



THE MICROBIOME OF HEALTHY HUMANS: COMMONALITIES AND VARIATIONS

Makedonka MITREVA, PhD

OUTLINE

- **GOAL of HMP:** Comprehensive characterization of the human microbiota and analysis of its role in human health and disease.
- **DATSETS:** i) reference genomes; ii) 16S rRNA marker gene and iii) shotgun metagenomic sequences.
- **OUTCOME:** i) organismal characterization of the community structure; ii) absolute comparisons of communities without reference to biased databases; iii) genetic variations among strains; iv) analysis of the metabolic potential of a community through gene identification and comparative metabolomics; v) identification of novel organisms.



ANALYSES ARE BASED ON:

Body site	# samples			Total reads (#)
	Visit 1	Visit 2	Visit 3	
Anterior_nares	57	30	1	141,007,722
Buccal_mucosa	69	37	3	1,344,299,975
Supragingival_plaque	71	43	2	6,651,280,717
Tongue_dorsum	73	50	2	10,630,491,838
Stool	81	54	4	14,471,969,023
Posterior_fornix	33	20	1	250,071,193
Total	384	234	13	33,489,120,468

3.3 TB

- Illumina 100bp pair-end reads



- Reads were aligned to reference genomes with 80% ID and 75% FoL
- Breadth and depth of coverage per genome
- Due to the different number of input bp per sample the depth of coverage was normalized per 100 million bp (DCPM, (depth of coverage * 100Mb / #aligned bps);
- A strain was considered to be present within a community when 1% of the genome is covered at 0.01X depth of coverage (Presence/absence).
Additional normalization only when needed

Kingdom	Genera	Species	Strains
Bacteria	533	1253	1751
Archaea	57	97	131
Viruses	918	1420	3683
Eukaryotes	237	326	326
Total	1745	3096	5891

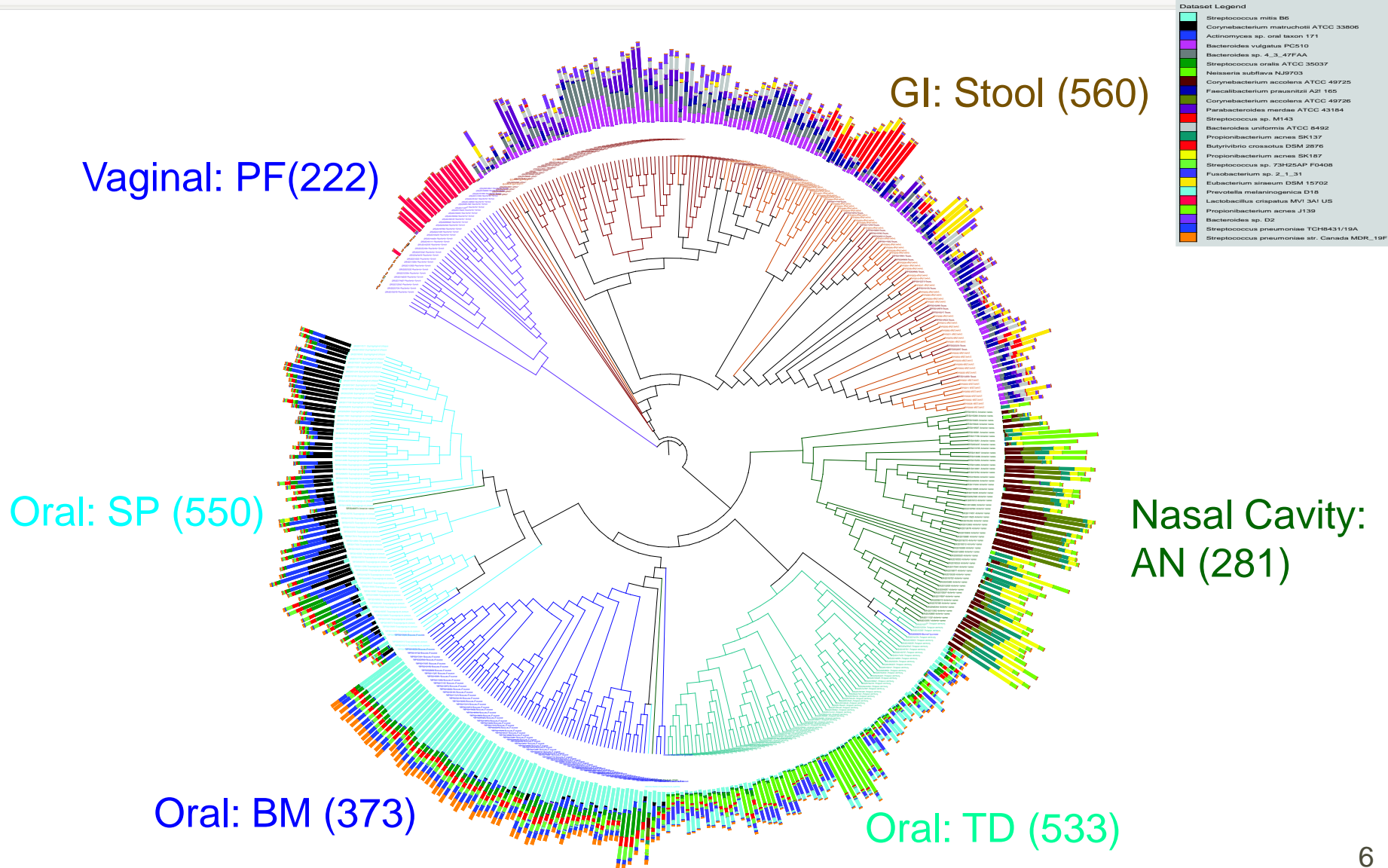


% Read characterized by the reference genomes

Body site	Sample used for stats	Total reads (#)	Aligned (%)
Anterior nares	88	141,007,722	33.2
Buccal mucosa	109	1,344,299,975	60.8
Supragingival plaque	116	6,651,280,717	54.8
Tongue dorsum	125	10,630,491,838	53.1
Stool	139	14,471,969,023	61.2
Posterior fornix	54	250,071,193	77.3
Total / Avg	631	33,489,120,468	56.7

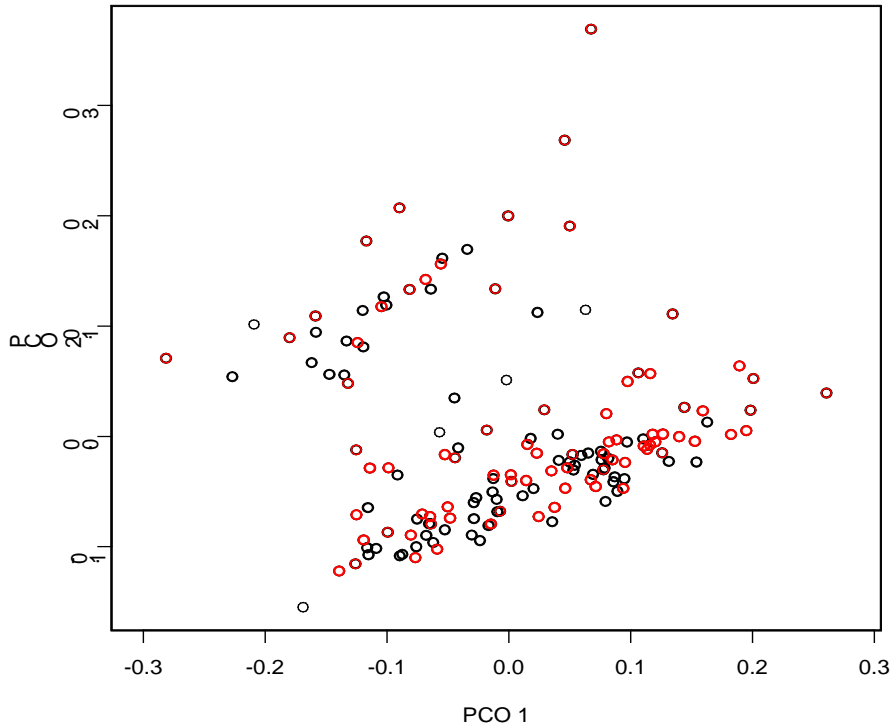


Community structure based clustering (DCPM)

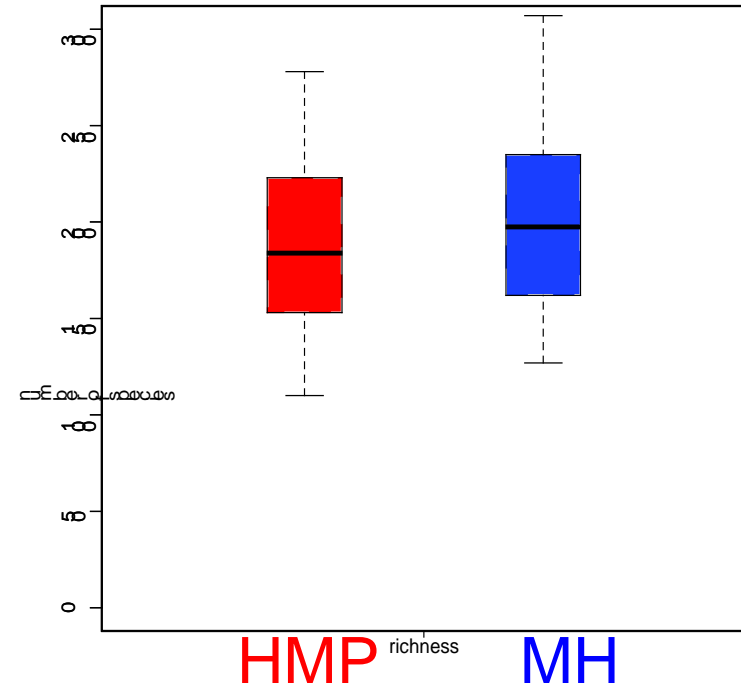


Presence / Absence: no separation between HMP and MH

hmp and methit dis=sorensen



richness

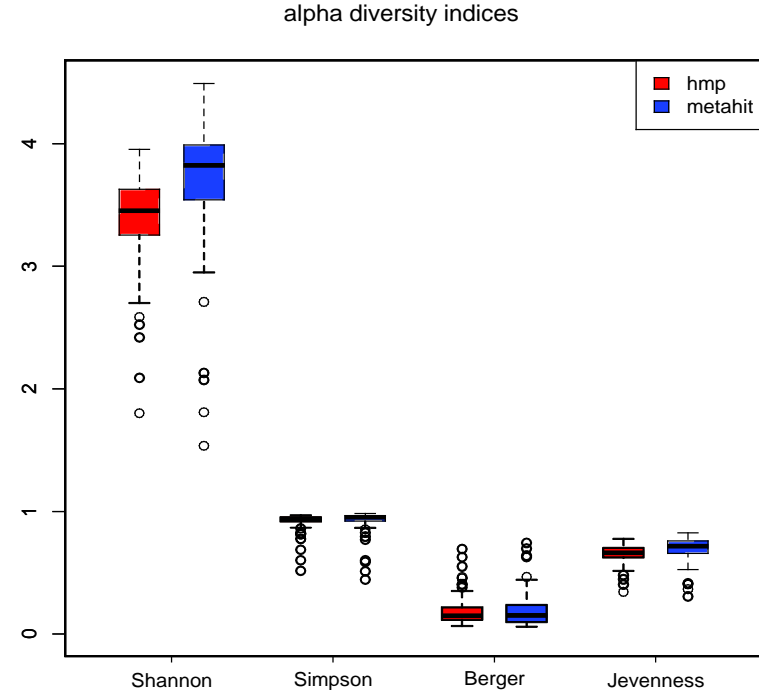
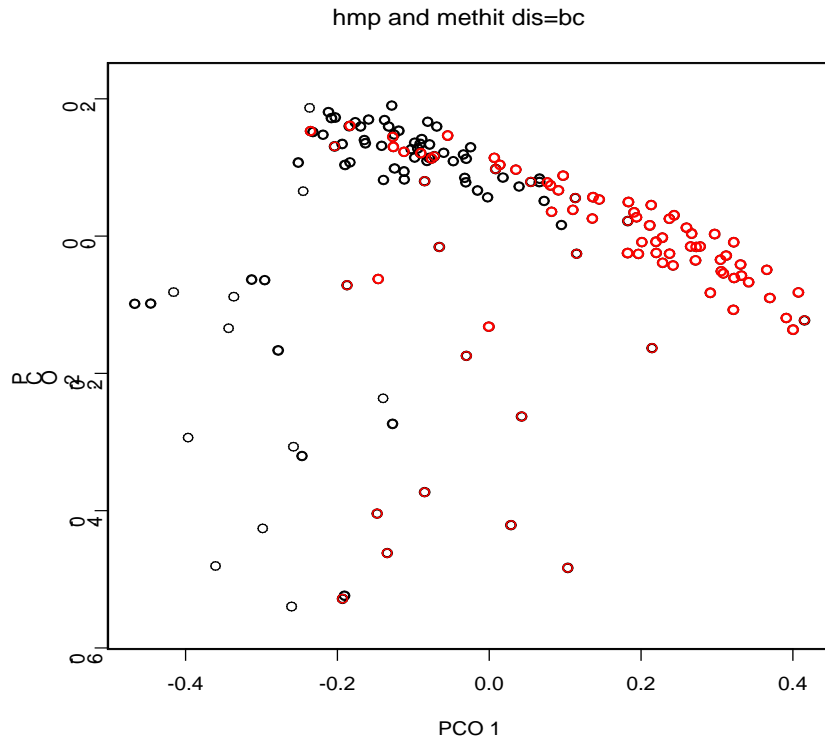


MH on average captured more species than HMP and has higher alpha diversity (not statistically different).

Sorensen distance matrix; NMCD based on Sorensen with 1st dimension using Bray Curtis.



When abundance (DCPM), richness & evenness is considered



Using abundance data lead to a separation pattern

Significantly different: Shannon and Jevnness

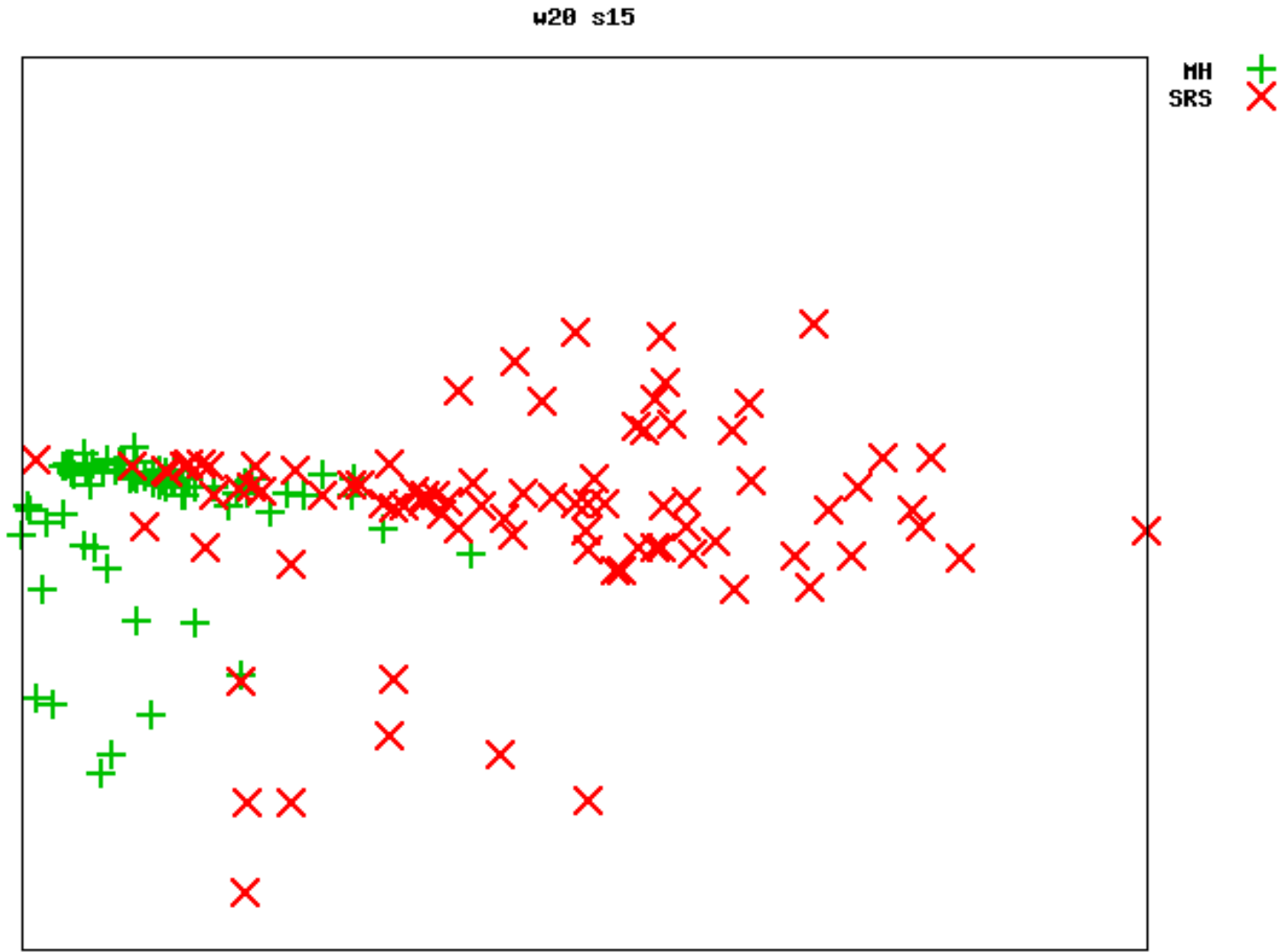


Sequence Phylogeny - taxonomy free comparisons

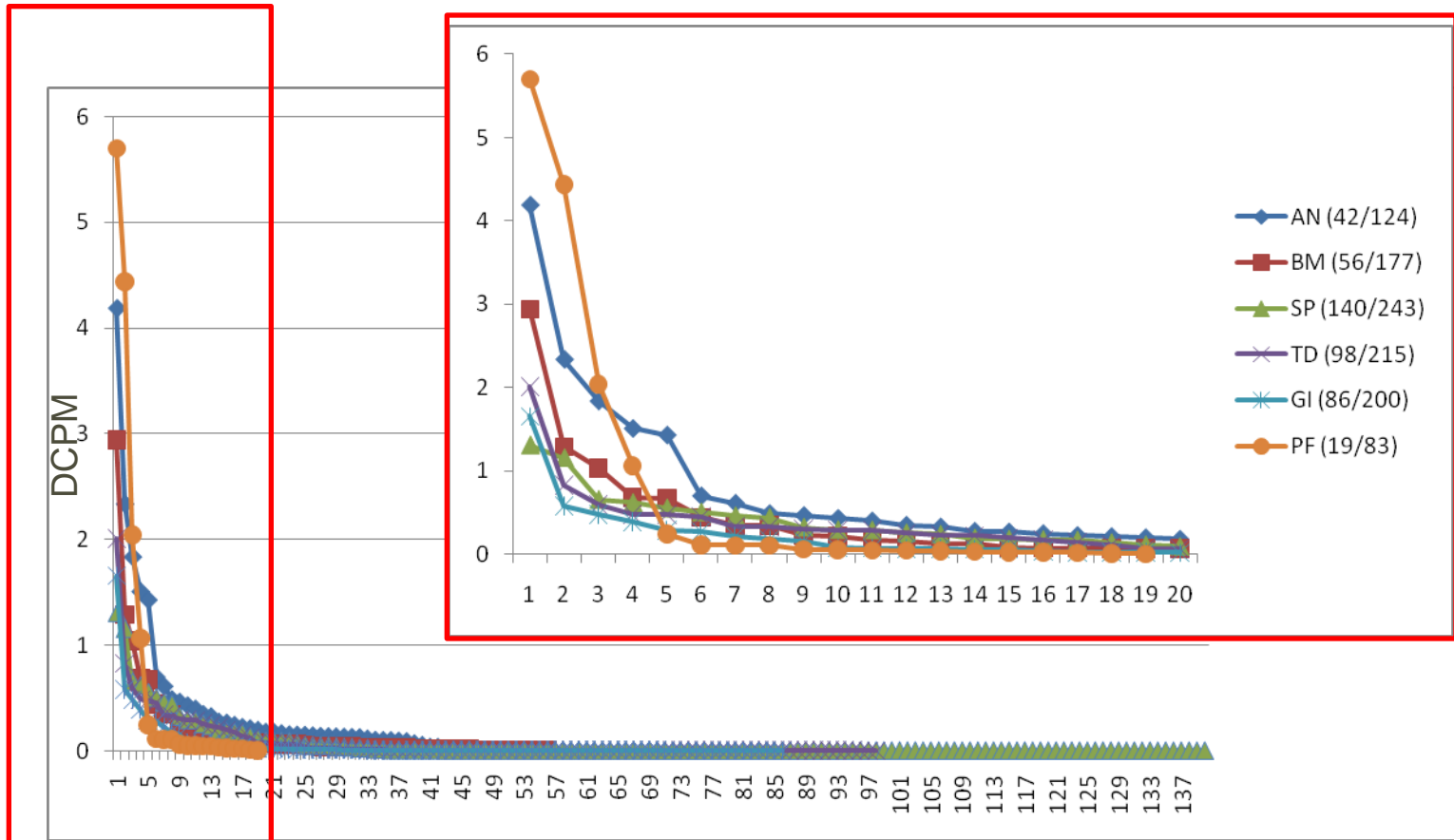
- Phylogenetic approach solves problems with fragmentary, non- overlapping reads;
- Absolute view of community complexity without biases from existing databases.
- Approach:
 - based on sequence composition, w20-30, s15-30 (RTG phylogeny module);
 - metagenomics read trees, similarity distance and hierarchical clustering;
 - generate k-mer profile, exclude k-mers that occur more than once;
 - compare unique k-mers per sample among samples.



Clustering retains when changing parameters



Strains that are at least at 10% of all the samples

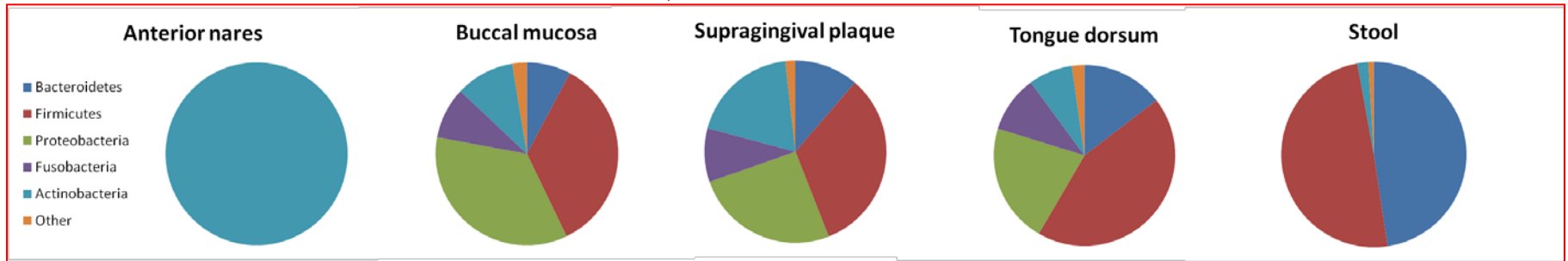
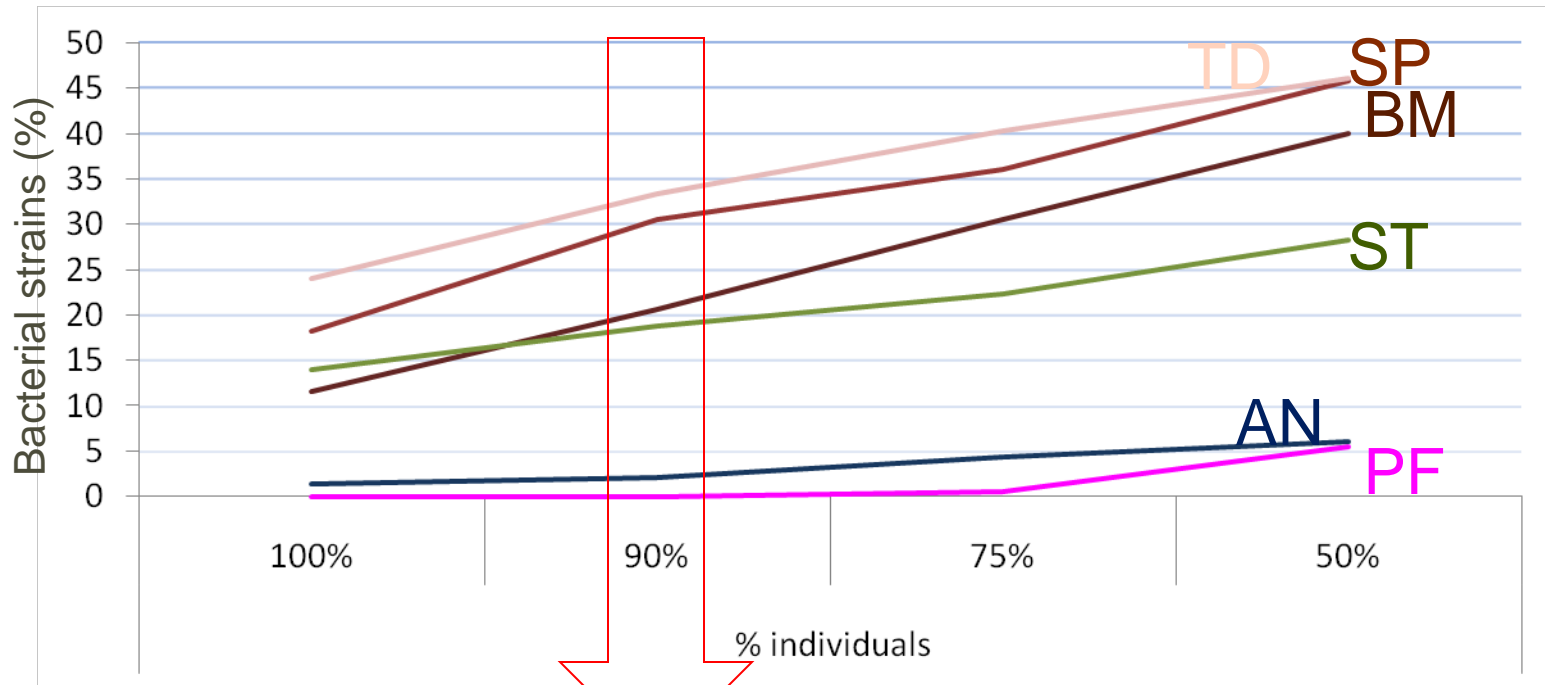


The most abundant:
 Nasal: Propionibacterium
 GI: Faecalibacterium

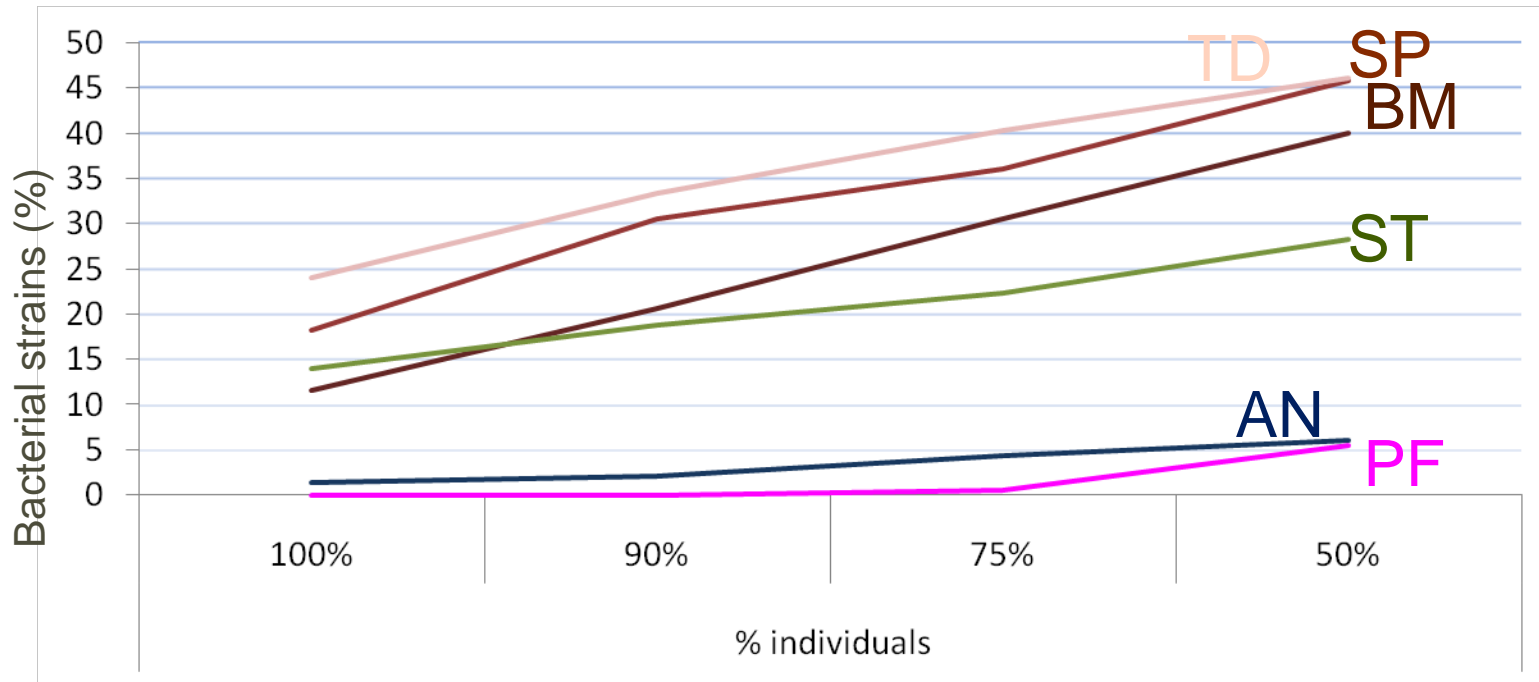
Oral: Streptococcus
 Vaginal: Lactobacillus



Common strains



Common strains



Anterior nares

Buccal mucosa

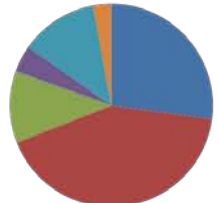
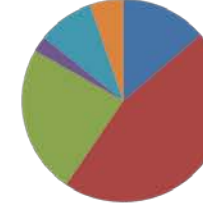
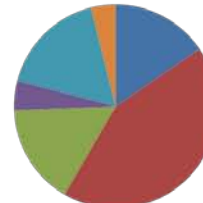
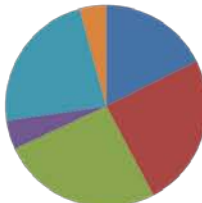
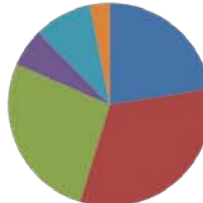
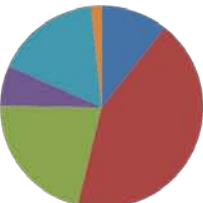
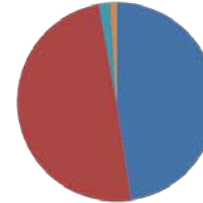
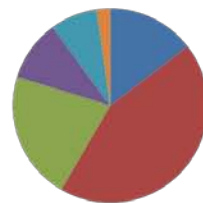
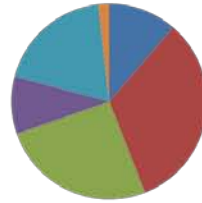
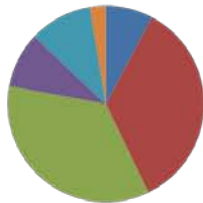
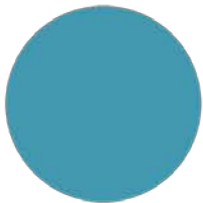
Supragingival plaque

Tongue dorsum

Stool

Posterior fornx

- Bacteroidetes
- Firmicutes
- Proteobacteria
- Fusobacteria
- Actinobacteria
- Other



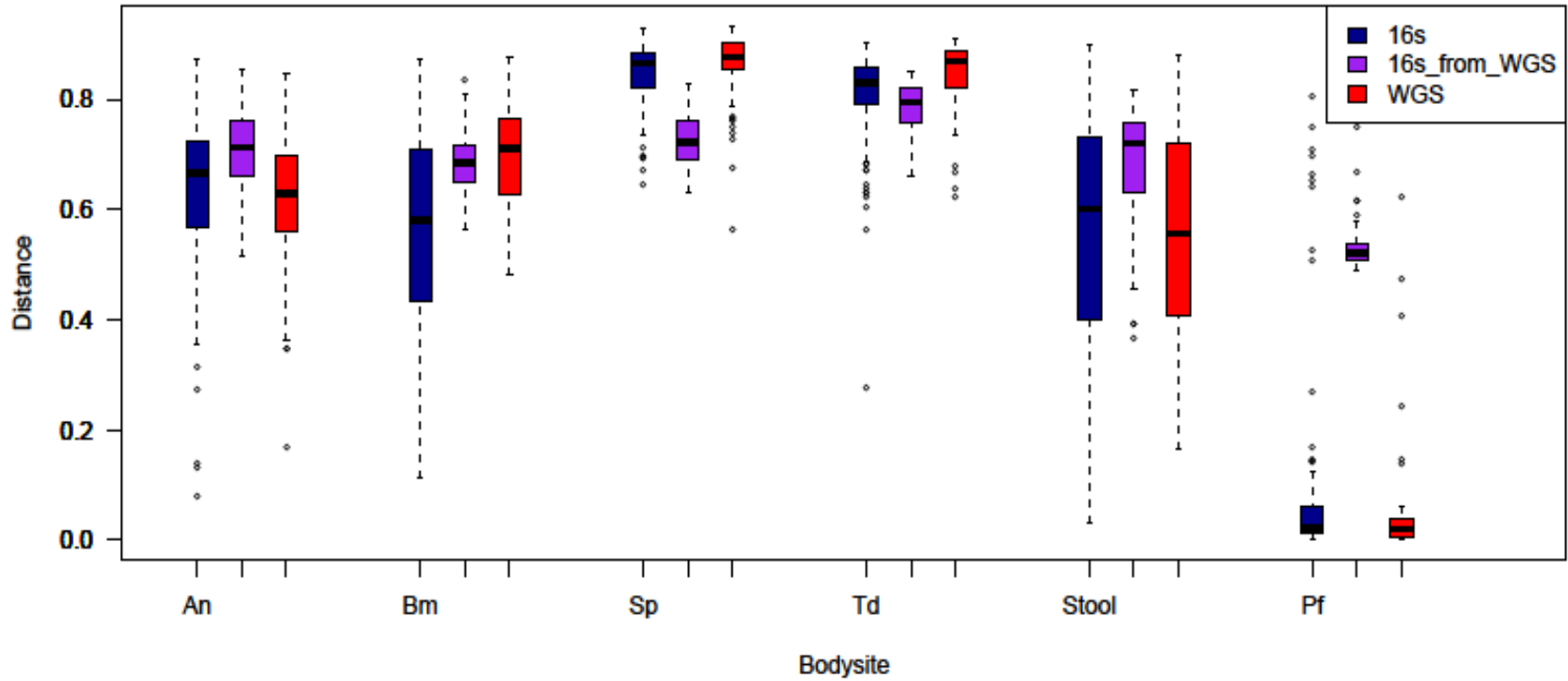
Community membership and structure: three approaches

1. 16S: Roche/454 16S V35, normalized to 7000k, RDP @ 0.8 confidence;
2. 16S from WGS: using crossmatch 16S were identified in WGS Illumina sequences (70% identity over 90bp of the length. Accept alignments only to V35 region);
3. WGS: Alignments to Reference Genomes (DCPM) Depth of coverage normalization per 100M bps (depth of coverage*100Mb/#aligned bps);

Normalization was done to scale all the methods.



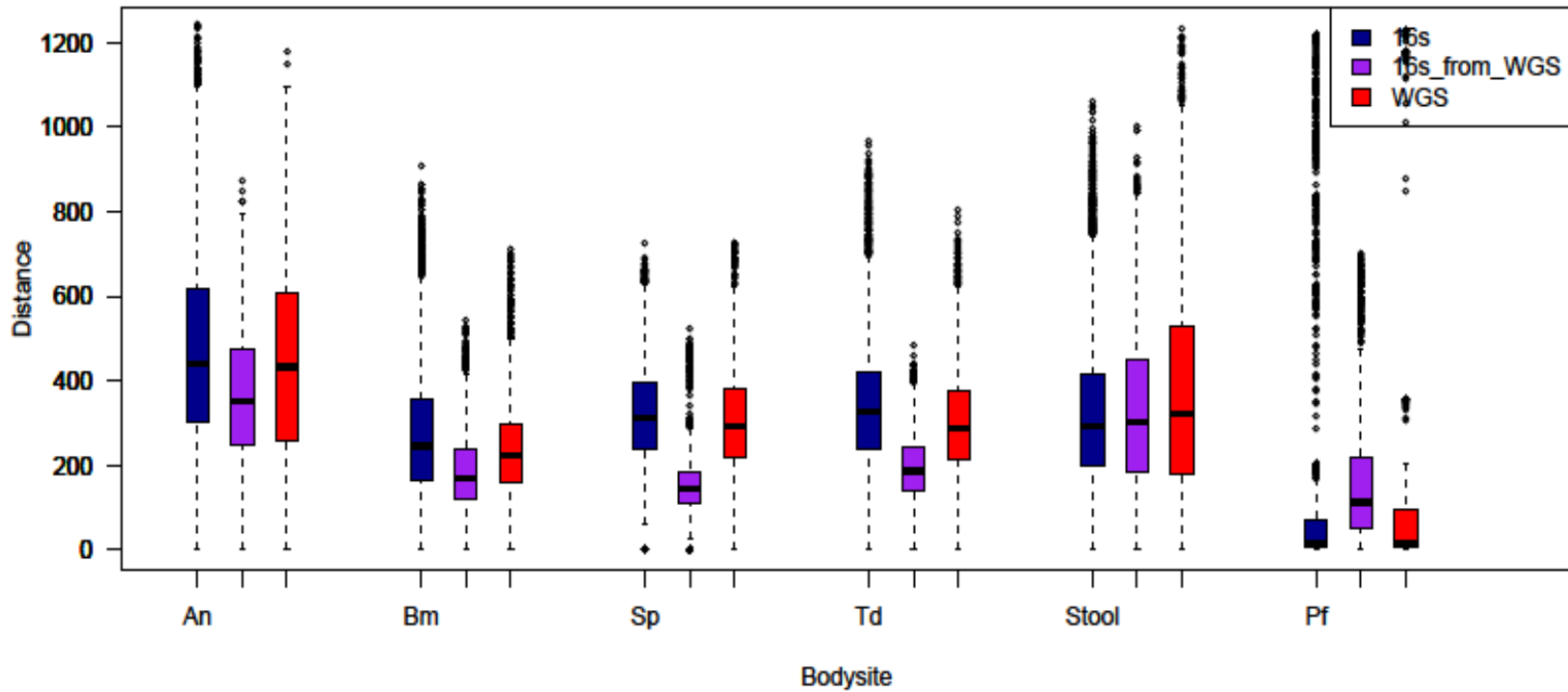
Alpha diversity



Simpson index



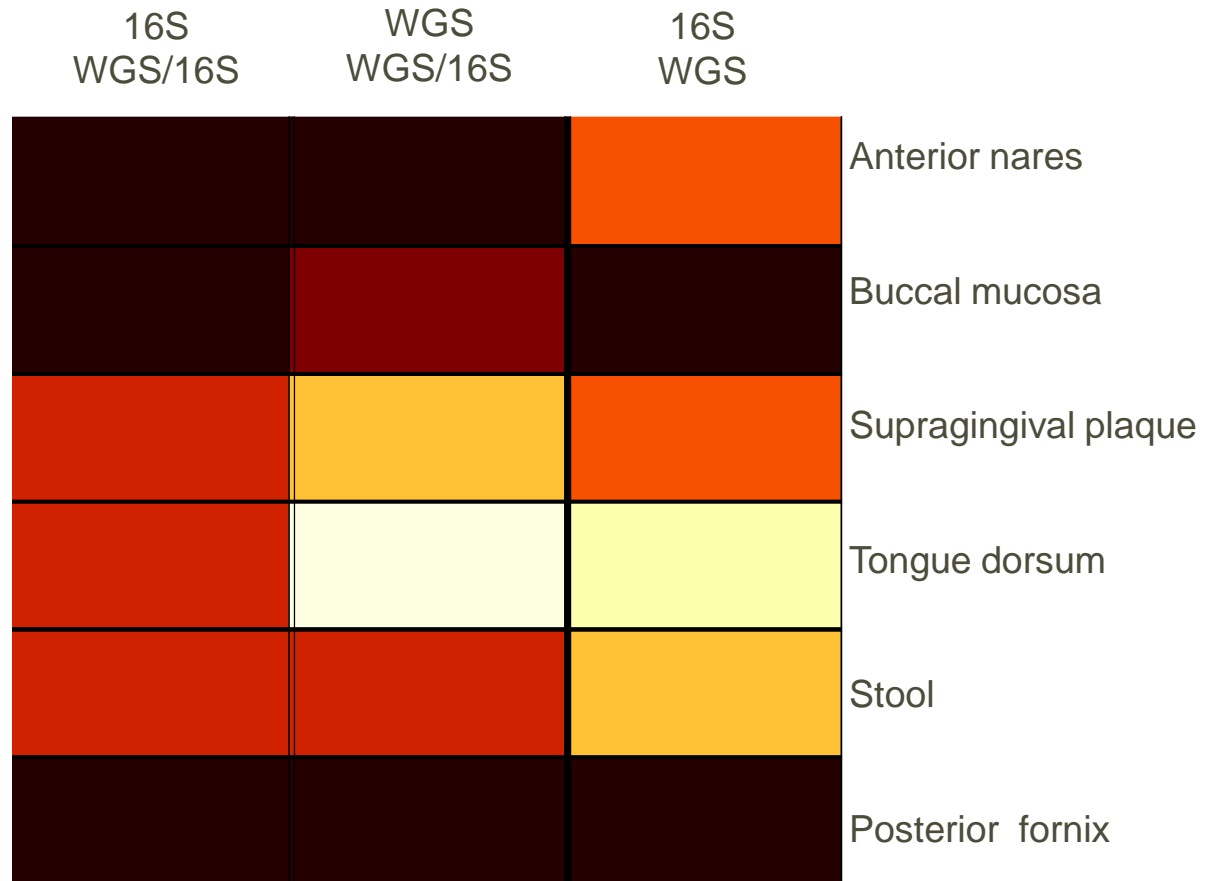
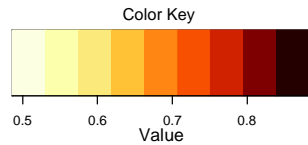
Beta diversity



Euclidean distance matrix



Concordance among approaches



Pearson's correlation coefficient is plotted: degree of linear relationship between the genera that overlap based on 16S, WGS/16S and WGS, and number of species sharing that genera in each of the datasets.



Genetic variations in strain shared among the oral sites

- Alignments 95%id + 90%length;
- Streptococcus oralis ATCC 35037

Body site	Samples (#)	Aligned reads (#)	Genome Coverage	
			Breadth (%)	Depth (X)
Buccal mucosa	42	268,850,352	90%	17--192
Supragingival Plaque	37	993,792,356	92%	21
Tongue dorsum	29	1,236,436,803	70%	7--13

- Filtering of loci before being considered a SNP:
 - min. consensus quality 20
 - min. read depth 3
 - window size for filtering out dense SNPs 10
 - max. no. SNPs allowed in window 2
- Major allele: was identified by summing up the occurrence of each base across all samples, and selecting the most frequent one. Then the frequency of that base/allele was reported for that loci for each sample.



Deviation vs. Variation and Shared vs. Union

	LOCI #1	LOCI #2	LOCI #3	LOCI #4
Reference	T	T	T	T
BODY_SITE_A				
	variation	not a snp	not a snp	deviation
sample1 read1	T	T	T	G
sample1 read2	A	T	T	G
sample1 read3	A	T	T	G
	deviation	not a snp	variation	variation
sample2 read1	A	T	A	G
sample2 read2	A	T	A	C
sample2 read3	A	T	T	C
sample2 read4	A	T	A	T
sample2 read5	A	T	T	A
	variation	not a snp	deviation	deviation
sample3 read1	G	T	C	A
sample3 read2	T	T	C	A
sample3 read3	C	T	C	T
BODY_SITE_B				
	deviation	not a snp	deviation	not a snp
sample4 read1	C	T	A	T
sample4 read2	C	T	A	T
sample4 read3	C	T	A	T
sample4 read4	C	T	A	T
sample4 read5	C	T	A	T
	not a snp	variation	variation	not a snp
sample5 read1	T	A	A	T
sample5 read2	T	A	T	T
sample5 read3	T	T	T	T
	Shared loci		Shared loci	
	Union of all loci present in at least 1 body site			

Deviation: all reads within a sample conform the same SNP;

Variation: the underlying reads conform SNP but different base;

Union of SNPs: one single sample had to show that position as a SNP;

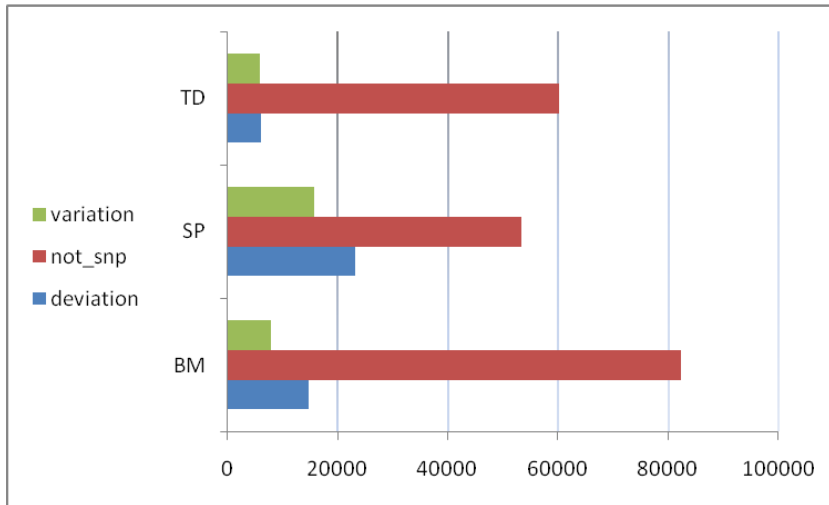
Shared SNPs: at least 1 sample from each body site had to have that position as a SNP.

The top 1000 most shared SNP loci per body site

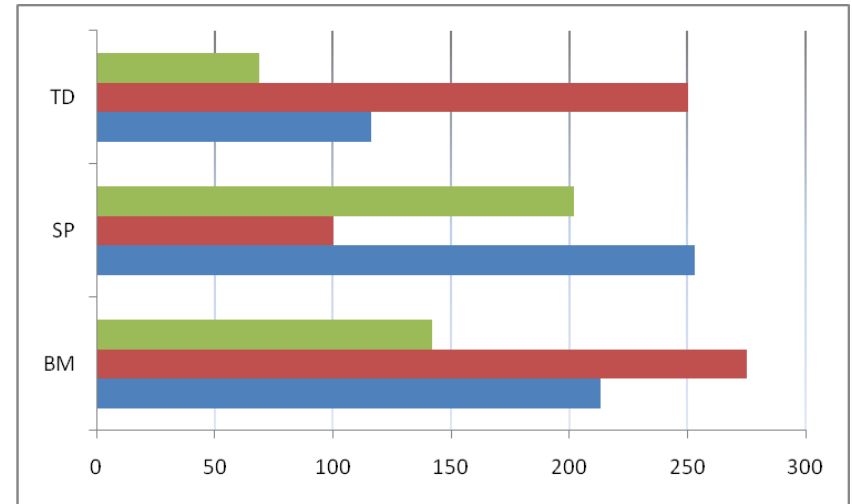


Genetic variants per body site (108 samples)

Union 269,460



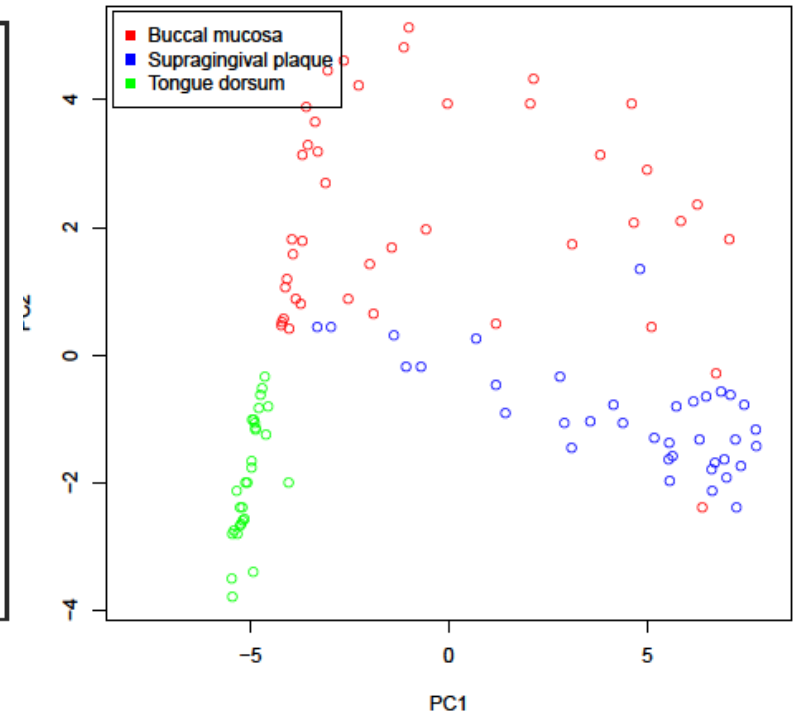
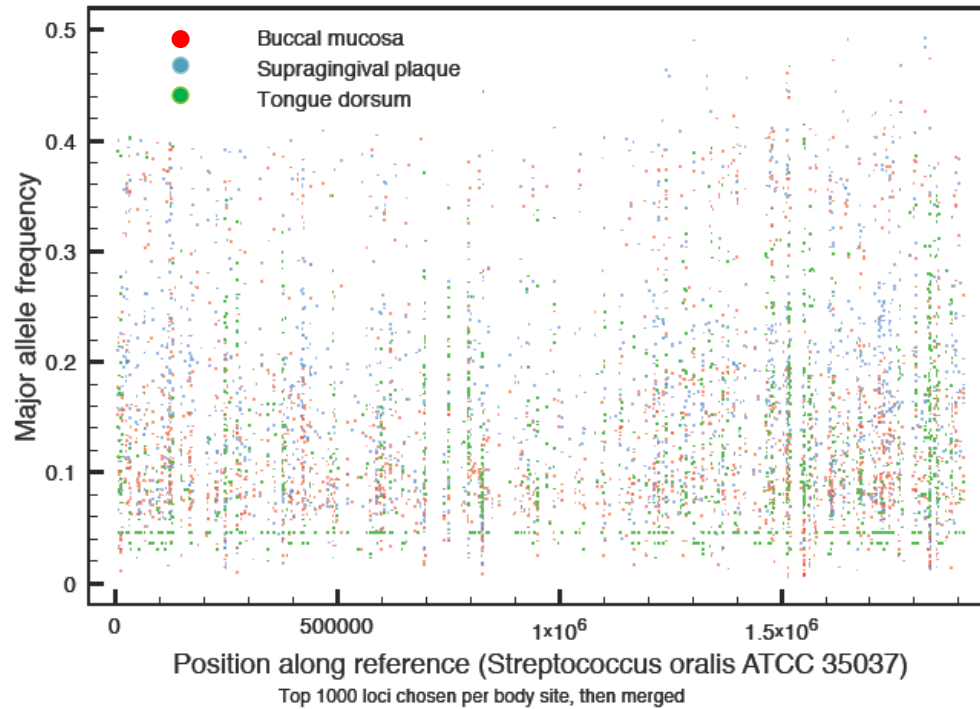
Shared 1,620



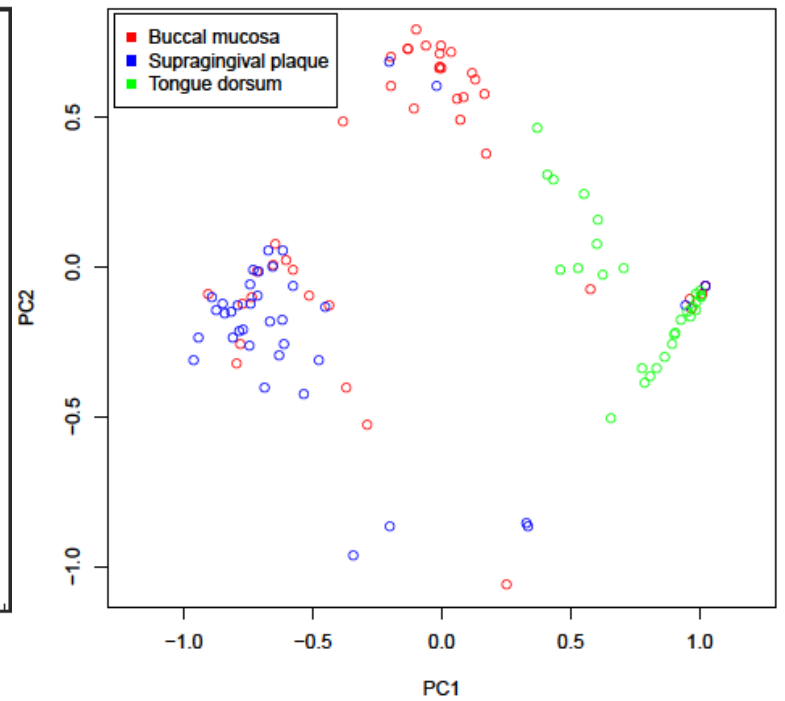
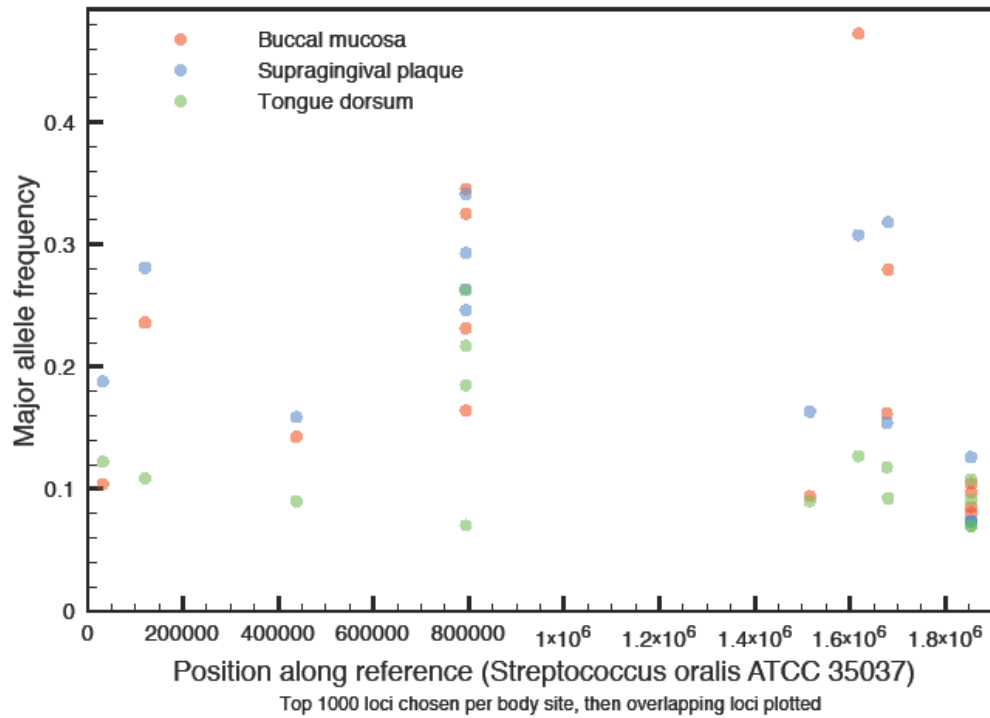
Not-SNPs: all the samples across all loci per body-site that were not SNP

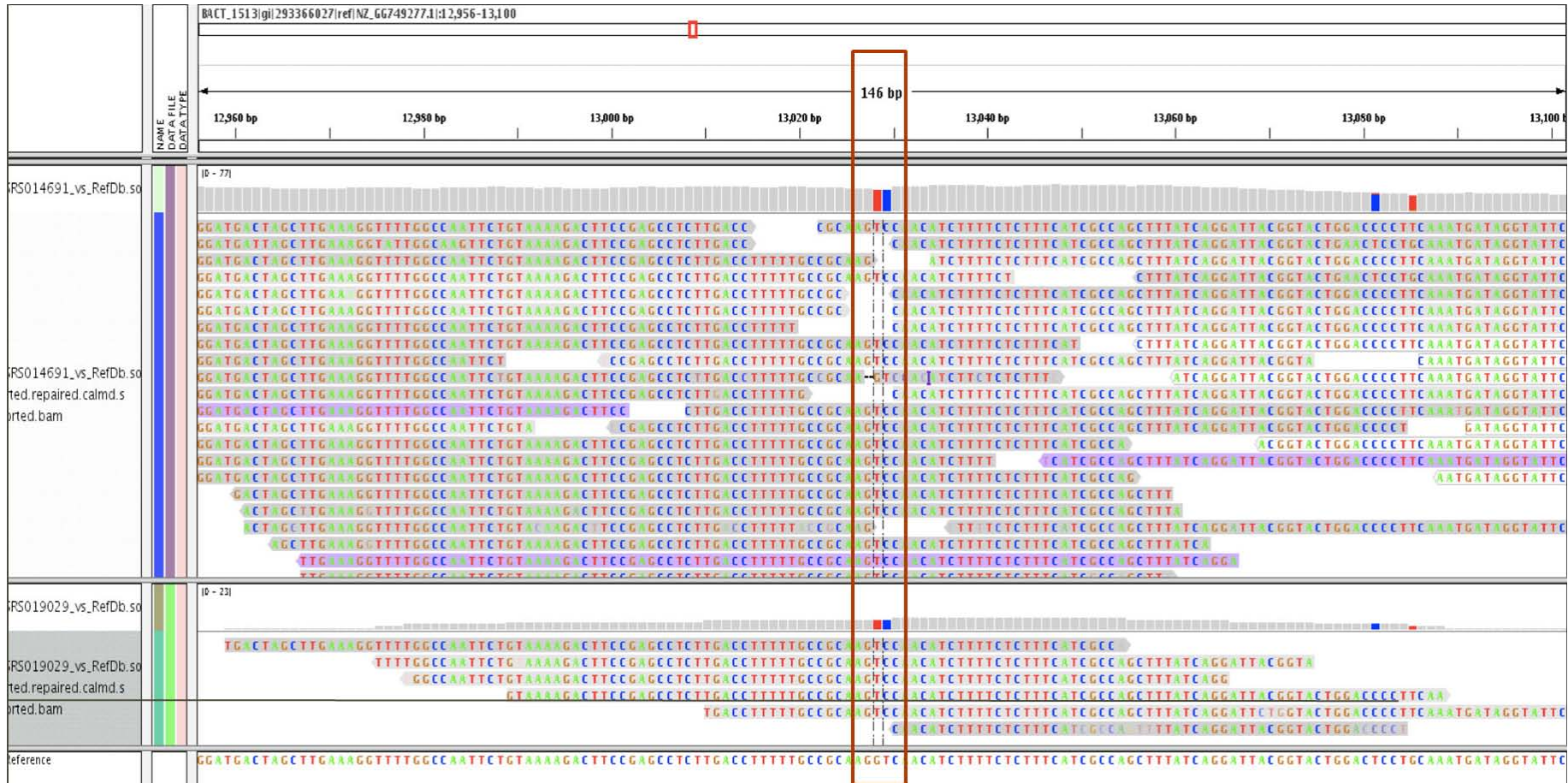


Union SNPs



Shared SNPs





- *Streptococcus oralis*.
- Analysis of SNPs that are conserved between the 2 visits in a Supragingival Plaque samples.
- Deviation from the reference strain is consistent between the 2 visits.



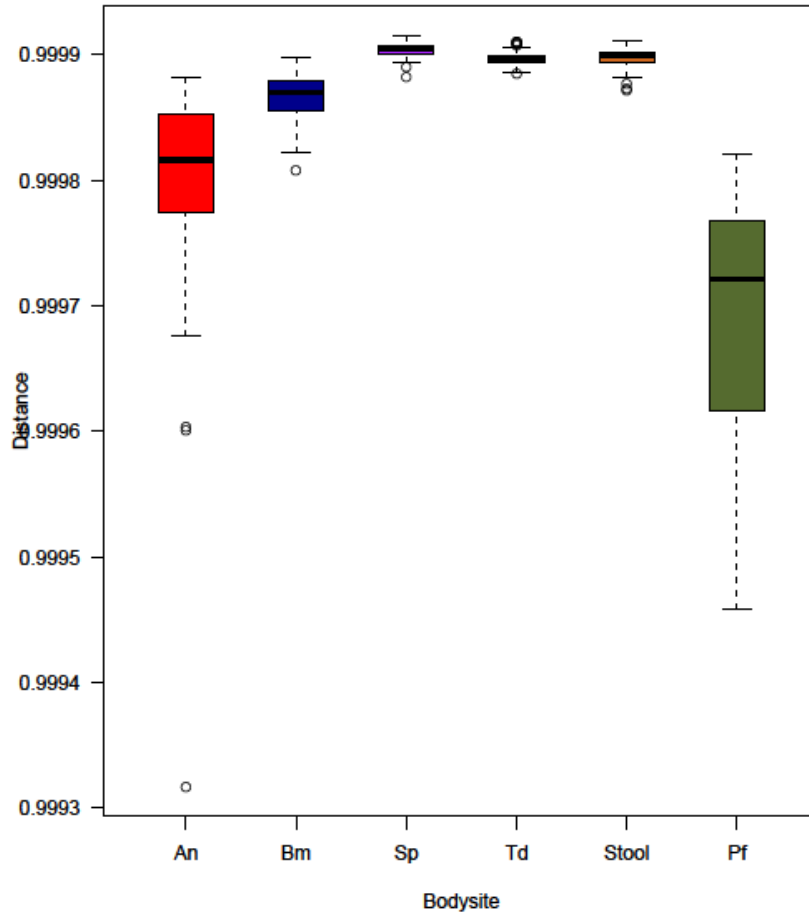
Metabolic profiling

- Blastx of the metagenomic shotgun reads vs the KEGG
- **HUMANn**: HMP Unified Metabolic Analysis Network (Poster#3)
 1. Genes to pathways (MinPath (Ye et al., 2009))
 2. Taxonomic limitations (Rem pathways in taxa , ave)
 3. Smoothing (Witten-Bell)
 4. Gap filling ($c(g) = \max(c(g), \text{median})$)
 5. Xipe (Distinguish zero/low; Rodriguez-Mueller in review)
- KO & Pathways and Module Coverage
- KO& Pathway and Module Abundance

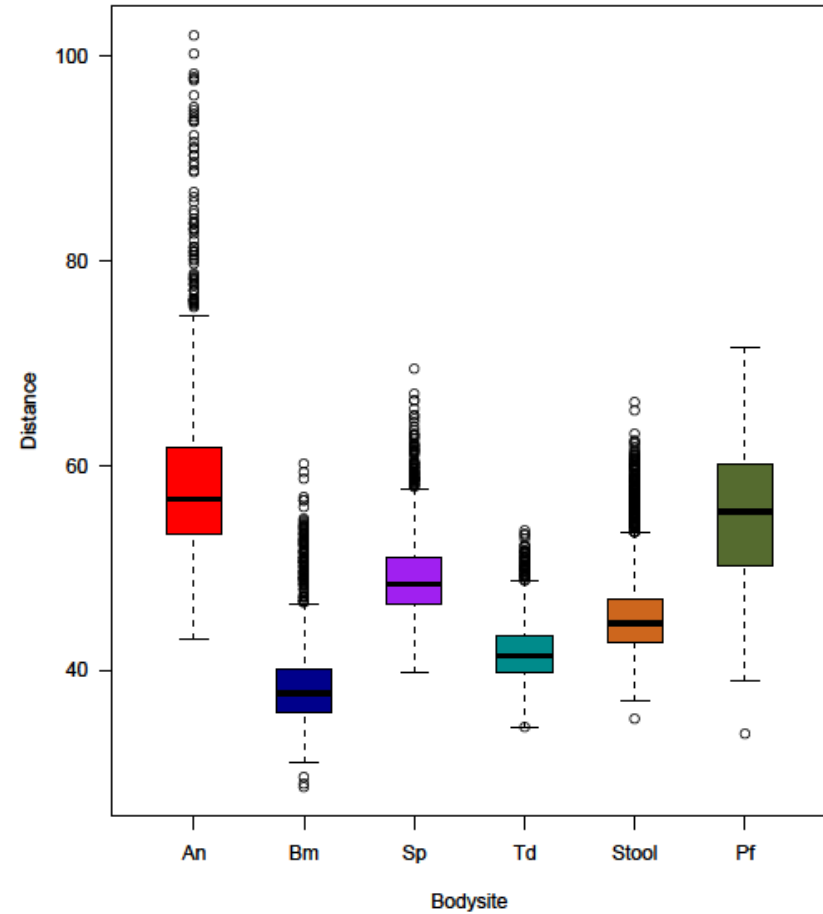


KO diversity

Alpha Diversity



Beta Diversity



Simpson / Euclidean

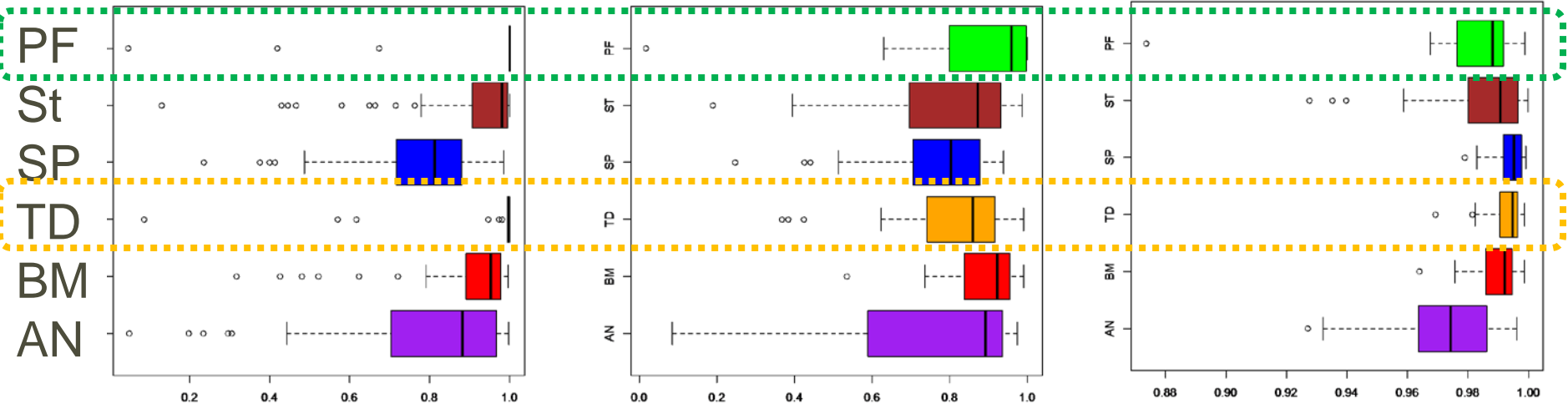


Correlations between visits

16S

WGS

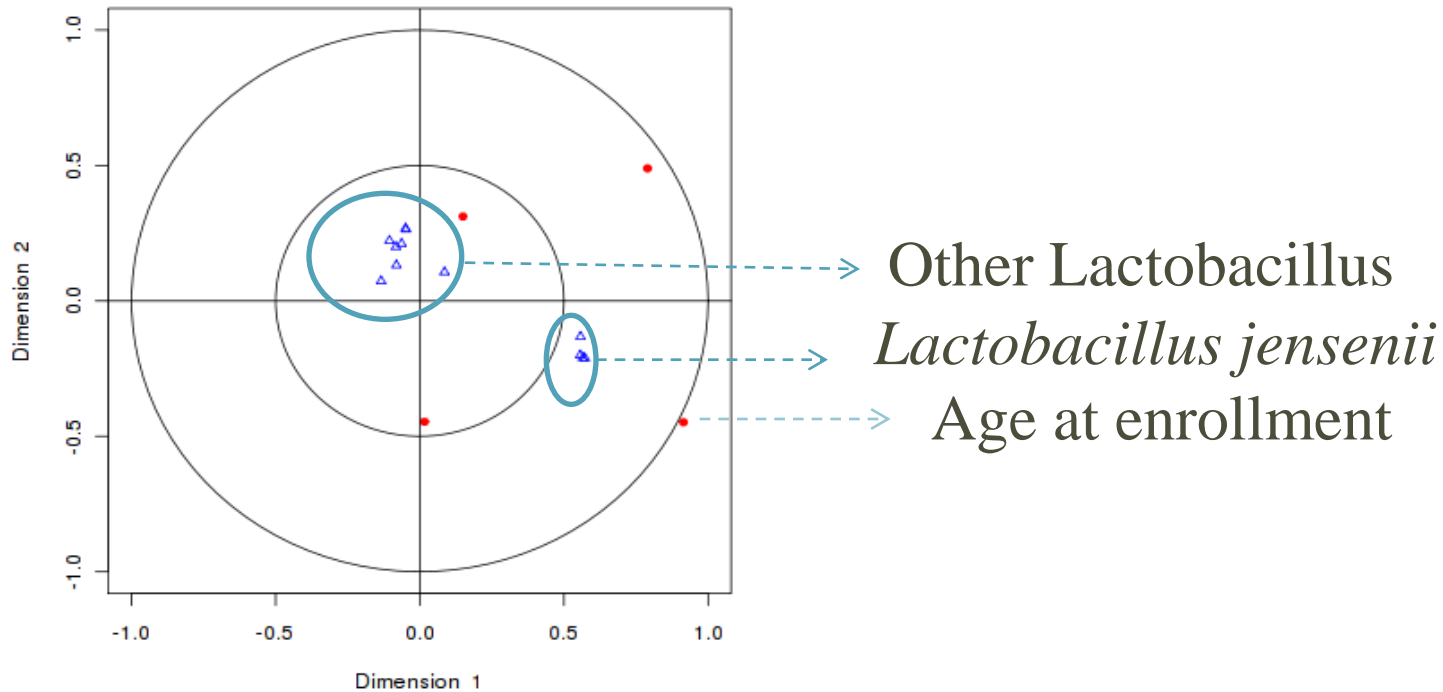
Function



Plotting the Pearson correlation coefficients between visit 1 and 2.



Correlation between diversity and metadata

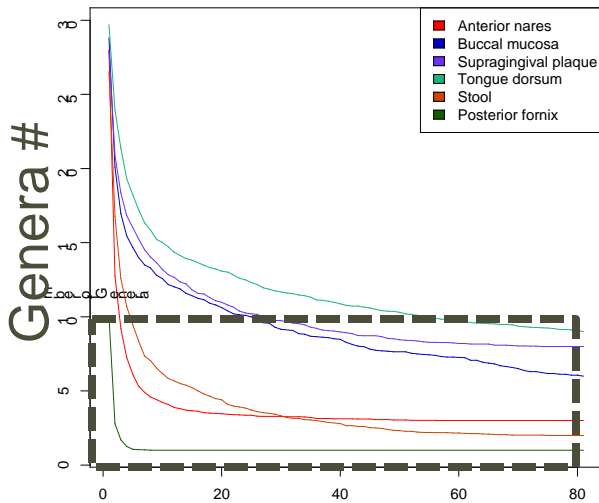


Canonical Correlation Analysis (CCA) plot of the species represented by Posterior Fornix (shown in blue) with the age, region, ethnicity and BMI (shown in red). The area outside the inner black circle represents strong correlations.

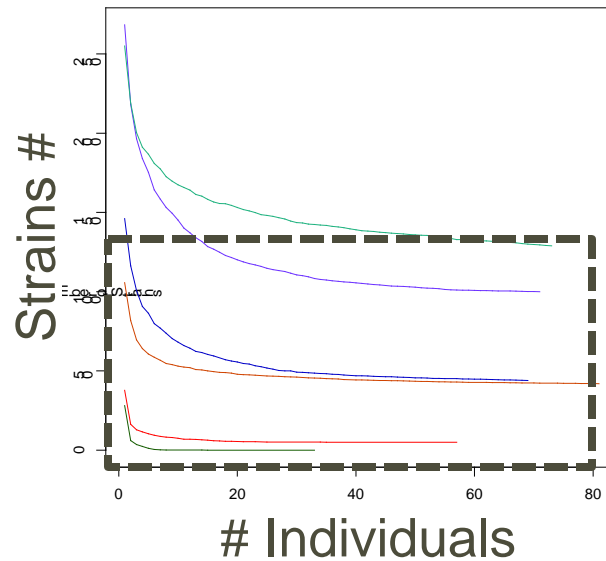


Core Biome

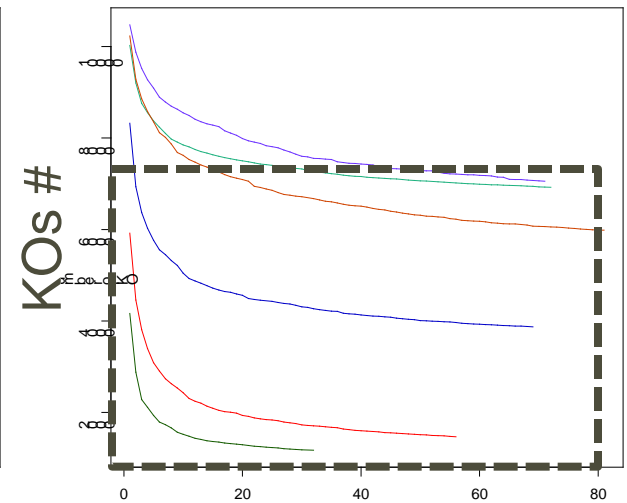
16S



Mapping to Reference Genomes



Metabolic Capabilities

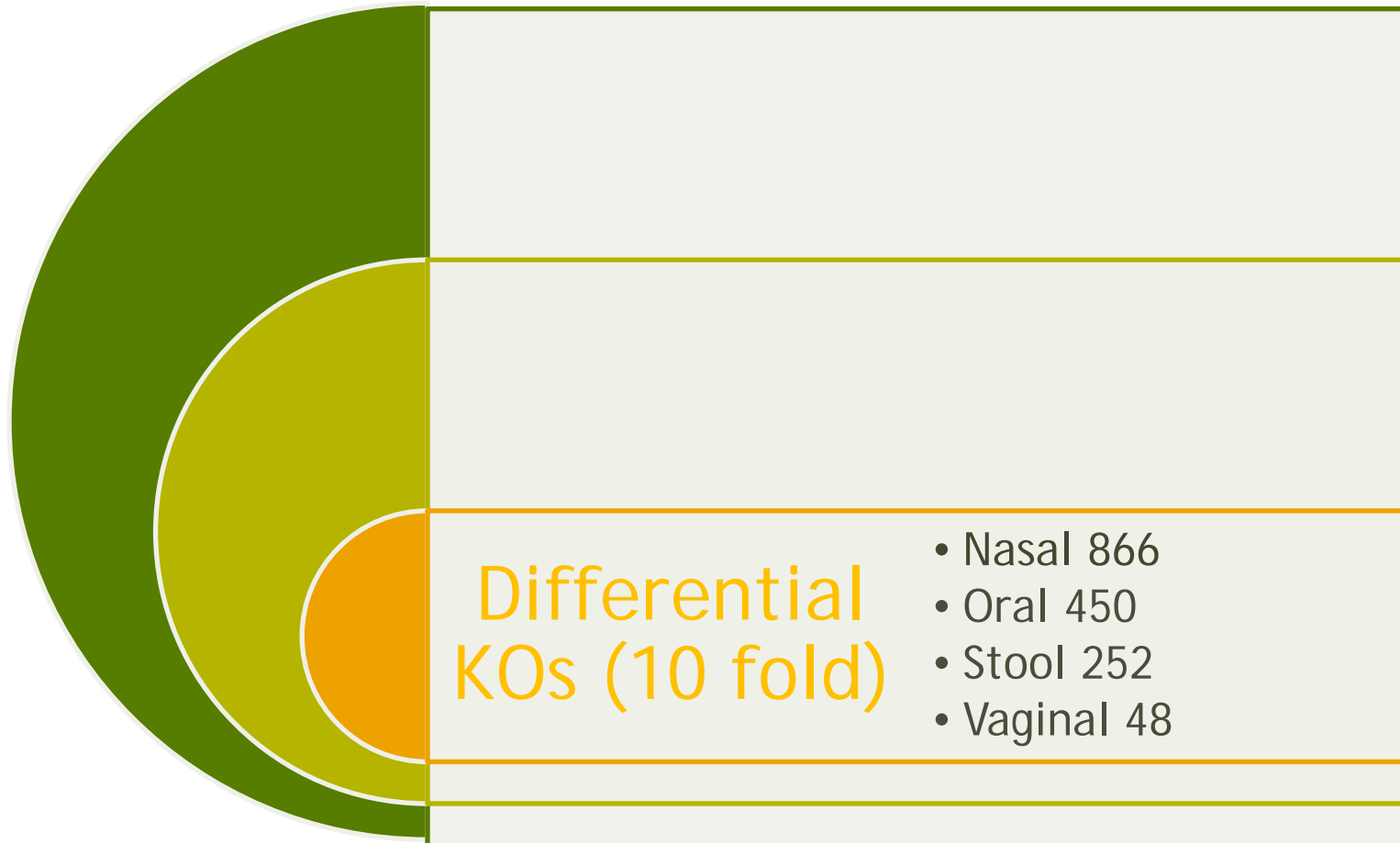


Decreasing taxonomic plots with increasing number of Individuals, but....

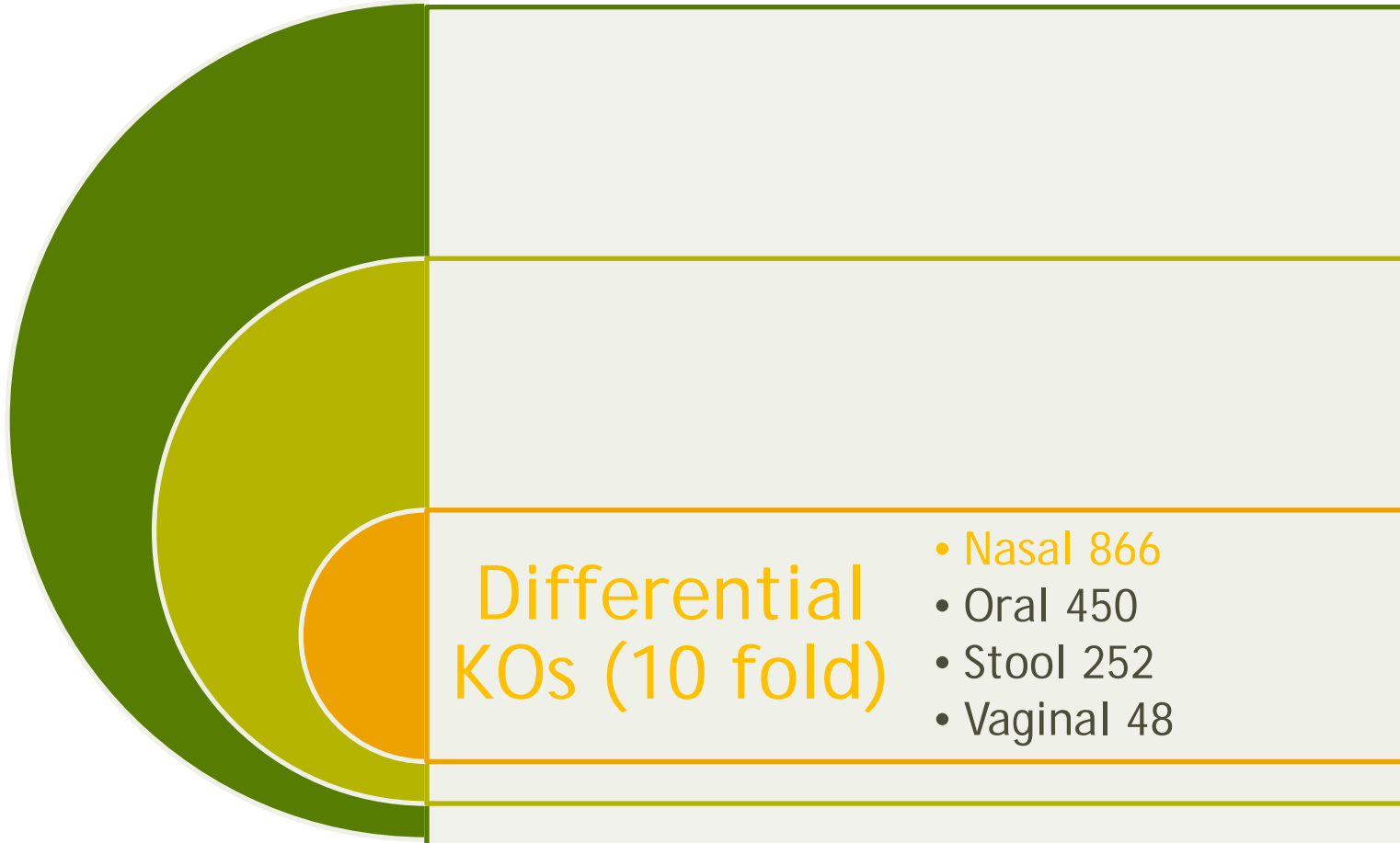
29



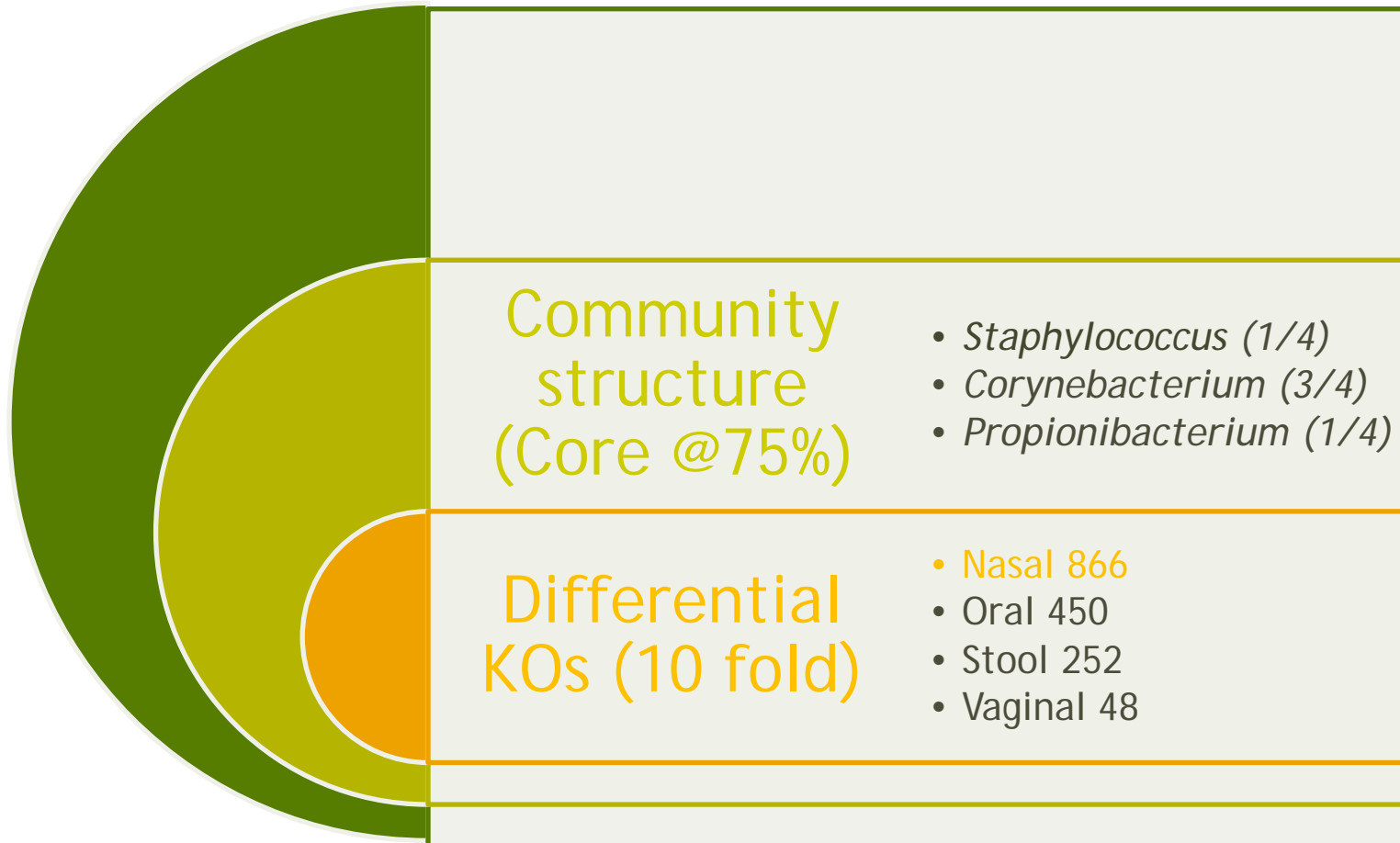
Interrelated and overlapping information



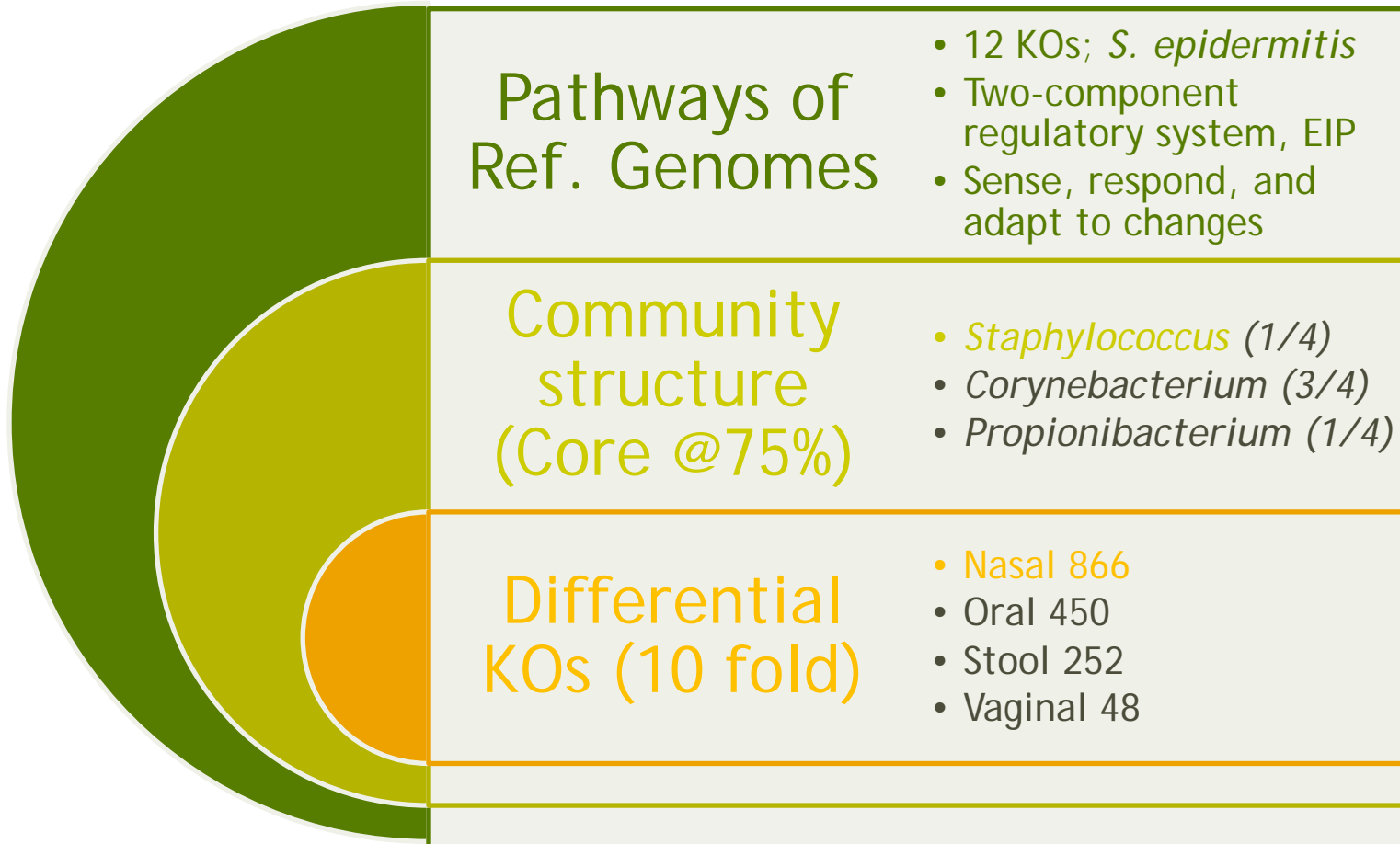
Interrelated and overlapping information



Interrelated and overlapping information



Interrelated and overlapping information



Conclusions

- Community structure membership and complexity varies among body sites (high concordance when different approaches are applied);
- Oral sites have more strains in common than the nasal cavity or the vaginal site;
- USA and EU stool samples have similar strains but different abundance and sequence structure;
- The function is less diverse than the organismal structure and more stable over time;
- The genetic variations can be used to separate body sites from the same region, and to monitor strains dynamics within a body site over time or among different conditions;
- Interrelated and overlapping information from the shotgun metagenomics reads, 16S and reference genomes is a powerful combination for studies on human and other metagenomes.



Acknowledgements

HMP DAWG:

- Data Processing and Mapping (Sarah Young, John Martin)
- Metabolic reconstruction (Sahar Abubucker, Curtis Huttenhover)
- 16S WG (Erica Sodergren, Dirk Gevers)

Washington University Genome Institute:

- George Weinstock
- Hongyu Gao
- Kathie Mihindukulasuriya
- Kristine Wylie
- Yanjiao Zhou
- Sahar Abubucker
- Karthik Kota
- John Martin
- Zhengyuan Wang
- Guohui Zhao

