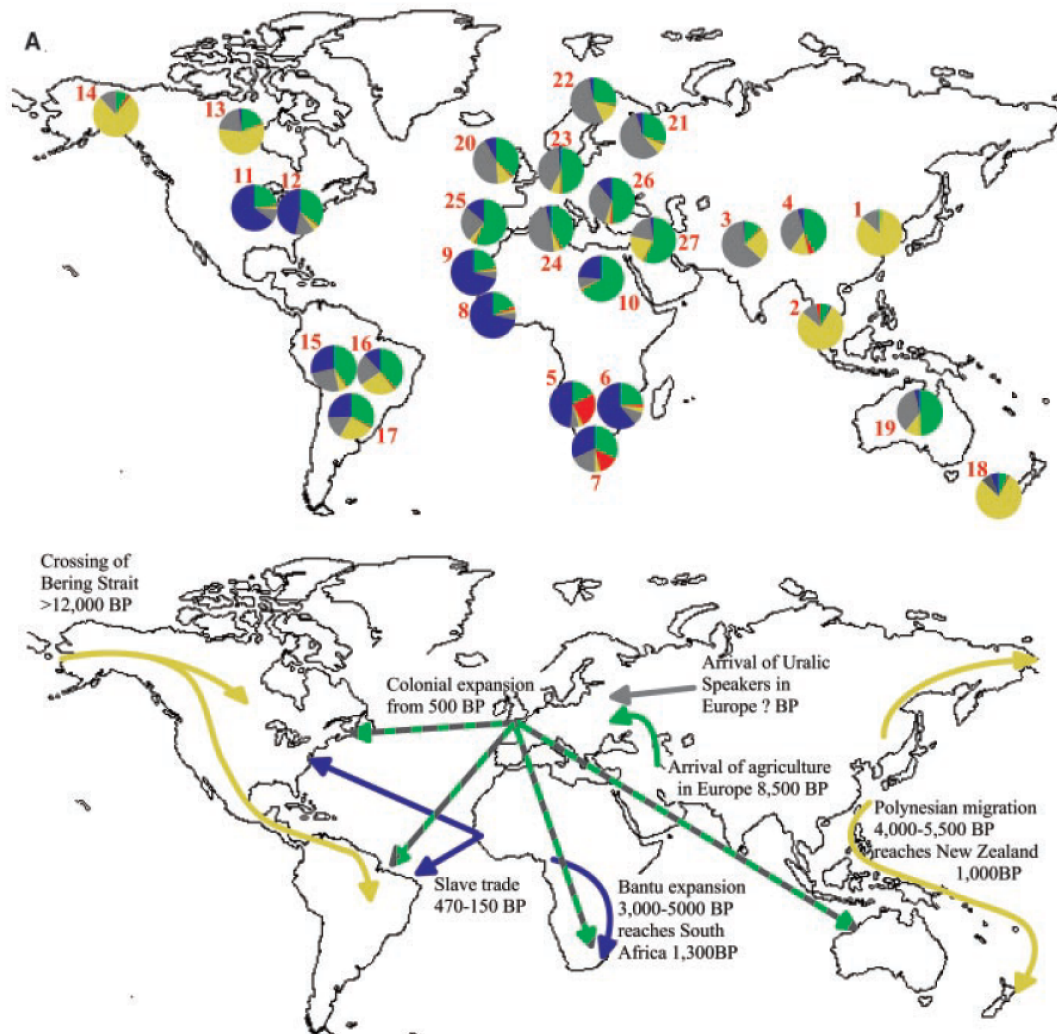


- Part I: A method to estimate the abundance, gene content, and phylogeny of bacterial species from metagenomic data
- Part II: Uncovering the gene content and biogeography of microbial species from the global human gut microbiome
- Part III: A web server for rapid annotation and quantitative analysis of specific genes in the human gut microbiome

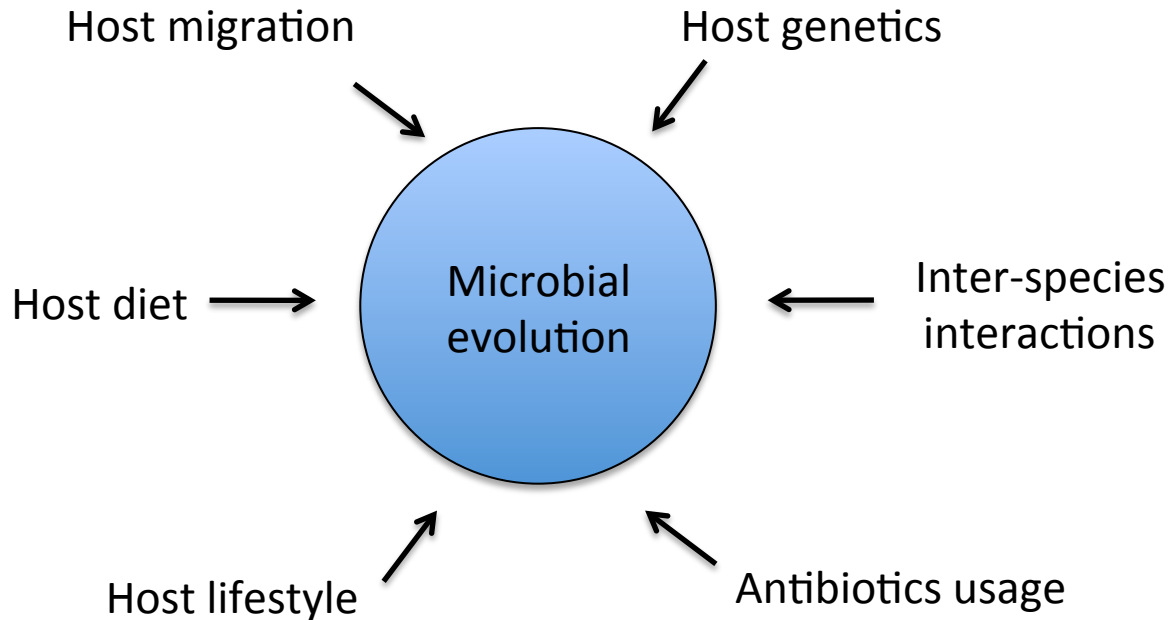
Motivation

- Large differences in the gene content and ecology of closely related bacteria (i.e. strain-level variation)
- How does this vary between individuals? Do these differences correspond with host geography?
- Are these differences driven by historical events (e.g. human migration) or environment (e.g. diet).
- Can we leverage existing shotgun metagenomes to address these questions?

Traces of Human Migrations in *Helicobacter pylori* Populations



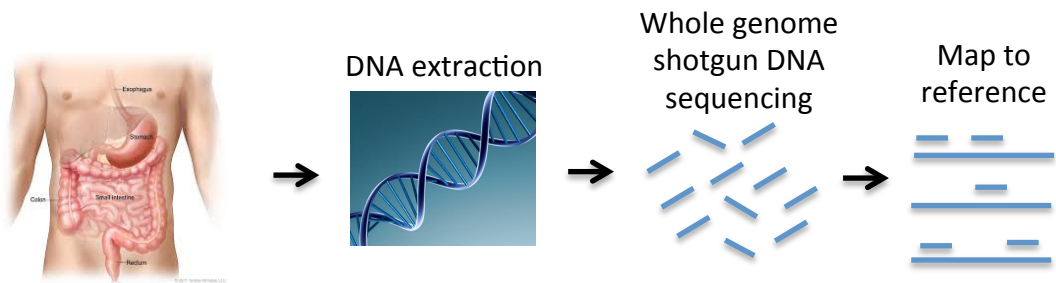
Potential forces shaping recent evolution in the human microbiome



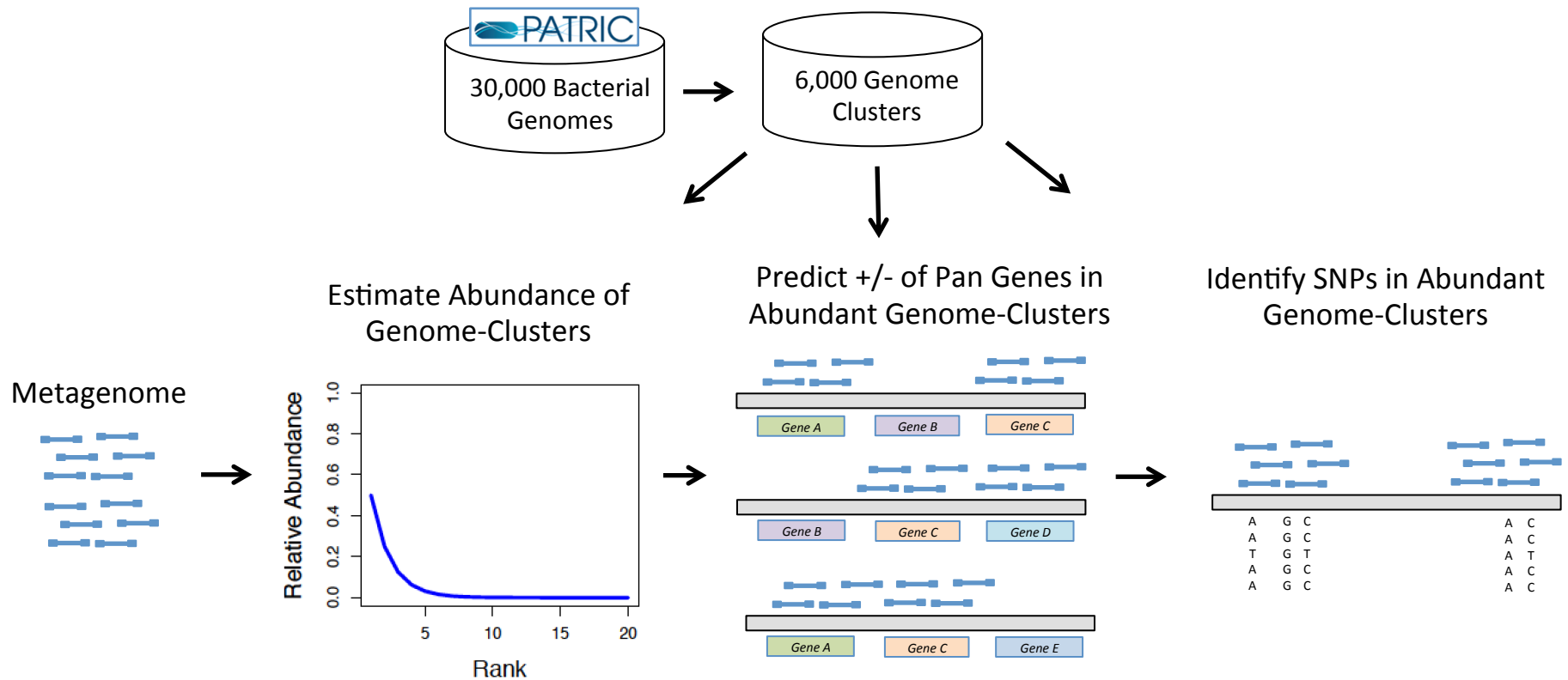
Shotgun metagenomes are a source of strain-level diversity in the human gut microbiome

- Growing number of publicly available, deeply sequenced metagenomes
- Span human populations, phenotypes, gender, age
- Central access to data via the NCBI sequence read archive

Study	Samples	Continent	Country
HMP 2012	337	North America	United States
Qin 2012	359	Asia	China
Le Chatelier 2012	292	Europe	Denmark
Karlsson 2013	147	Europe	Sweden
Qin 2010	124	Europe	Spain, Denmark
Obregon-Tito 2015	57	South America	Peru
Rampelli 2015	27	Africa	Tanzania

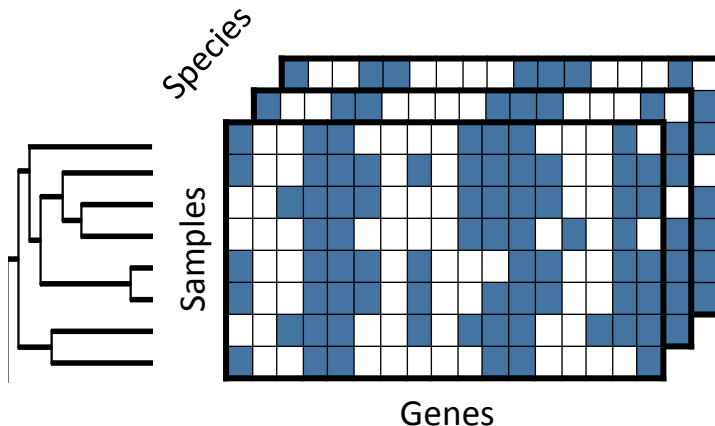


PhyloCNV: estimating the abundance, gene content, and phylogeny of microbial species from shotgun metagenomes



PhyloCNV: estimating the abundance, gene content, and phylogeny of microbial species from shotgun metagenomes

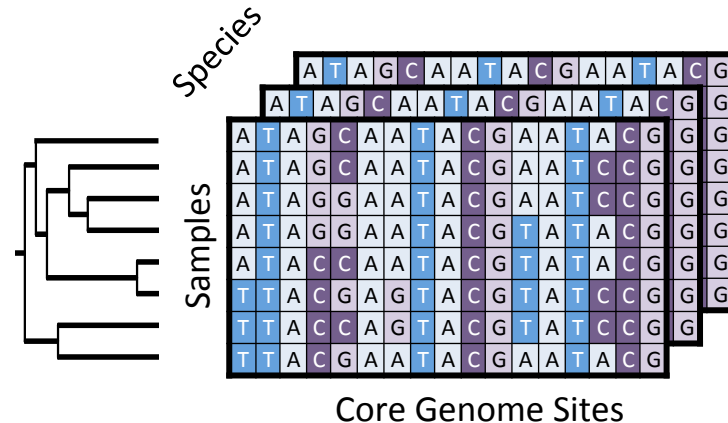
Predict +/- of Genes in
Abundant Bacterial Species



Strain-Level Functional Profiles

- Presence/absence, or
- Copy number

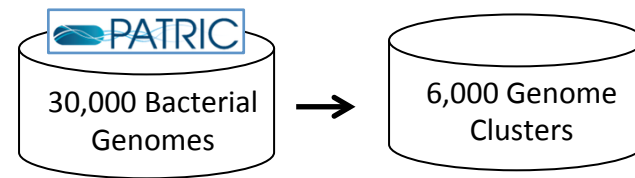
Identify SNPs in Abundant
Bacterial Species



Strain-Level Phylogenies

- Consensus alleles, or
- Allele frequencies

A database of genome-clusters based on a consistent definition of bacterial species

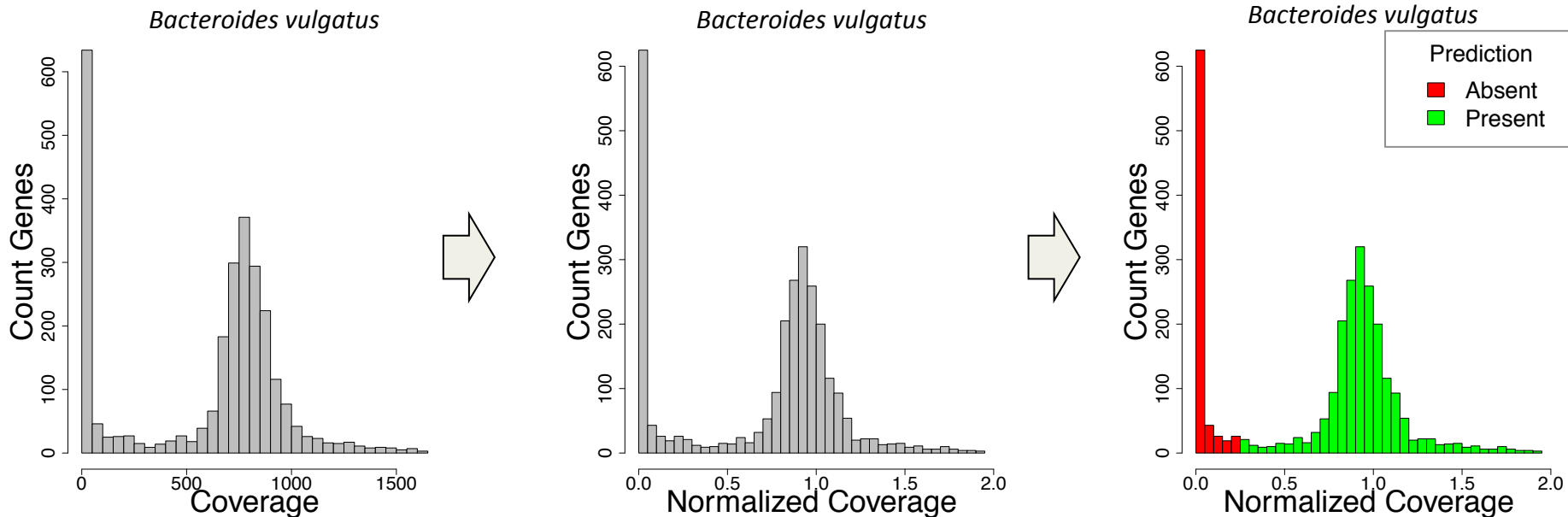


- Based on % ID across 30 universal single-copy genes
- Reflects 95% ANI cutoff commonly used for defining species
- Enabled classification of 2,428 unannotated genomes
- Classified 1,400 genomes differently from NCBI
- Split 579 NCBI species into two or more groups
- Top heavy: top 10 clusters represent >50% of total genomes

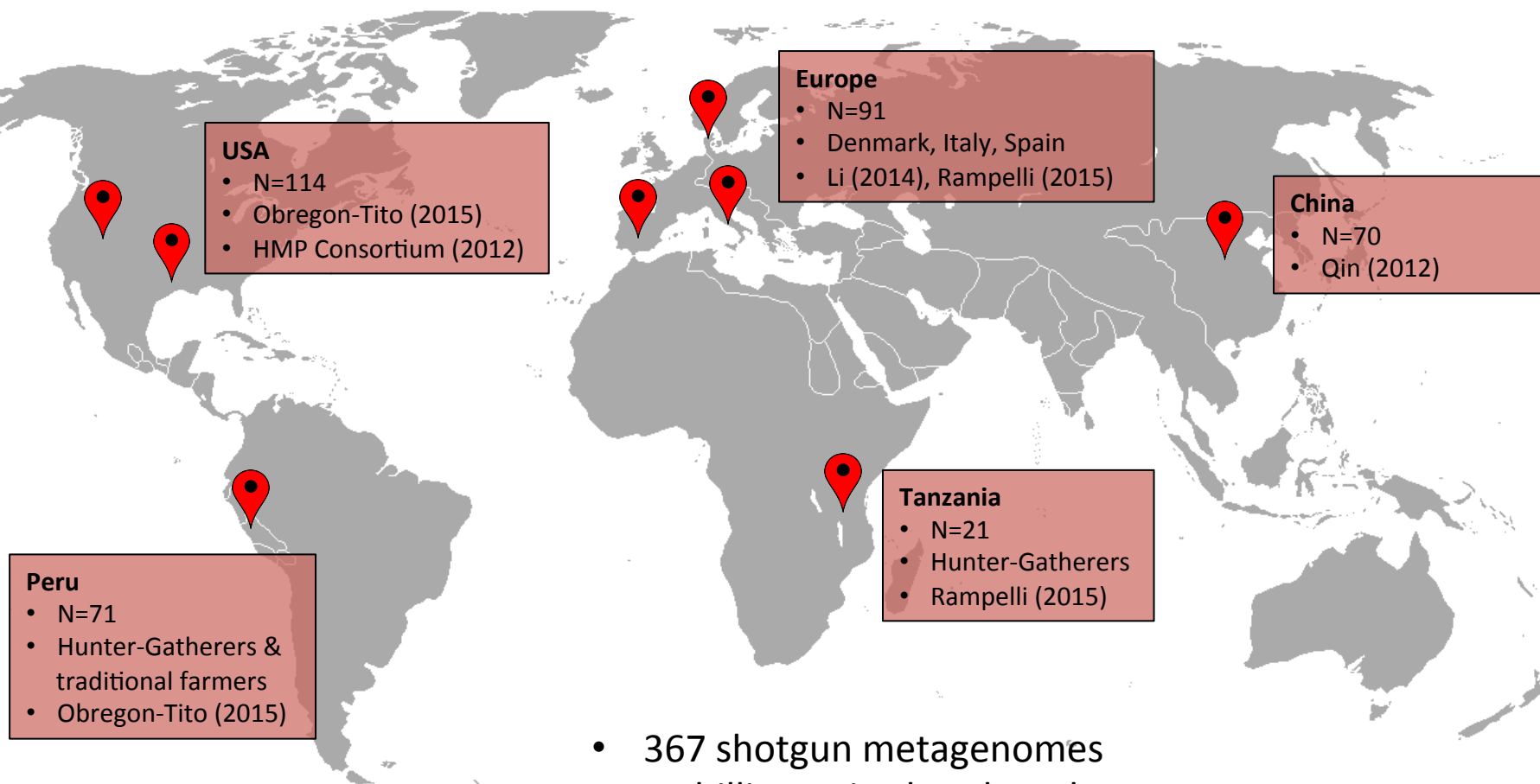
cluster_id	consensus_name	genome_name
56463	Eggerthella lenta	Eggerthella lenta 1_1_60AFAA
56463	Eggerthella lenta	Eggerthella lenta DSM 2243
56463	Eggerthella lenta	Eggerthella sp. 1_3_56FAA
56463	Eggerthella lenta	Eggerthella sp. HGA1
61662	Eggerthella sp.	Eggerthella sp. YY7918
62698	Eggerthia catenaformis	Eggerthia catenaformis OT 569 = DSM 20559

Estimation of Gene +/- From Shotgun Metagenomes

- Map reads to clusters of genomes
- Compute coverage of pan-genes (non-redundant genes)
- Normalize coverage using universal single-copy genes
- Apply cutoff to predict gene +/-

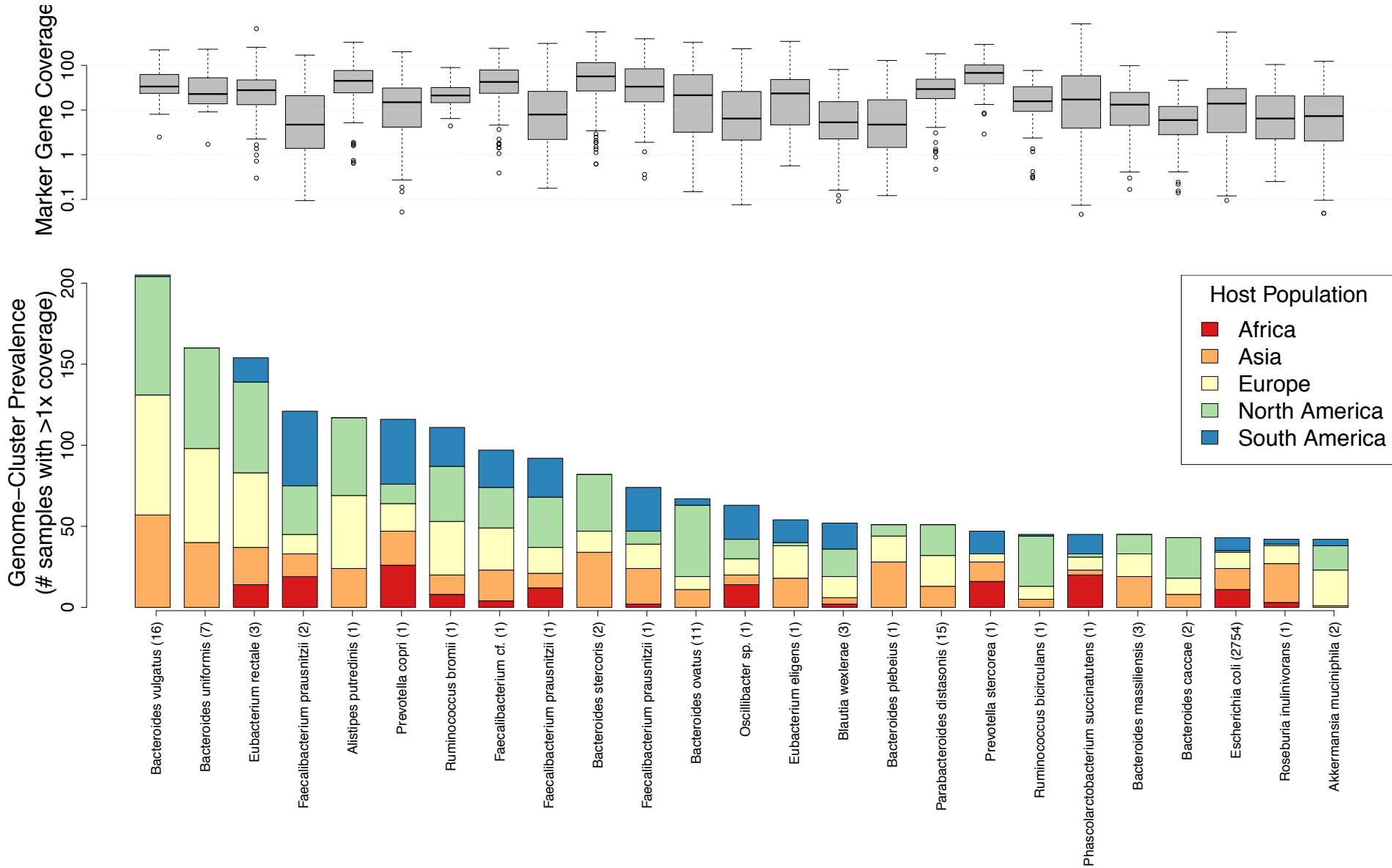


Faecal Metagenomes from Healthy, Unrelated Individuals from 5 Continents

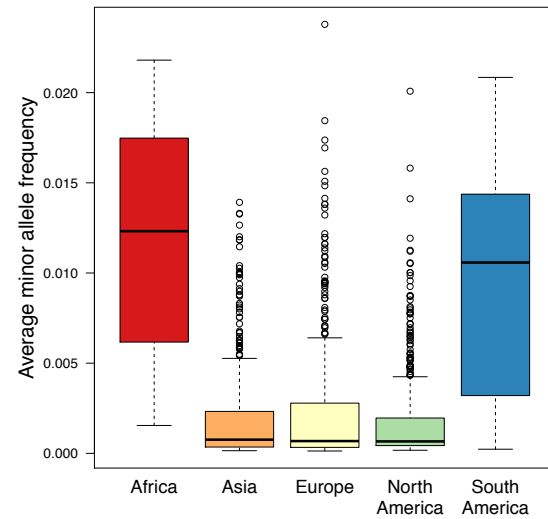
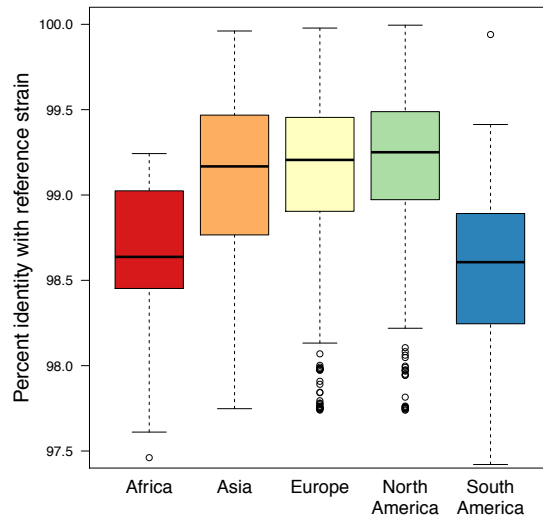
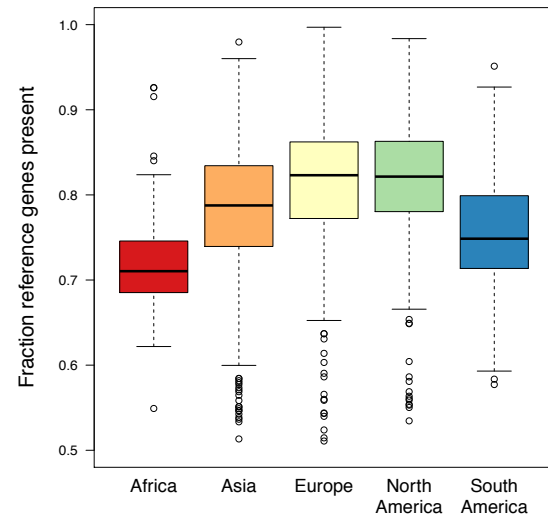
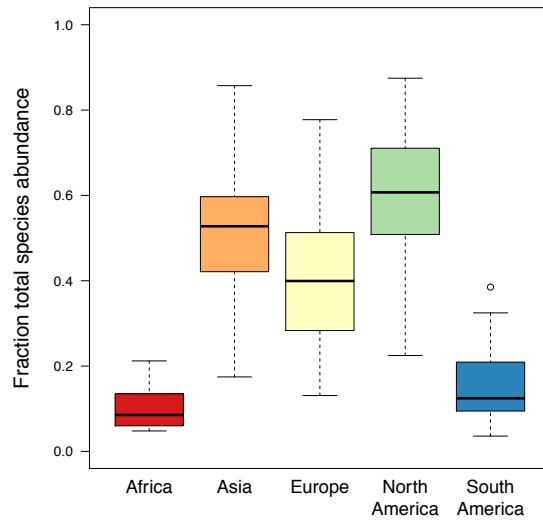


- 367 shotgun metagenomes
- ~7 billion paired-end reads
- ~1.4 tera base-pairs

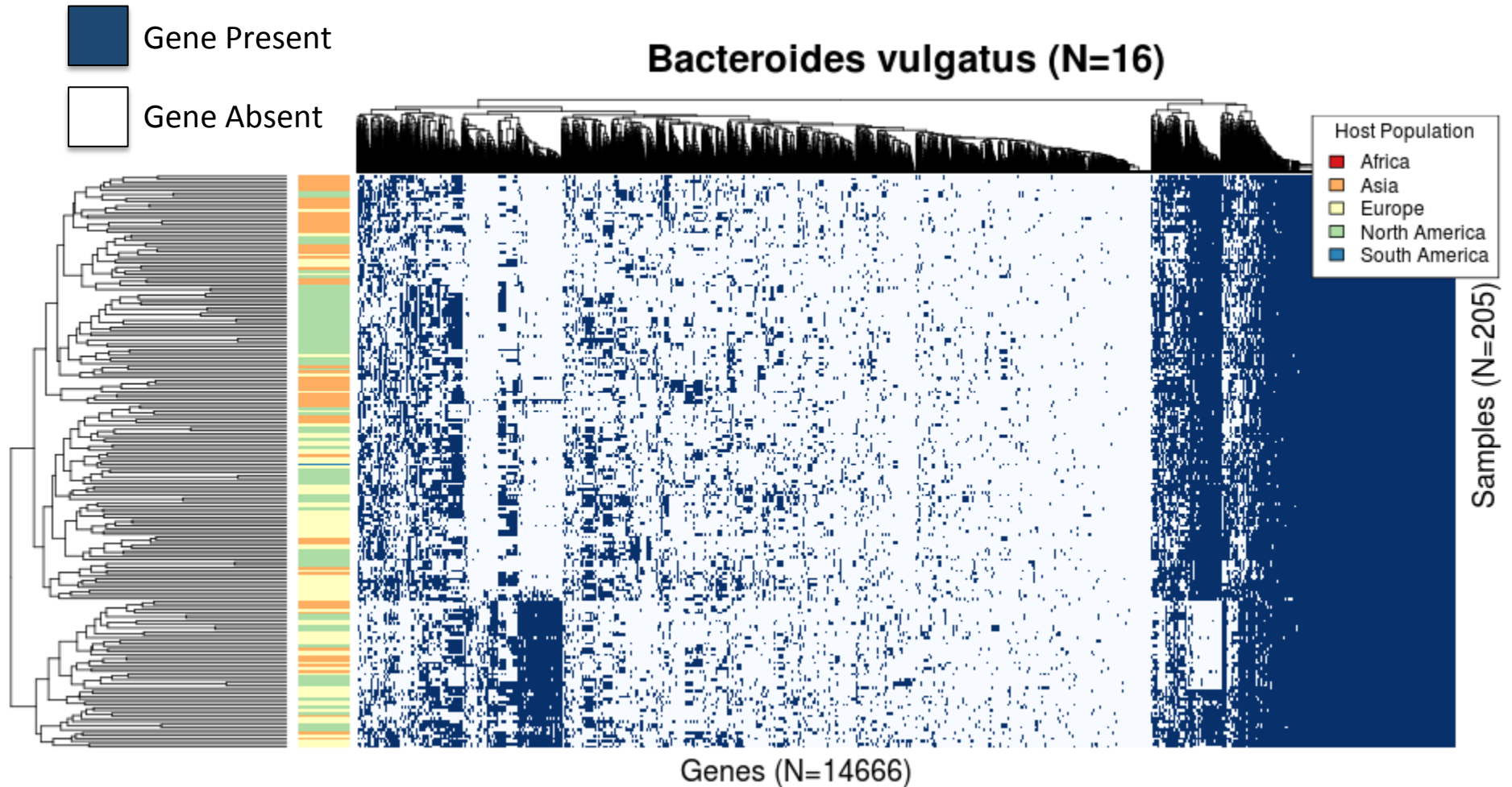
Identification of prevalent genome-clusters that span multiple host populations



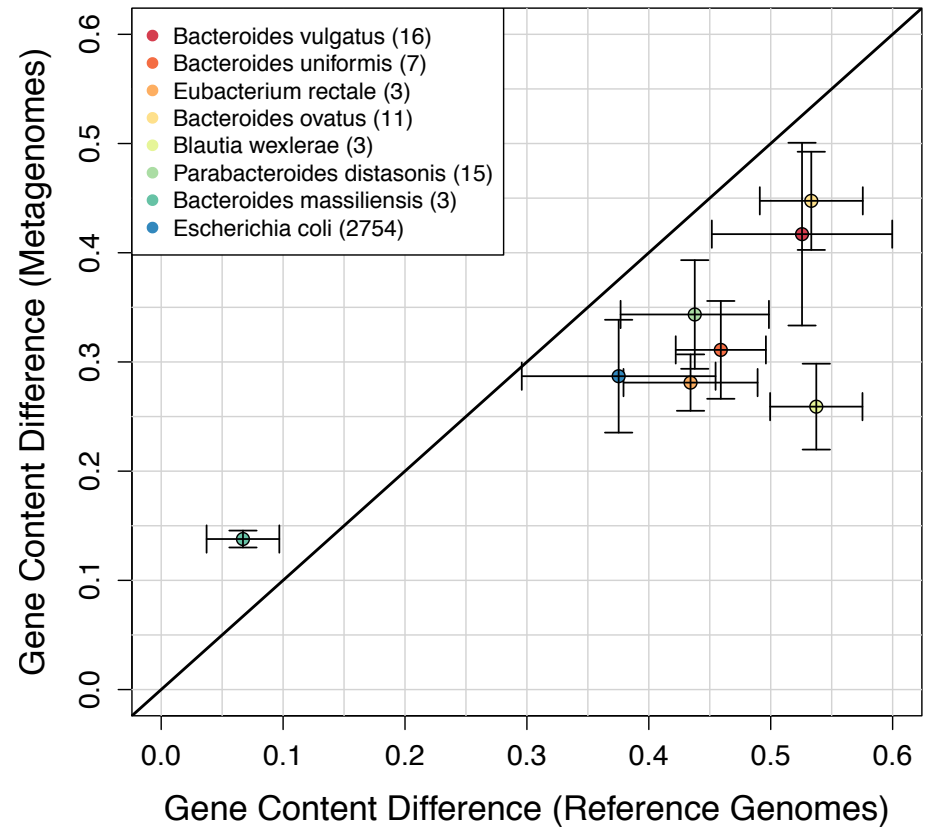
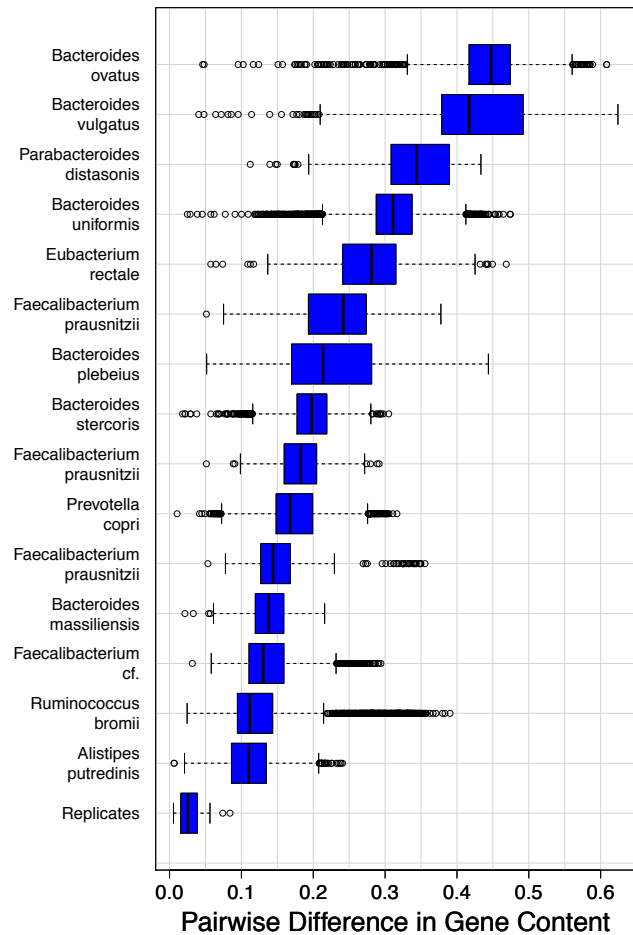
Summary of Differences Across Continents



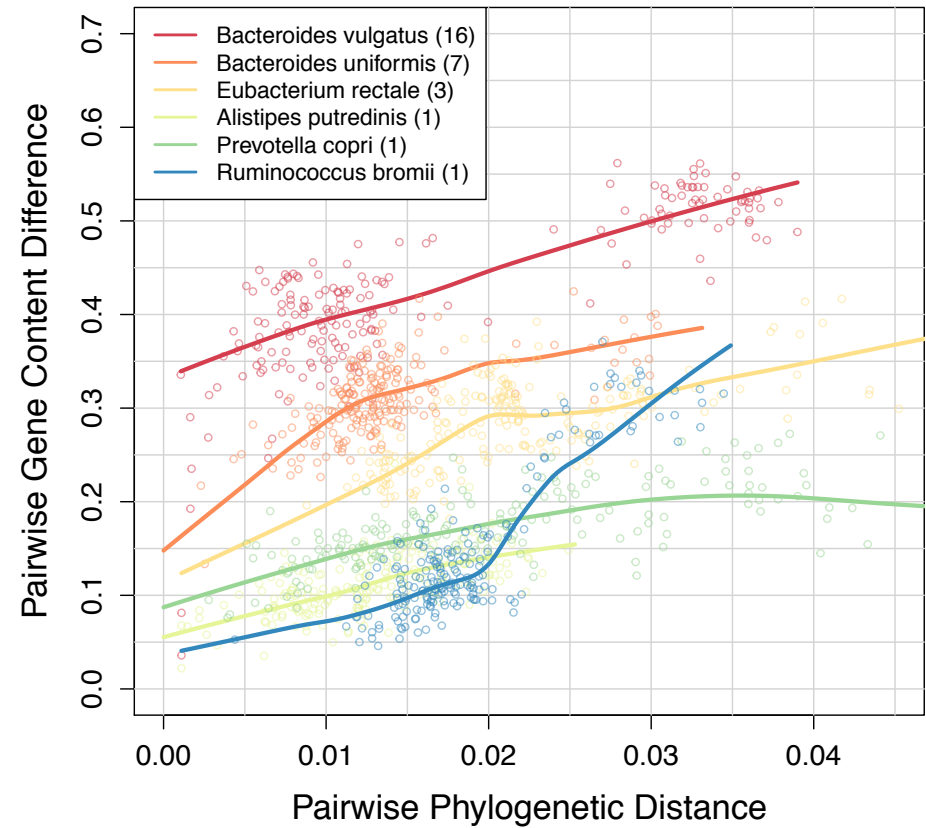
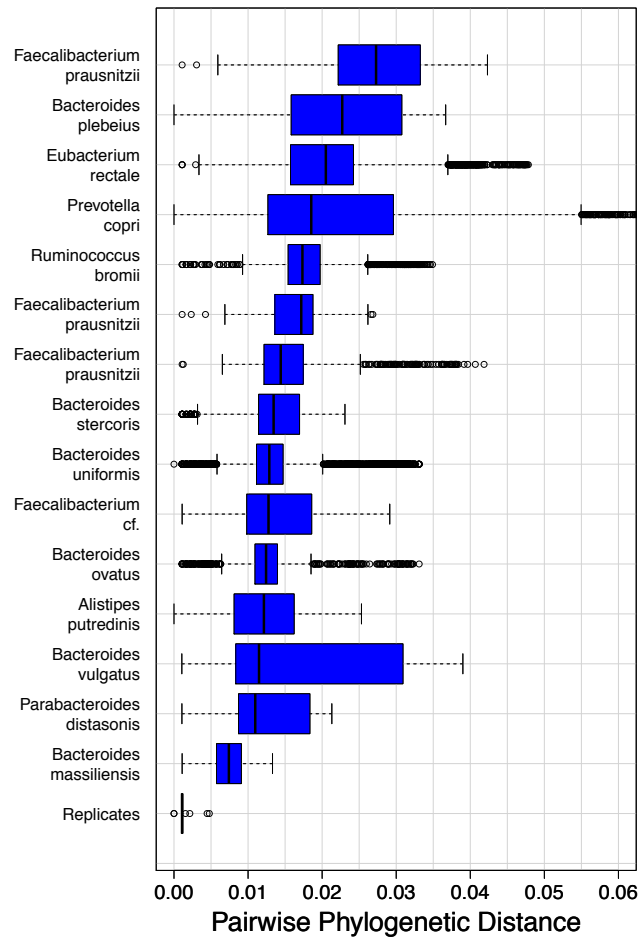
Extensive inter-sample gene content variability of *Bacteroides vulgatus*



Extensive Inter-Sample Differences in the Gene Content of Bacterial Species

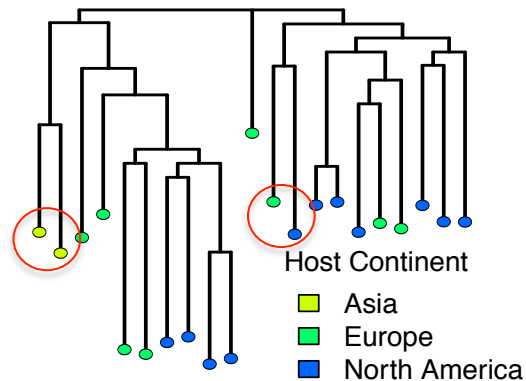


Inter-Sample Differences in the Phylogeny of Bacterial Species

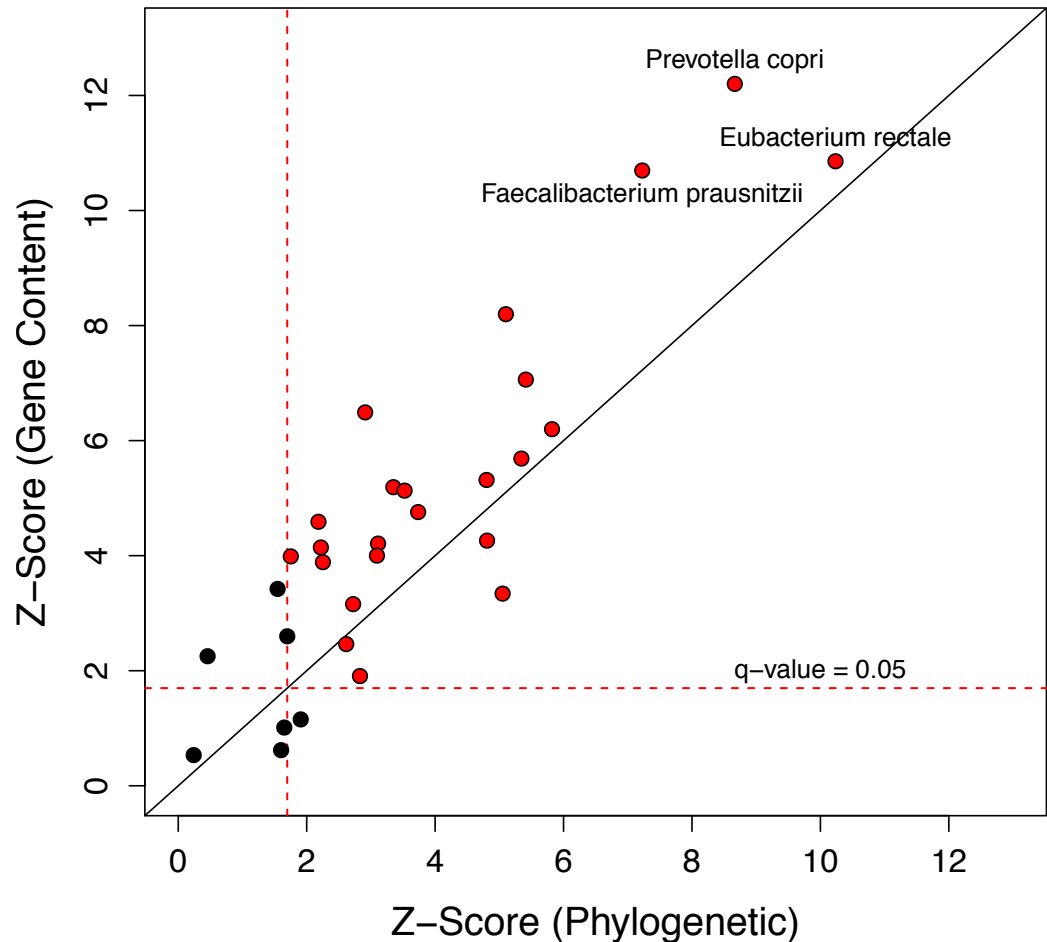
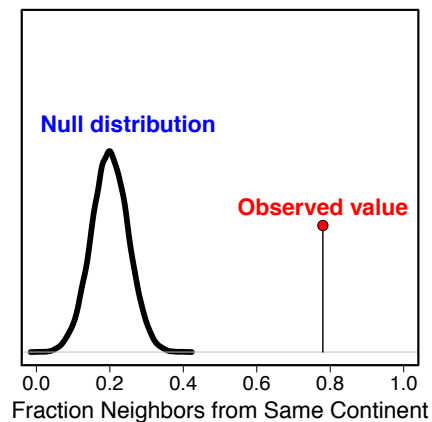


Gene content and phylogeny reveals widespread biogeography of bacterial species

Compute % of nearest-neighbors from same continent

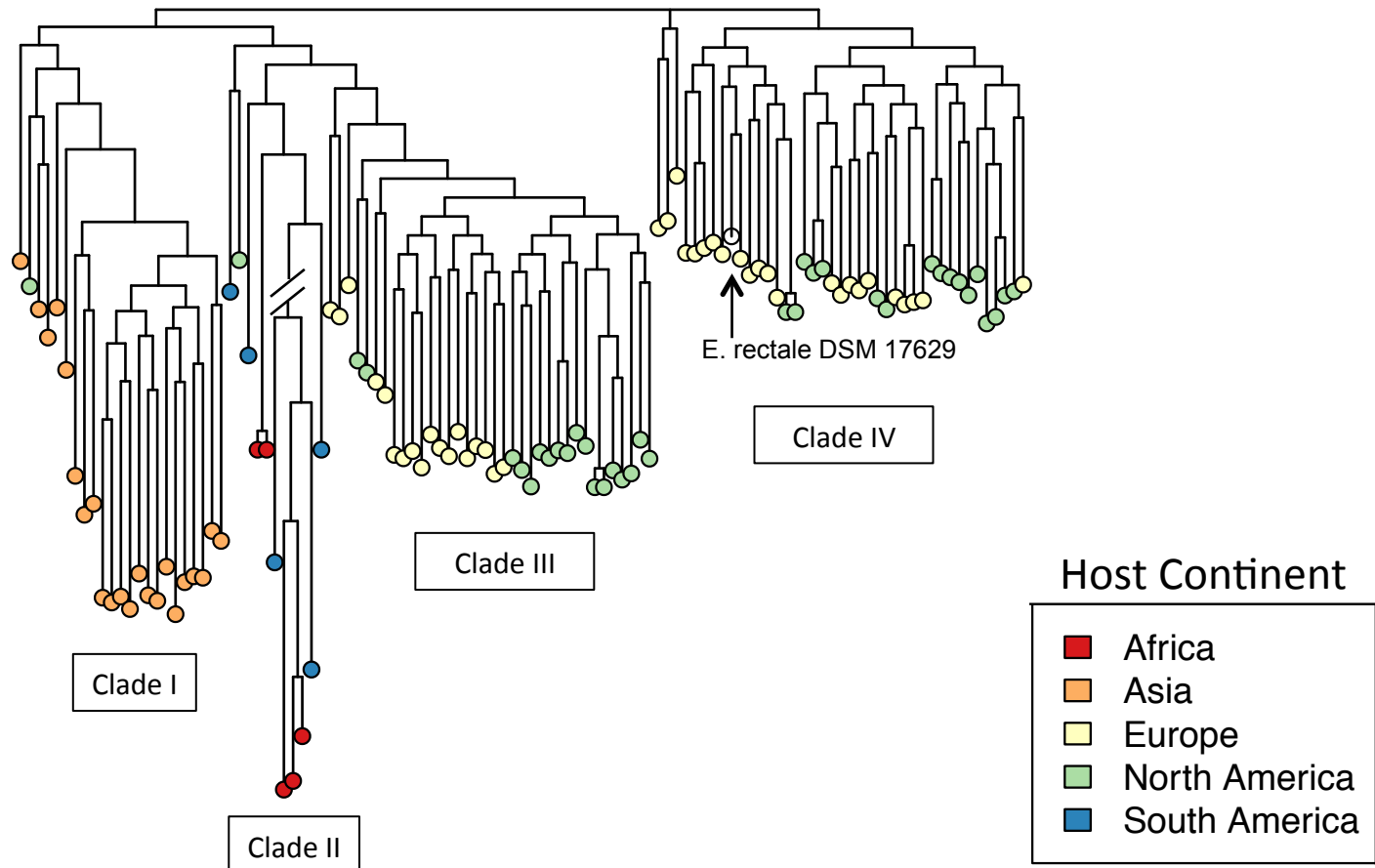


Compare to null distribution

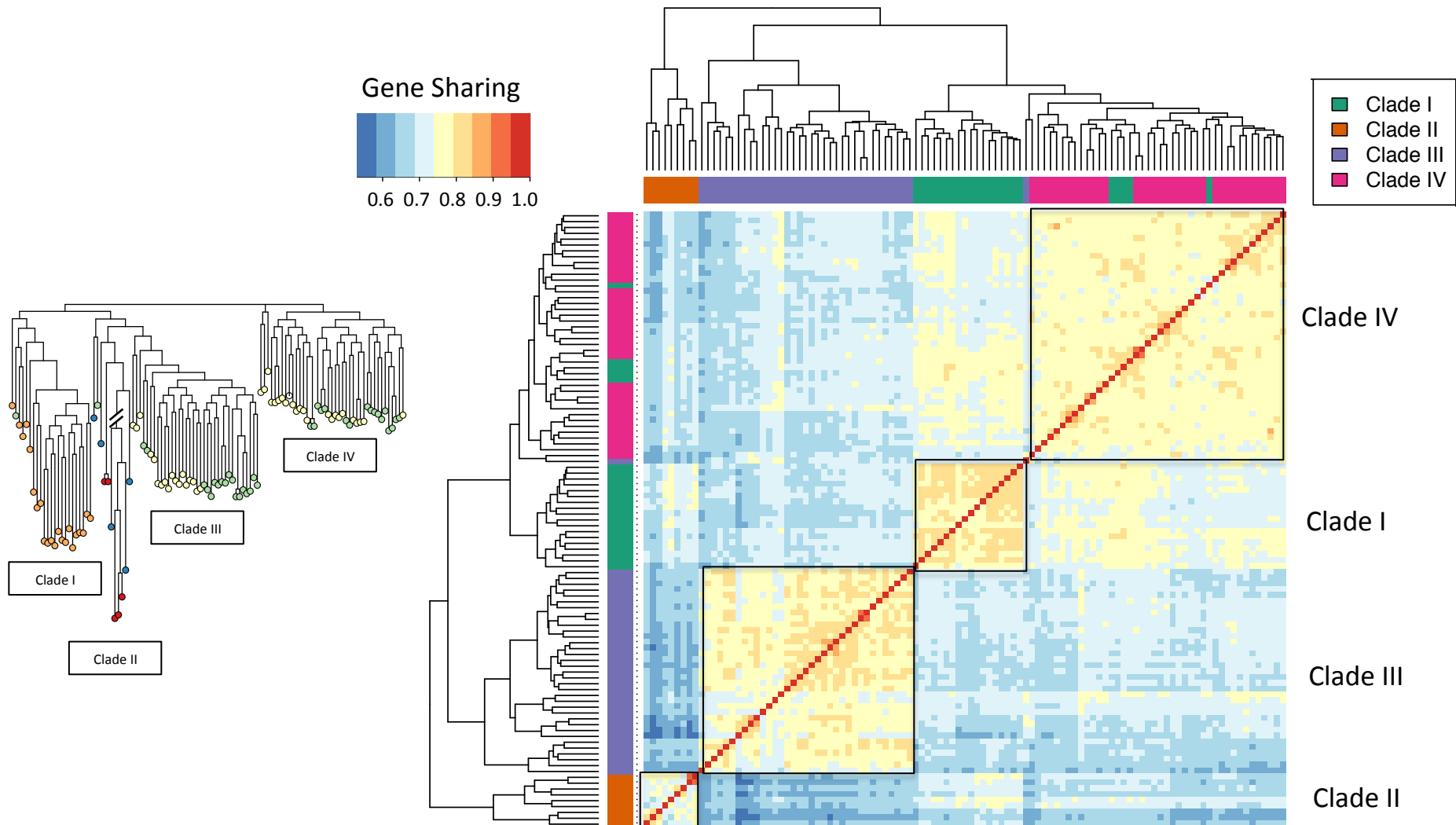


Phylogeny of *Eubacterium rectale* strains from shotgun metagenomes

- 741,602 core genome sites
- 86% of neighbors from same continent
- Z-Score=10.4 ($p < 1e-16$)

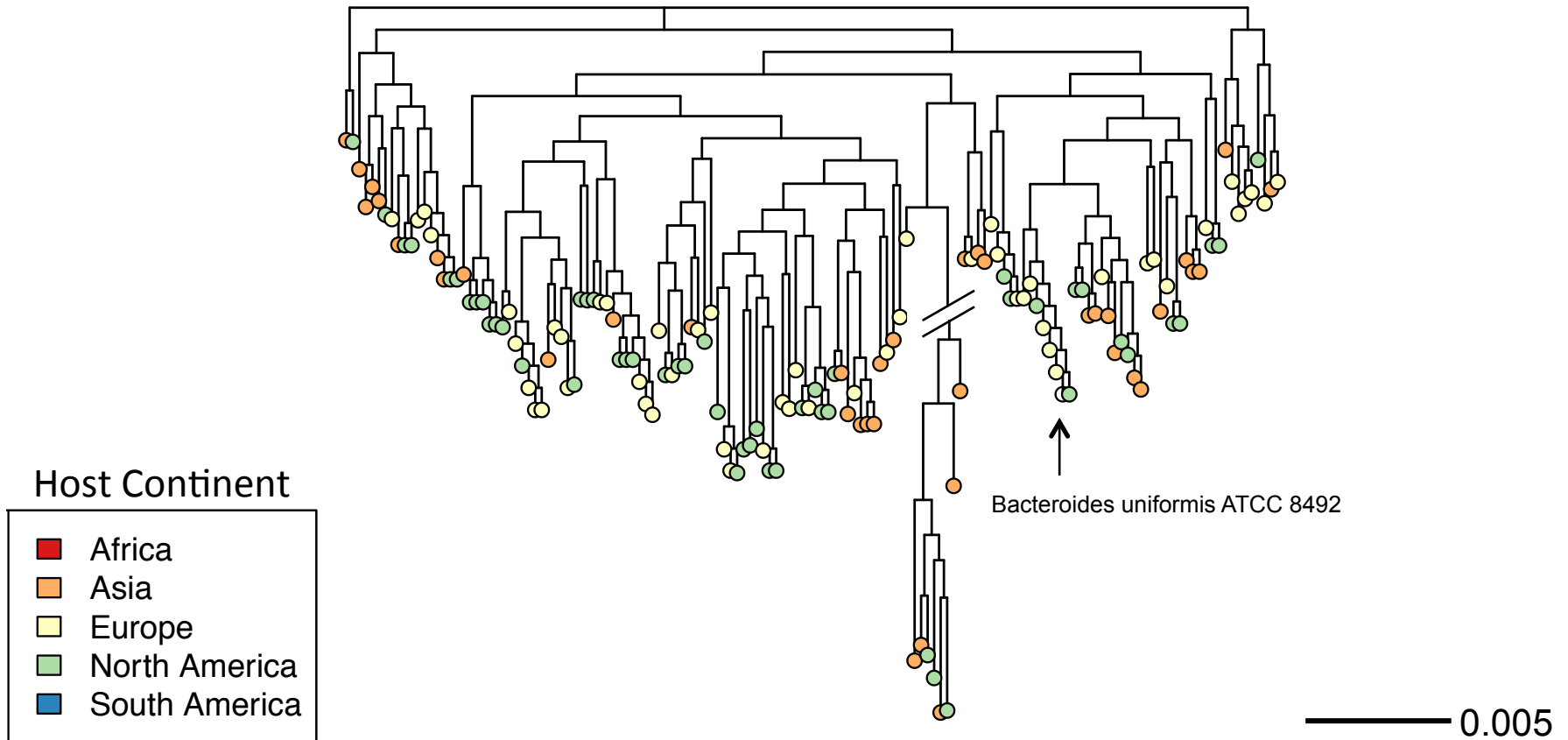


Gene sharing of *Eubacterium rectale* strains from shotgun metagenomes



Lack of strong biogeography for *Bacteroides uniformis*

- 1,129,356 core genome sites
- 50% of neighbors from same continent
- Z-score=3.2



Computational barriers prevent most researchers from easily leveraging publicly available shotgun metagenomes:

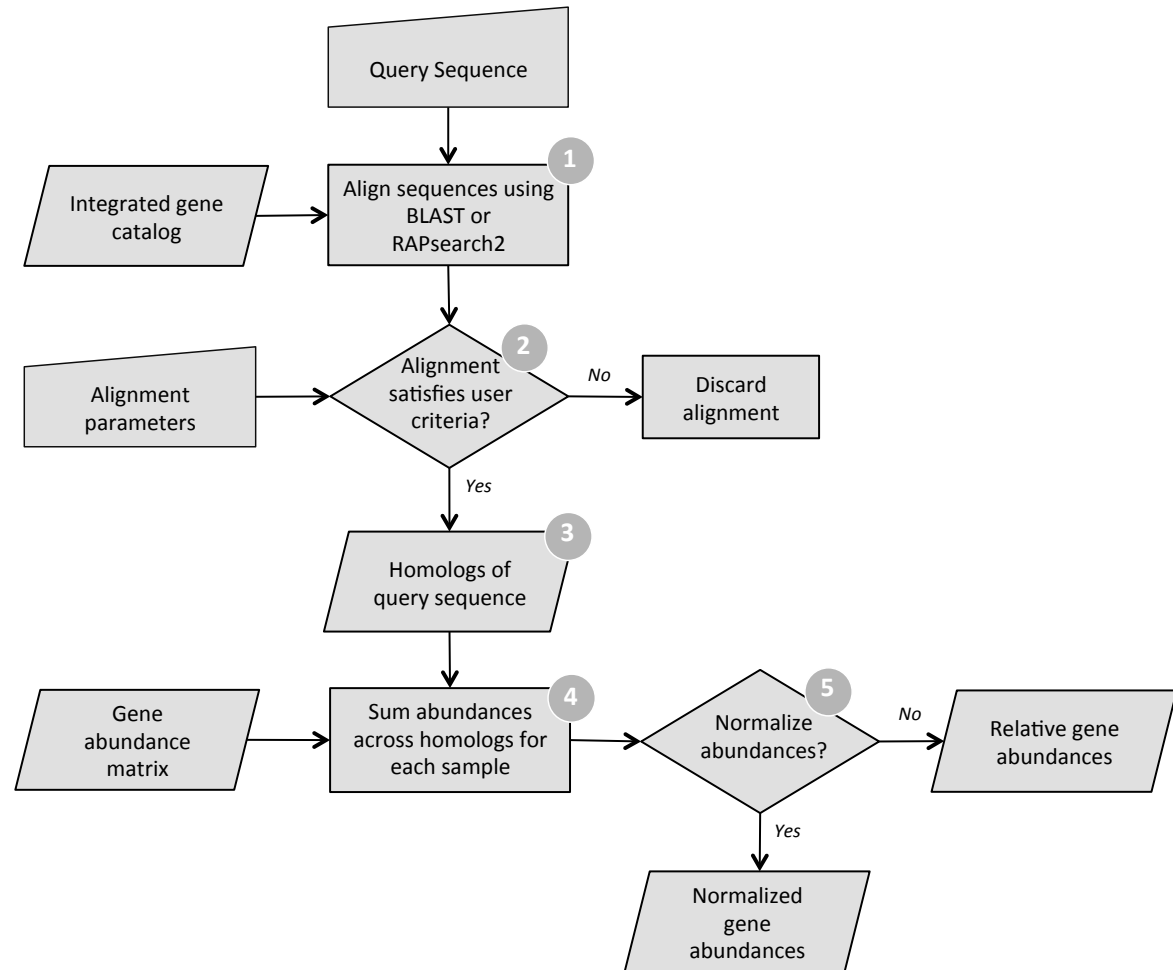
- How abundant is a particular microbial gene in the human gut microbiome?
- Is my gene associated with any clinical variables? With certain taxonomic groups? With other genes?

MetaQuery is a web server that addresses this problem

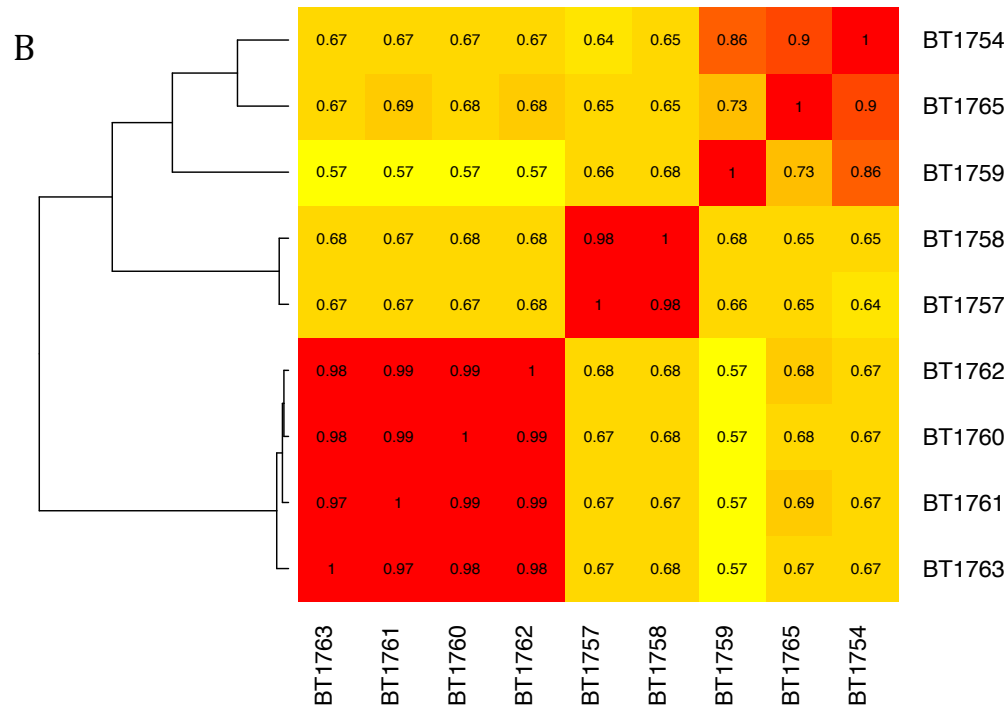
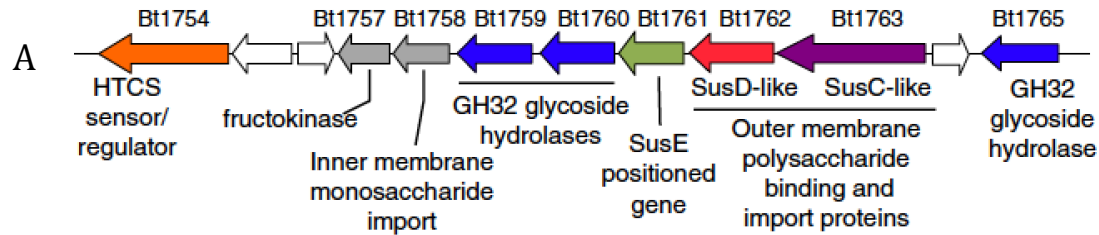
- Leverages comprehensive gene catalog and precomputed abundance matrix (9.9M genes x 1,267 samples)
- The user enters one or more query genes & MetaQuery returns the estimated abundance of query homologs across 1,267 faecal metagenomes
- Alignment parameters allow the user to define the relationship between sequence similarity and function
- Gene abundances normalized using a panel of 30 universal, single-copy genes
- Statistical test performed to identify features that co-vary with gene abundance

<http://metaquery.docpollard.org/>
<http://dx.doi.org/10.1093/bioinformatics/btv382>

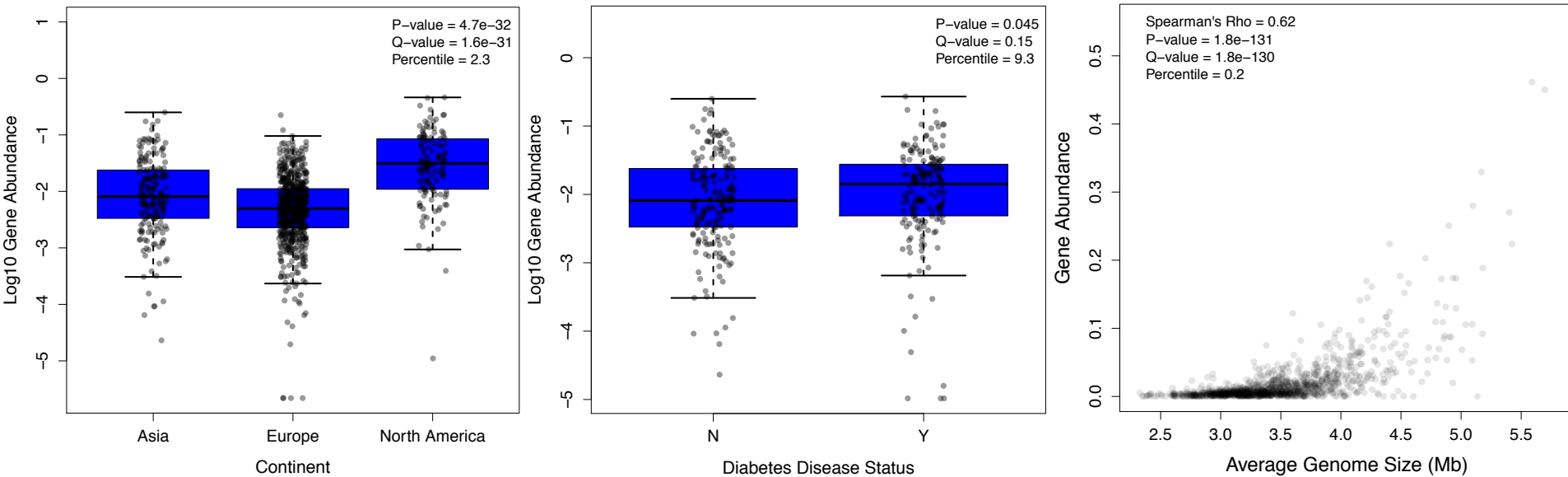
Flowchart for estimating query gene abundance



Case Study: Metagenomic Variation of the Fructan Polysaccharide Utilization Locus



Case Study: Metagenomic Variation of the Fructan Polysaccharide Utilization Locus



tax_name	rho	p.value	q.value
Bacteroides_ovatus	0.53	9.30E-91	3.60E-88
Bacteroides_xylanisolvens	0.52	4.30E-89	8.40E-87
Bacteroides_helcogenes	0.42	1.50E-53	2.00E-51
Bacteroides_thetaiotaomicron	0.4	2.10E-49	2.00E-47
Bacteroides_unclassified	0.4	6.70E-49	5.20E-47