informed when making medical decisions.

Contents of the report may be distributed further so long as the authorship of the AAM is acknowledged and this disclaimer is included.

FAQ

HUMAN MICROBIOME

https://www.asm.org/images/stories/documents/ColloquiaDoc/ASM%20Trifold.pdf

YOUR MICROBIOME

The human body contains trillions of microbes, including bacteria, fungi and viruses. The community of microbes living in intimate association with our bodies, and the genes they contain, make up the **human microbiome**.

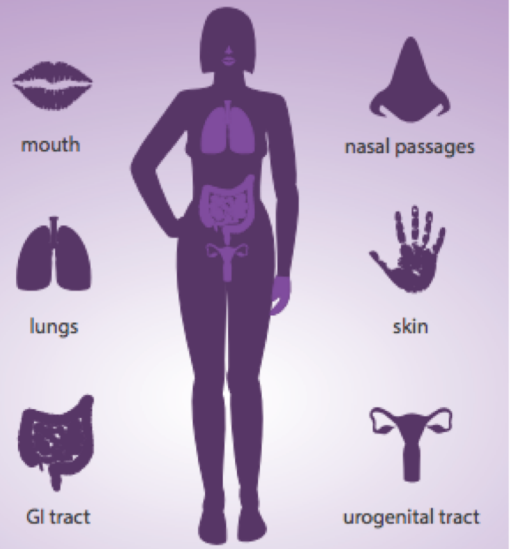
The collection of microbes living in and on our bodies isn't random. Our microbiome has co-evolved with humans over millions of years and plays many essential roles.

MICROBIOME CROSSTALK

Even though your gut microbiota seems to be very far removed from other body parts, these helpful microbes make signalling molecules called neurotransmitters that affect brain development and behavior.

WHERE ARE THEY?

Wherever the human body is exposed to the outside world, there is a microbial community. Microbes live on the entire surface of our skin, and the linings of our nasal passages, lungs, digestive systems and urogenital tracts.



WHO'S THERE?

A human body is actually only about 25% human cells, with many thousands of species of bacteria and other microbes.

BUILDING BLOCKS

The gut microbiome acts as a highly effective bioreactor. Our gut microbes not only extract energy and nutrients from the food we eat, but they also make essential molecules and vitamins that we cannot make for ourselves.

WHERE DOES OUR MICROBIOME COME FROM?

BIRTH:
We get our microbiome mostly from other humans. A newborn gets its microbes from:

- its mother's birth canal
- skin of its mother and other care-givers

BREAST MILK:
Breast milk has been fine-tuned over millions of years to provide:

- diverse microbes to populate the baby's gut
- nutrients for the baby and baby's microbes
- vitamins and antibodies

ENVIRONMENT:
For the rest of the baby's life, it will continuously encounter new microbes from:

- soil and water
- people, pets, plants
- new and diverse foods

https://www.asm.org/images/stories/documents/ColloquiaDoc/ASM%20Trifold.pdf

Metagenomics sampling projects

- Human Microbiome Project (HMP)
  - https://www.hmpdacc.org/
- Earth Microbiome Project
  - http://www.earthmicrobiome.org/
- Tara Oceans Expedition
  - https://doi.pangaea.de/10.1594/PANGAEA.840721
- Hospital Microbiome Project
  - http://hospitalmicrobiome.com/
- DIABIMMUNE Microbiome Project
  - https://pubs.broadinstitute.org/diabimmune

16S vs Metagenomics

- 16S is targeted sequencing of a single gene which acts as a marker for identification
- Pros
  - Well established
  - Sequencing costs are relatively cheap (~50,000 reads/sample)
  - Only amplifies what you want (no host contamination)
- Cons
  - Primer choice can bias results towards certain organisms
  - Usually not enough resolution to identify to the strain level
  - Need different primers usually for achaea & eukaryotes (18S)
  - Doesn't identify viruses

16S vs Shotgun Metagenomics

- Metagenomics: sequencing all the DNA in a sample
- Pros
  - No primer bias
  - Can identify all microbes (euks, viruses, etc.)
  - Provides functional information ("What are they doing?")
- Cons
  - More expensive (millions of sequences needed)
  - Host/site contamination can be significant
  - May not be able to sequence "rare" microbes
  - Complex bioinformatics

16S rRNA Tools

Very standard pipeline.

Two commonly used software for taxon identifications

- QIIME (2) (Quantitative Insights Into Microbial Ecology)
- mothur
- Two microbial gene data bases (GreenGenes and SILVA) for 16S rRNA gene analysis

- PICRUSt (2) (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) is a software for predicting functional abundances based only on marker gene sequences.

=> More details in the next lecture with lab!

Shotgun Metagenomics: Who is there?

- Goal: Identify the relative abundance of different microbes in a sample given using metagenomics
- Problems:
  - Reads are all mixed together
  - Reads can be short (~100bp)
  - Lateral gene transfer
- Three Broad Approaches (with or without references)
  - Assembly (to reconstruct genome and genome annotation, for binning, and so on)
  - Binning (mostly after assembly)
  - Annotation (for taxonomy profiling, for functional analysis, and so on)

Assembly

- SPAdes (http://bioinf.spbau.ru/spades)
  - Short reads assembler
- MEGAHIT (https://github.com/voutcn/megahit)
  - For assembling large and complex metagenomics data
  - MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph (https://www.ncbi.nlm.nih.gov/pubmed/25609793)
- SSPACE (http://www.baseclear.com/bioinformatics-tools)
  - Scaffolding pre-assembled contigs using SSPACE
  - SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.
- Velvet (https://github.com/dzerbino/velvet)
  - Fast assembler for small genomes
- SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html)
  - Designed for assembling large plant and animal genomes
- Omega (https://omega.omicsbio.org/)
  - Designed for mid range reads (like Illumina MiSeq)

Merge Overlapping Paired-end Reads

- Typical Illumina fragment libraries would use fragment length (F) ~ 450bp but this is variable. For paired-end reads, you want to make sure that F is long enough to fit two reads. This means you need F to be at least 2L where L is a read length.
  - As L=100 or 150bp these days for most people, using F~450bp is fine, there is a still a safety margin in the middle.
- Illumina read lengths are now moving to >150bp on the HiSeq (and have already been on the GAIIx), and to >250bp on the MiSeq.
- This means that the standard library size F~450bp has become too small, and paired end reads will overlap. Secondly, the new enyzmatic Nextera library preparation system produces a wide spread of F sizes compared to the previous TruSeq system. With Nextera, we see F ranging from 100bp to 900bp in the same library. So some reads will overlap, and others won't. It's starting to get messy.

Merge Overlapping Paired-end Reads

- Merging paired-end shotgun reads generated on high-throughput sequencing platforms can substantially improve various subsequent bioinformatics processes, including genome assembly, binning, mapping, annotation, and clustering for taxonomic analysis.

Merge Overlapping Paired-end Reads Tools

- iTag, BIPES, SHERA, PANDASeq, COPE, FLASH, PEAR, BBMerge
- FLASH (2011, 1636 citations)
  - https://ccb.jhu.edu/software/FLASH/
- PEAR (2014, 660 citations)
  - https://sco.h-its.org/exelixis/web/software/pear/doc.html
- BBMerge (2017, new)
  - https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmerge-guide/

Merge Overlapping Paired-end Reads Tools

BBMerge Guide

BBMerge is designed to merge two overlapping paired reads into a single read. For example, a 2x150bp read pair with an insert size of 270bp would result in a single 270bp read. This is useful in amplicon studies, as clustering and consensus are far easier with single reads than paired reads, and also in assembly, where longer reads allow the use of longer kmers (for kmer-based assemblers) or fewer comparisons (for overlap-based assemblers). And in either case, the quality of the overlapping bases is improved. BBMerge is also capable of error-correcting the overlapping portion of reads without merging them, as well as merging nonoverlapping reads, if enough coverage is available. BBMerge is the fastest and by far the most accurate overlap-based read merger currently in existence.

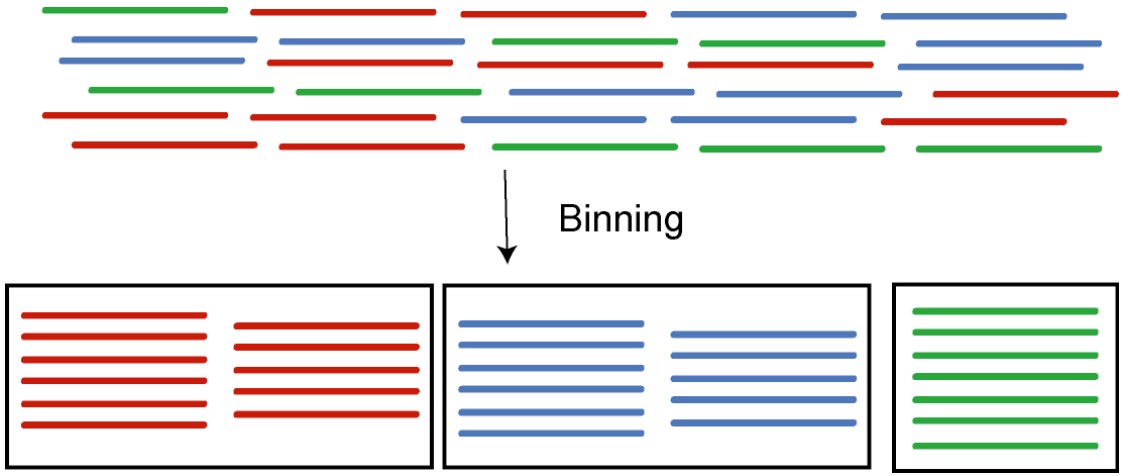
BBMerge's parameters are described in its shell script (bbmerge.sh). This file provides usage examples of various common tasks.

Broad Binning

- Attempts to group or “bin” reads into the genome from which they originated
- Composition-based
  - Uses sequence composition such as GC%, k-mers (e.g. Naïve Bayes Classifier)
  - Fast
- Sequence-based (This can be classified differently)
  - Compare reads to large reference database using BLAST (or some other similarity search method)
  - Reads are assigned based on “Best-hit” or “Lowest Common Ancestor” approach

Binning

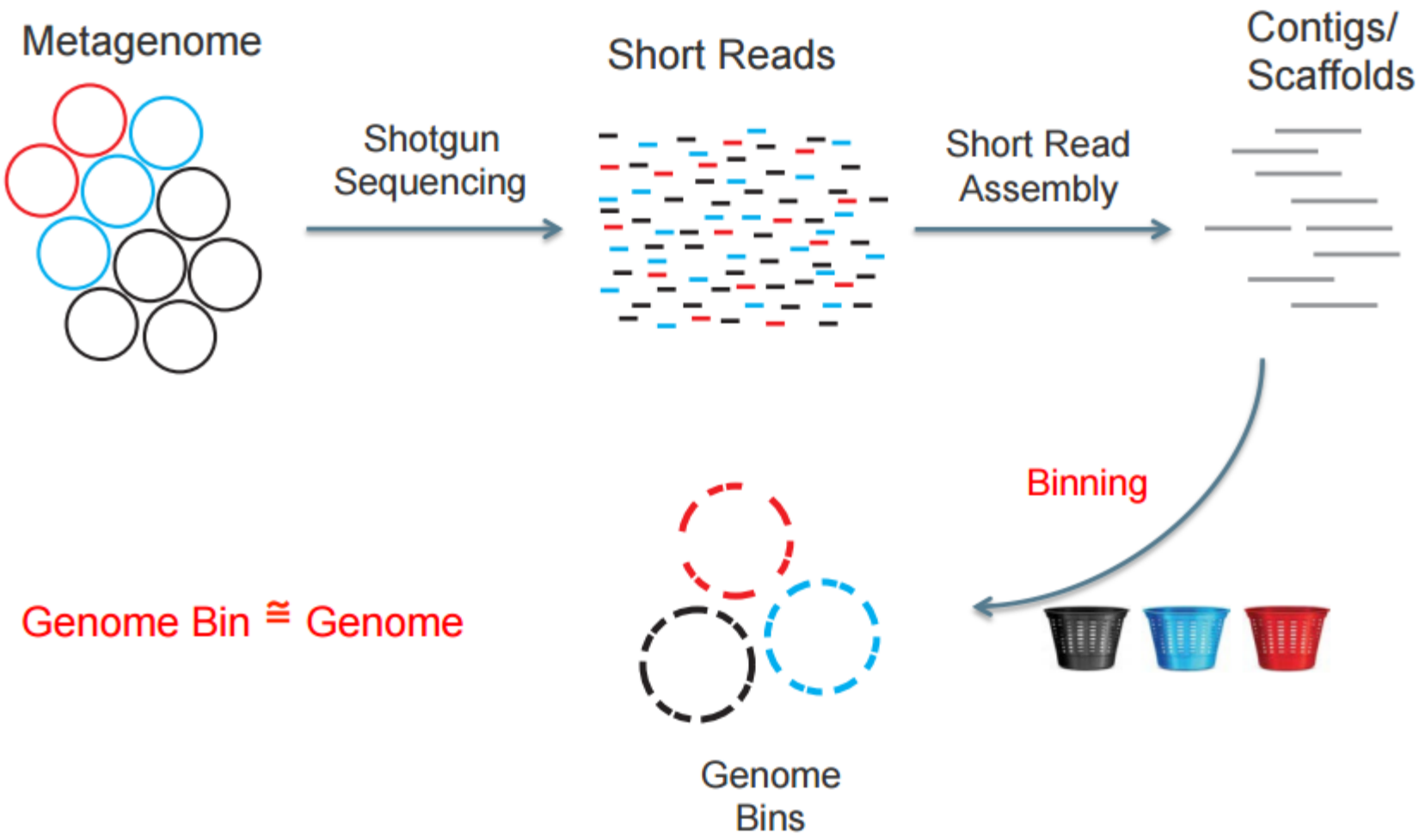
- In metagenomics, binning is the process of grouping reads or contigs and assigning them to operational taxonomic units (OTUs).
- Group DNA fragments from similar species



Composition-based methods

- Group DNA fragments using genetic features such as genome structure or composition
- Low availability and reliability of taxonomic markers
- Some species may share multiple marker with other species

Composition-based binning



l-mer Based Methods

- Tetra (Teeling et al., 2004)
- MetaCluster (Yang et al., 2009)
- MetaCluster 2.0 (Yang et al., 2010)
- AbundanceBin (Wu and Ye, 2010)
- MetaCluster 3.0 (Yang et al., 2011)
- MetaCluster 4.0 (Yang et al., 2012)
- MetaCluster 5.0 (Yang et al., 2012)
- MetaBat (2014)

DNA-composition metrics

AGCTTTTCATTCTGACTGCAACGG…

|AG |CT |TT |TC | ……..
|GC |TT |TT |CA | ……..

Measure dimer frequencies

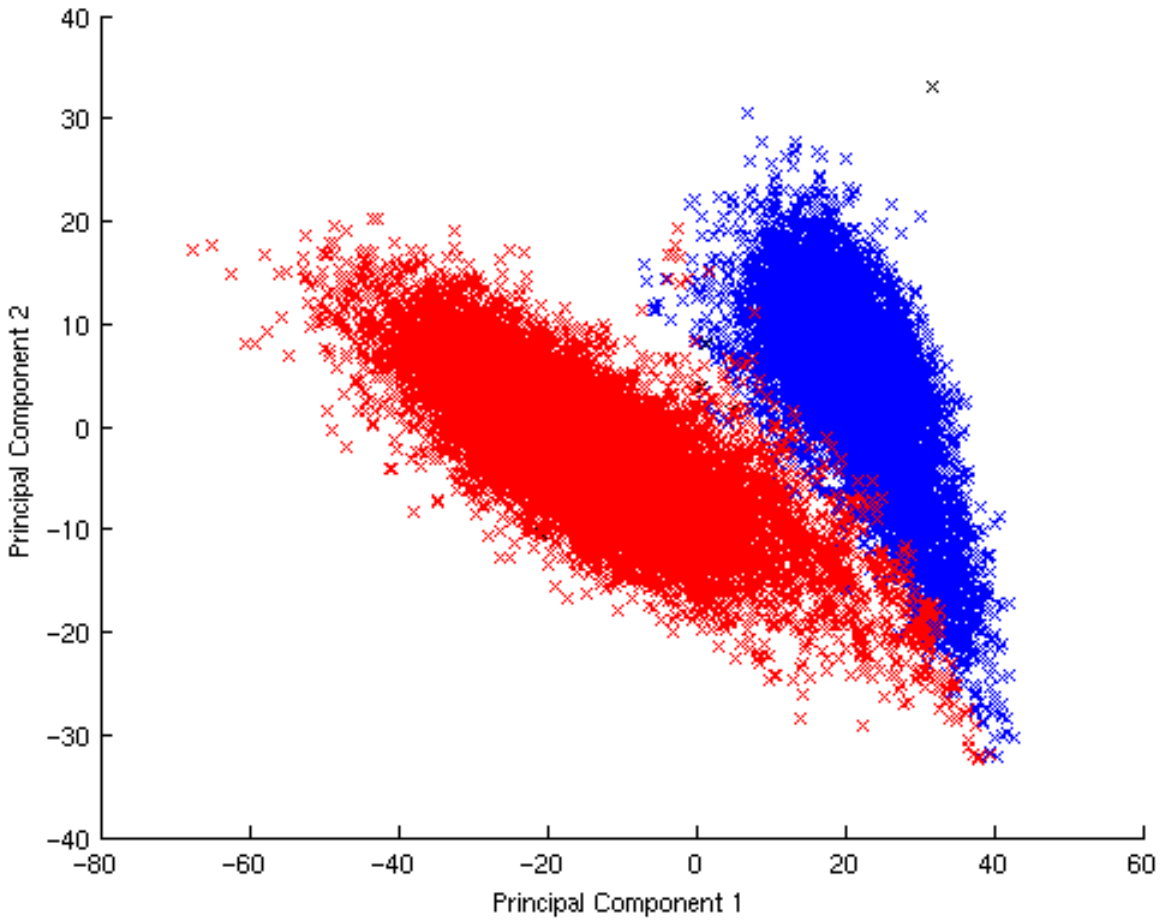
[f(AA) f(AC) f(AG) f(AT) f(CA) ..]

The K-mer Frequency Metric
CompostBin uses hexamers

DNA-composition metrics

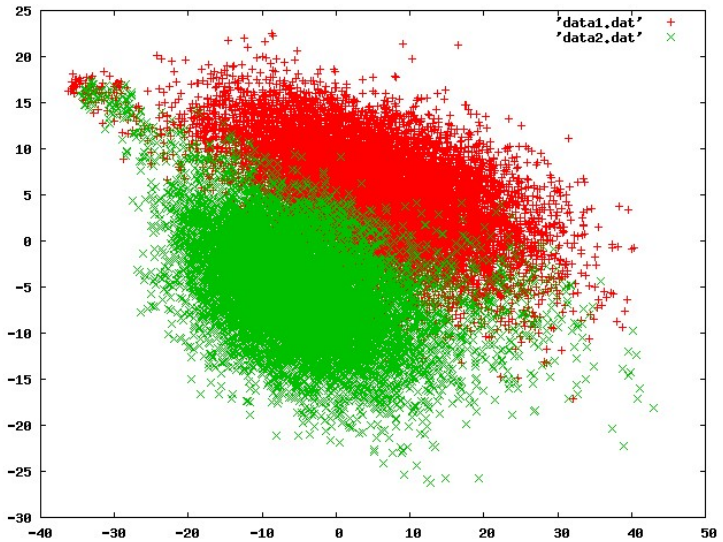
- Working with K-mers for Binning.
  - *Curse of Dimensionality* : $O(4^K)$ independent dimensions.
  - Statistical noise increases with decreasing fragment lengths.
- Project data into a lower dimensional space to decrease noise.
  - Principal Component Analysis (PCA)
  - Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

PCA separates species

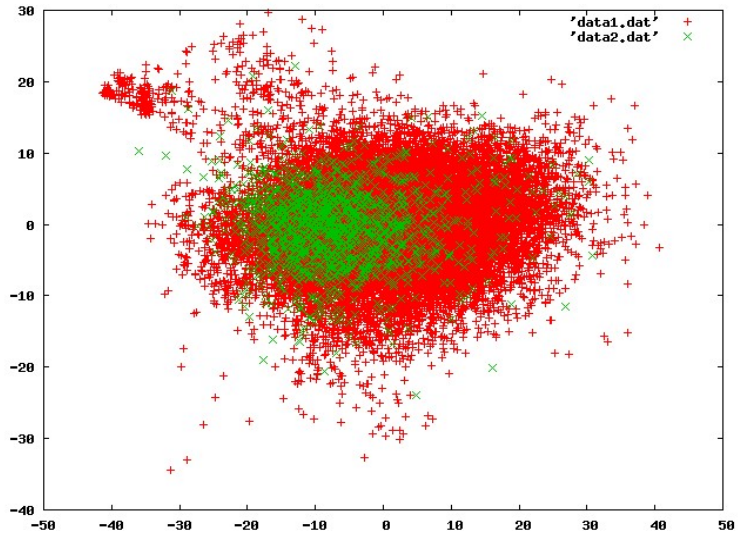


Gluconobacter oxydans[65% GC] and *Rhodospirillum rubrum*[61% GC]

Effect of Skewed Relative Abundance



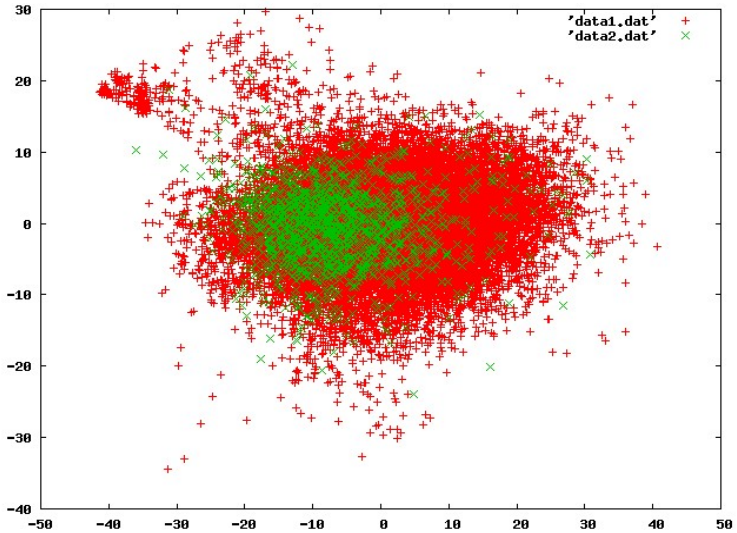
Abundance 1:1



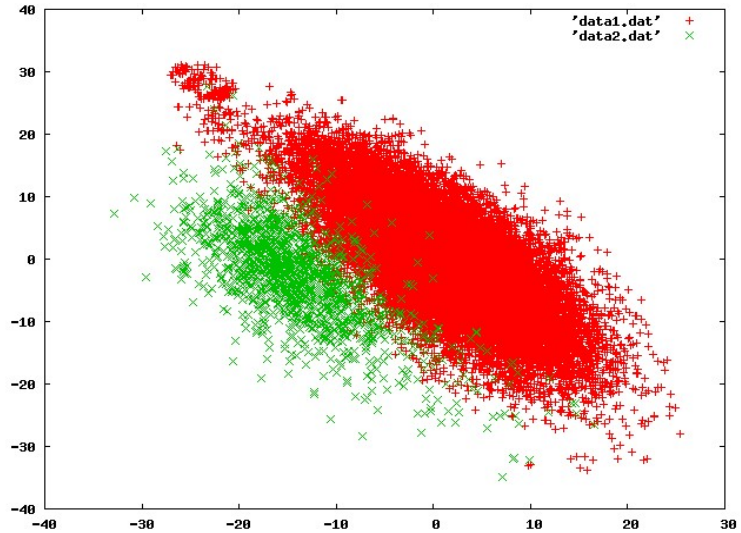
Abundance 20:1

B. anthracis* and *L. monogocytes

Weighted PCA separates species



PCA



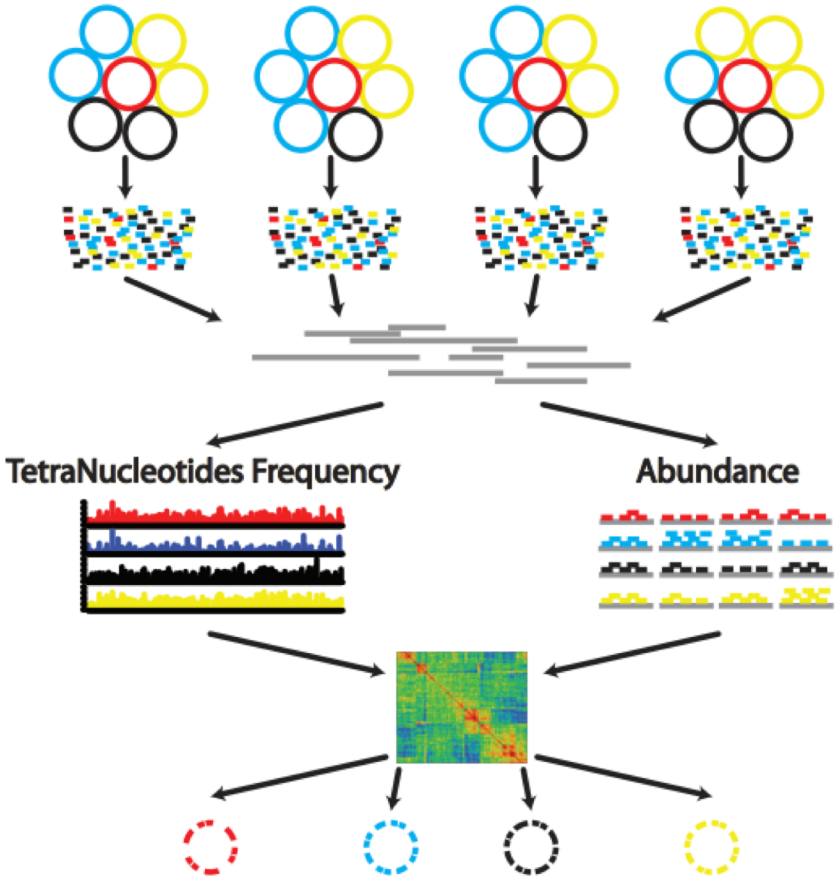
Weighted PCA

B. anthracis and *L. monogocytes : 20:1*

MetaBat

- As a pre-requisite for binning, the user must create BAM files by aligning the reads of each sample separately to the assembled metagenome.
- MetaBAT takes an assembly file (fasta format, required) and sorted bam files (one per sample, optional) as inputs.
- For each pair of contigs in a metagenome assembly, MetaBAT calculates their probabilistic distances based on tetranucleotide frequency (TNF) and abundance (i.e., mean base coverage), then the two distances are integrated into one composite distance.
- All the pairwise distances form a matrix, which then is supplied to a modified k-medoid clustering algorithm to bin contigs iteratively and exhaustively into genome bins.

MetaBAT Pipeline

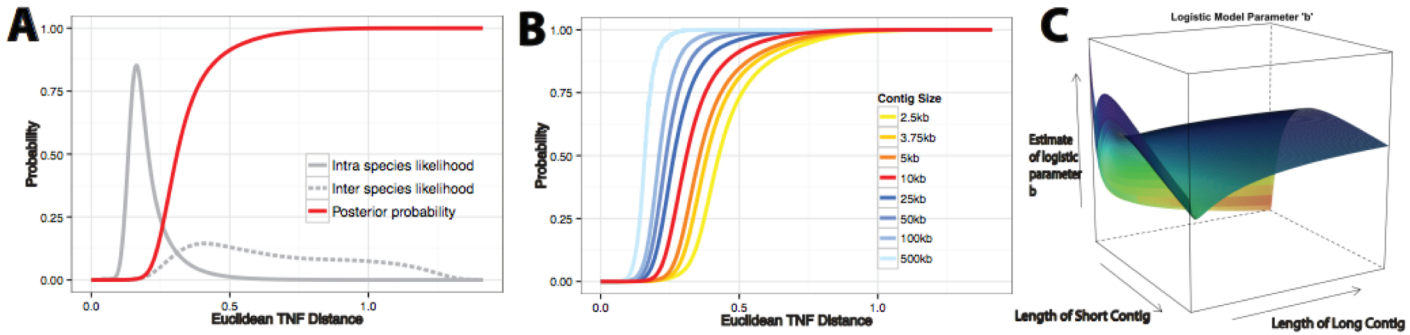


1. Metagenomic samples from multiple sites or times
2. Metagenome shotgun sequencing
3. Initial de-novo assembly using the combined library
4. Calculate TNF for each contig
5. Calculate Abundance (ABD) per sample for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Briefly, MetaBAT works as the following. For each pair of contigs, it first calculates two probabilities of pairwise distances from genome signatures and abundances, and it integrates all pairwise probabilities into a composite distance matrix. It then employs a modified k-medoid clustering algorithm to iteratively cluster the contigs into genome bins, each of which corresponds to a single genome.

http://1ofdmq2n8tc36m6i46scovo2e-wpengine.netdna-ssl.com/wp-content/uploads/2013/11/Kang_JGIUM-14-Poster-w-Title-Page.pdf

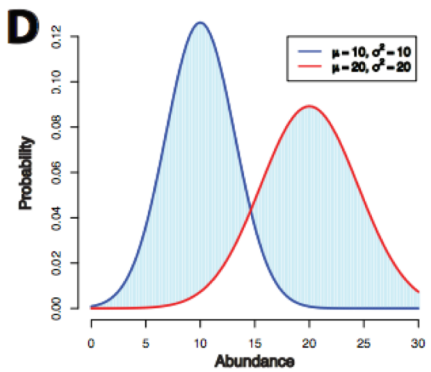
Probablistic Modeling of TNF and ABD



A. Converting Euclidean TNF distance to empirical probability by statistical modeling. The likelihood of inter- and intra-species distance using unique, complete genomes from NCBI, and posterior probability distribution of inter-species distance for a pair of genomic fragments of fixed size (10kb).

B. Dynamics of posterior probabilities depend on contig sizes. Better inter-species separation is achieved with larger fragment sizes. And each line can be fit to a logistic curve, F(*d*) = 1/(1+exp(-(*b*+*c***d*))), where parameters b and c are functions of contig sizes.

C. Parameter estimation of a dynamic logistic curve based on two different contig lengths. We modeled the posterior inter-species probability as a logistic curve that changes depending on two different contig lengths.



D. Probabilistic abundance distance between two contigs. The shaded area represents the probability of two contigs originating from different genomes calculated as follows:

$$P(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \quad = \quad \frac{1}{2} \sum_{i=0}^{\infty} |\phi_{\mu_1,\sigma_1^2}(i) - \phi_{\mu_2,\sigma_2^2}(i)|$$

The combined probability of multiple abundance distances are calculated as weighted geometric mean of each probabilities:

$$P_{ij} = \sum_n \mathbb{1}\{\mu_{in} > c \text{ OR } \mu_{jn} > c\} \sqrt{\prod_n P_{ijn}(\mu_{in}, \sigma_{in}^2, \mu_{jn}, \sigma_{jn}^2) * \mathbb{1}\{\mu_{in} > c \text{ OR } \mu_{jn} > c\}}$$

http://1ofdmq2n8tc36m6i46scovo2e-wpengine.netdna-ssl.com/wp-content/uploads/2013/11/Kang_JGIUM-14-Poster-w-Title-Page.pdf

Similarity Based

- LCA (Lowest Common Ancestor): Use all BLAST hits above a threshold and assign taxonomy at the lowest level in the tree which covers these taxa.

- Notable Examples:
  - MEGAN:
    - http://ab.inf.uni-tuebingen.de/software/megan5/
    - One of the first metagenomic tools
    - Does functional profiling too!
  - MG-RAST: https://metagenomics.anl.gov/
    - Web-based pipeline (might need to wait awhile for results)
  - Kraken: https://ccb.jhu.edu/software/kraken/
    - Fastest binning approach to date and very accurate.
    - Large computing requirements (e.g. >128GB RAM)

- Search strain-level resolution metagenomics tools
  - Pathoscope, Sigma, PanPhlAn, metaSNV,…

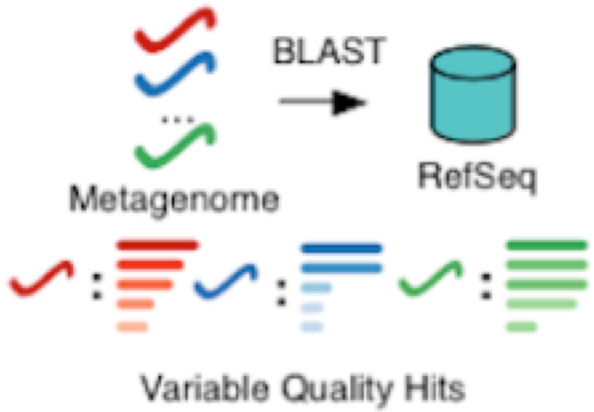
The LCA algorithm = "Lowest Common Ancestor" algorithm

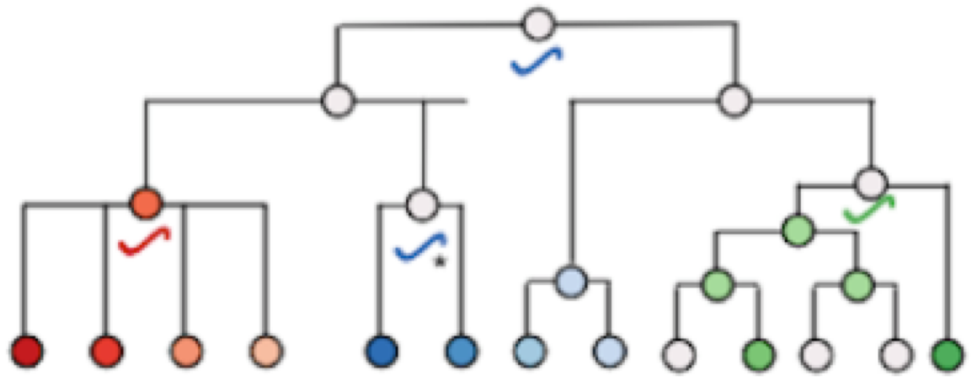
"In this approach, every read is assigned to some taxon. If the read aligns very specifically only to a single taxon, then it is assigned to that taxon. The less specifically a read hits taxa, the higher up in the taxonomy it is placed. Reads that hit ubiquitously may even be assigned to the root node of the NCBI taxonomy."

(the MEGAN manual)

LCA

- Lowest Common Ancestor (LCA) algorithm are used in MEGAN and Kraken
- The taxonomic signal of many hits to an individual read are collapsed to their lowest common ancestor (LCA) on the NCBI Taxonomy phylogenetic 'tree-of-life.'
- This LCA algorithm is the central method incorporated into the highly popular Metagenome Analyzer (MEGAN) software and Kraken.
- A slightly tempered version of the LCA method based on similarity of the top-two BLAST hits is the default 'MEGAN taxonomy' method, and is currently one of the most popular ways to extract taxonomic signal from a metagenome (Figure 2b).





https://github.com/hallamlab/mp_tutorial/wiki/Taxonomic-Analysis

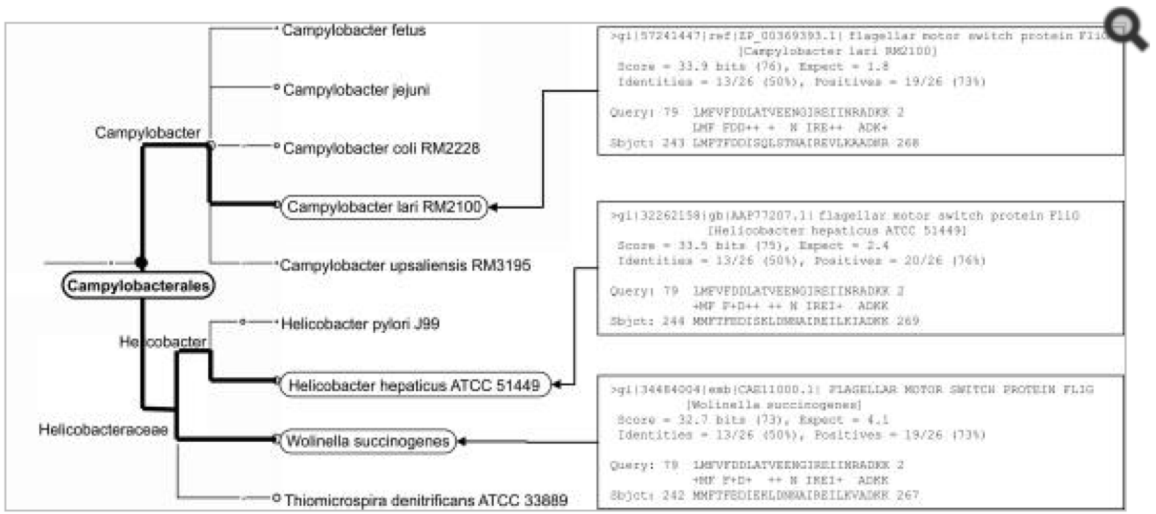
LCA

GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

PMC full text: Genome Res. 2007 Mar; 17(3): 377–386.
doi: 10.1101/gr.5969107
Copyright/License ► Request permission to reuse

<< Prev Figure 2. Next >>

Figure 2.

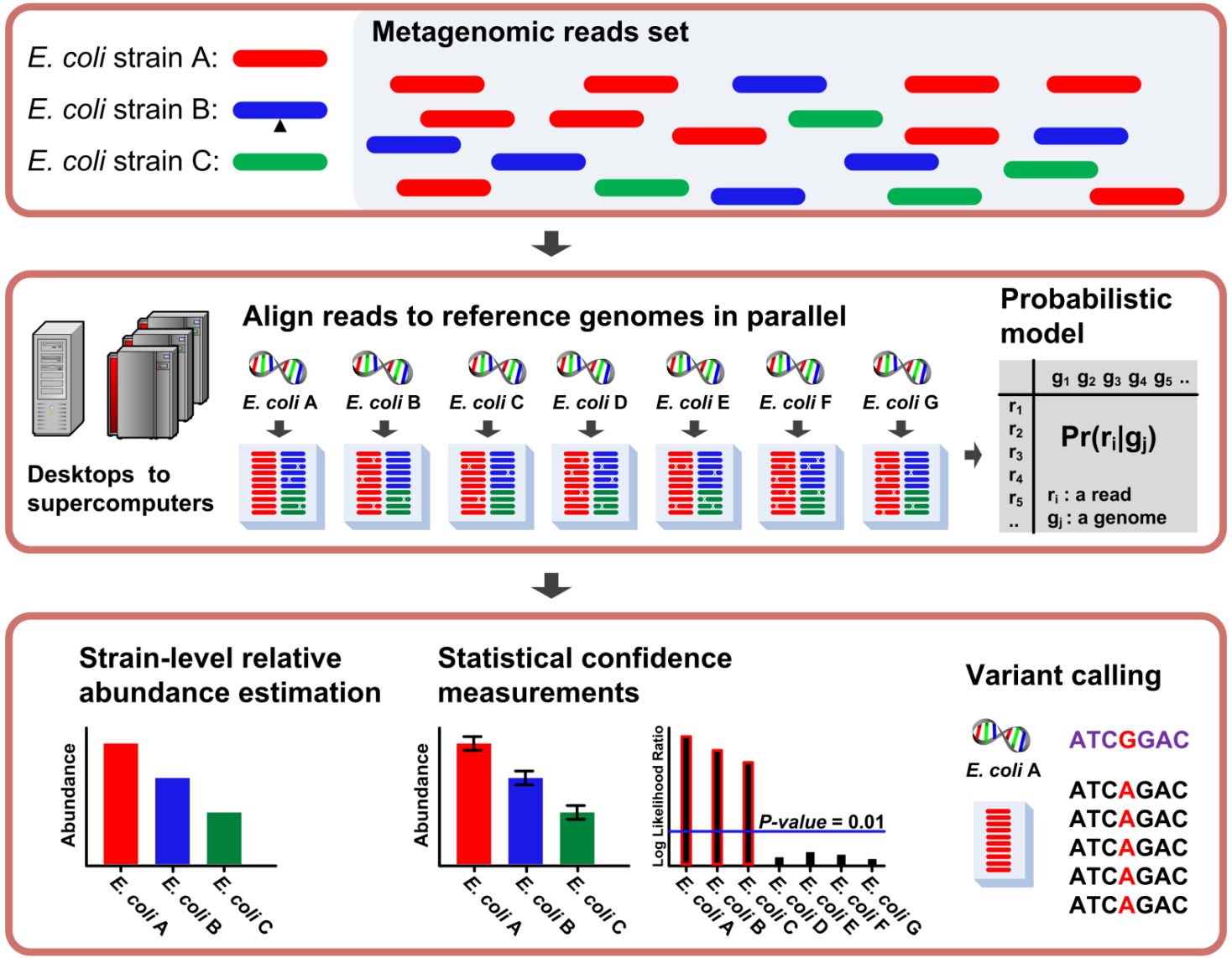


On the *right*, we list the three BLASTX matches obtained for a specific read *r* from the mammoth data set, to sequences representing *Campylobacter lari*, *Helicobacter hepaticus*, and *Wolinella*, respectively. The LCA-assignment algorithm assigns *r* to the taxon *Campylobacterales*, shown on the *left*, as it is the lowest-common taxonomical ancestor of the three matched species.

Generic Pipeline using DIAMOND and MEGAN6

- http://megan.informatik.uni-tuebingen.de/t/generic-pipeline-using-diamond-and-megan6/50
- DIAMOND (https://ab.inf.uni-tuebingen.de/software/diamond/)
  - DIAMOND is a new high-throughput program for aligning DNA reads or protein sequences against a protein reference database such as NR, at up to 20,000 times the speed of BLAST, with high sensitivity.
- MEGAN6 (https://ab.inf.uni-tuebingen.de/software/megan6)
  - A powerful interactive microbiome analysis tool

Sigma Overview



Marker Based

- Single Gene
  - Identify and extract reads hitting a single marker gene (e.g. 16S, cpn60, or other "universal" genes)
  - Qiime2 (http://qiime.org/)
- Multiple Gene
  - Several universal genes
  - PhyloSift (Darling et al, 2014)
  - Uses 37 universal single-copy genes
  - Clade specific markers
  - MetaPhlAn2 (https://bitbucket.org/biobakery/metaphlan2)

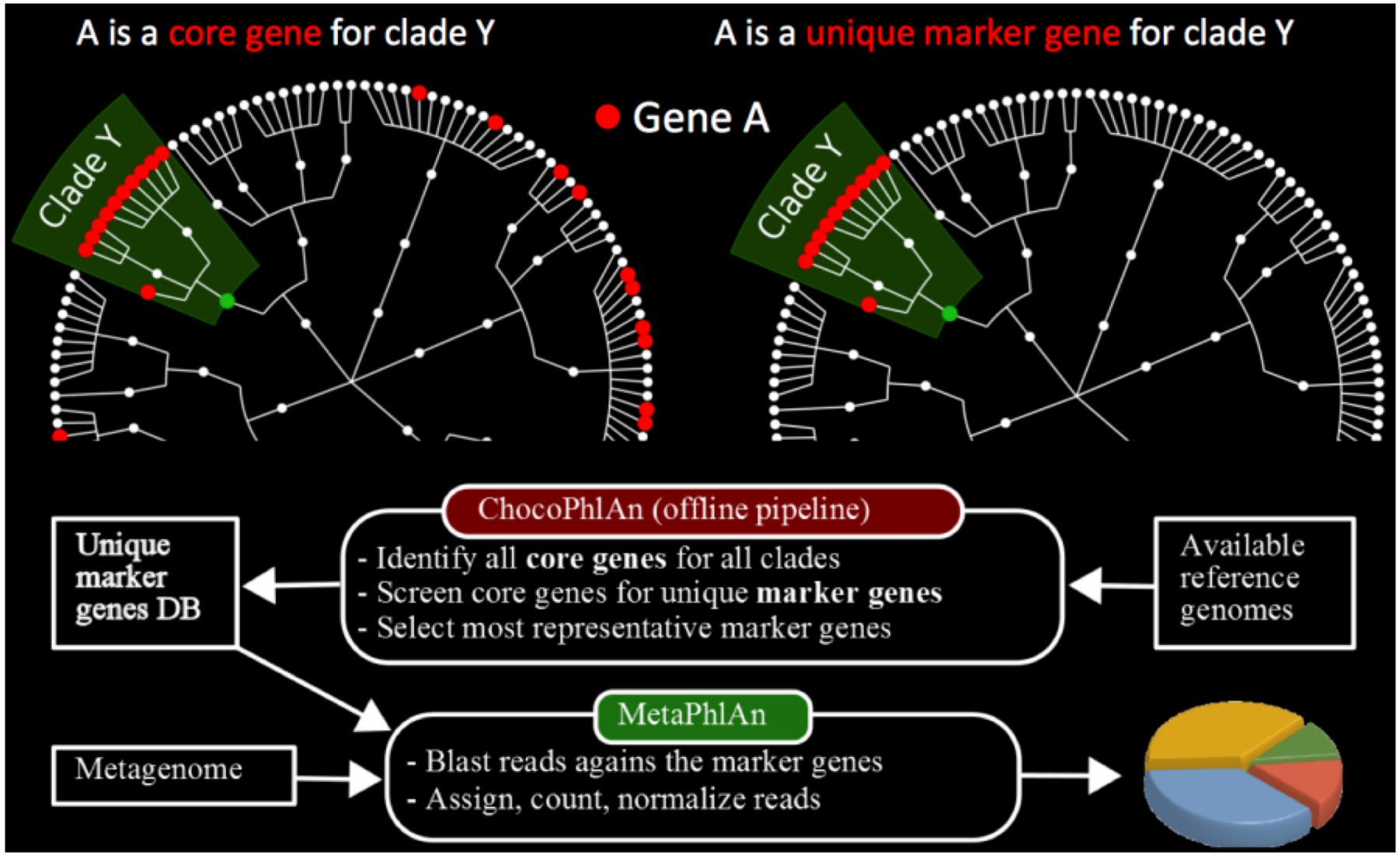
Why MetaPhlAn2?

- Fast (marker database is considerably smaller)
- Markers for bacteria, archaea, eukaryotes, and viruses (since MetaPhlAn2 was released)
- Being continuously updated and supported
- Used by the Human Microbiome Project
- Generally accepted as a robust method for taxonomy assignment
- Main Disadvantage: not all reads are assigned a taxonomic label

MetaPhlAn

- Uses “clade-specific” gene markers
- A clade represents a set of genomes that can be as broad as a phylum or as specific as a species
- Uses ~1 million markers derived from 17,000 genomes
  - ~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic
- Can identify down to the species level (and possibly even strain level)
- Can handle millions of reads on a standard computer within a few minutes

MetaPhlAn Marker Selection



Using MetaPhlan

- MetaPhlan uses Bowtie2 for sequence similarity searching (nucleotide sequences vs. nucleotide database)
- Paired-end data can be used directly (but are treated as independent reads)
- Each sample is processed individually and then multiple sample can be combined together at the last step
- Output is **relative abundances** at different taxonomic levels

Absolute vs. Relative Abundance

- Absolute abundance: Numbers represent real abundance of thing being measured (e.g. the actual quantity of a particular gene or organism)

- Relative abundance: Numbers represent proportion of thing being measured within sample

- In almost **all cases** microbiome studies are measuring relative abundance
  - This is due to DNA amplification during sequencing library preparation not being quantitative

Relative Abundance Use Case

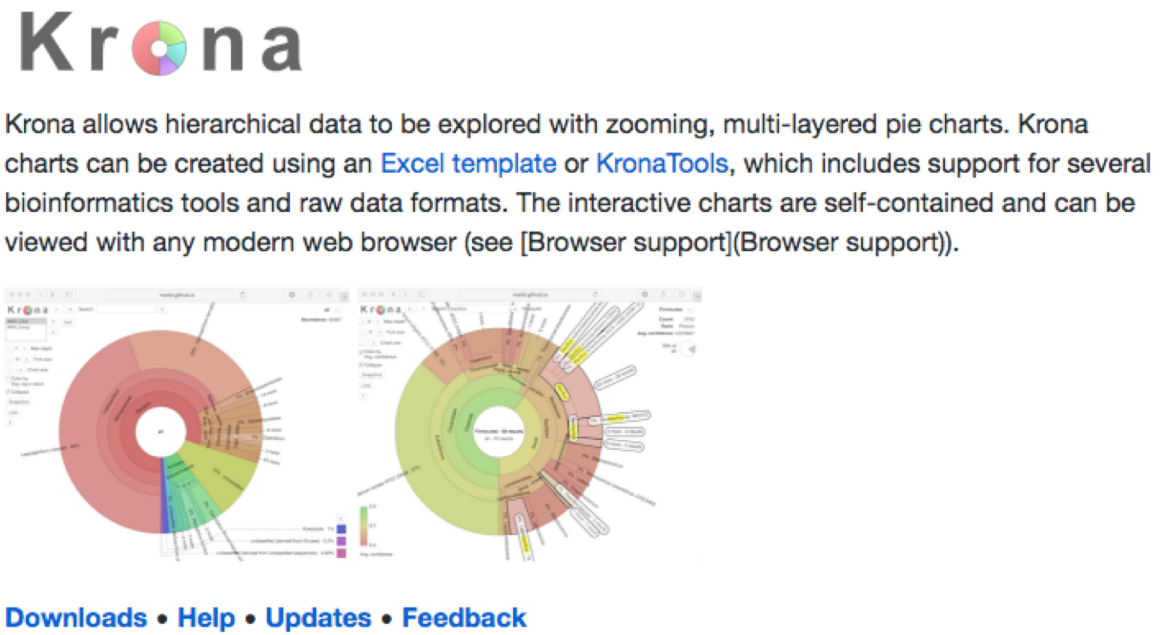
- Sample A:
  - Has 10^8 bacterial cells (but we don't know this from sequencing)
  - 25% of the microbiome from this sample is classified as Shigella
- Sample B:
  - Has 10^6 bacterial cells (but we don't know this from sequencing)
  - 50% of the microbiome from this sample is classified as Shigella
- "Sample B contains twice as much Shigella as Sample A"
  - WRONG! (If quantified it we would find Sample A has more Shigella)
- "Sample B contains a greater proportion of Shigella compared to Sample A"
  - Correct!

Visualization and Statistics

- Various tools are available to determine statistically significant taxonomic differences across groups of samples
  - Excel
  - SigmaPlot
  - Past
  - R (many libraries)
  - Python (matplotlib)
  - **STAMP**

Krona

- Interactive metagenomic visualization in a Web browser
- https://github.com/marbl/Krona/wiki
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-385



STAMP

Oxford Journals › Science & Mathematics › Bioinformatics › Volume 30, Issue 21 › Pp. 3123-3

STAMP: statistical analysis of taxonomic and functional profiles

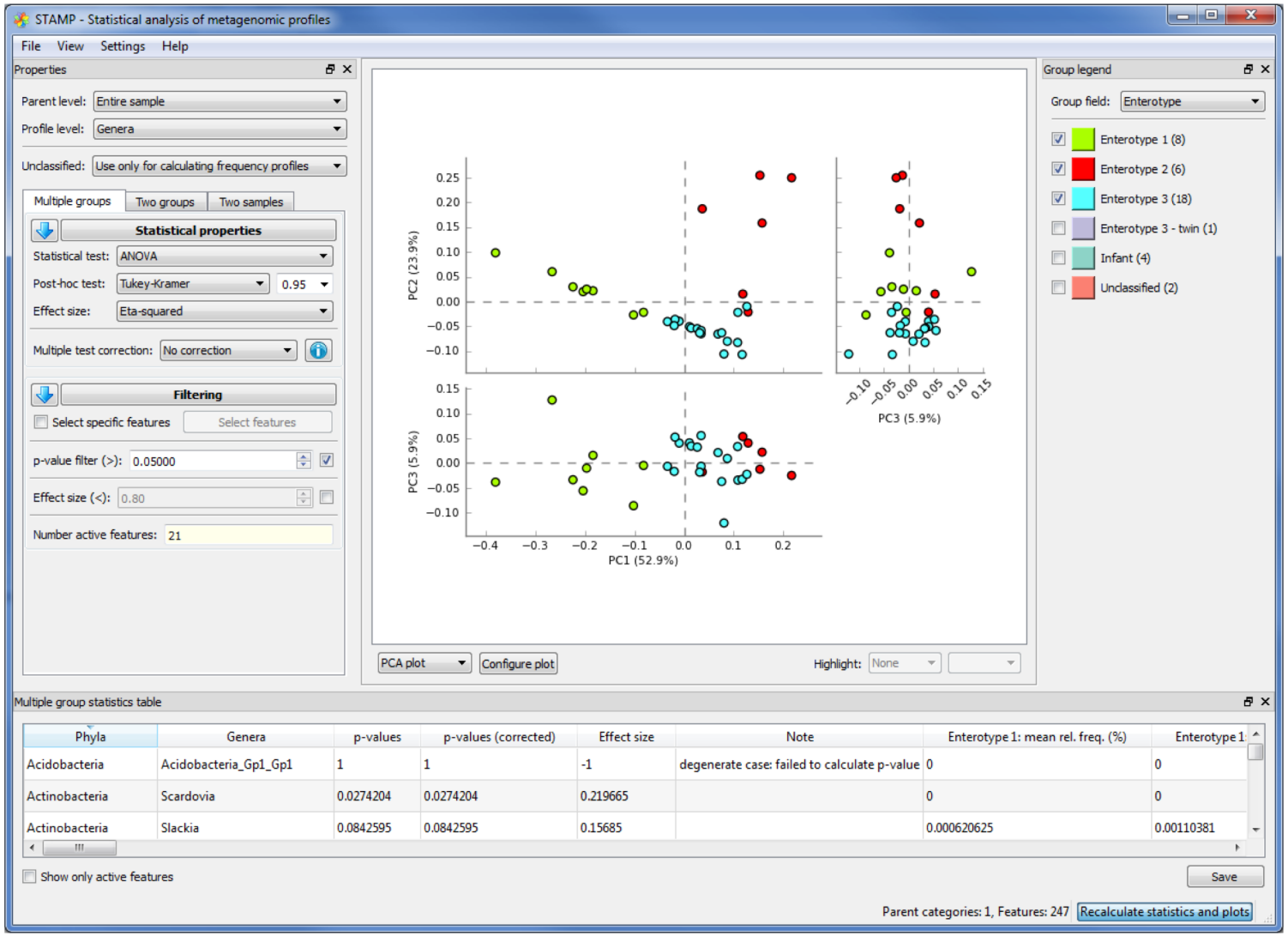
Donovan H. Parks[1,*], Gene W. Tyson[1,2], Philip Hugenholtz[1,3] and Robert G. Beiko[4]

+ Author Affiliations

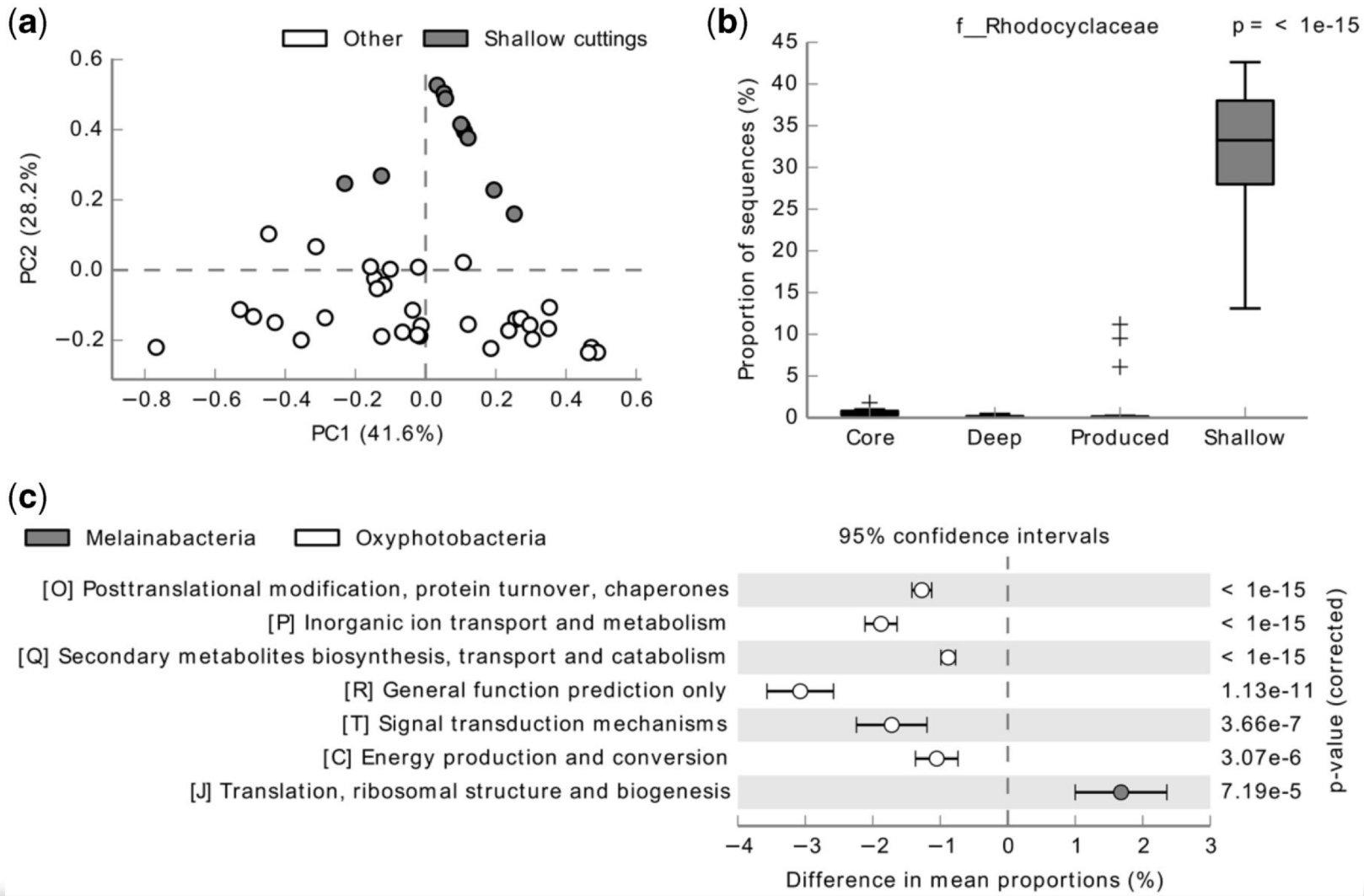
*To whom correspondence should be addressed.

Received June 4, 2014.
Revision received July 11, 2014.
Accepted July 15, 2014.

STAMP


STAMP - Statistical analysis of metagenomic profiles
File View Settings Help
Properties
Parent level: Entire sample
Profile level: Genera
Unclassified: Use only for calculating frequency profiles
Multiple groups Two groups Two samples
Statistical properties
Statistical test: ANOVA
Post-hoc test: Tukey-Kramer 0.95
Effect size: Eta-squared
Multiple test correction: No correction
Filtering
Select specific features Select features
p-value filter (>): 0.05000
Effect size (<): 0.80
Number active features: 21
PC2 (23.9%)
PC3 (5.9%)
PC1 (52.9%)
PC3 (5.9%)
Group legend
Group field: Enterotype
Enterotype 1 (8)
Enterotype 2 (6)
Enterotype 3 (18)
Enterotype 3 - twin (1)
Infant (4)
Unclassified (2)
PCA plot
Configure plot
Highlight: None
Multiple group statistics table
Phyla Genera p-values p-values (corrected) Effect size Note Enterotype 1: mean rel. freq. (%) Enterotype 1
Acidobacteria Acidobacteria_Gp1_Gp1 1 1 -1 degenerate case: failed to calculate p-value 0 0
Actinobacteria Scardovia 0.0274204 0.0274204 0.219665 0 0
Actinobacteria Slackia 0.0842595 0.0842595 0.15685 0.000620625 0.00110381
Show only active features
Save
Parent categories: 1, Features: 247
Recalculate statistics and plots

STAMP Plots

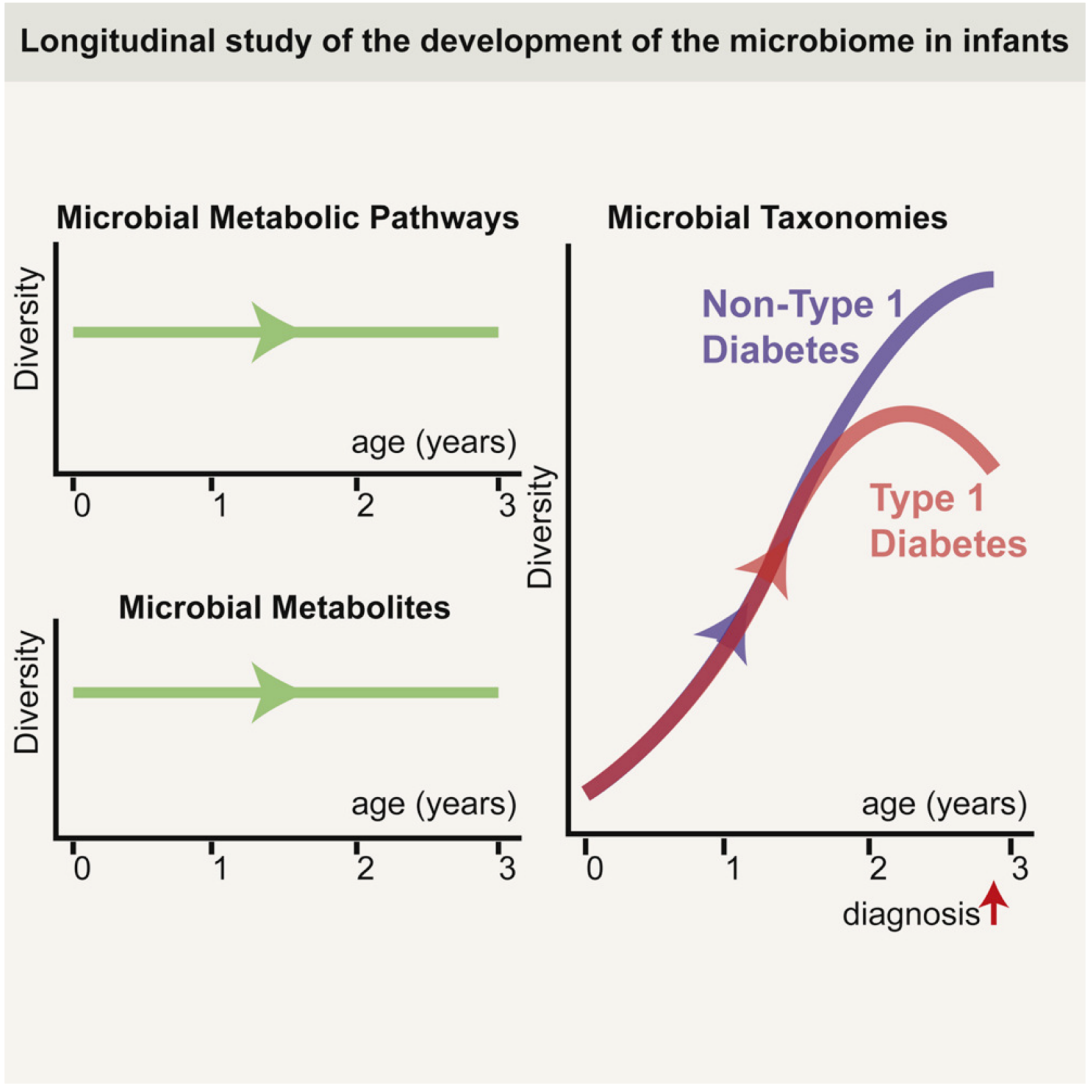

(a)
Other
Shallow cuttings
PC2 (28.2%)
PC1 (41.6%)
(b)
f__Rhodocyclaceae
p = < 1e-15
Proportion of sequences (%)
Core
Deep
Produced
Shallow
(c)
Melainabacteria
Oxyphotobacteria
95% confidence intervals
[O] Posttranslational modification, protein turnover, chaperones
[P] Inorganic ion transport and metabolism
[Q] Secondary metabolites biosynthesis, transport and catabolism
[R] General function prediction only
[T] Signal transduction mechanisms
[C] Energy production and conversion
[J] Translation, ribosomal structure and biogenesis
< 1e-15
< 1e-15
< 1e-15
1.13e-11
3.66e-7
3.07e-6
7.19e-5
p-value (corrected)
Difference in mean proportions (%)

Metagenomics Research Trend Shifting

More time-series analysis

Kostic et. al., The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes, *Cell Host & Microbe*, Vol 17, Issue 2, 2015

- Gut microbial metabolic pathways but not taxonomies are stable throughout infancy
- Strain composition of high-abundance species remains constant throughout infancy
- Decreased community diversity occurs after seroconversion but before onset of T1D
- T1D onset is preceded by increased inflammation-associated organisms and pathways

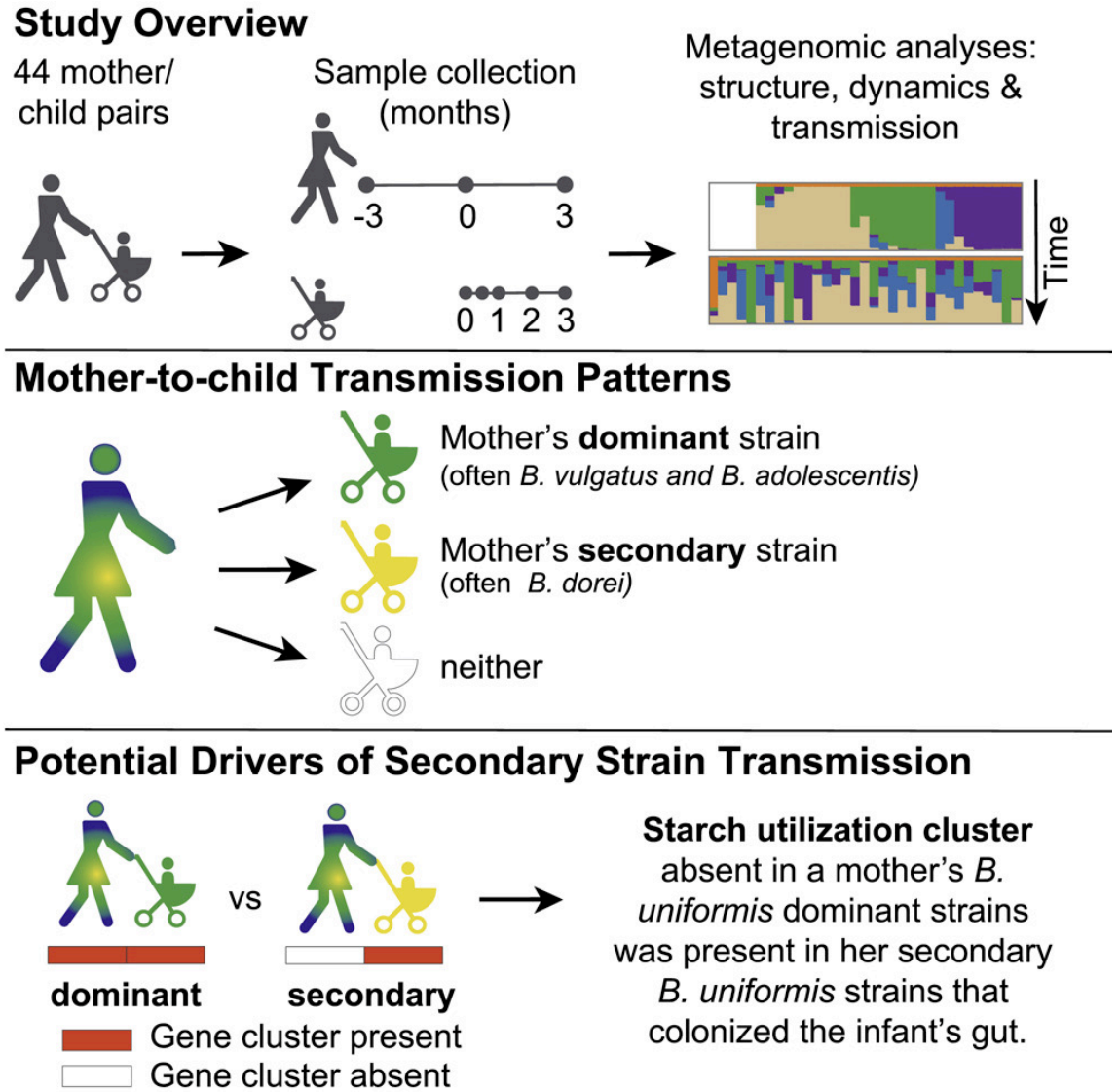


Metagenomics Research Trend Shifting

More strain-level analysis

Yassour et. al., Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life, *Cell Host & Microbe*, Vol 24, Issue 1, 2018

- Gut bacterial transmission patterns assessed longitudinally in 44 mother-infant pairs
- Metagenomic sequencing reveals transmission patterns beyond dominant strains
- Mother's minor strain sometimes colonizes infant, likely driven by functional selection
- Some antibiotic resistance genes co-occur in families, suggesting their inheritance



Strain Resolution Metagenomics Profilers

Tool	Approach
MetaPhlAn2 (2015) - expanded version of MetaPhlAn (2012) https://www.nature.com/articles/nmeth.3589.pdf https://www.nature.com/articles/nmeth.2066	Uses clade specific markers to detect the taxonomic clades present in a microbiome sample. Expanded set of ~1 million markers from >7,500 species. Subspecies markers enable strain-level analyses. Estimates the relative abundance of microbial cells by mapping reads against reduced set of marker reads.
PanPhlAn (2016) https://www.nature.com/articles/nmeth.3802	Uses metagenomic data to achieve strain-level microbial profiling resolution. Compares gene sets/families across samples to enable population genomics tasks such as subspecies and novel strain identification, strain tracking, and association of strain-specific genes
StrainPhlAn (2017) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5378180/	Maps metagenomic reads against species-specific marker sequences (up to 200/species from total set of ~ 1 million markers). Strain-specific consensus sequences can be identified from as few as a single reference genome (the reconstructed strain-specific consensus is independent from the sequence of the marker used for mapping). Provides a strain-level phylogeny of each analyzed species from the concatenated alignment of the markers

Strain Resolution Metagenomics Profilers

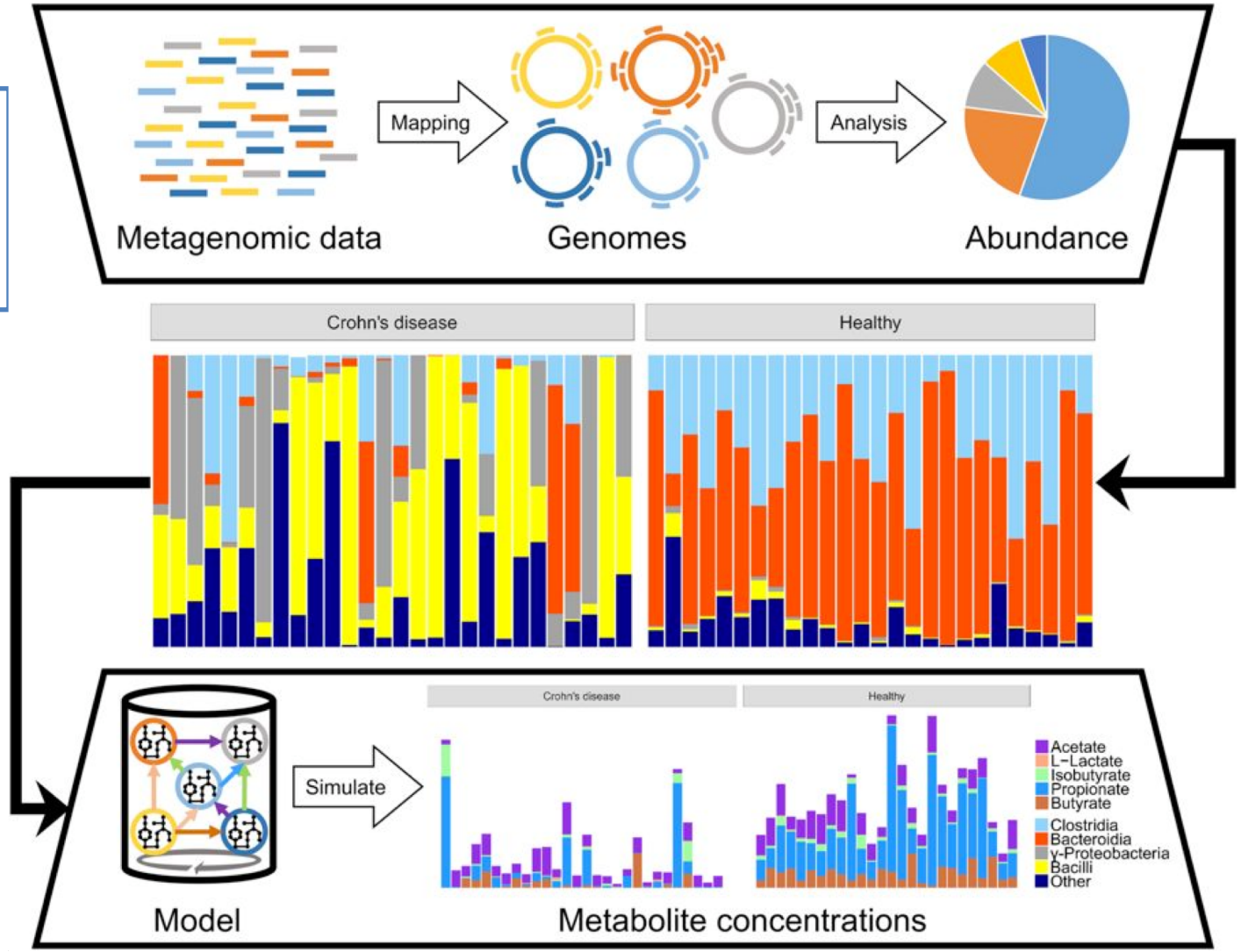
Tool	Approach
PathoScope 2.0 (2014) https://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-2-33 Expanded on PathScope1.0	Allows users to generate custom reference genome libraries for specific datasets. Extracts reference library. Aligns reads to the target library and removes any sequences that align to filter library (includes BowTie2 wrapper). Reassigns ambiguous reads to the most likely source genome in the library (based on penalized statistical score)
Sigma (2015) https://academic.oup.com/bioinformatics/article/31/2/170/2366214	Map all reads in a metagenomic dataset onto a user-defined database of reference genomes. Probabilistic model permits maximum likelihood estimation of the relative abundances of all genomes measured by the percentages of reads sampled from these genomes
ConStrains (2015) https://www.nature.com/articles/nbt.3319	An open-source algorithm that identifies conspecific strains from metagenomic sequence data and reconstructs the phylogeny of these strains in microbial communities. Algorithm uses SNP patterns in a set of universal genes to infer within-species structures that represent strains
MIDAS (2016) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088602/	An integrated computational pipeline for quantifying bacterial species abundance and strain-level genomic variation, including gene content and SNPs, from shotgun metagenomes
StrainEst (2017) https://www.nature.com/articles/s41467-017-02209-5	Uses the single-nucleotide variants (SNV) profiles of the available genomes of selected species to determine the number and identity of coexisting strains and their relative abundances in mixed metagenomic samples. Concentrates on species of interest by defining their population structure through a clustering of the SNV profiles.

Metagenomics Research Trend Shifting

More disease analysis

Bauer et. al., From metagenomic data to personalized in silico microbiotas: predicting dietary supplements for Crohn's disease, *Systems Biology and Applications*, 4, 27, 2018

- Crohn's disease (CD) is associated with an ecological imbalance of the intestinal microbiota
- Predicted short chain fatty acid (SFCA) levels for patients and controls
- Low concentrations of SCFA were predicted for CD patients and the SCFA signatures were unique to each patient
- Consequently, suggest personalized dietary treatments that could improve each patient's SCFA levels



Good Tutorials of Metagenomics

- https://galaxyproject.github.io/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html
- https://www.mothur.org/wiki/MiSeq_SOP
- https://metagenomics-workshop.readthedocs.io/en/latest/index.html#
- https://mgm.jgi.doe.gov/
- https://usda-ars-gbru.github.io/Microbiome-workshop/