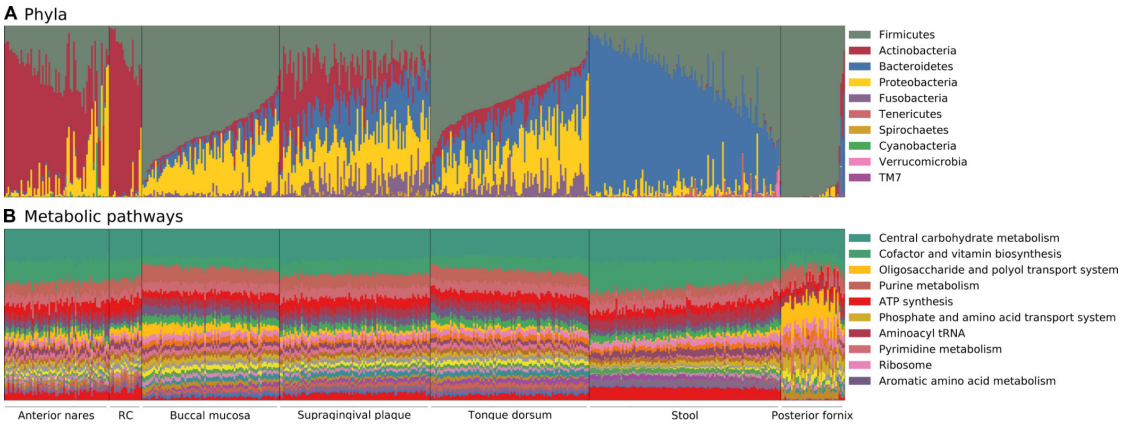
Lecture 8: Predicting metagenomic composition from 16S survey data

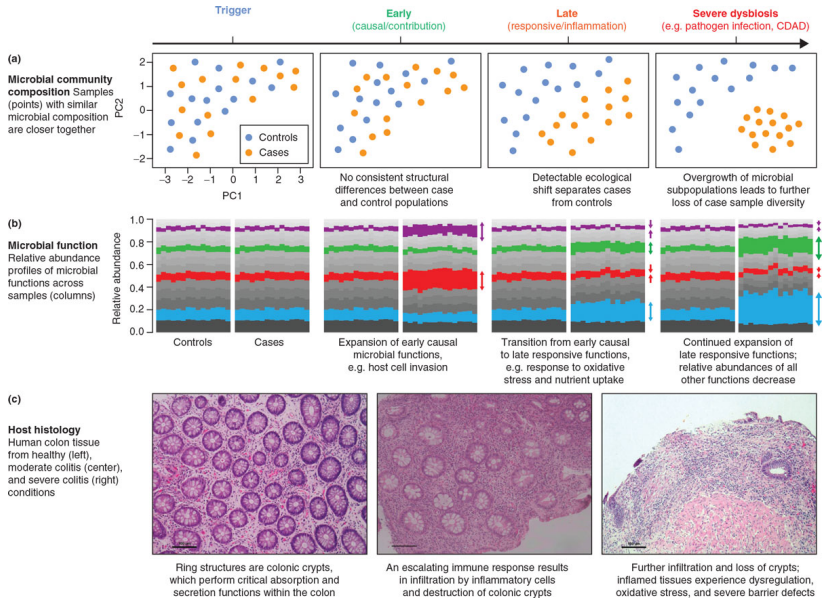
Taxonomic and functional stability of microbiota



Nature. 2012; 486(7402): 207–214. doi:10.1038/nature11234

2

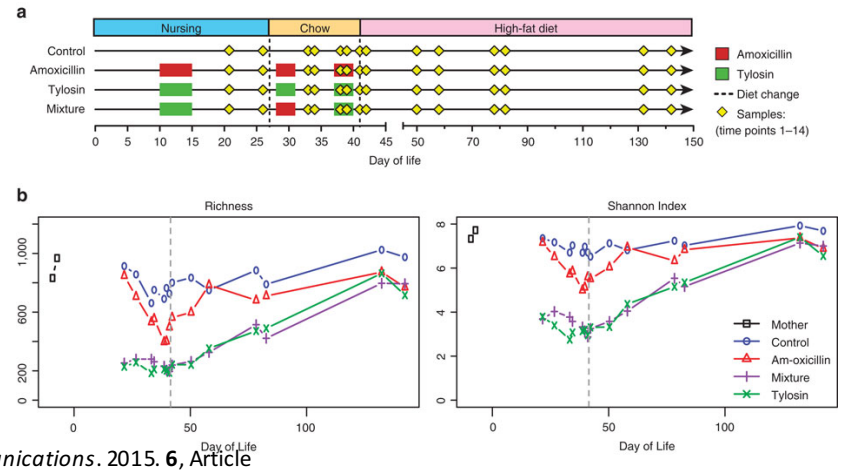
A model of functional dysbiosis in the human gut microbiome during initiation and progression of complex disease.



Genome Medicine, 2013; **5**:65
DOI: 10.1186/gm469

3

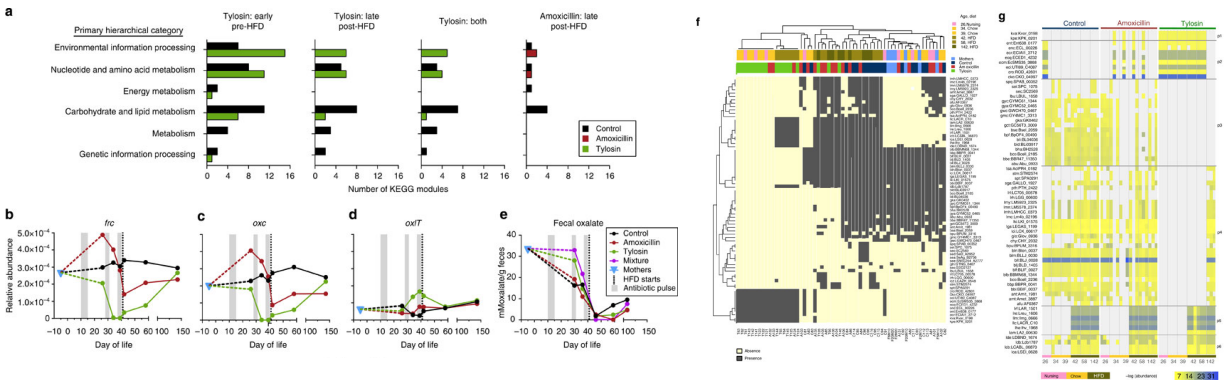
Functional differentiation in pulsed antibiotic treatment



Nature Communications. 2015. **6**, Article number: 7486. doi:10.1038/ncomms8486

4

Functional differentiation in pulsed antibiotic treatment



5

How to measure metagenomes?

- Unlike 16S rRNA gene sequencing, metagenomic sequencing is not targeted to a specific gene, but does an unbiased sample of the entire (bacterial) genomic DNA in a specimen.
- Typically shorter sequence reads are used to obtain >5Gb of data per sample.
- HiSeq instruments are typically more cost effective for metagenomic sequencing.
- This approach is also called shotgun whole metagenome sequencing or WmGS or WGS.

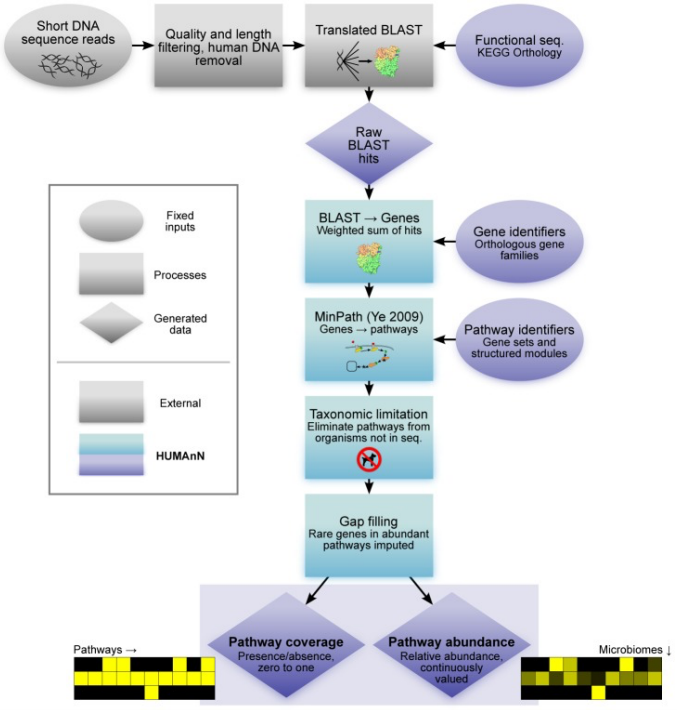
6

Metagenomic processing pipelines

- MG-RAST: *BMC Bioinformatics*, 2008; **9**:38. DOI**:** 10.1186/1471-2105-9-386
- SUPER-FOCUS: Bioinformatics, 2016; 32 (3): 354-361. DOI: 10.1093/bioinformatics/btv584
- HUMAnN: *PLoS Comput Biol*, 2012; 8(6):e1002358. DOI: 10.1371/journal.pcbi.1002358

7

Example pipeline: HUMAnN



8

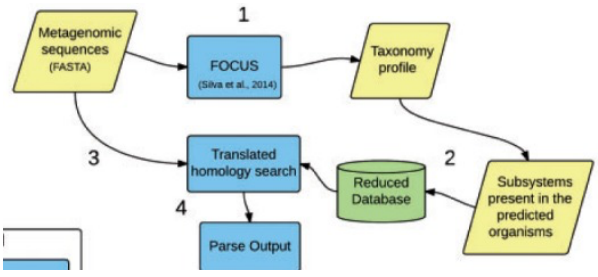
How is metagenome data representation different from 16S rRNA gene sequencing data?

- Taxonomic information is typically discarded, only functional data remain.
- The most fundamental unit of analysis is an individual gene, or orthologous group of genes.
- Genes may be grouped by pathways, systems, diseases, etc.
- Abundance or presence/absence of genes, pathways, etc. is captured in the data matrix.

9

Predicting metagenomes

- Metagenomic sequencing is considerably more expensive.
- The informatics processing is much more complicated.
- In the end, we rely on reference databases of known genes; no true de novo functional information is discovered.
- Some pipelines (e.g. SUPER FOCUS) require taxonomic information.



10

Idea: We can predict metagenomes from the 16S rRNA gene bacterial identification data

- 16S rRNA gene sequencing allows for indentification of microbiota.
- If we know the organism, we may have gene content of that organism or a related organism.
- We can use the information to infer the metagenomic content and use the abundances to reconstruct the metagenomic abundances.

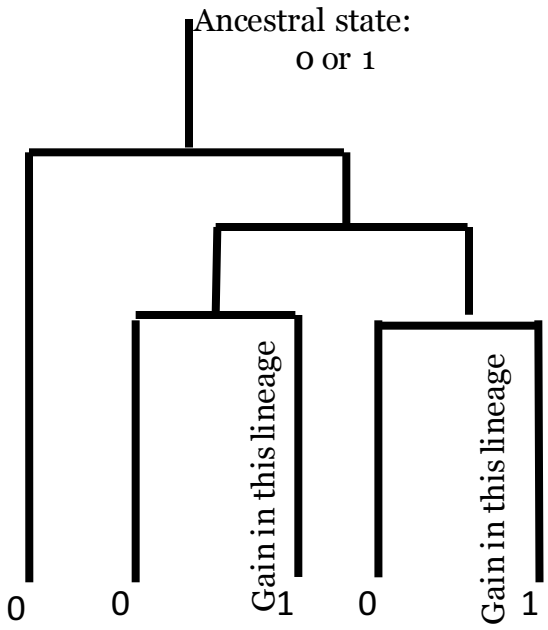
11

Key issues to address

- 16S rRNA gene may have multiple copies in some genomes
  - Solution: normalize the 16S data by multiplicity
- How do we infer metagenomic content of related organisms?
  - Solution: ancestral reconstruction

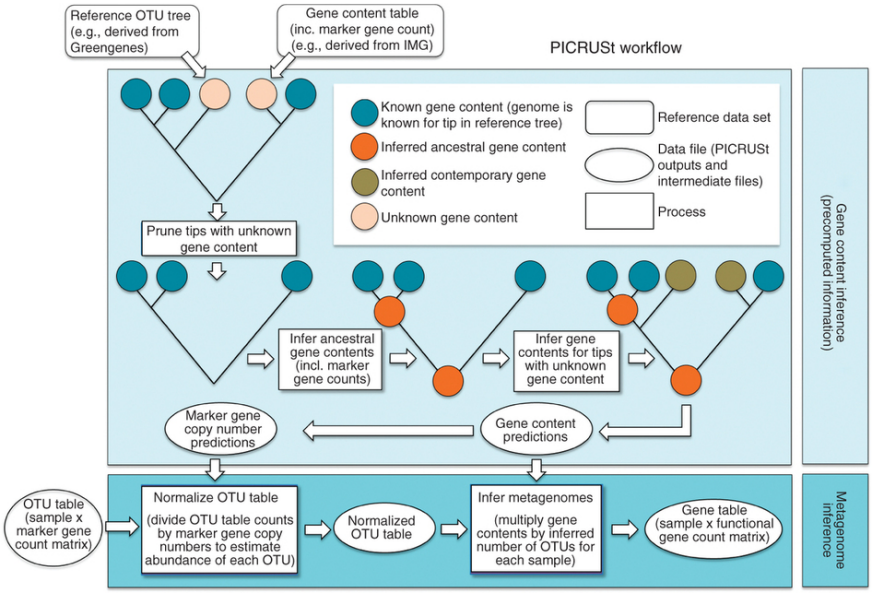
12

Ancestral reconstruction



13

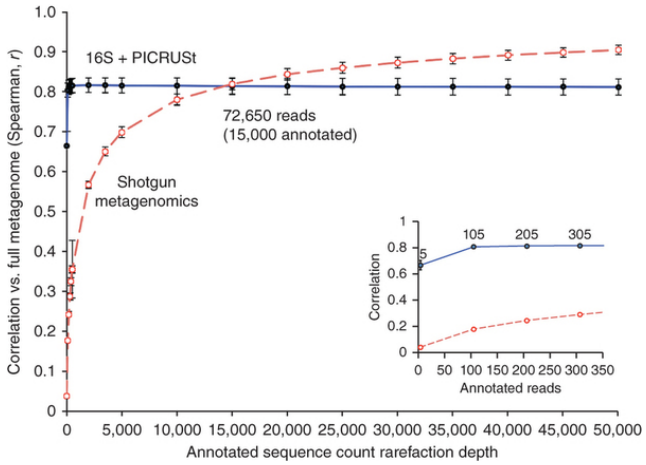
PICRUSt workflow



Nature Biotechnology. 2013; **31**, 814–821.

14

Performance of PICRUSt



Nature Biotechnology. 2013; **31**, 814–821.

15

How do we analyze the metagenome data?

- Analyses we discussed before are still applicable
  - Multivariate analysis
  - Hypothesis testing
  - Machine learning approaches
- Gene set enrichment analysis:
  - Determine if the number of significant genes within a category is
  - Can be accomplished with Fisher-exact test, hypergeometric test

16