Microbiomes:
Large datasets and LGT

Conor Meehan
conor.meehan@dal.ca

What is a microbiome?

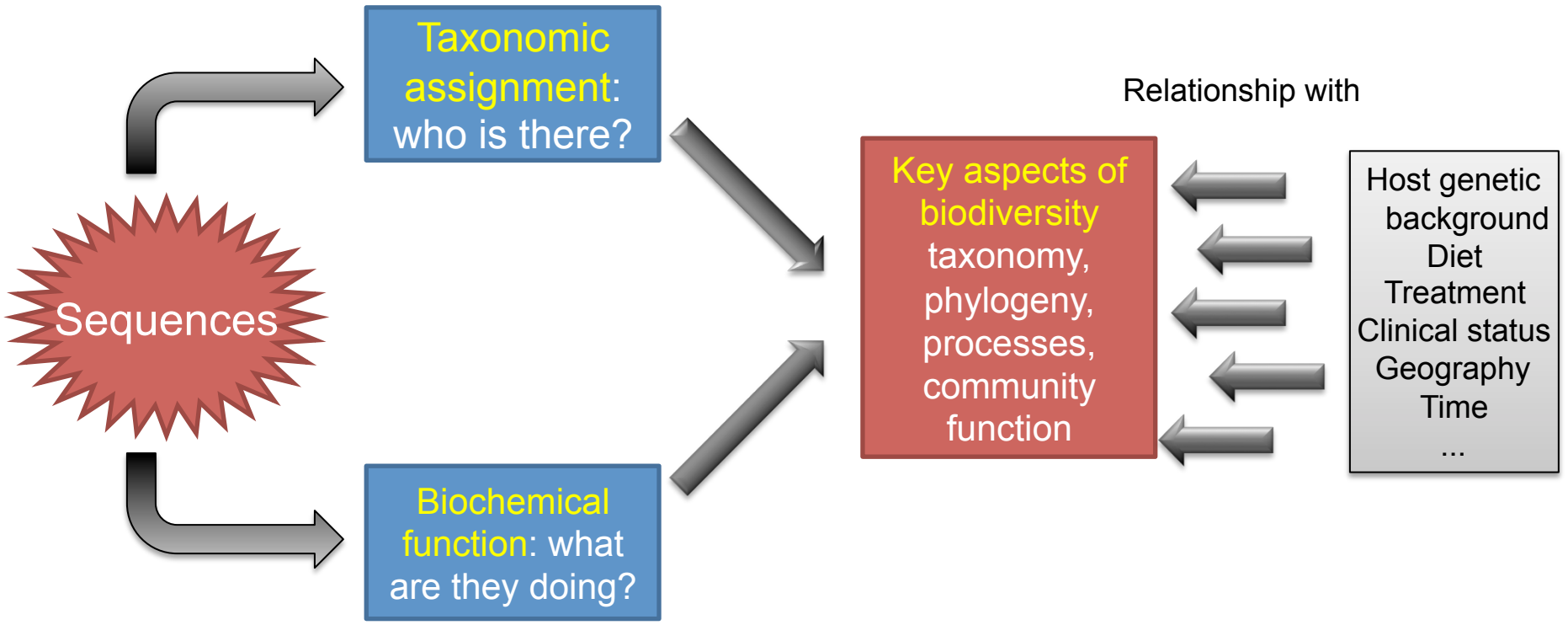
Microbiomes

- A microbiome is the collection of all micro-organisms that live in a given environment
- Metagenomics is the study of the genomic material recovered from a microbiome sample
- Allows us to investigate communities in habitats

How do we sample a habitat?

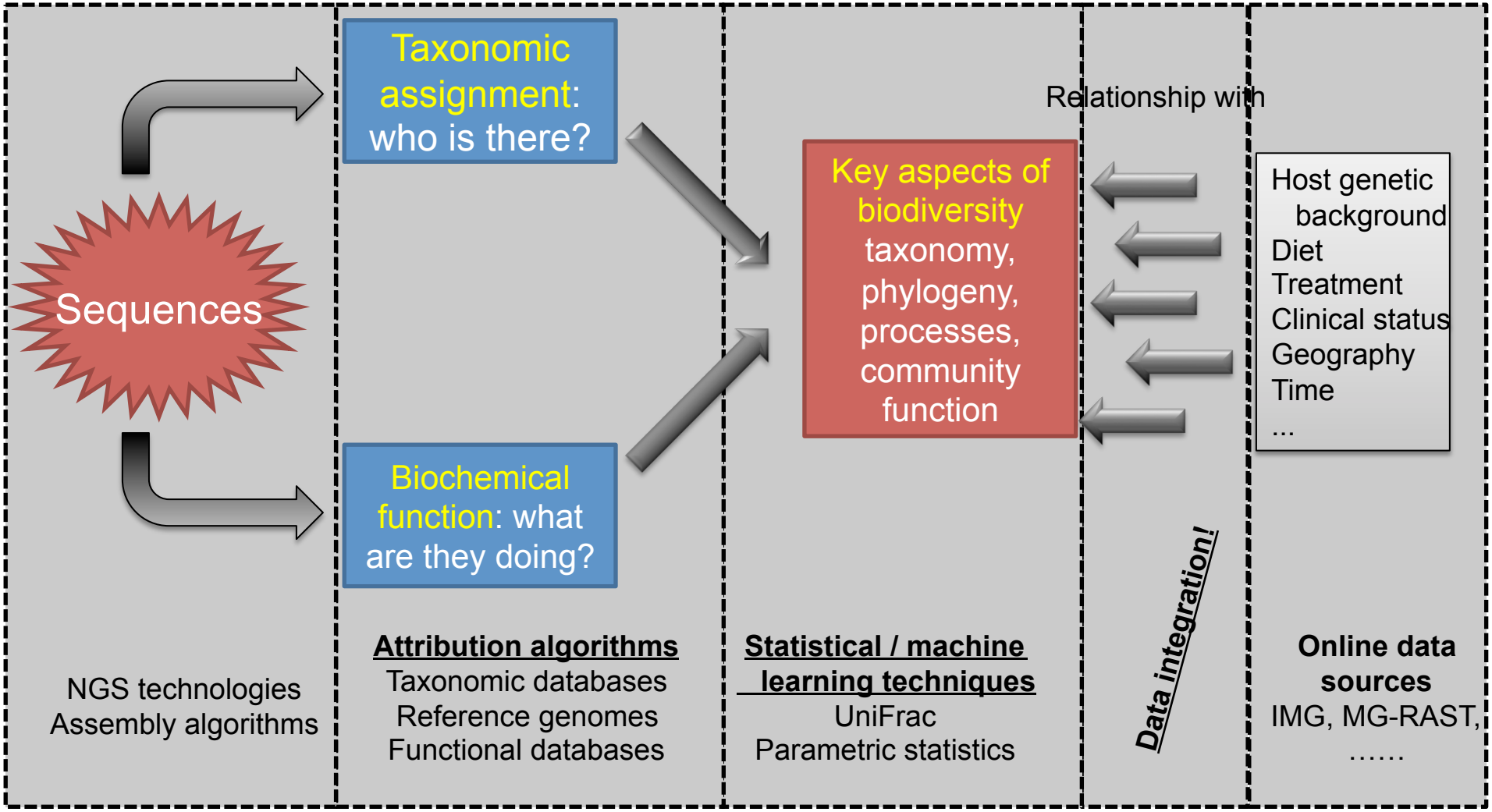
- Originally culture and sequence strategies were employed
  - Found that many (if not most) microbes were unculturable
- Often 2nd generation sequencing is used in one of two manners:
  - Direct sequencing of a (part of a) marker gene
  - Random sequencing of whole genomes
    - Often results in capturing ~80% of total metagenome
  - Can lead to large datasets that are difficult to process
    - Recent marker gene study resulted in 1 billion reads from over 500 samples
    - (this is where all that UNIX training can come in handy)
- Meta-data is normally also collected. E.g.:
  - Body Mass Index (BMI)
  - pH
  - Salinity

What’s involved in studying microbiomes?



Created by R. Beiko

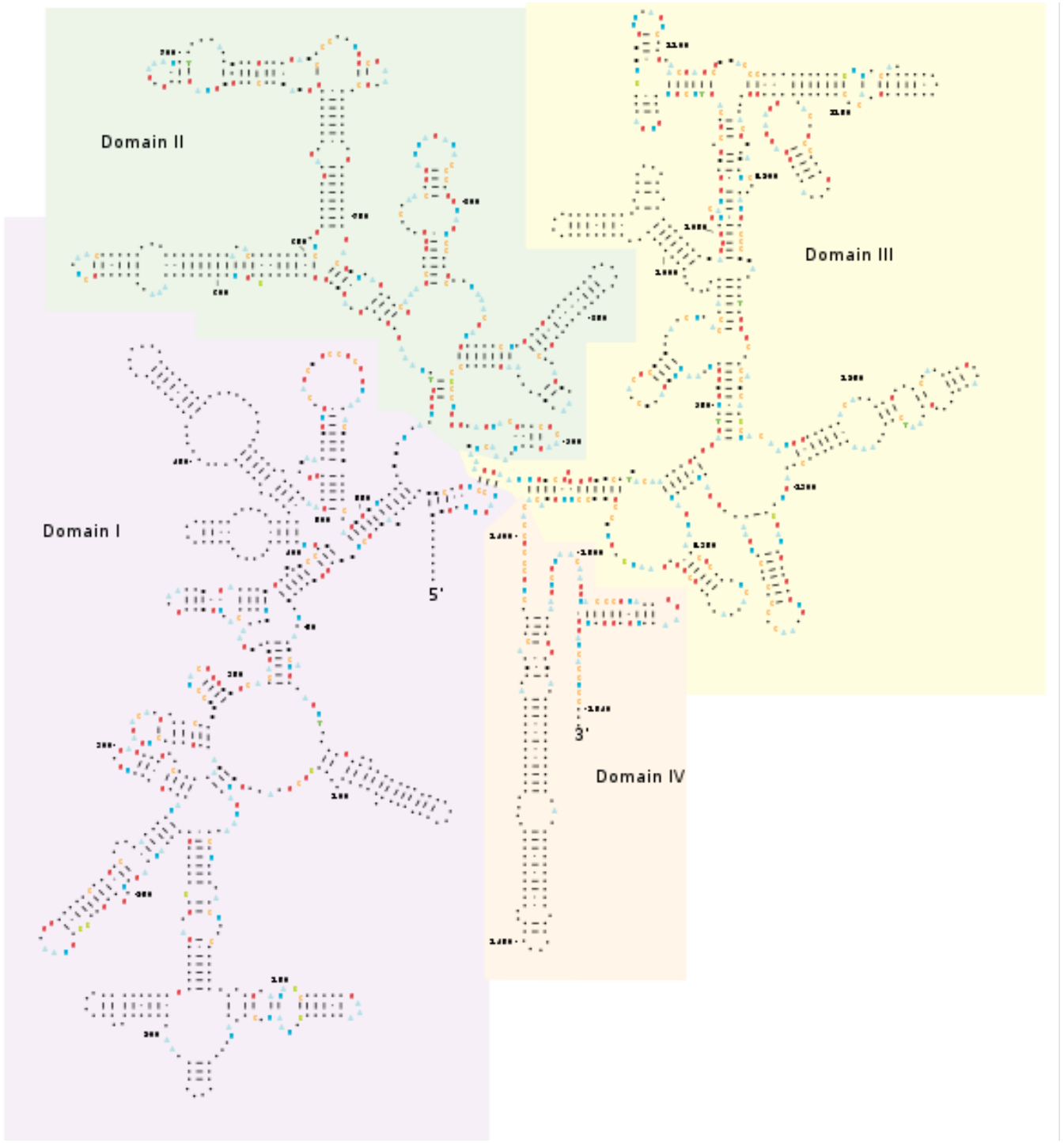
Tools at our disposal



Computer hardware: important when you have >1TB of sequence data!!

Who is there?

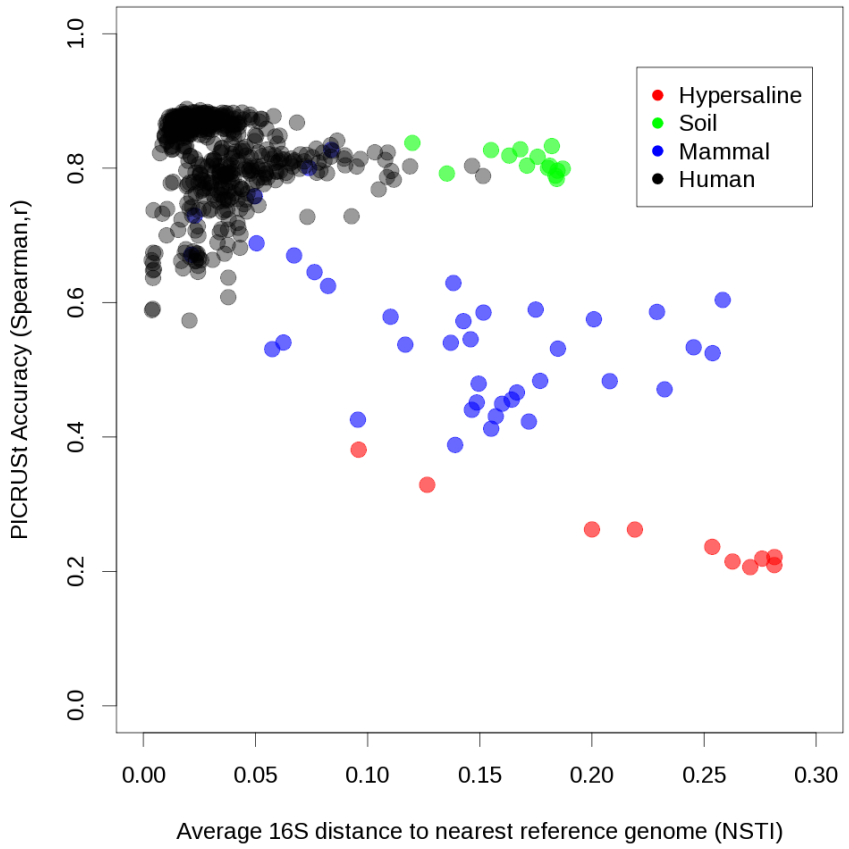
- We have a sample and want to know the species abundance
  - Can then relate this to metadata and environmental changes
- Basic procedure
  - Sequence a marker gene from your sample
  - Compare sequenced reads to a database of full length marker gene sequences
    - Homology or composition based
    - Phylogenetic based
  - Assign reads to a given species or higher taxonomic rank
- Taxonomic assignment of a metagenome is not trivial
  - Need good reference database
  - Need diverse reads/deep sequencing to cover all microbiome
- Approaches
  - Unsupervised (e.g. Qiime[1])
    - Place into bins but don't know if these bins relate to 'species'
  - Supervised homology/composition (e.g. RITA[2])
    - Reliant on a good reference database
  - Supervised phylogenetic insertion (e.g. pplacer[3])
    - Reliant on a good reference database
- All these require a good marker gene to be chosen

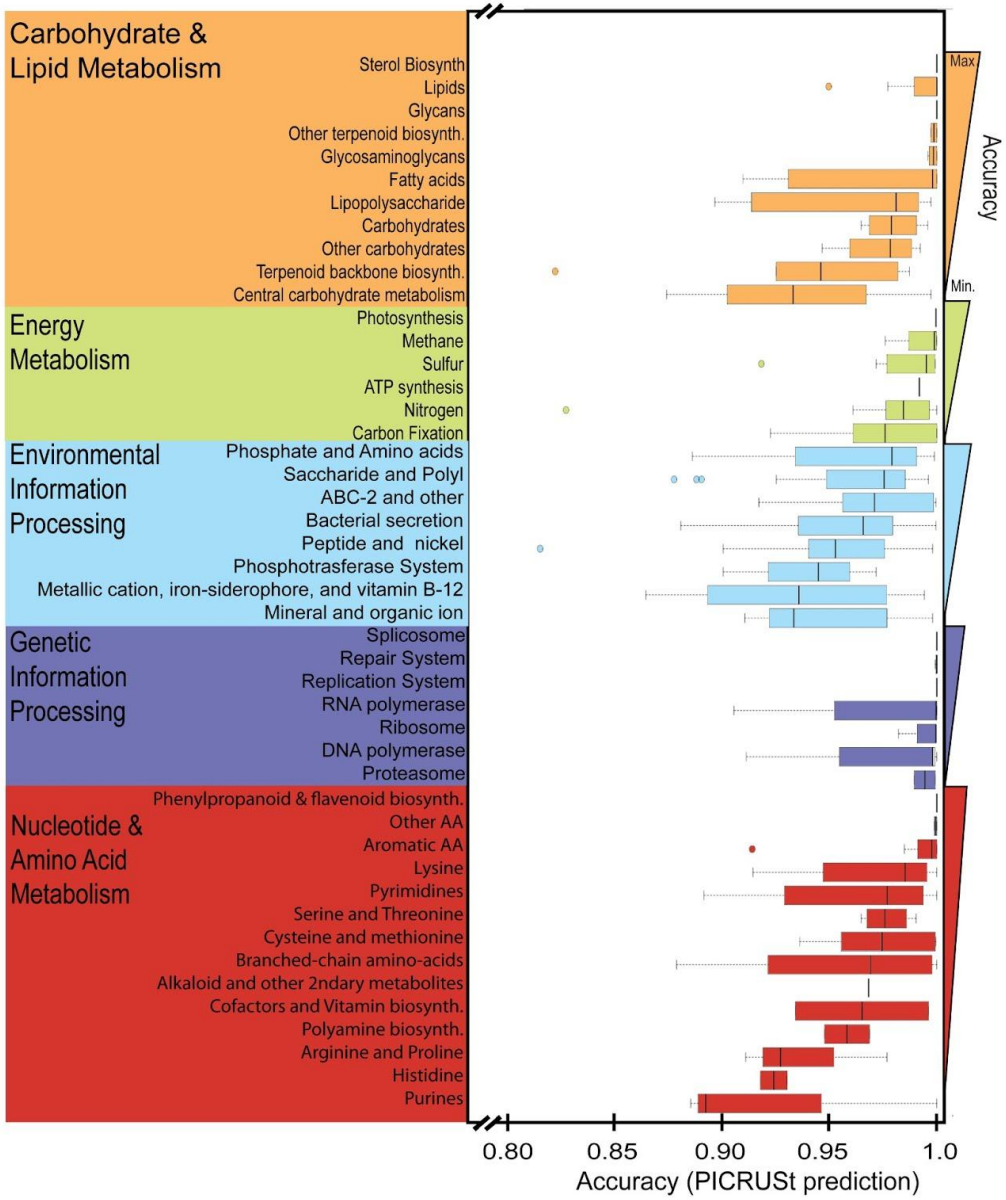


From ref 4

16S as a marker

- Component of the 30S small subunit in prokaryotes
- First used by Carl Woese
  - Was used to define the Archaea as a separate domain and introduce the 3 domain system [5]
- Often used as a phylogenetic marker due to high conservation across prokaryotes
- Near universal primers can be used for sequencing in microbiome studies
- Can tell us who is there but not what they do…mostly





Inference of function from 16
Undertaken using PI-CRUST[6]

16S as a marker

- Component of the 30S small subunit in prokaryotes
- First used by Carl Woese
  - Was used to define the Archaea as a separate domain and introduce the 3 domain system [5]
- Often used as a phylogenetic marker due to high conservation across prokaryotes
- Near universal primers can be used for sequencing in microbiome studies
- Can tell us who is there but not what they do…mostly
- Multiple copies are present in each prokaryotic cell
  - Makes quantification of relative abundances difficult
  - Not all copies are identical

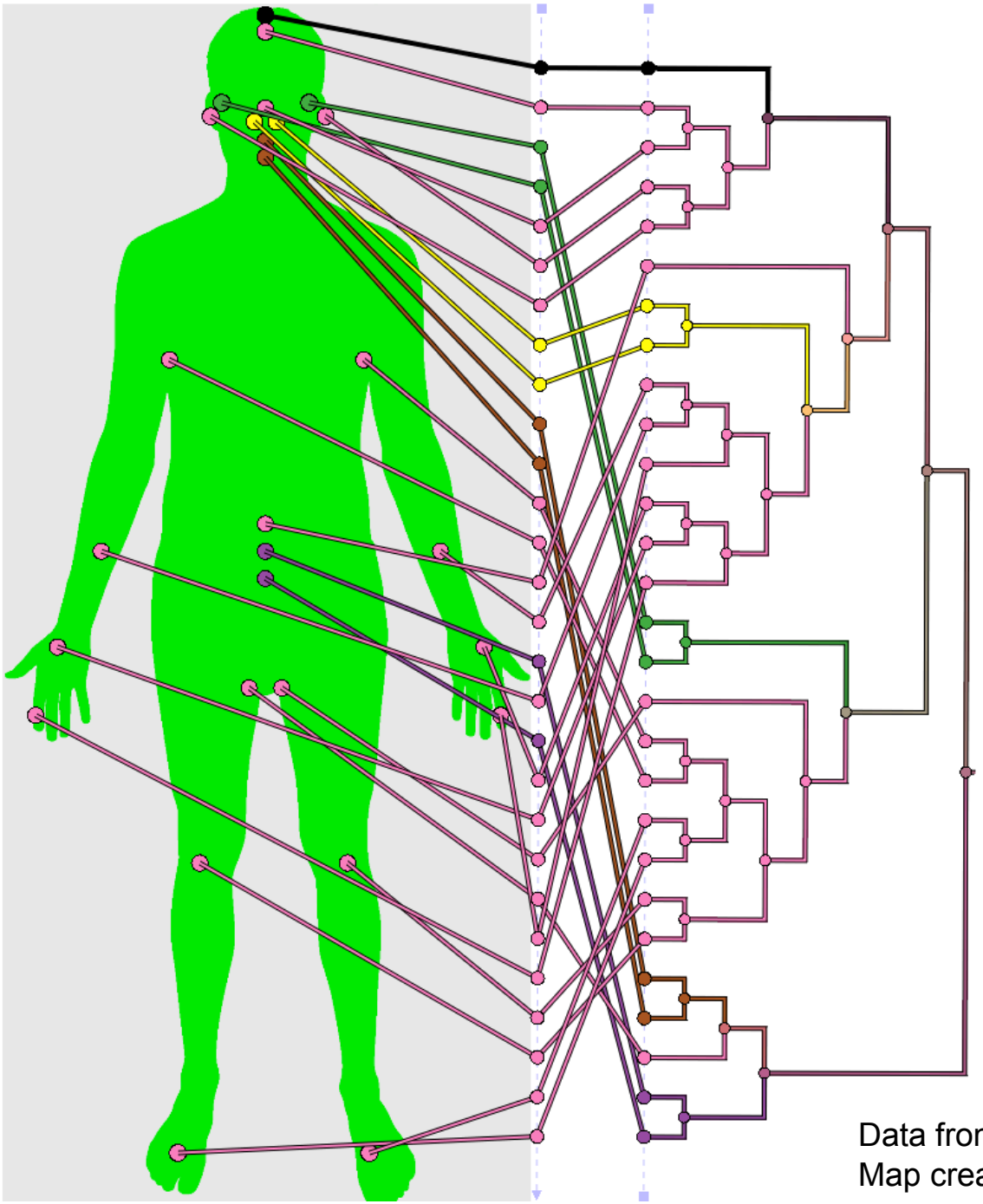
Human microbiome:
Creation, shaping and variation

Human microbiome video (that probably wont work)

- http://www.wimp.com/yourmicrobes/
- Jessica Green, University of Oregon

Human microbiome

- The collection of microbes which live in or on the human body
- Approximately 10 times as many microbial cells as human cells[7]
- Generally studied as multiple environments:
  - Skin
  - Vaginal
  - Oral
  - Digestive tract



Data from Costello *et al*[13]
Map created with GenGIS[14]

Human microbiome

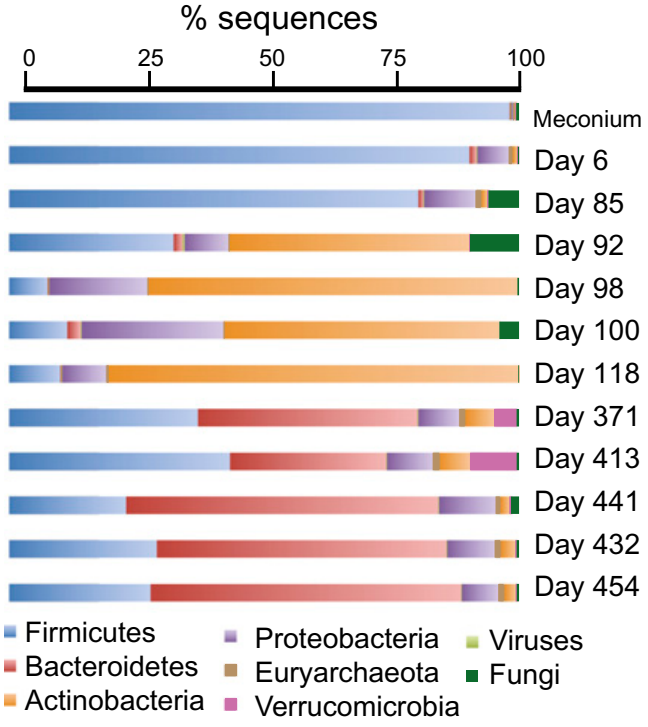
- The collection of microbes which live in or on the human body
- Approximately 10 times as many microbial cells as human cells[7]
- Generally studied as multiple environments:
  - Skin
  - Vaginal
  - Oral
  - Digestive tract
- Undertakes several functions
  - Modulates immune system
  - Breaks down glycans
  - Creates energy sources for host and community
  - Protects against pathogens

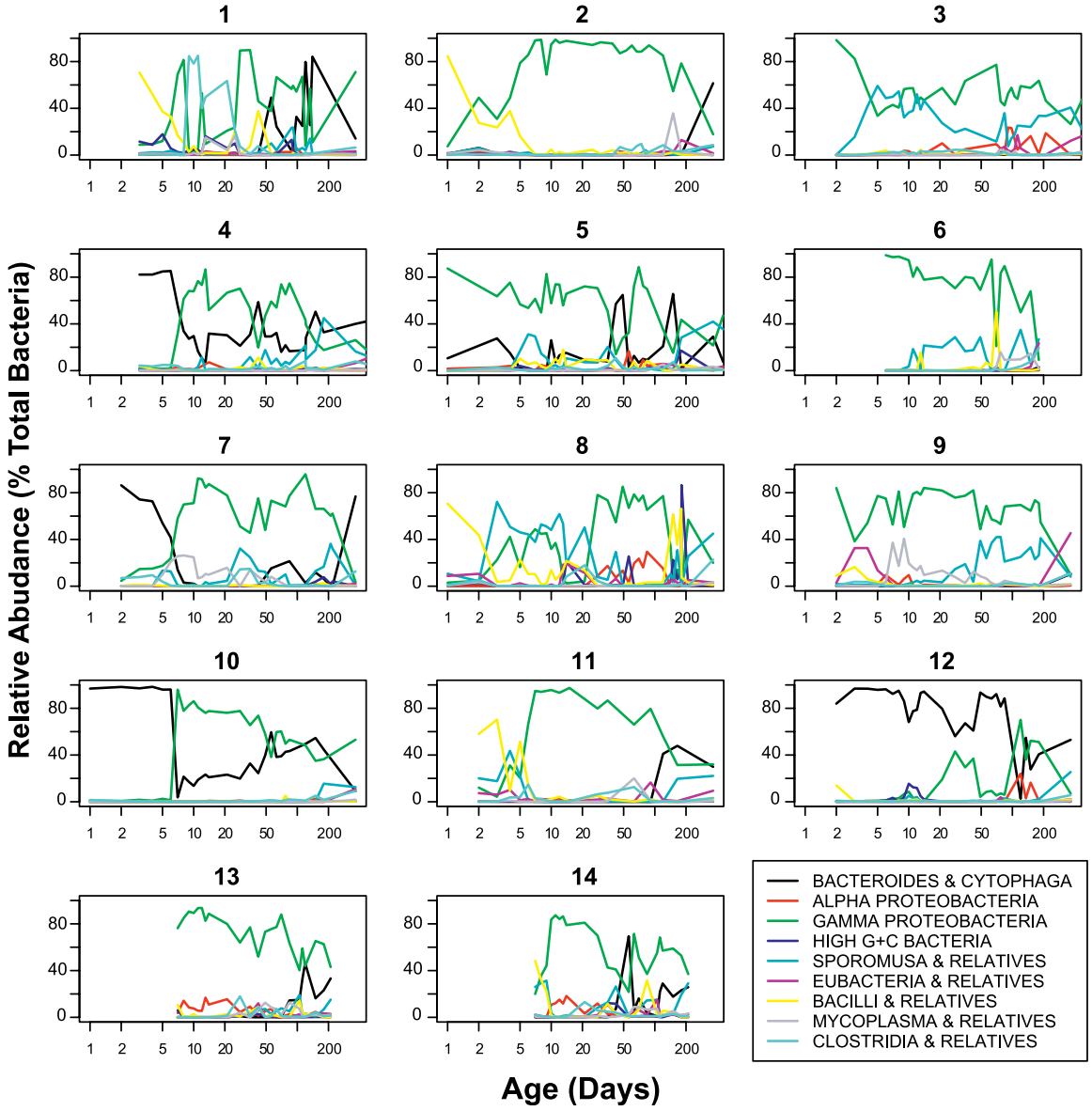
Human microbiome: Cool Facts

- Genetic pool of the gut microbiome in Japanese populations contains a gene that can break down porphyran found in seaweed[8]
  - LGT from a seaweed-residing bacterium *Zobellia galactanivorans*
- Strains of bacteria present in the gut may affect levels of stress and anxiety (shown in mice)[9]
  - Strains of Lactobacillus may stimulate production of GABA receptors in brain
- The composition of the skin microbiome can influence the likelihood of being bitten by a mosquito[10]
  - Higher levels of Staphylococcus and Variovorax
- Faecal transplants from healthy patients into those who suffer from Crohn's disease can cause disease to go into remission[11]
  - Also works for curing *C. difficile* infections
  - Now have a synthetic version called 'RePOOPulate'
- Changes in penile microbiome in circumcised men may affect the likelihood of being infected by HIV[12]
  - Reduced numbers of anaerobic bacteria reduces the number of immune cells in the area

Beginnings of a human microbiome

- A foetus in the womb is (likely) sterile[15]
- Initial colonisation of a baby occurs the moment it is born
  - Microbes adhere to child from birth canal
  - Environmental exposure seeds further microbes
- Establishment of this community in early years is influenced by many factors
  - Breast-fed/formula-fed
  - Antibiotic usage
  - Home environment such as plants and animals
- These initial inhabitants are not the ones that progress in later life[16]
  - May exist as ‘primers’ for environment
  - Diversity increases and relative abundances of phyla shifts dramatically

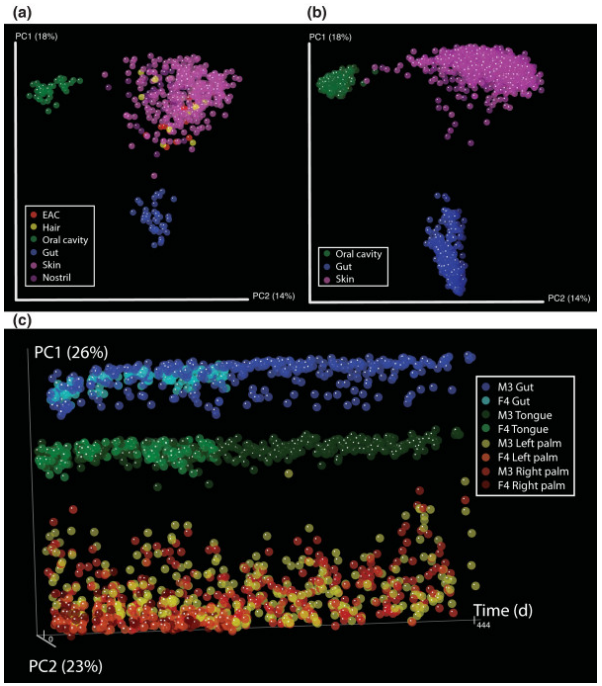


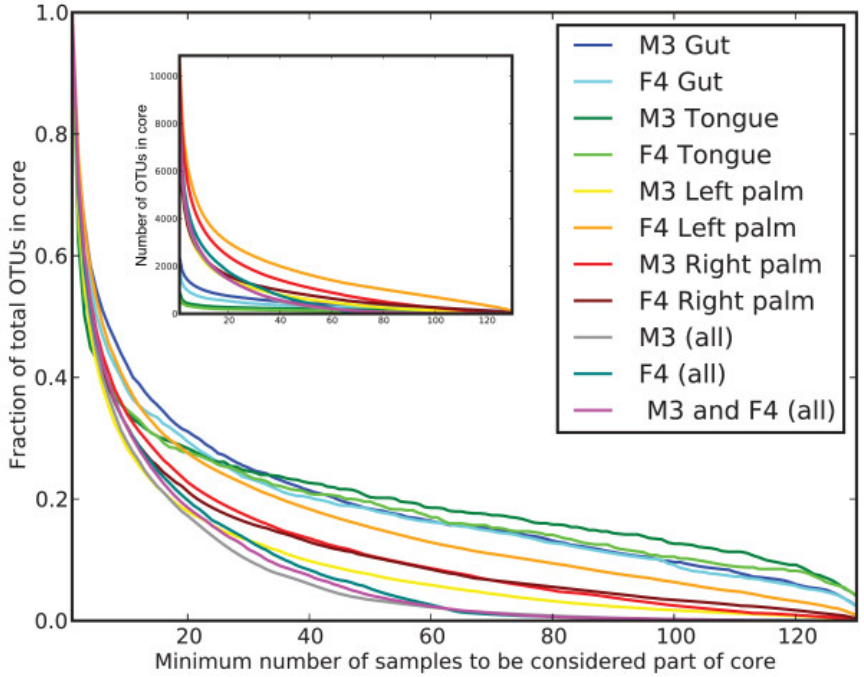


Adapted from refs 16 and 17

Temporal fluctuations

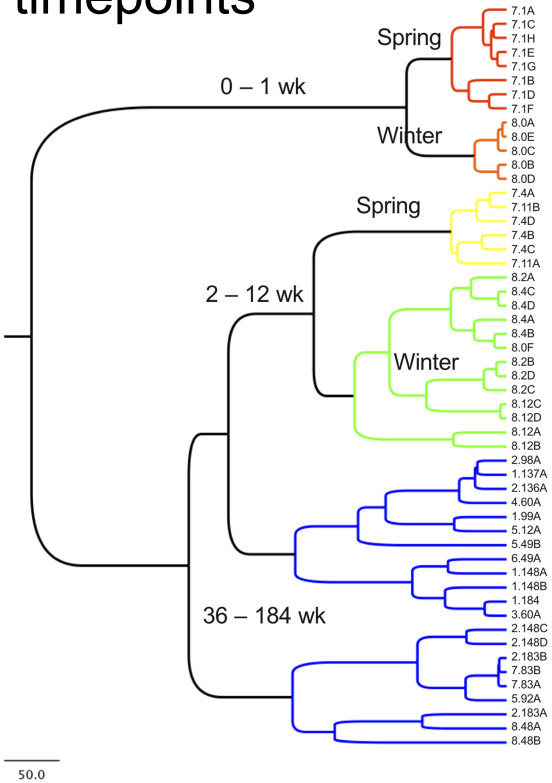
- The human microbiome appears to change daily
  - Fluctuations in environmental interactions
  - Changes in diet can reshape microbiome in a single day[18]
  - New microbes entering body
- Different body sites are distinguishable but no core species consistently present within a site[19]

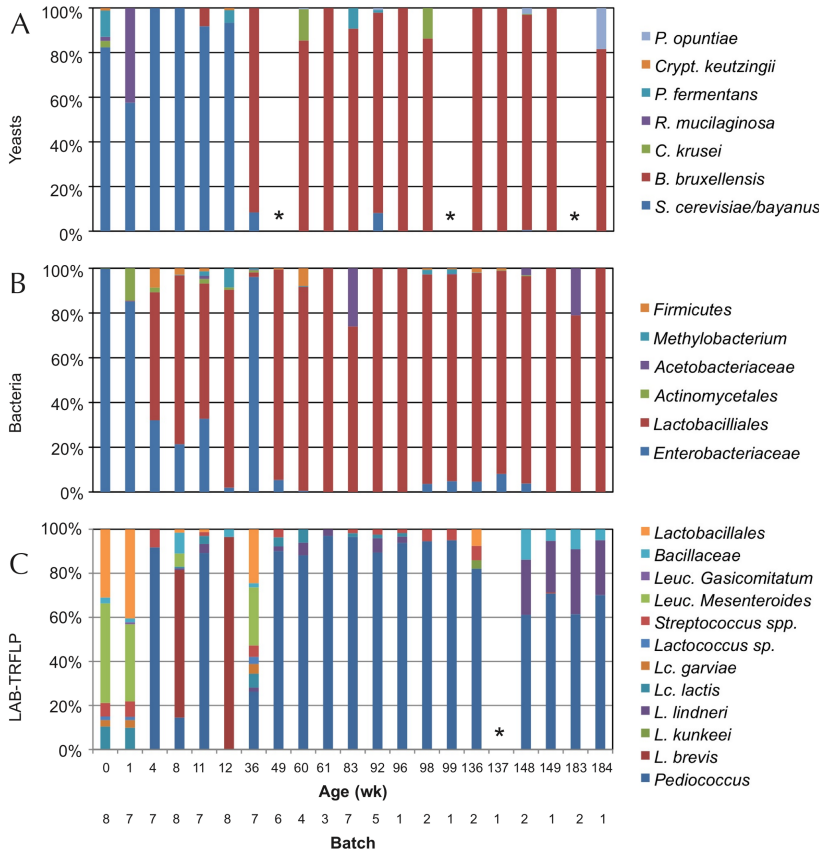




Temporal fluctuations

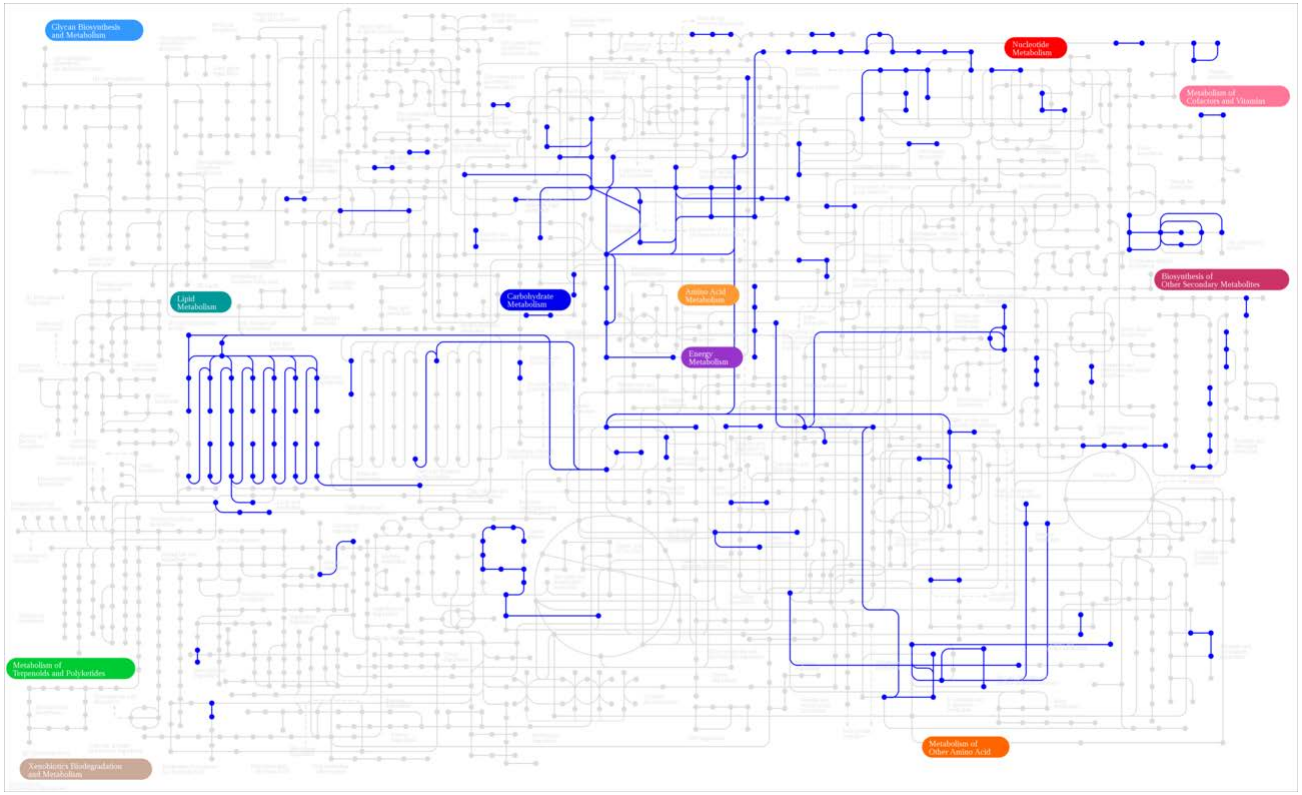
- Communities shown to fluctuate in other habitats too
- Microbiomes of microbrewery has different community profiles at different points in the brewing[20]
- These communities occur naturally with little overlap between timepoints





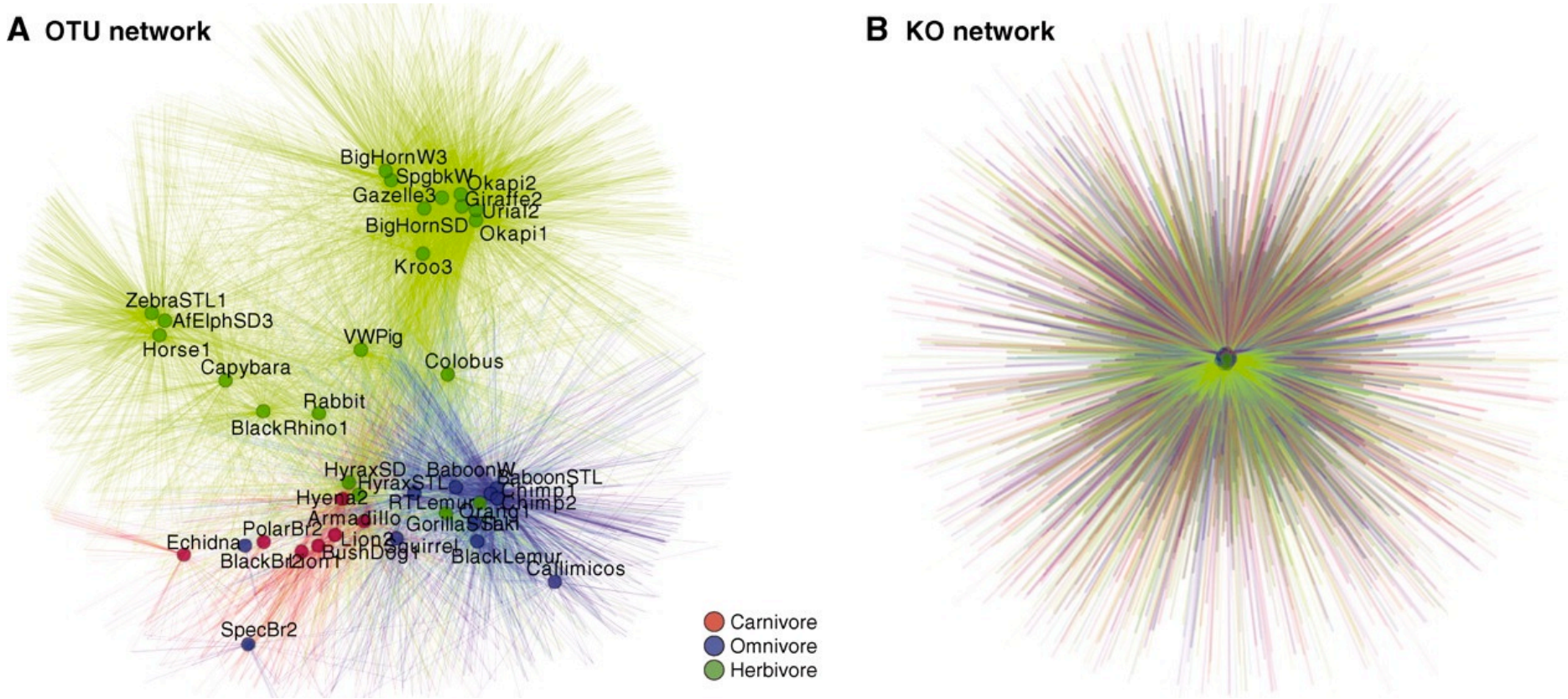
Functional core

- Despite lack of species continuity, functional cores appear to be present[21]
- Varying species undertake similar metabolic pathways
- Allows for transient population without large fluctuations in community function



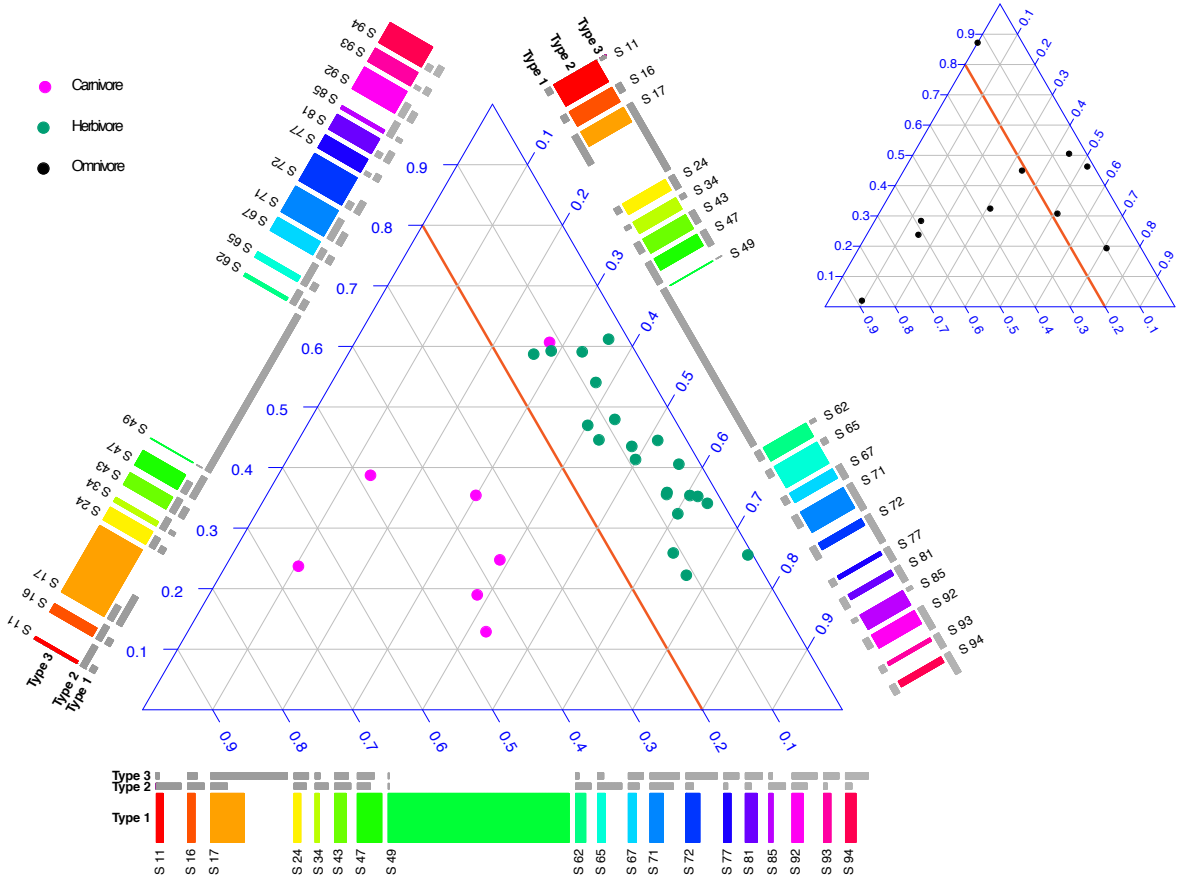
Discriminating by cores

- Many studies segregate samples based on species abundances
  - Fluctuations in species make this unlikely to work over time
- Difficult to find functions that discriminate between samples



Discriminating by cores

- Many studies segregate samples based on species abundances
  - Fluctuations in species make this unlikely to work over time
- Difficult to find functions that discriminate between samples
- However, subsystems may discriminate, not individual functions[21]



Discriminating by cores

- Many studies segregate samples based on species abundances
  - Fluctuations in species make this unlikely to work over time
- Difficult to find functions that discriminate between samples
- However, subsystems may discriminate, not individual functions
- The pathways that form these discriminating ‘metabotypes’ may not be completed within same organism (metabolic handovers) or may be transferred within a habitat (lateral gene transfer) to allow for environmental adaptation

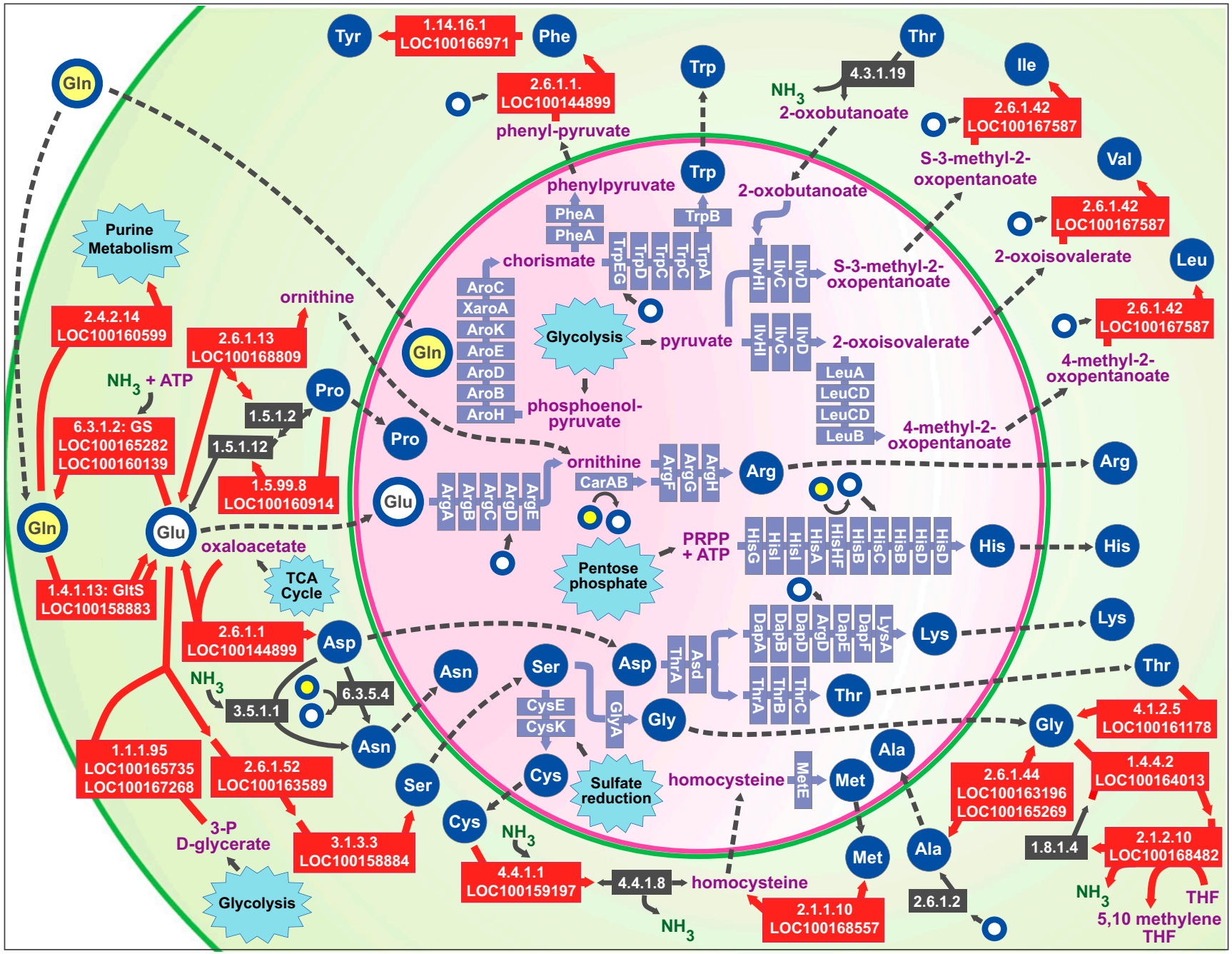
Habitual sharing:
diverse communities and LGT

Levels of diversity

- Microbiomes can vary greatly in the level of diversity
- Simple communities such as the Aphid microbiome
  - Highly integrated relationship between host and a few endosymbionts[22]
- Moderately diverse communities such as KB1
  - A dechlorinating community consisting of 13 species with moderate to high levels of integration[23]
- Highly diverse communities such as the human microbiome
  - Many different species in one habitat[24]
  - Unknown levels of integration

Community integration and evolution

- Microbes interact with each other constantly within microbiome
- Symbiotic relationships arise within such communities
- Symbiosis can turn into dependencies
  - Prochlorococcus reliant on marine community for hydrogen peroxide removal[25]
  - Dehalococcoides reliant on KB1 community for cobalamin and methionine production[23]
  - Black queen hypothesis[25]
- Host/microbe symbiosis often occurs in tightly integrated communities



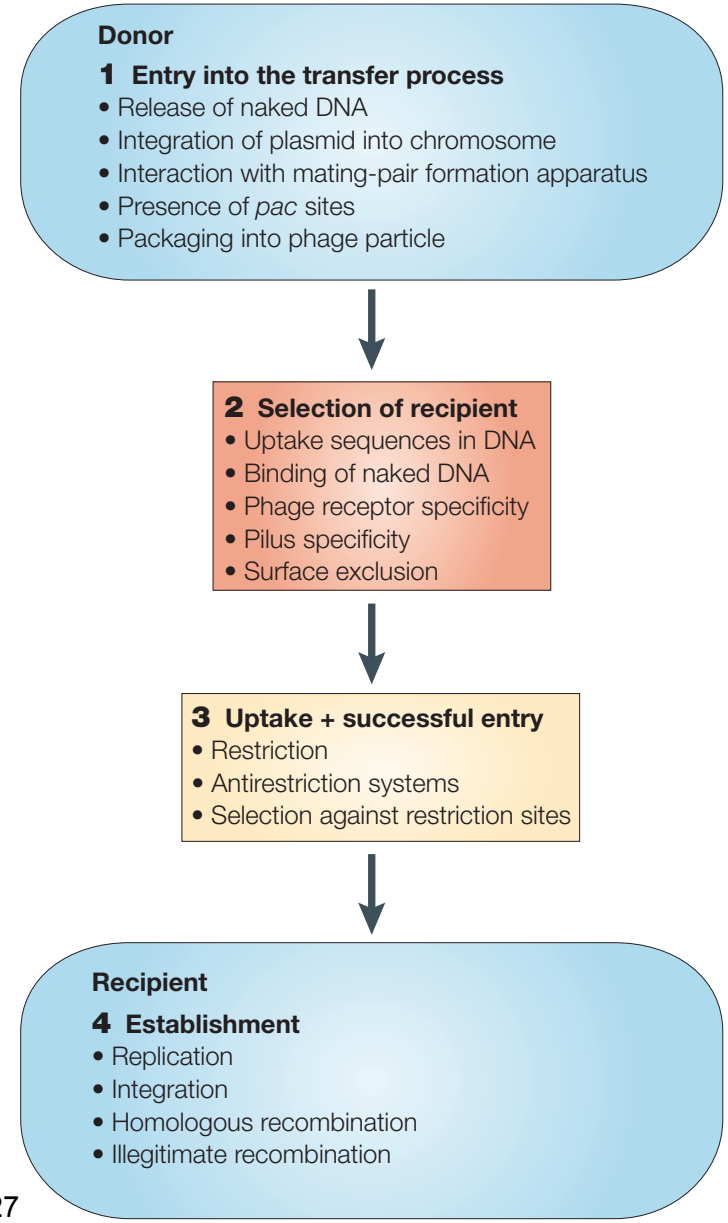
From reference 22

Community integration and evolution

- Microbes interact with each other constantly within microbiome
- Symbiotic relationship arise within such communities
- Symbiosis can turn into dependencies
  - Prochlorococcus reliant on marine community for HOOH removal[25]
  - Dehalococcoides reliant on KB1 community for cobalamin and methionine production[23]
  - Black queen hypothesis[25]
- Host/microbe symbiosis can also occur in tightly integrated communities
- Where does one organism end and the other begin?

Lateral Gene Transfer

- Also called horizontal gene transfer
- First observed between pneumococci in mice[26]
- 3 main ways:
  - Transformation
    - Uptake of naked DNA
    - Often limited to specific environmental cues
    - Estimated ~1% of known species
  - Conjugation
    - Involves the transfer of plasmids
    - Many plasmids are highly promiscuous
  - Transduction
    - Involves an intermediate phage
    - Rampant evidence in nearly all prokaryotic genomes



From ref 27

Studying LGT

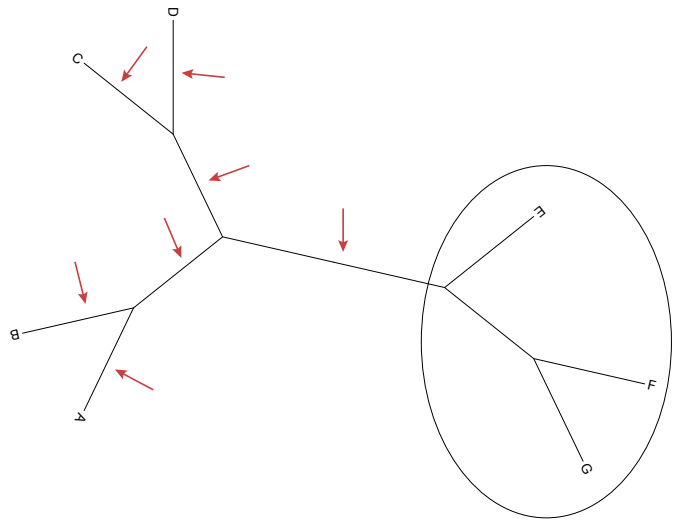
- Detection of LGT is a very difficult problem
- Most common method:
  - Take gene of interest
  - BLAST against a database
  - Build tree of hits and construct phylogeny
  - Find closest relative to gene of interest
  - Is it the taxonomically closest partner?
    - If not, likely LGT
  - Confirm with composition approaches such as codon usage, flanking genetic content, genomic island analysis etc
- How do we determine the likely donor?
  - Cannot root the tree as potential LGT means there is no reliable outgroup
  - Need to deal with unrooted trees and clanistics[28]

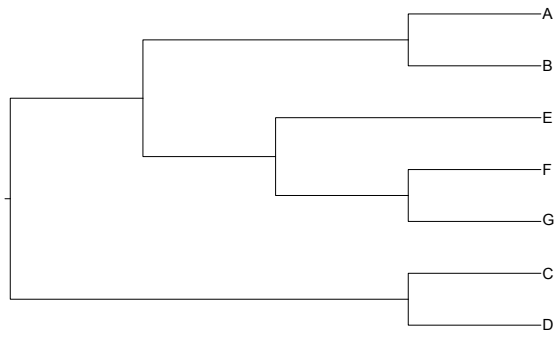
Clanistics

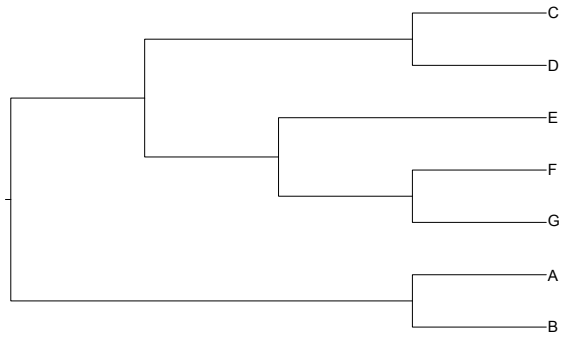
- 2 taxa are more closely related to each other than to others if they have a common ancestor that excludes all others
  - Requires a root
- Clades are groups of taxa that are more closely related to each other than all others
  - Requires a root
- Unrooted trees can tell us likely clades and sister groups but not definitive
- Suggested that unrooted clade equivalents be called ‘clans’ and equivalent sister groups be called ‘adjacent groups’ [27]

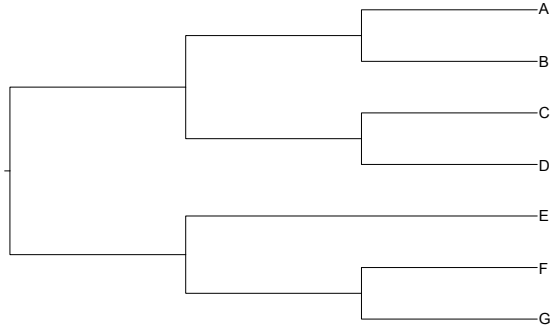
Determining an adjacent group

- Assume E, F, G form a clan
- They form a clade under 7 possible rootings (red arrows)
- There are 3 adjacent groups that would be sister groups in a rooted tree





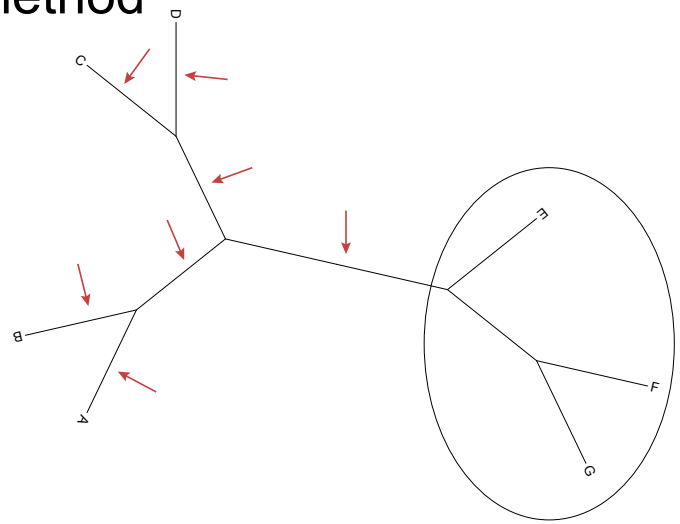




Adapted from ref 27

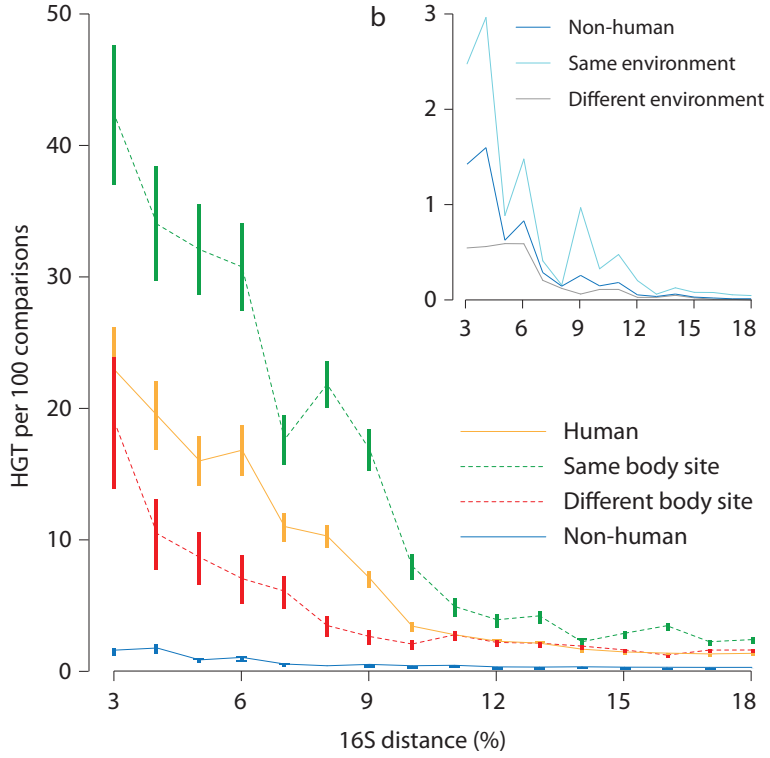
Determining an adjacent group

- Assume E, F, G form a clan
- They form a clade under 7 possible rootings (red arrows)
- There are 3 adjacent groups that would be sister groups in a rooted tree
- For a single gene it is simpler as there are only 2 adjacent groups (For E it would be (F,G) or (A,B,C,D))
- Can use dating information, compositional analysis or maximum distance methods to attempt to resolve
- No sure-fire method



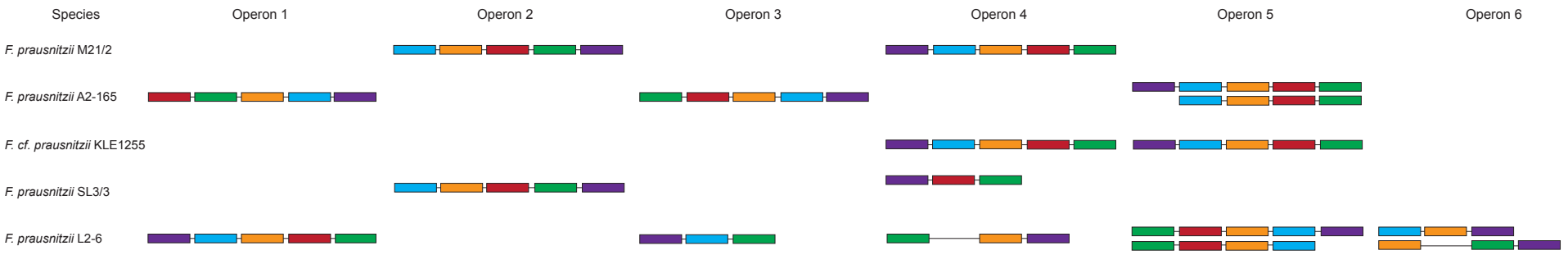
Frequency and Distance of LGT

- Estimated that at least 18% of bacterial genome contents are derived from LGT events[29]
- LGT was found to be rampant in the human microbiome
  - Occurred more often between closely related species[30]



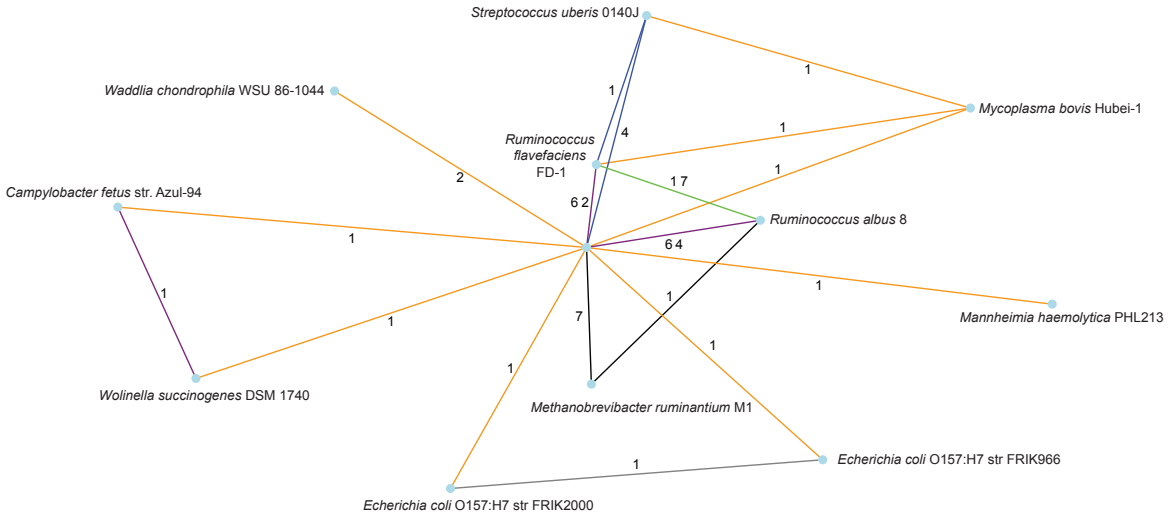
Frequency and Distance of LGT

- Estimated that at least 18% of bacterial genome contents are derived from LGT events[29]
- LGT was found to be rampant in the human microbiome
  - Occurred more often between closely related species[30]
  - Can greatly modify the genomes of related strains[31]



Frequency and Distance of LGT

- Estimated that at least 18% of bacterial genome contents are derived from LGT events[29]
- LGT was found to be rampant in the human microbiome
  - Occurred more often between closely related species[30]
  - Can greatly modify the genomes of related strains[31]
- Can occur between distantly related species
  - Several examples of inter-domain transfers



Green: species
Purple: order
Blue: phylum
Orange: domain
Black: cellular

Frequency and Distance of LGT

- Estimated that at least 18% of bacterial genome contents are derived from LGT events[29]
- LGT was found to be rampant in the human microbiome
  - Occurred more often between closely related species[30]
  - Can greatly modify the genomes of related strains[31]
- Can occur between distantly related species
  - Several examples of inter-domain transfers
- If genes can be identical between evolutionarily distant ‘species’, how do we define the boundaries?

LGT

+

Community evolution

=

New species concepts?

Do we need species?

- Perhaps not
- Useful for clinicians for treatment
- Useful for counting organisms in an environment or relating abundances to changes
- Useful for discussing projects etc.
- “I look at the term species as one arbitrarily given for the sake of convenience to a set of individuals resembling each other” (Darwin, 1859)

Species as clusters

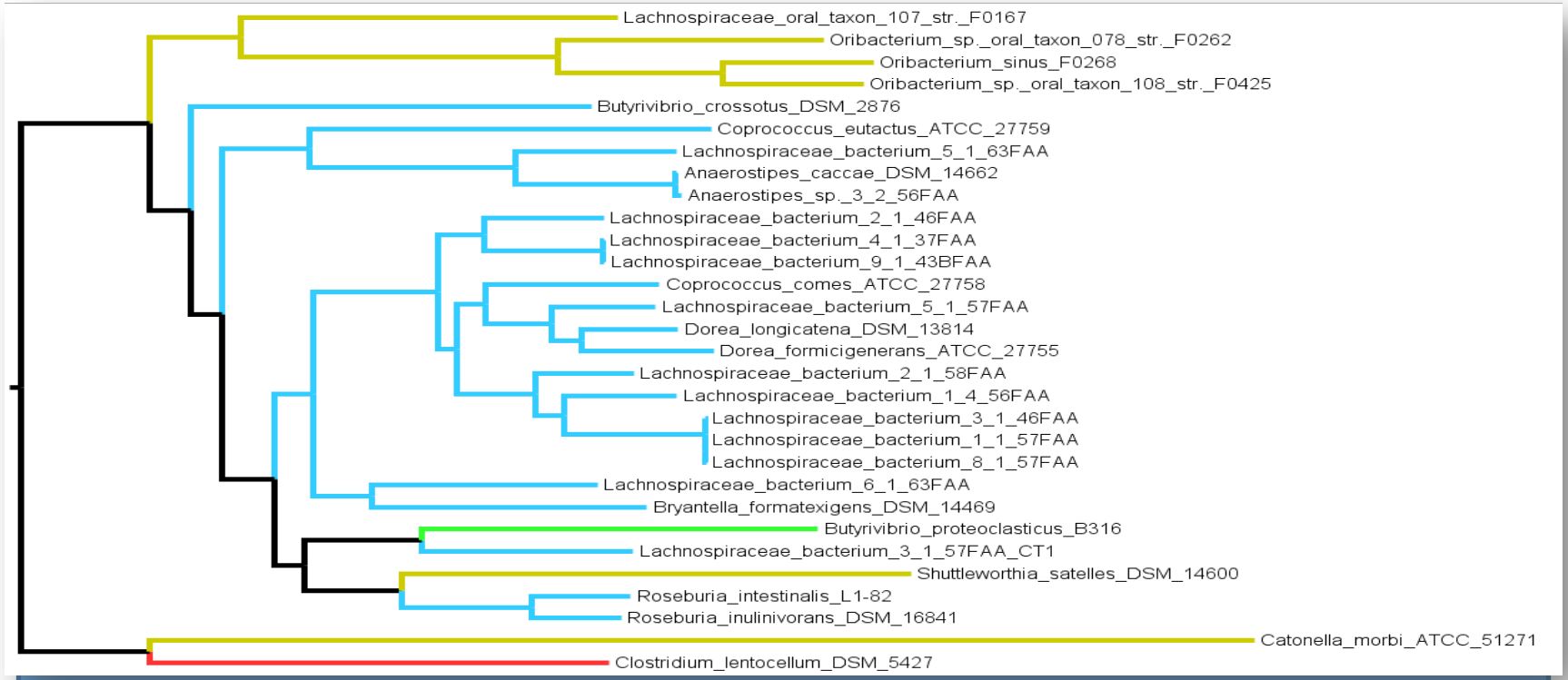
- Species concepts are generally based on the notion that organisms comprise distinct clusters in nature
- This means that there is not a continuum of genotypes and/or phenotypes
- However, clusters can form under random birth/death models[32]
  - A species should presumably be a cluster that is formed by some process, not just random drift
- Any gaps between clusters should not be due to sampling bias or error
  - Not slices from a loaf
  - Probably the biggest problem for proving clustering

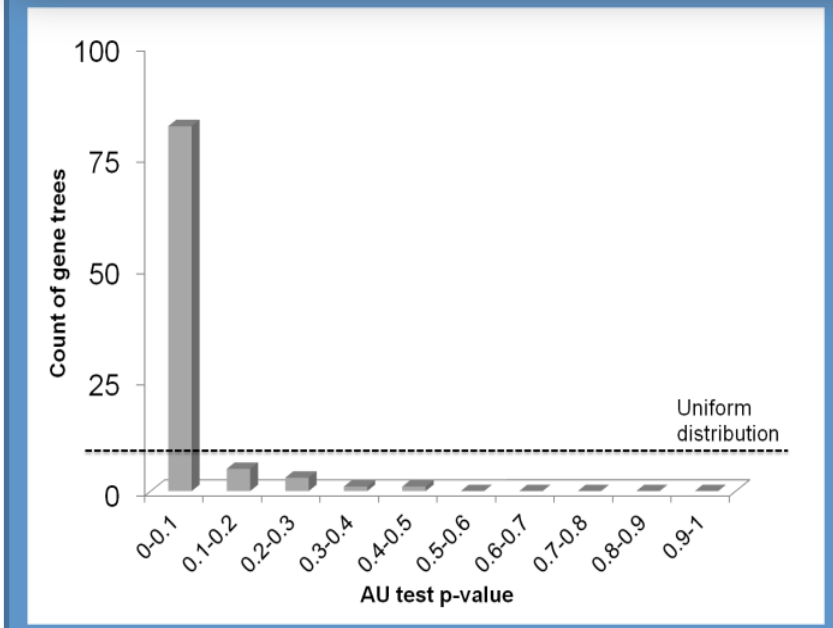
How can we define a species?

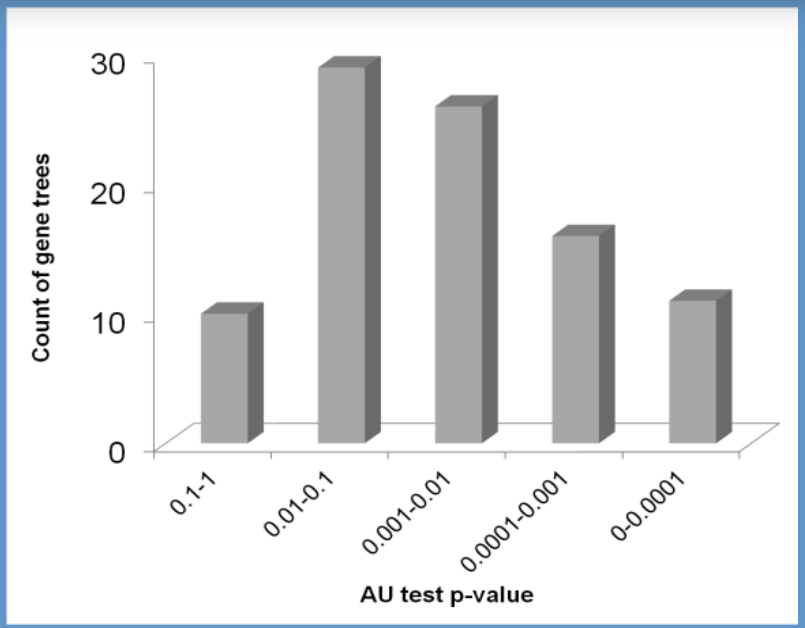
- The Biological Species Concept is the most often used definition of a species
  - States that a species is a group where members can produce fertile offspring through mating[33]
  - Works for (most) animals and plants
  - Excludes all asexual organisms
- Cohans ecotype model[34]
  - States that an asexual clonal species can form by a mutations that allow it to outcompete others and thus selective sweeps occur
  - LGT is allowed in model to initiate a selective sweep but not to shape long-term cohesiveness
    - Recombination has been shown to contribute more to diversification than point mutations in some bacteria

Defining species from genetic data

- In prokaryotes, species were originally defined by >70% in a standardized DNA–DNA hybridisation experiment[35]
  - Makes some bacterial species as diverse as vertebrate orders
- Now, often a species is defined as having within 97% identical 16S sequences between the two organisms[36]
  - This bases the whole organisms classification on one gene's evolutionary history
  - Can have 16S copies that are up to 20% different[37] within the same organism, likely due to LGT[38]
- Can also use shared orthologous genes using:
  - an average nucleotide identity cut-off (usually 95%) [39]
  - concatenated trees
    - Often incongruent with individual gene trees

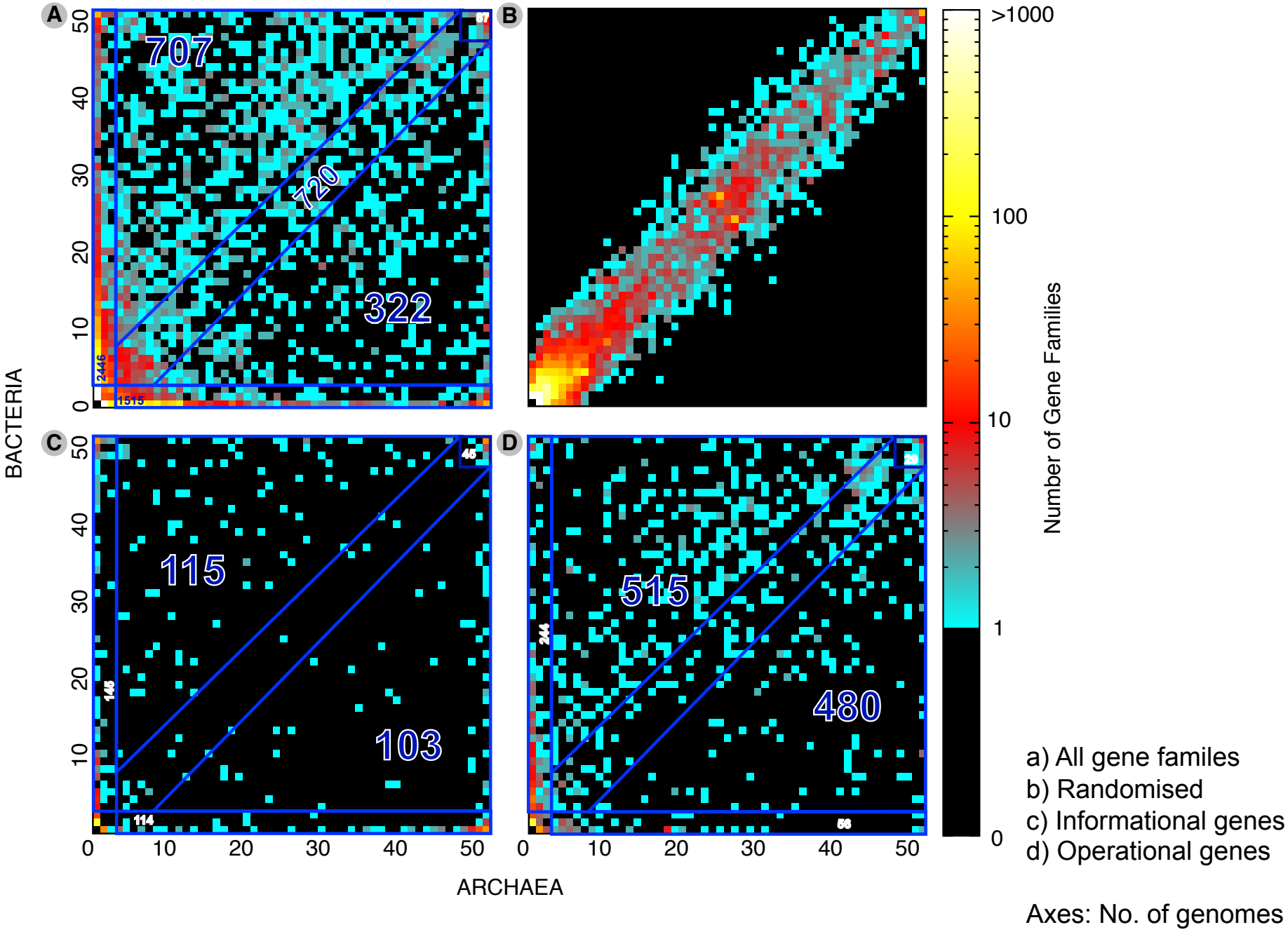

Lachnospiraceae_oral_taxon_107_str._F0167
Oribacterium_sp._oral_taxon_078_str._F0262
Oribacterium_sinus_F0268
Oribacterium_sp._oral_taxon_108_str._F0425
Butyrivibrio_crossotus_DSM_2876
Coprococcus_eutactus_ATCC_27759
Lachnospiraceae_bacterium_5_1_63FAA
Anaerostipes_caccae_DSM_14662
Anaerostipes_sp._3_2_56FAA
Lachnospiraceae_bacterium_2_1_46FAA
Lachnospiraceae_bacterium_4_1_37FAA
Lachnospiraceae_bacterium_9_1_43BFAA
Coprococcus_comes_ATCC_27758
Lachnospiraceae_bacterium_5_1_57FAA
Dorea_longicatena_DSM_13814
Dorea_formicigenerans_ATCC_27755
Lachnospiraceae_bacterium_2_1_58FAA
Lachnospiraceae_bacterium_1_4_56FAA
Lachnospiraceae_bacterium_3_1_46FAA
Lachnospiraceae_bacterium_1_1_57FAA
Lachnospiraceae_bacterium_8_1_57FAA
Lachnospiraceae_bacterium_6_1_63FAA
Bryantella_formatexigens_DSM_14469
Butyrivibrio_proteoclasticus_B316
Lachnospiraceae_bacterium_3_1_57FAA_CT1
Shuttleworthia_satelles_DSM_14600
Roseburia_intestinalis_L1-82
Roseburia_inulinivorans_DSM_16841
Catonella_morbi_ATCC_51271
Clostridium_lentocellum_DSM_5427


100
75
50
25
0
Count of gene trees
Uniform distribution
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
AU test p-value


30
20
10
0
Count of gene trees
0.1-1
0.01-0.1
0.001-0.01
0.0001-0.001
0-0.0001
AU test p-value

Defining species from genetic data

- In prokaryotes species were originally defined by >70% in a standardized DNA–DNA hybridisation experiment[35]
  - Makes some bacterial species as diverse as vertebrate orders
- Now, often a species in a prokaryote is defined as having within 97% identical 16S sequences between the two organisms[36]
  - This bases the whole organisms classification on one gene's evolutionary history
  - Can have 16S copies that are up to 20% different[37] within the same organism, likely due to LGT[38]
- Can also use shared orthologous genes using:
  - an average nucleotide identity cut-off (usually 95%) [39]
  - concatenated trees
    - Often incongruent with individual gene trees
  - both disregard the variable part of genome
    - Can have *E. coli* strains that differ by ~50% of genome but same 'species'[40]


A
B
C
D
707
720
322
115
103
515
480
2446
1515
97
45
145
114
244
29
56
BACTERIA
ARCHAEA
0
10
20
30
40
50
>1000
100
10
1
0
Number of Gene Families
a) All gene familes
b) Randomised
c) Informational genes
d) Operational genes
Axes: No. of genomes

From ref 41

Microbiomes and Species

- We cannot ask the simple question ‘Who is there?’ without defining species (the who)
- Metagenome data has allowed us to somewhat overcome the sampling bias
  - Can now see the minor populations
  - Can observe if there is more of a continuum
- Studying prokaryotes as a community has raised many extra questions
  - Do they evolve as a community?
    - Community evolution often thought of just as sum of individual evolutionary paths
  - What role do host-associated microbiomes play in their evolution/ speciation?
- Can begin to ask what a unit of diversity is
  - Are the clusters we see distinguishable from random birth/death models?
  - Are microbial ecotypes and communities species or what is the base unit of diversity?

- 1 http://www.qiime.org/
- 2 http://kiwi.cs.dal.ca/Software/RITA
- 3 http://matsen.fhcrc.org/pplacer/
- 4 http://en.wikipedia.org/wiki/16S_ribosomal_RNA
- 5 http://www.pnas.org/content/87/12/4576.full.pdf
- 6 http://sourceforge.net/projects/picrust/
- 7 http://www.annualreviews.org/doi/abs/10.1146/annurev.mi.31.100177.000543
- 8 http://www.nature.com/nature/journal/v464/n7290/full/nature08937.html
- 9 http://www.pnas.org/content/108/38/16050
- 10 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028991
- 11 http://www.pnas.org/content/106/40/17187.full
- 12 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0008422
- 13 http://www.sciencemag.org/content/326/5960/1694.full
- 14 http://kiwi.cs.dal.ca/GenGIS/Main_Page
- 15 http://www.ajcn.org/content/69/5/1035s.full.html
- 16 http://www.pnas.org/content/108/suppl.1/4578.full
- 17 http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0050177
- 18 http://www.nature.com/nrmicro/journal/v10/n5/full/nrmicro2746.html
- 19 http://genomebiology.com/2011/12/5/R50/
- 20 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0035507
- 21 M. Shafiei, K. Dunn, H. Chipman, H. Gu, J. Bielawski, "Bayesian Inference of Metabolic Divergence among Microbial Communities", under review
- 22 http://www.pnas.org/content/early/2011/01/25/1013465108
- 23 L.A. Hug, R.G. Beiko, A.R. Rowe, R.E. Richardson, E.A. Edwards, "Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community", BMC Genomics (in press)
- 24 http://www.nature.com/nature/journal/v464/n7285/full/nature08821.html
- 25 http://mbio.asm.org/content/3/2/e00036-12
- 26 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2167760/
- 27 http://www.nature.com/nrmicro/journal/v3/n9/full/nrmicro1234.html
- 28 http://www.sciencedirect.com/science/article/pii/S0169534707000195
- 29 http://www.pnas.org/content/95/16/9413.full
- 30 http://www.nature.com/nature/journal/v480/n7376/fig_tab/nature10571_F2.html
- 31 http://www.biomedcentral.com/1471-2180/12/248
- 32 http://www.cell.com/trends/genetics/abstract/S0168-9525(04)00042-3
- 33 "Systematics and the Origin of Species from the Viewpoint of a Zoologist", Ernyst Mayr, 1942
- 34 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1764936/
- 35 http://ijsb.sgmjournals.org/content/37/4/463.full
- 36 http://ijs.sgmjournals.org/content/44/4/846.full.pdf
- 37 http://aem.asm.org/content/76/12/3886.full
- 38 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC296235/
- 39 http://ijsb.sgmjournals.org/content/57/1/81.full.pdf
- 40 http://jb.asm.org/content/190/20/6881.long
- 41 What Is a Prokaryote? W. Ford Doolittle and Olga Zhaxybayeva,in E. Rosenberg et al. (eds.), The Prokaryotes, Springer-Verlag Berlin Heidelberg 2012