Metagenomics: Characterization of Microbial Communities using NGS

WGS

UT Southwestern Medical Center | BICF

Brandi Cantarel, PhD
10/19/2016

- Community Structure Based on WGS
  - 16S VS WGS
  - Least Common Ancestor
  - Reference Genome databases
  - Assignment Stratgies
- Data Processing Workflow
  - QC
  - Functional Databases
  - Analysis Platforms
  - DYI Analysis
- Non-DNA based technologies: Metatranscriptomics, Metaproteomics and Metabolomics
  - Expression (RNASeq)
  - Translation (Proteomics)
  - Metabolites (Metabolomics)

- Community Structure Based on WGS
  - 16S VS WGS
  - Least Common Ancestor
  - Reference Genome databases
  - Assignment Strategies
- Data Processing Workflow
  - QC
  - Functional Databases
  - Analysis Platforms
  - DYI Analysis
- Non-DNA based technologies: Metatranscriptomics, Metaproteomics and Metabolomics
  - Expression (RNASeq)
  - Translation (Proteomics)
  - Metabolites (Metabolomics)

Taxonomic Assessment using 16S

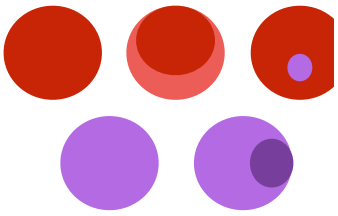
- 16S is targeted sequencing for a single gene which acts as a marker for organisms
- Pros
  - Well established
  - Relatively inexpensive $50-$100/sample
  - Amplifies only bacteria not host or environmental fungi, plants, etc
- Cons
  - Amplifies only bacteria not viruses, microbial fungi, archaea, etc
    - Although can be paired with 18S and archaeal specific 16S
  - Is based on a very well conserved gene, making it hard to resolve species and strains
  - V-region choice can bias results

Taxonomic Assignment using WGS

- WGS (whole genome shotgun) aims to sequence the “whole” metagenome
- Pros
  - Not biased by amplicon primer set
  - Not limited to by conservation of the amplicon
  - Can also provide functional information
- Cons
  - Environmental contamination, including host
  - More expensive - $1000+/sample
  - Complex data analysis
    - Requires high performance computing, high memory, high compute capacity

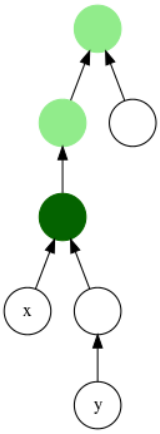
Taxonomic Assignment: Complex Analysis

- All of the organism mixed together
  - It’s hard to bin all of the reads from one organism (strain or species) for deconvolution
  - Reads are short
  - Reads can potentially share similarity to multiple taxa
- Lateral gene transfer
  - Not all of the genes in a genome “shares” the same evolutionary history

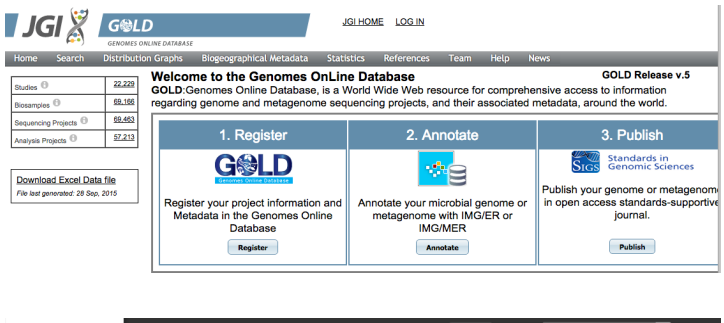


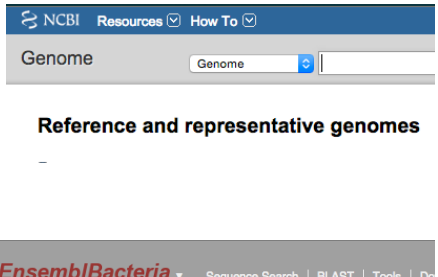
Least Common Ancestor Taxonomic Assignment

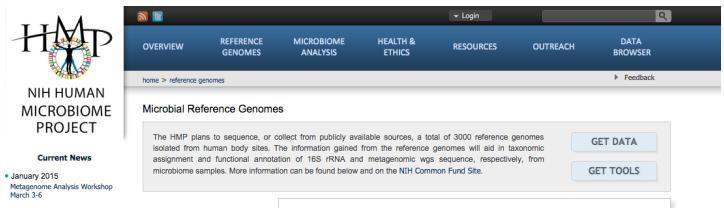
- Reads can potentially share similarity to multiple taxa
- Least Common Ancestor allows for the taxonomic assignment when similarity is shared to multiple taxa
- Dependent on the taxonomic tree and similarity to genomes
  - Remember there are different versions of bacterial taxonomy

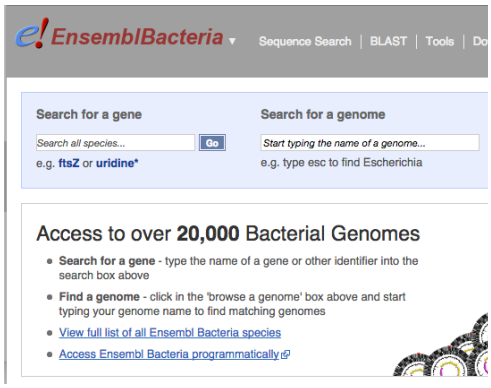


Sources of Reference Genomes for Comparison









Strategies for Taxonomic Assignment of WGS

- Compositional Based Taxonomic Assignment
  - This is assignment based on “base content”
- Sequence Alignment Based Taxonomic Assignment
  - This assignment is based on an alignment
- Maker Gene Based Taxonomic Assignment
  - This assignment is based similarity on a subset of the reads to conserved genes.

Composition Based Taxonomic Assignment

- GC content (TETRA)
- K-mer based (naïve Bayes classifier)
- Pros
  - Speed
  - Require less compute power compared to alignment-based methods.
- Cons
  - Requires query sequences of sufficient length
  - Genomes in the same clade (genera, family, etc) can be quite heterogenous in some regions

ATTGCC	17
AGTGCC	10
CCGTGA	25
TTGTGA	57
CCGTGA	12

Sequence Alignment Based Taxonomic Assignment

- BLAST/Megablast
- Malt/Diamond
- Kraken
- Pros
  - Higher assignment accuracy and specificity
- Cons
  - These methods are computationally intense because they either:
  - Require a high memory machine to generate the database and complete the searches
  - Require high number of cpus to complete the searches

Marker Gene Based Composition

- MetaPhlAn2
  - Relies on ~1M unique clade-specific marker genes
- PhyloSift
  - Uses a reference database of protein and RNA sequences
- Taxonomer
  - Based on 16S rRNA
- Pros
  - Less computationally intensive
  - Accurate for the marker gene composition
- Cons
  - Only assigns a subset of of the data ie can't determine taxonomy of certain function.

WGS Taxonomy Assignment and Visualization

- Taxonomer
  - http://taxonomer.iobio.io
- Megan
  - http://ab.inf.uni-tuebingen.de/software/megan5/
  - Tool with WGS taxonomic assignment (based on BLAST) and functional assignment
- MG-RAST
  - http://metagenomics.anl.gov/
  - Online tools with WGS taxonomic assignment and functional assignment
- MetaCRAM

- Community Structure Based on WGS
  - 16S VS WGS
  - Least Common Ancestor
  - Reference Genome databases
  - Assignment Strategies
- Data Processing Workflow
  - QC
  - Functional Databases
  - Analysis Platforms
  - DYI Analysis
- Non-DNA based technologies: Metatranscriptomics, Metaproteomics and Metabolomics
  - Expression (RNASeq)
  - Translation (Proteomics)
  - Metabolites (Metabolomics)

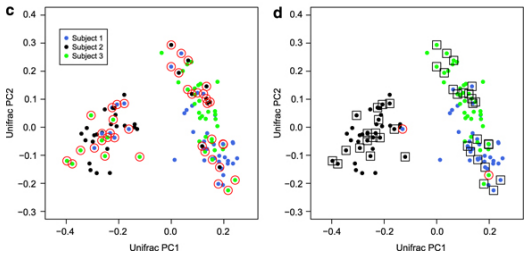
What is the Functional Capability

- Identify organisms present — if closely related to organisms with sequenced genomes
- Identify gene families present — if homologs have been functionally characterized
- Identify functional pathways present — if homologs have been annotated to gene pathways
- Identify new species/strains — if assemblies are of sufficient depth

Extract DNA → WGS → QC → Compare Gene Content

Quality Control

- Negative Controls are the best way to identify microbial lab contamination
- Sequencing Errors
  - Low Quality Bases
  - Homopolymer Strings
  - Too short trimmed reads
- Biological and Technical Replicates
  - Helps to ensure group trends and identify sample mislabeling and possible “compromised” samples



Knights D, Kuczynski J, Koren O, Ley RE, Field D, Knight R, DeSantis TZ, Kelley ST. Supervised classification of microbiota mitigates mislabeling errors. ISME J. 2011 Apr;5(4):570-3. doi: 10.1038/ismej.2010.148. Epub 2010 Oct 7. PubMed PMID: 20927137; PubMed Central PMCID: PMC3105748.

Host/Environmental Contamination

- In the human body — in human stool composes < 5% of reads, but the skin can be > 80% human reads
- Fungal, plant and soil bacteria can contaminate environmental samples.
- When you are collecting samples from “inside” of a habitat, it can be easy to contaminate the site with another site ie a colon biopsy with rectal microbiome.
- The natural environment can also contaminate samples, even the lab.

Metagenome Databases

Comprehensive Functional Databases

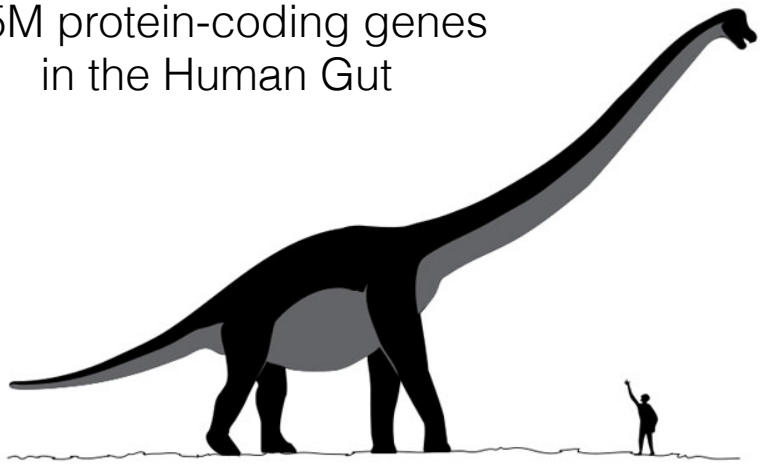
- KEGG
- eggNOG/COG
- PFAM
- SEED used by MG-RAST
- MetaCyc
- Uniref

Specialized Functional Databases

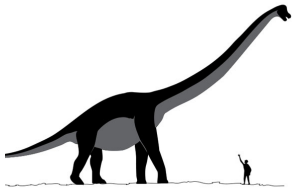
- Antibiotic resistant genes
  - http://ardb.cbcb.umd.edu/
  - https://card.mcmaster.ca/
- Virulent factors
  - http://www.mgc.ac.cn/VFs/main.htm
- Carbohydrate Active Enzymes
  - www.cazy.org
- Phage
- Proteases
  - http://merops.sanger.ac.uk/
- Transporters
  - http://www.membranetransport.org/

Microbial Gene Content

3-5M protein-coding genes
in the Human Gut



~25K Genes in the Human Genome

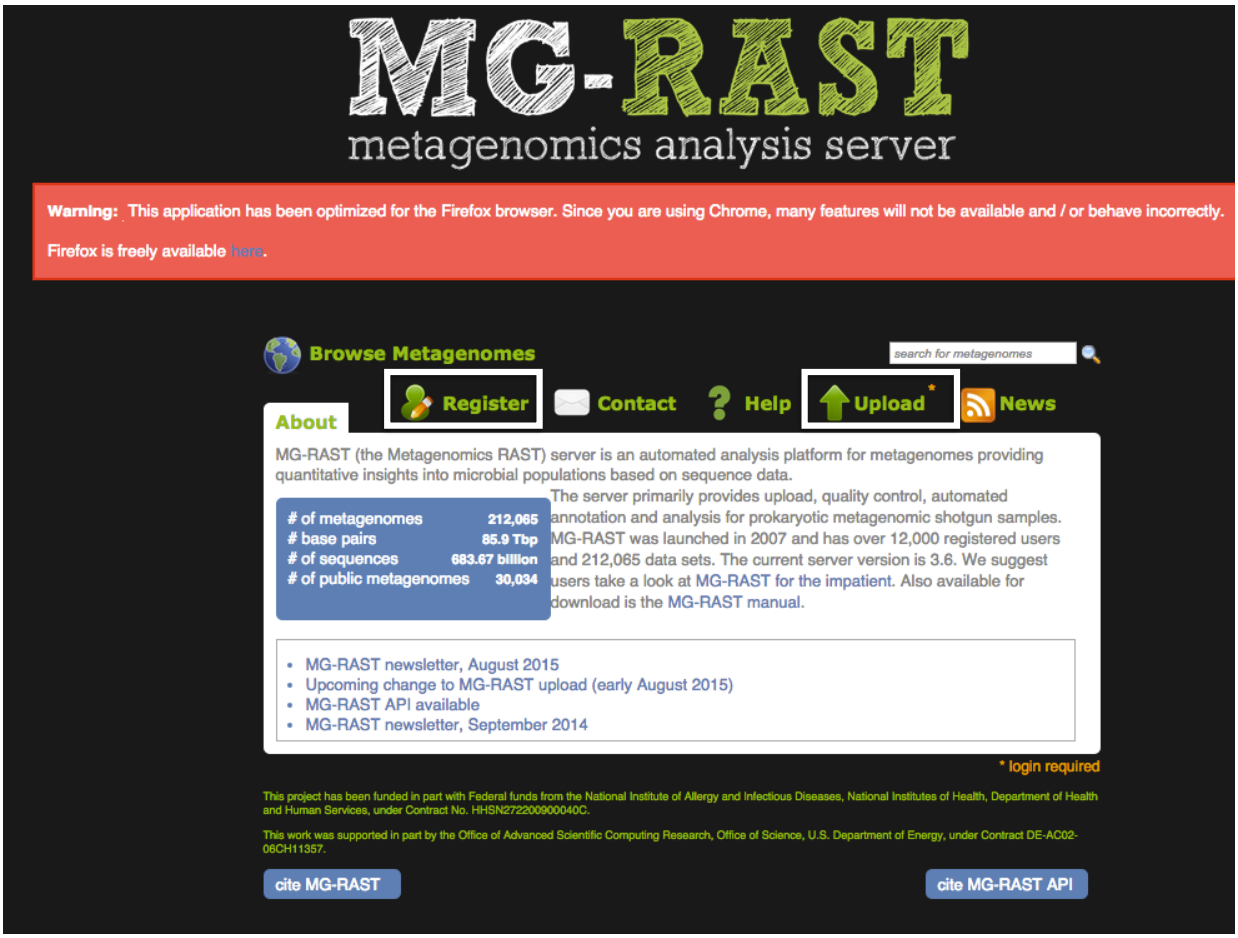


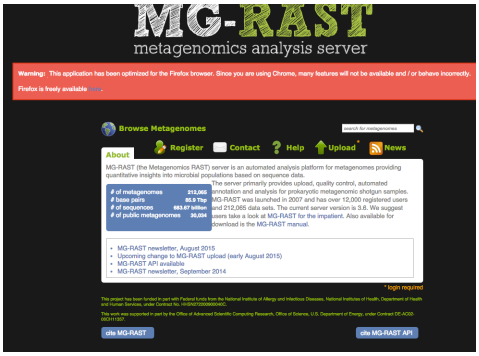
Metagenomic Datasets Tend to Be Big

- Depending on taxonomic diversity, sequencing depth for each sample averages from 1M - 100M reads
- Analysis programs such as assembly and some alignment algorithms require >100 GB of RAM
- High performance computing platform is necessary
  - There are some publicly available resources for analysis

Available Web-based Analysis Pipelines

- MG-RAST
  - Preference given to “public” datasets
  - Every easy to use
- EBI Metagenomics
  - Includes data visualization and customizable samples comparisons
    - DIAG
- JGI Integrated Microbial Genomes
  - Includes data visualization and customizable samples comparisons
- CloVR
  - Cloud-based workflow manager
  - Can run pipelines on your desktop
  - Available on the Academic Cloud



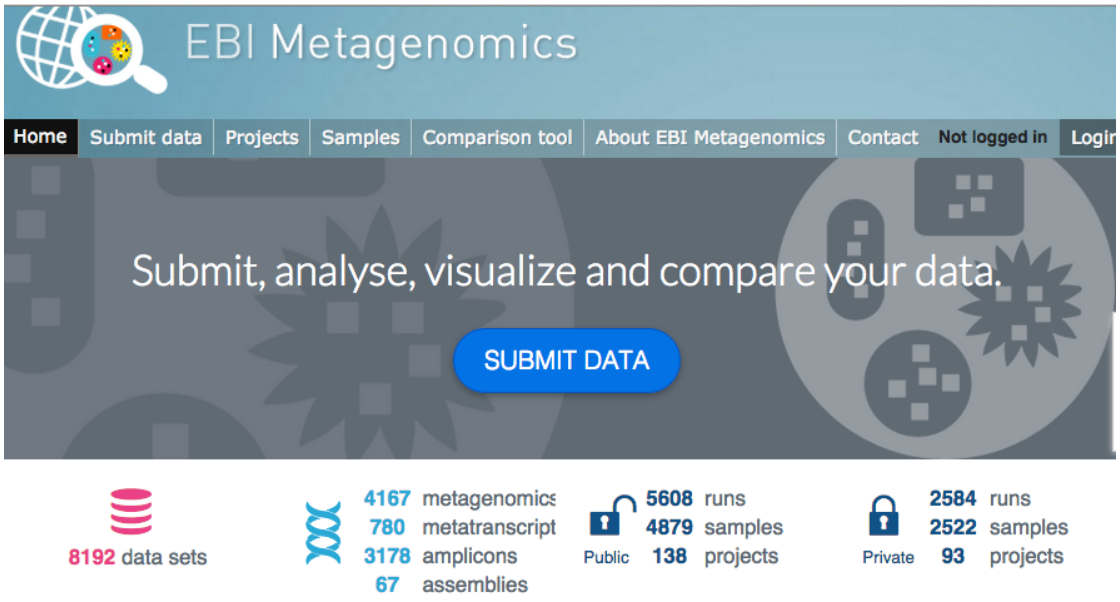


✔ qc_stats	10/7/2015, 9:19:23 AM
✔ preprocess	10/7/2015, 9:16:29 AM
✔ dereplication	10/7/2015, 9:16:46 AM
✔ screen	10/7/2015, 9:16:54 AM
✔ rna detection	10/7/2015, 9:17:11 AM
✔ rna clustering	10/7/2015, 9:17:29 AM
✔ rna sims blat	10/7/2015, 9:17:48 AM
✔ genecalling	10/7/2015, 9:17:48 AM
✔ aa filtering	10/7/2015, 9:17:53 AM
✔ aa clustering	10/7/2015, 9:19:26 AM
✔ aa sims blat	10/7/2015, 10:00:58 AM
✔ aa sims annotation	10/7/2015, 10:09:27 AM
✔ rna sims annotation	10/7/2015, 9:17:54 AM
✔ index sim seq	10/7/2015, 10:16:36 AM
✔ md5 annotation summary	10/7/2015, 10:19:23 AM
✔ function annotation summary	10/7/2015, 10:10:47 AM
✔ organism annotation summary	10/7/2015, 10:10:29 AM
✔ lca annotation summary	10/7/2015, 10:10:54 AM
✔ ontology annotation summary	10/7/2015, 10:10:51 AM
✔ source annotation summary	10/7/2015, 10:10:03 AM
✔ md5 summary load	10/7/2015, 10:32:53 AM
✔ function summary load	10/7/2015, 10:21:30 AM
✔ organism summary load	10/7/2015, 10:15:54 AM
✔ lca summary load	10/7/2015, 10:16:01 AM
✔ ontology summary load	10/7/2015, 10:17:31 AM
✔ done stage	10/7/2015, 10:35:28 AM
✔ notify job completion	10/7/2015, 10:35:31 AM

ıe job **pig59_colon (234800)** was submitted as part of the project PigColon at **10/7/2015, 16:17 AM**.

ıe current status is **completed**, the computation is finished. It took **1 hours 19 minutes** ɔm job submission until completion.

ıe result data is available for download on the download page. You can take a look at the ɾerview analysis data on the metagenome overview page.



Back to query page

Sample comparison tool: GO term annotation (functional analysis)

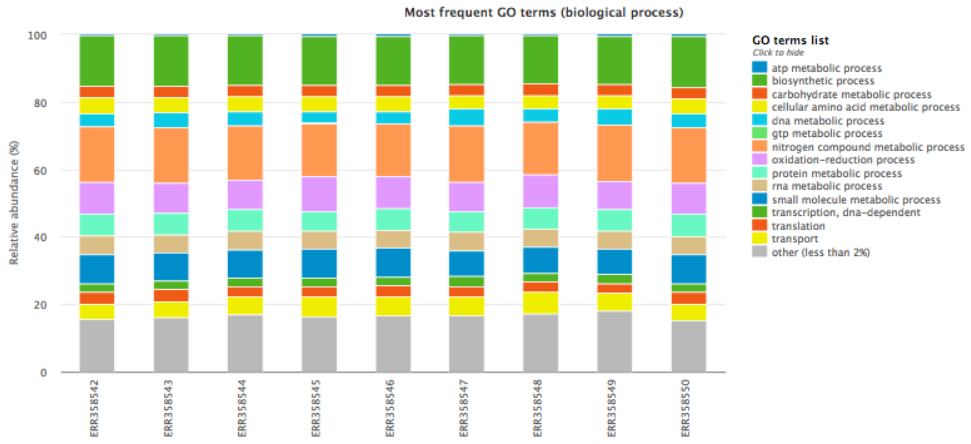
- Runs: ERR358542, ERR358543, ERR358544, ERR358545, ERR358546, ERR358547, ERR358548, ERR358549, ERR358550
- Project: Comparative freshwater metagenomics of Swedish and American lakes (ERP004168)

Barcharts | Stacked columns | Heatmap | Principal Component Analysis | Table

Jump to: Biological process | Molecular function | Cellular component

Export

Biological process



CloVR | Automated Sequence Analysis from Your Desktop

Welcome | Protocols | Getting Started | Download | Developers | Blog

Try CloVR

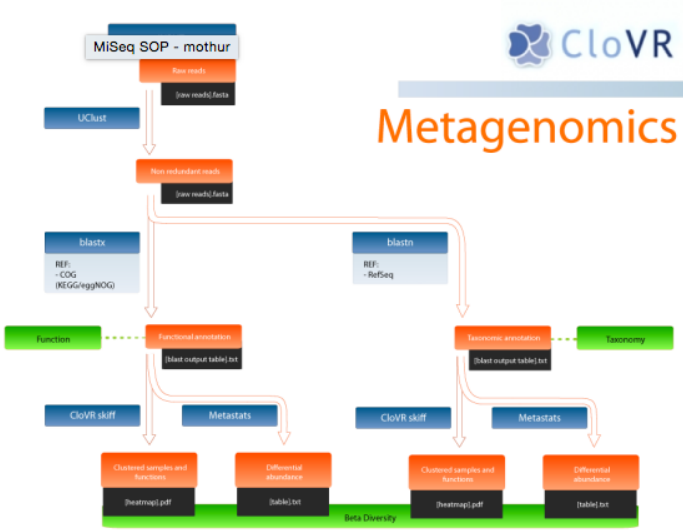
Read CloVR tutorials and run test applications on the DIAG cloud.

Get CloVR

Download and install CloVR to run supported microbial sequence analysis locally or on the cloud.

About CloVR

The Cloud Virtual Resource supports user-friendly automated microbial sequence analysis applications.



We will sometimes refer to the protocol described above as *CloVR-Metagenomics (no-orfs)*, which is our default. For users who wish to first call open reading frames (ORFs) on their sequences, we provide an alternative metagenomic analysis protocol that utilizes MetaGene for ORF-calling prior to functional assignment.

JGI IMG/EDU
INTEGRATED MICROBIAL GENOMES / EDUCATION SITE

ALL Genomes | Quick Genome Search: | Go

IMG Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | Analysis Cart | My IMG | Data Marts | Using

IMG Content

Datasets	
Bacteria	25671
Archaea	532
Eukarya	190
Plasmids	1186
Viruses	3888
Genome Fragments	1192
Total Datasets	32859

Genome by Metadata
Project Map
Metagenome Projects Map
System Requirements

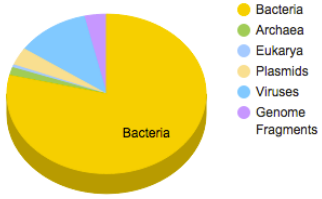
Hands on training available at the Microbial Genomics & Metagenomics Workshop

The **Integrated Microbial Genomes (IMG)** system (Nucleic Acids Research, Volume 42 Issue D1) serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in a uniquely integrated context. Plasmids that are not part of a specific microbial genome sequencing project and phage genomes are also included into IMG in order to increase its genomic context for comparative analysis.

Count	Total
DNA, number of bases	135,697,930,103
Total Genes	98,482,933
Total Genomes	32,859

IMG Statistics

All Genomes



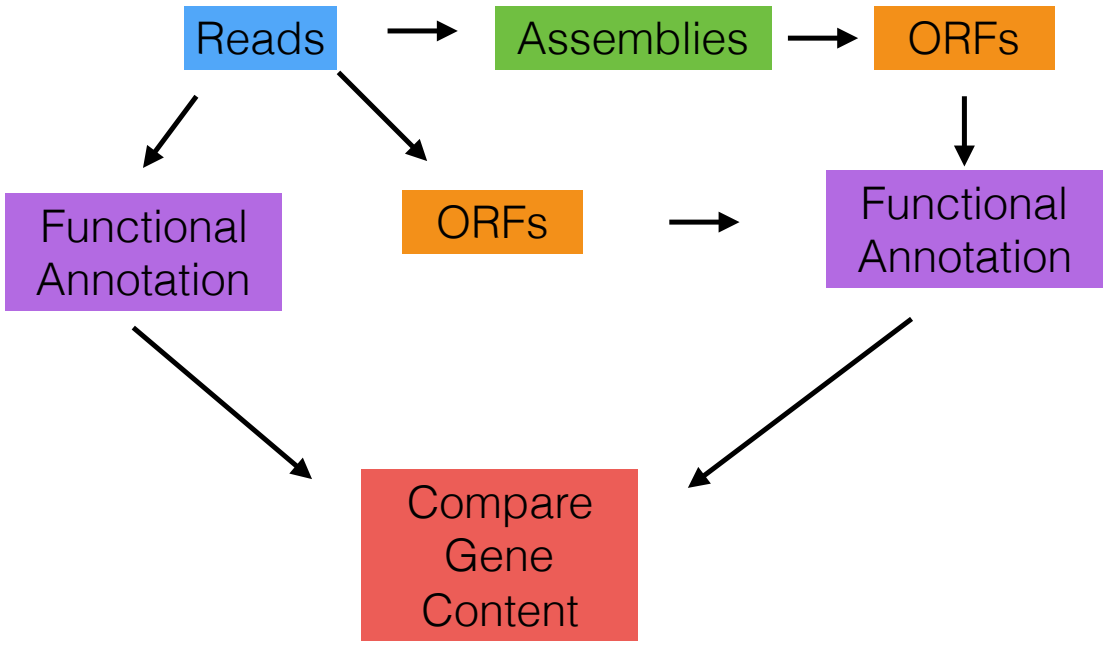
News

Oct 5 2015 After 10 Years, IMG Still Revolutionizing Genomics
Sep 2015 IMG ABC Data Mart
Sep 2015 MGM Workshops
Aug 11 2015 IMG Maintenance
July 9 2015 ANI News Release
July 8 2015 IMG Data Marts Changes
June 15 2015 ProDeGe News Release
June 11 2015 Plotting IMG's Next 10 Years
May 2015 IMG accounts deprecated
Apr 2015 BLAST in Workspace
Mar 2015 IMG using GOLD's new metadata
Read more...

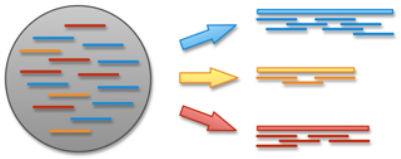
10 min break

Analysis Strategies

Many Paths for Functional Annotations



Assembly



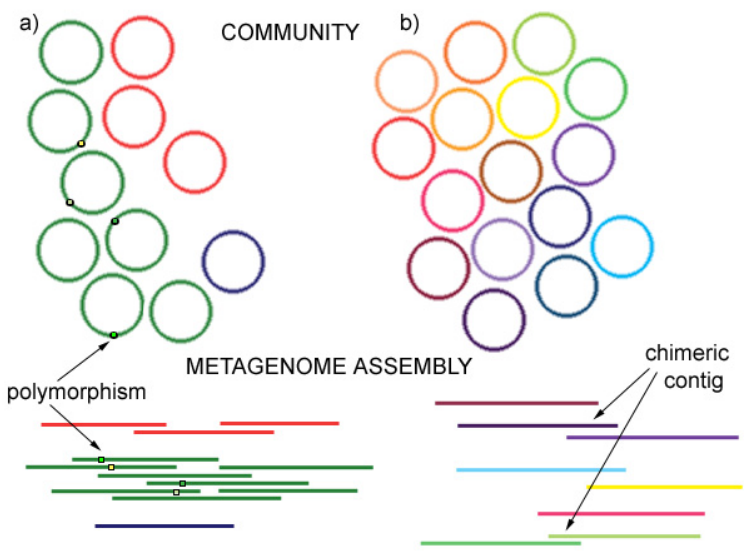
- Assembly can reduce the “amount of data” to optimize the annotation for function
- Assemblies in metagenomics can combine closely related strains or species
- Assemblies are high memory operations so there are some “pre-clustering” software to help reduce the data

khmer: A Data Reduction Strategy

- khmer is a k-mer based dataset analysis and transformation toolkit
- It can be used to reduce the size of a dataset by:
  - abundance filtering and error trimming
  - graph-size filtering by removing disconnected reads
  - partitioning by splitting reads into disjoint sets.

Assembly

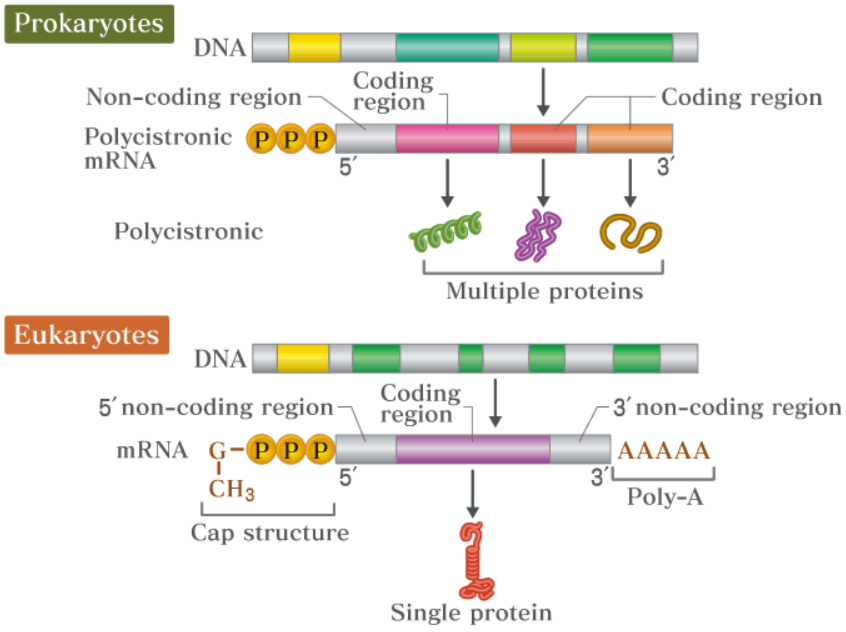
- Velvet/metaVelvet
- MetaAmos
- Mira
- Newbler (454 and hybrid assemblies)
- SOAPdenovo
- Meta-IDBA
- SPAdes



ORF Detection

- Most aligners can perform translated alignment which can be more sensitive and “overcome” sequencing errors
- These alignments can be slower than protein alignments (6-frame translations)
- ORF detection can:
  - Reduce computations for functional profiling
  - Provide “de-novo” genes
  - Allow for a complete sets of genes for gene clustering and sample comparison

ORF Detection

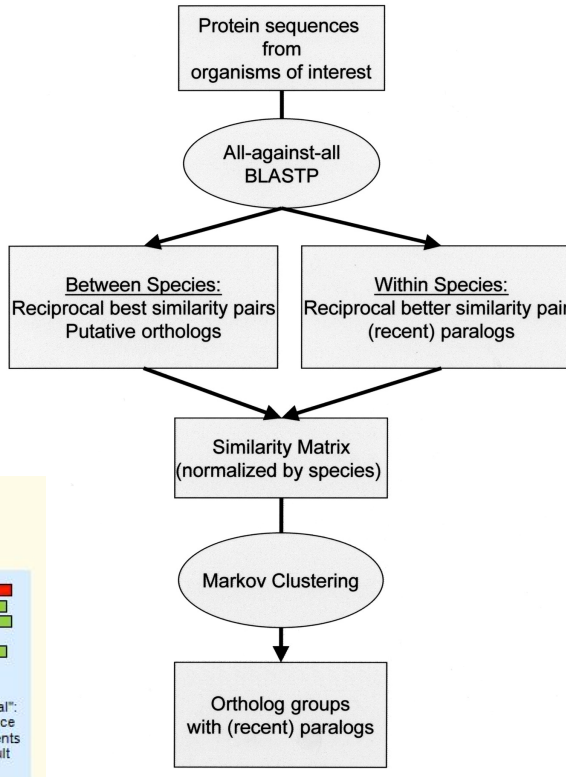


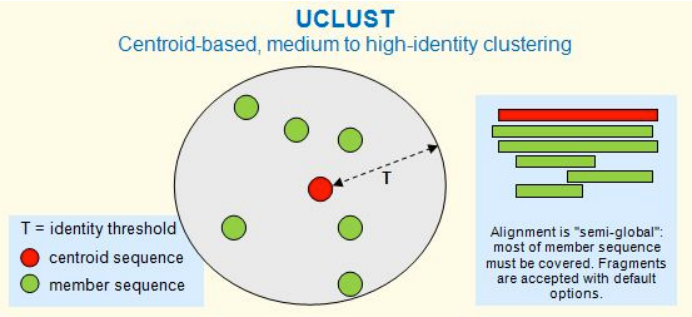
Gene Finding Packages

- Most Metagenomic gene finders are modified prokaryotic gene finders

- MetaGeneMark
- FragGeneScan (on reads)
- Glimmer MG
- Orphelia

Orthologous Clustering

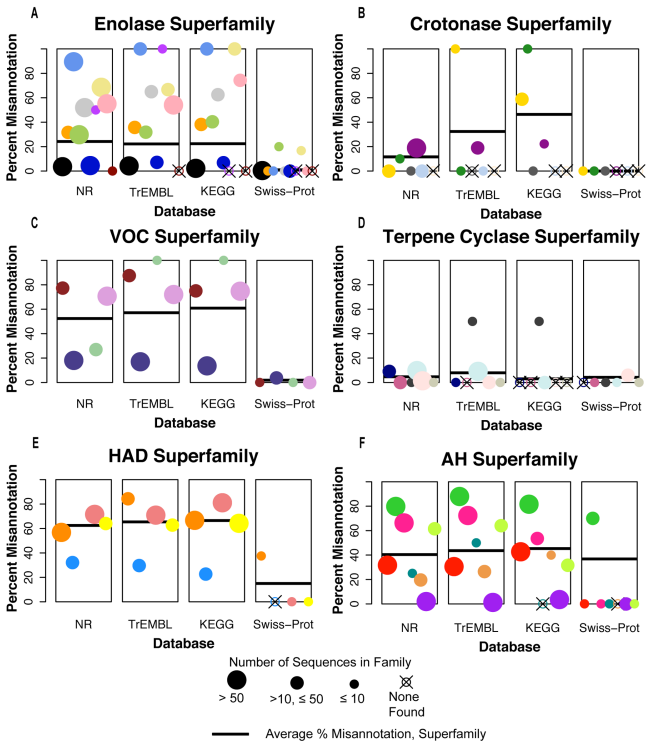




Functional Profiling

- High Throughput functional profiling comparison allows for gross comparisons of the functional capability of samples
  - Broad functional categories tend to be very similar in an ecological niche
- Profiling relies on alignments to functionally characterized proteins
- Homologous proteins tend to have similar broad “enzymatic function” i.e. kinase, hydrolase, transferase
  - However: Homology ≠ Same Biological Function

Functional Annotation Error are Common



Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009 Dec;5(12):e1000605. doi: 10.1371/journal.pcbi.1000605. Epub 2009 Dec 11. PubMed PMID: 20011109; PubMed Central PMCID: PMC2781113.

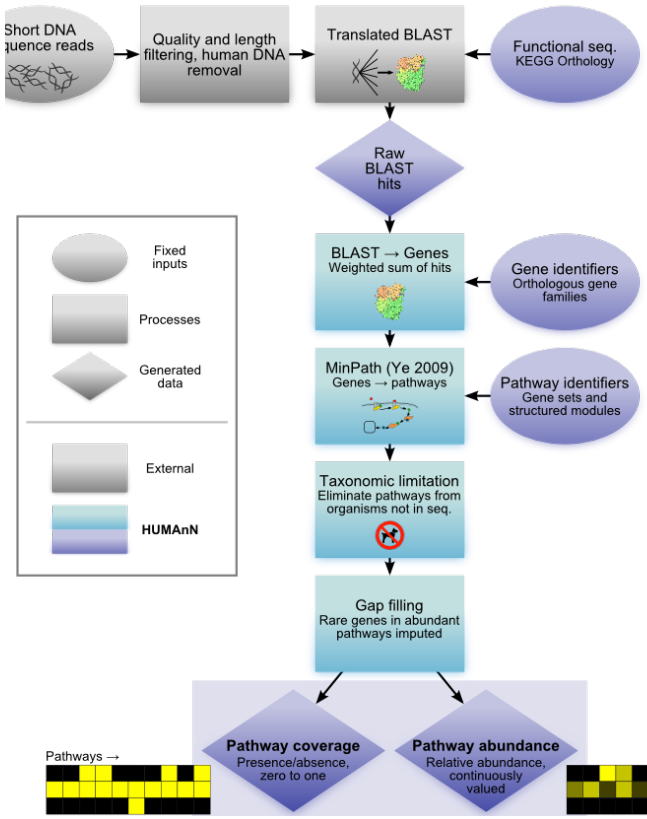
Alignment Strategies

- BLASTP or BLASTX — very slow
- MALT — Requires > 100GB of memory
- USEARCH — Requires paid license for 64 bit version; memory requirement too high for 32 bit version
- VSEARCH —Free version of USEARCH, lacks sensitivity
- DIAMOND — Much more sensitive than VSEARCH, low memory requirement and fast

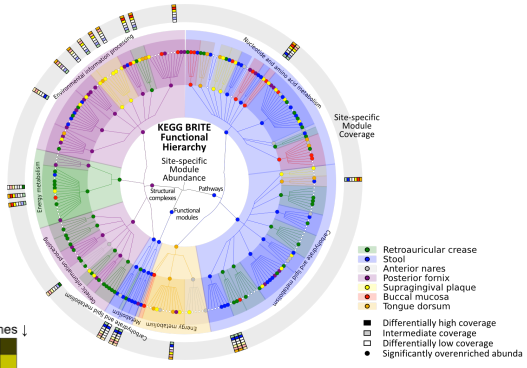
Post-Alignment

- Using Alignments (Translated or Protein) — functional assignment is based on broad functional categories or pathways of annotated hits.
- Available Packages for functional assignment and pathway profiling:
  - Humann
  - Megan

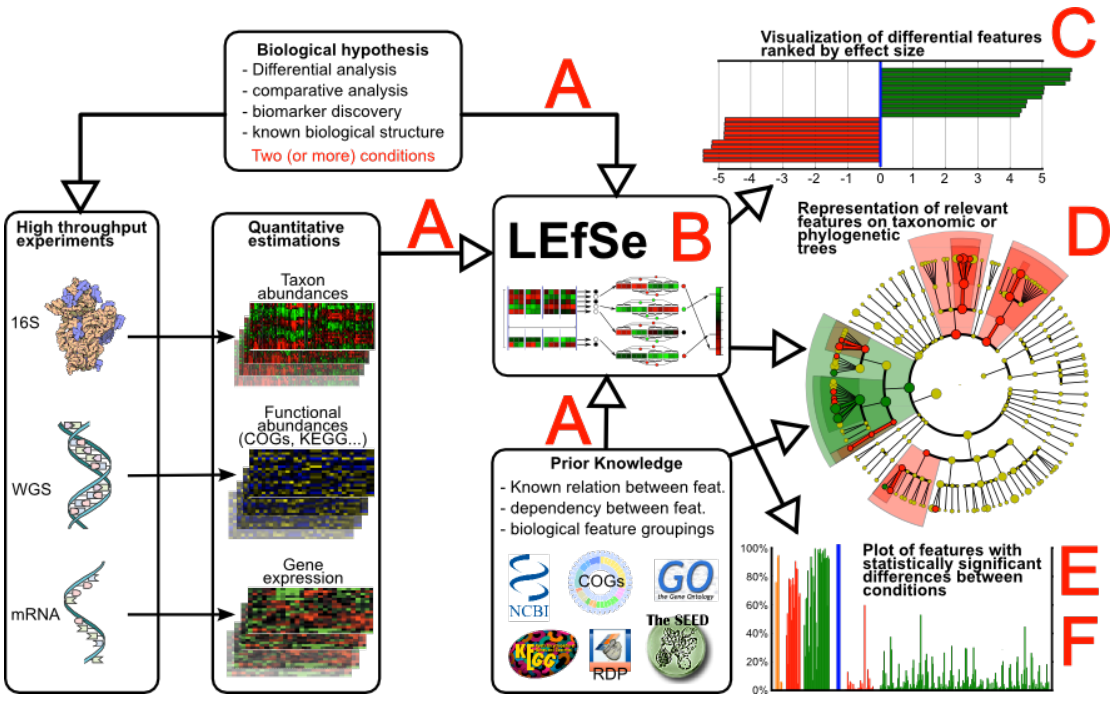
Humann



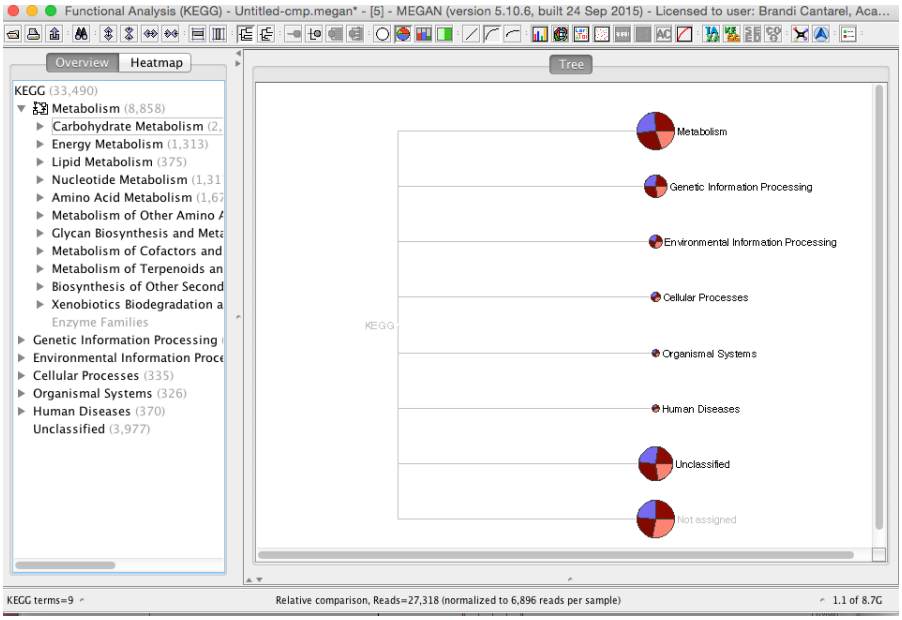
Pathway Analysis



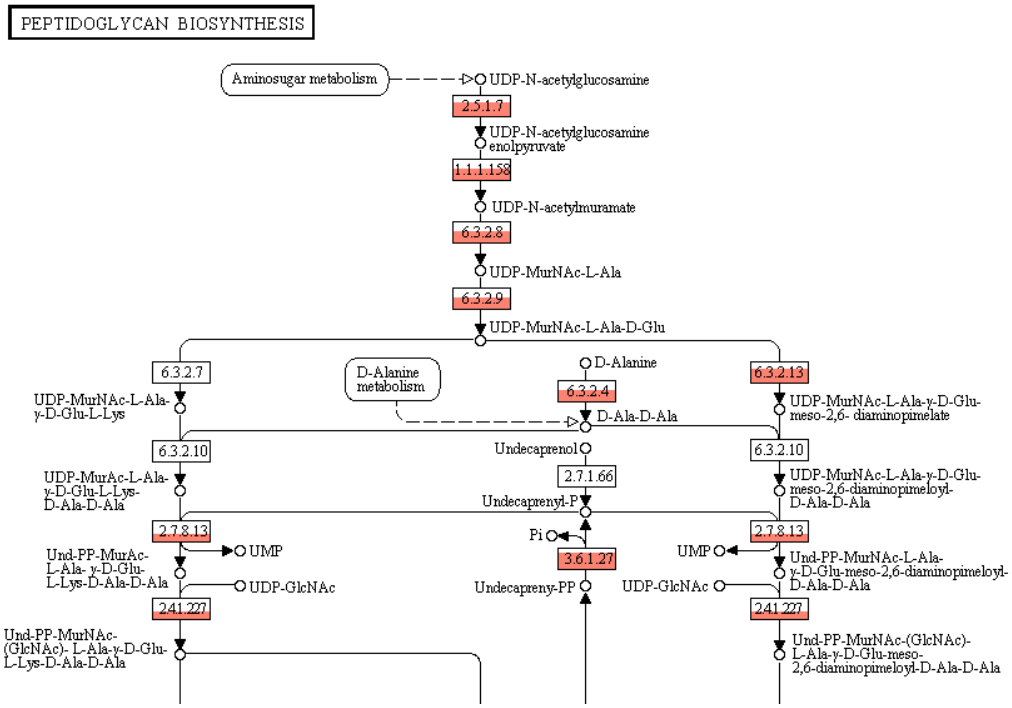
LefSe



MEGAN Broad Functional Comparisons



Pathway Exploration



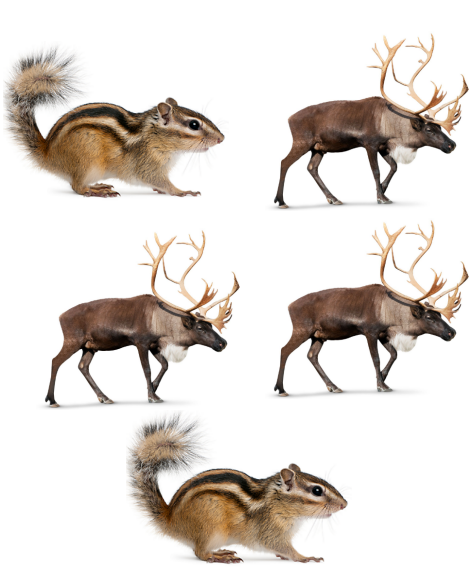
Statistical Analysis

- MetagenomeSeq
  - Normalization of counts based on relative abundance
  - 2-sample t-statistic
- STAMP
  - ANOVA for multiple groups differential testing
  - Kruskal-Wallis H-test — nonparametric method test to determine differences in medians

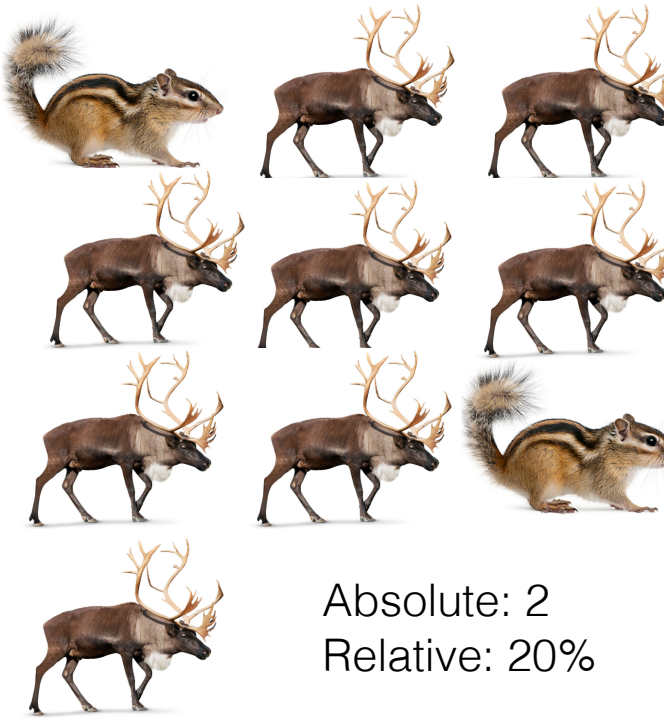
Relative Abundance vs Absolute Abundance

- Absolute abundance is a quantitate measure of the feature in the sample (qPCR)
- Relative abundance is the measure of a feature relative to all other features (sequencing)
- We can sequence every molecule in a sample, therefore abundances in microbiome studies are based on the number of measures (reads) and the proportion of the feature in the sample.

Relative Abundance vs Absolute Abundance



Absolute: 2
Relative: 40%



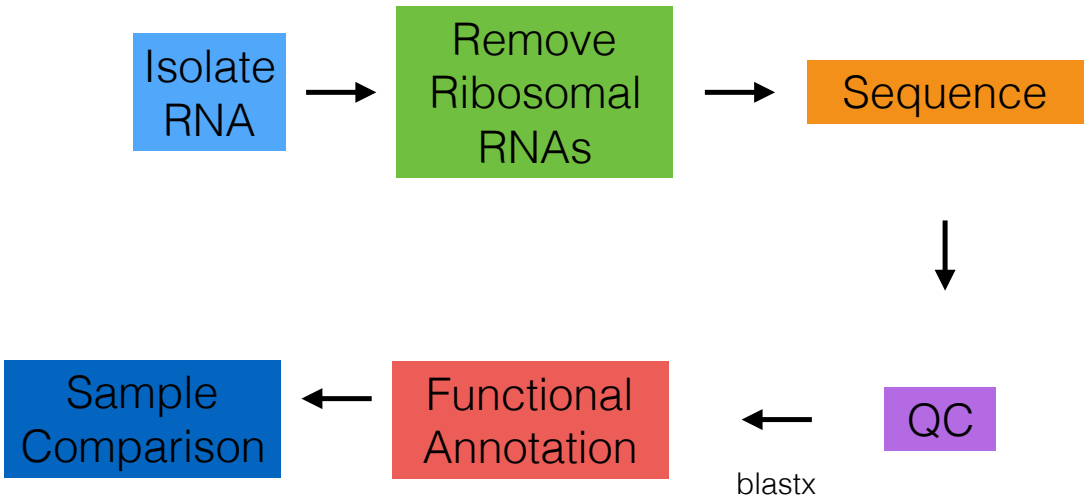
Absolute: 2
Relative: 20%

- Community Structure Based on WGS
  - 16S VS WGS
  - Least Common Ancestor
  - Reference Genome databases
  - Assignment Strategies
- Data Processing Workflow
  - QC
  - Functional Databases
  - Analysis Platforms
  - DYI Analysis
- Non-DNA based technologies: Metatranscriptomics, Metaproteomics and Metabolomics
  - Expression (RNASeq)
  - Translation (Proteomics)
  - Metabolites (Metabolomics)

Metagenomics vs Metatranscriptomics

- Metagenomics can give insight into gene content.
- Metatranscriptomics can measure how expression (functional potential) changes in response to the environment
- Metatranscriptomics can also show which organism are the most functionally active.

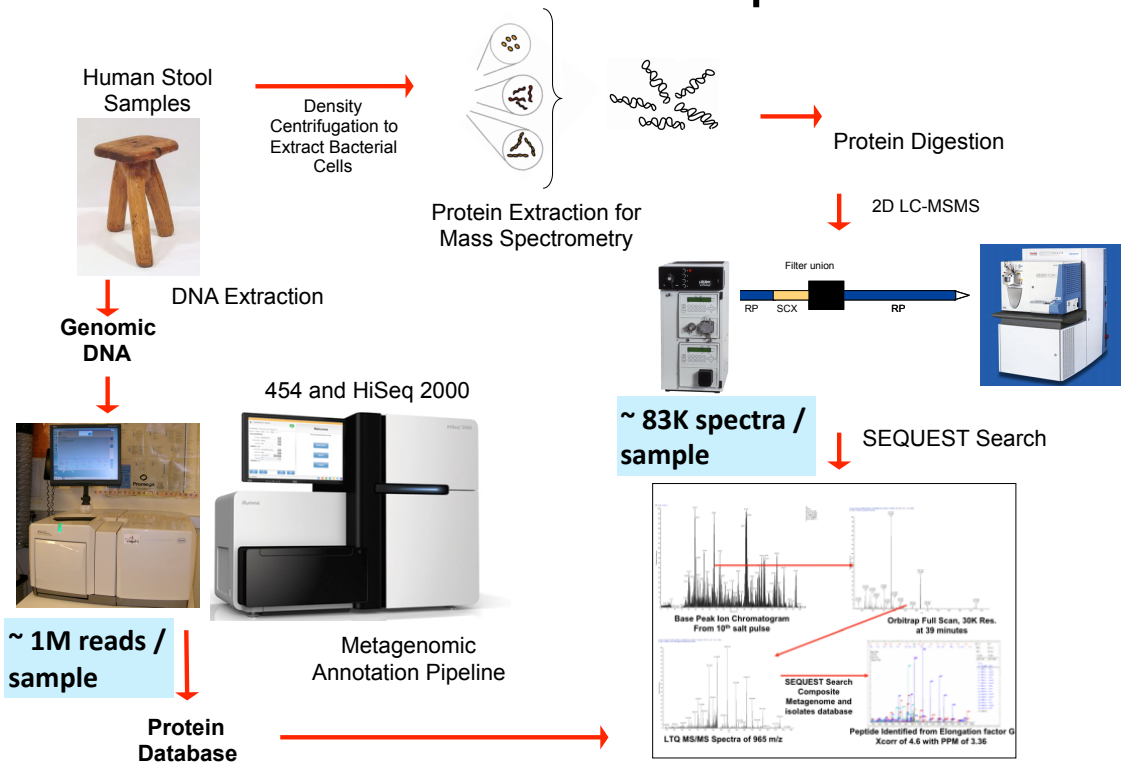
Metatranscriptomics



Metaprotomics

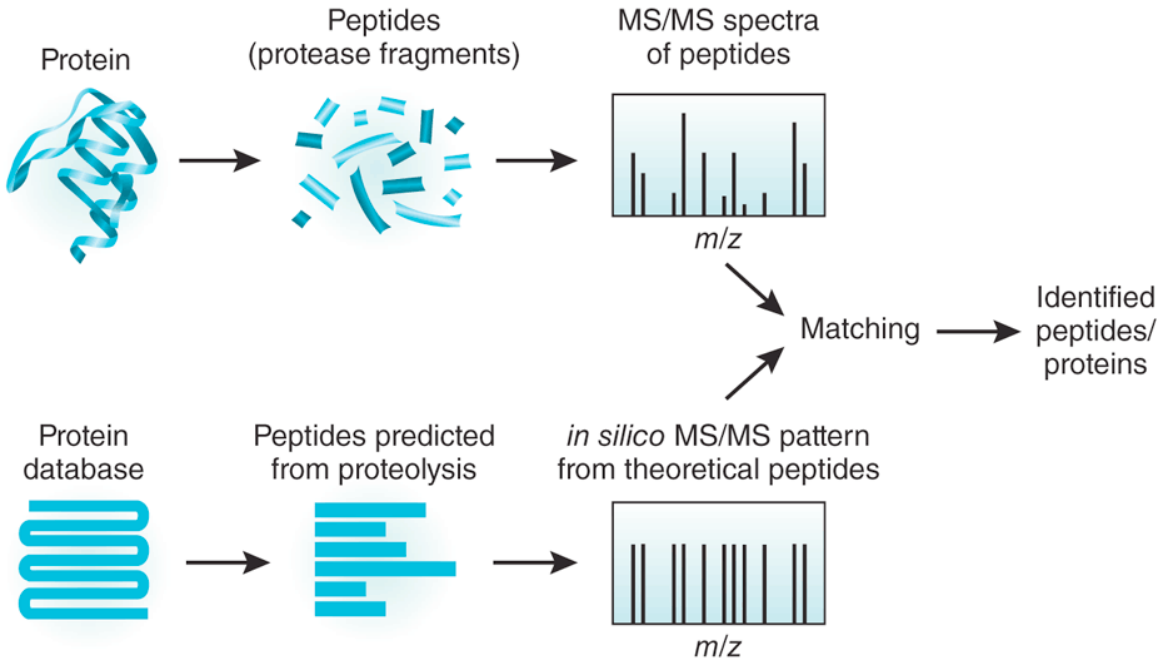
- Like metagenomics and metatranscriptomics, metaproteomics is complicated by the lack of a complete reference set
- In order to determine the protein sequence of peptide fragments, a metagenomic or reference genome database is necessary.
- Unlike sequencing, denovo protein prediction from MS/MS is not trivial.
- Contains a mixture of environmental and microbiome proteins

"Omics" Pipeline



Cantarel et al. (2011) PLoS One 6: e27173

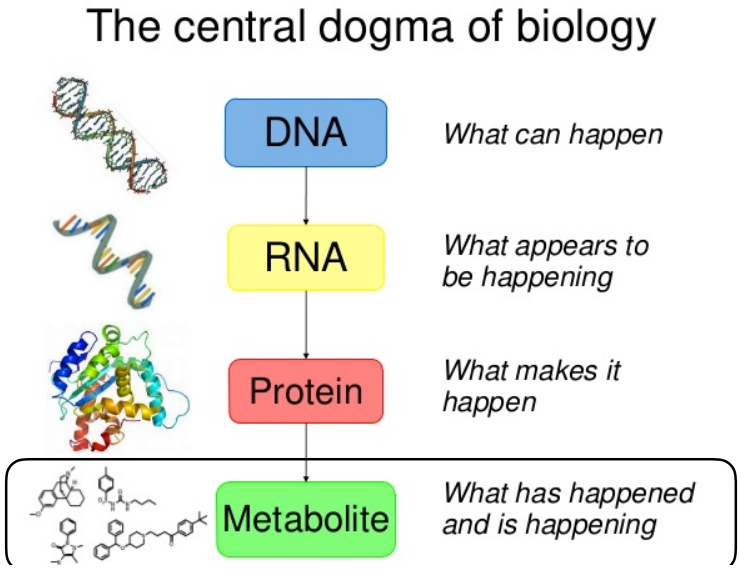
Peptide Spectral Matching



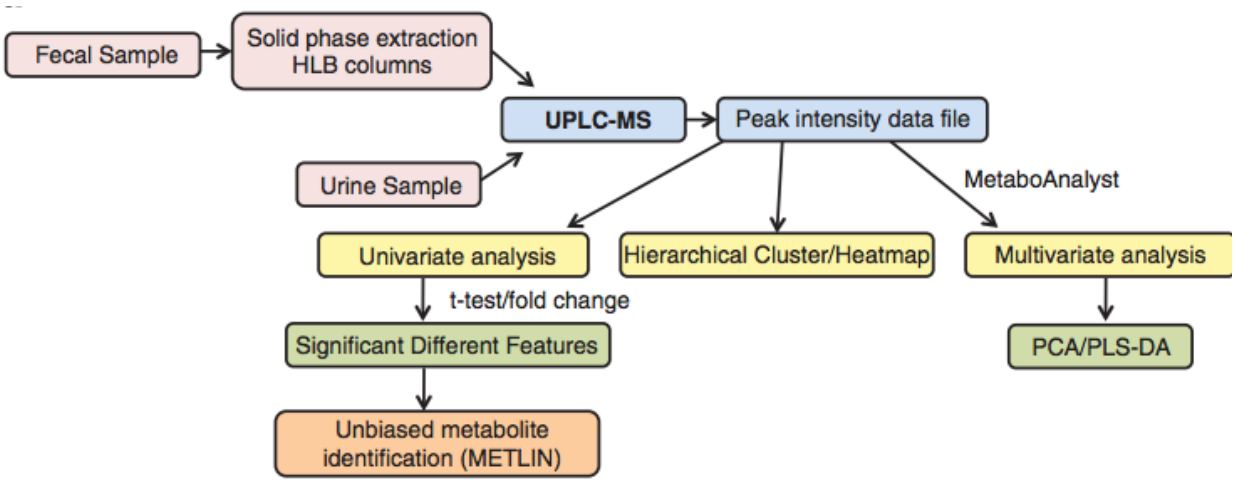
Duncan MW, Aebersold R, Caprioli RM. The pros and cons of peptide-centric proteomics. Nat Biotechnol. 2010 Jul;28(7):659-64.

meta-Metabolomics

Animal and environmental metabolomic studies are (meta)metabolomics — it is difficult to know "who" produced a particular metabolite.



meta-Metabolomics



Marcobal A, Kashyap PC, Nelson TA, Aronov PA, Donia MS, Spormann A, Fischbach MA, Sonnenburg JL. A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. ISME J. 2013 Oct;7(10):1933-43. doi: 10.1038/ismej.2013.89. Epub 2013 Jun 6. PubMed PMID: 23739052; PubMed Central PMCID: PMC3965317.

Workshop 2