THE Genome CENTER AT WASHINGTON UNIVERSITY

SHOTGUN METAGENOMIC SEQUENCING AND ANALYSIS AT THE WASHINGTON UNIVERSITY GENOME CENTER

Makedonka Mitreva, Ph.D.

HMP activities at Washington University

1. Generate Catalog of Microbial Reference Genomes (Poster #59)

2. Metagenomic analysis of the microbiota
   - 16S community profiling (Poster 123)
   - Shotgun metagenomics

3. Genome sequencing of viruses and eukaryotic microbes (Poster #117)

4. Other characterization of the microbiome
   - Functional emphasis, e.g. transcription

Progress in all – results from #2

THE Genome CENTER AT WASHINGTON UNIVERSITY

Metagenomic Shotgun vs. 16S rRNA sequencing

16S has biases:

- Degenerate primers
- PCR Amplification
- Databases
- Does not capture viruses and eukaryotes

Most useful is for binning

-Metagenomics:

- Excludes the 16S biases
- Shotgun bias is mainly from sequencing platform
- Provides absolute measurement

Five body sites (16 habitats) in 54 healthy adults

Site	Body habitat	Samples #	Microbial & Human Total Gb	Microbial & Human GB/sample	Microbial Only Total Gb	Microbial Only GB/sample
Oral cavity	Buccal Mucosa	47	512	11	143	3
	Supragingival Plaque	53	583	11	355	7
	Tongue Dorsum	63	729	12	635	10
	Subgingival plaque	8	103	13	34	4
	Palatine tonsils	6	75	12	24	4
	Throat	7	79	11	25	4
	Saliva	5	56	11	13	3
	Hard palate	1	11	11	7	7
	Attached gingivae	6	72	12	43	7
Nasal cavity	Anterior Nares	44	463	11	71	2
Skin	R_retroauricular crea	12	122	10	30	2
	L_retroauricular creas	2	19	9	4	2
Gut	Stool	59	664	11	650	11
Vagina	Mid vagina	2	23	11	4	2
	Vaginal introitus	3	36	12	8	3
	Posterior Fornix	28	329	12	69	2
Total / Avg	**16**	**346**	**3,875**	**11**	**2,114**	**5**

* 103 Illumina runs

THE Genome CENTER AT WASHINGTON UNIVERSITY

Today's analyses are on: 6 body sites in 16 individuals

		Samples	Microbial & Human		Microbial Only	
Site	Body habitat	#	Total Gb	GB/sample	Total Gb	GB/sample
Oral cavity	Buccal Mucosa	16	160	10	45	3
	Supragingival Plaque	16	171	11	104	7
	Tongue Dorsum	16	174	11	151	9
Nasal Cavity	Anterior Nares	16	180	11	27	2
Gut	Stool	16	180	11	176	11
Vagina	Posterior Fornix	8	94	12	20	2
Total	**6**	**88**	**959**	**66**	**523**	**33**

Recent MetaHIT paper* 124 576

* Qin et al., 2010 Nature

THE Genome CENTER AT WASHINGTON UNIVERSITY

Challenge: computational resources (protein searches)

Human screening w/ crossmatch	7 lanes vs HS36	~ 1 day
Blastx vs phylogenetic DBs	7 lanes vs prok, viral, euk, arch	13,000 days
Blastn vs bact genomes	7 lanes vs 3000 genomes	7.400 days
Blat vs RDP	7 lanes vs 450k 16S seqs	1.3 days
Blat reads vs reads	7 lanes vs 7 lanes	36 days
Crossmatch Illumina vs 454	7 lanes vs 2 runs' 16S reads	1000 days
Blastx vs KO set of KEGG	7 lanes vs KO DB	3700 days

Samples from 7 body sites:
7 Illumina lanes = 49 M reads
= 3.7 GB of paired 75 base reads

Times are for 7 samples using 1 core
Typically use >200 cores:

…this still takes >2 months for some tasks using data that took 0.3 months to produce

Possible approaches:

Hardware accelerations
- Accelerating BLASTx on GPUs

Software Accelerations
- Alternative algorithms to BLAST

Grids
- RENCI Teragrid Science Portal (UNC)

Clouds
- Internal cloud based on open standards (Hexagrid)

THE Genome CENTER AT WASHINGTON UNIVERSITY

Solution – Alternative algorithms (Collaboration with Real Time Genomics; Poster #15)

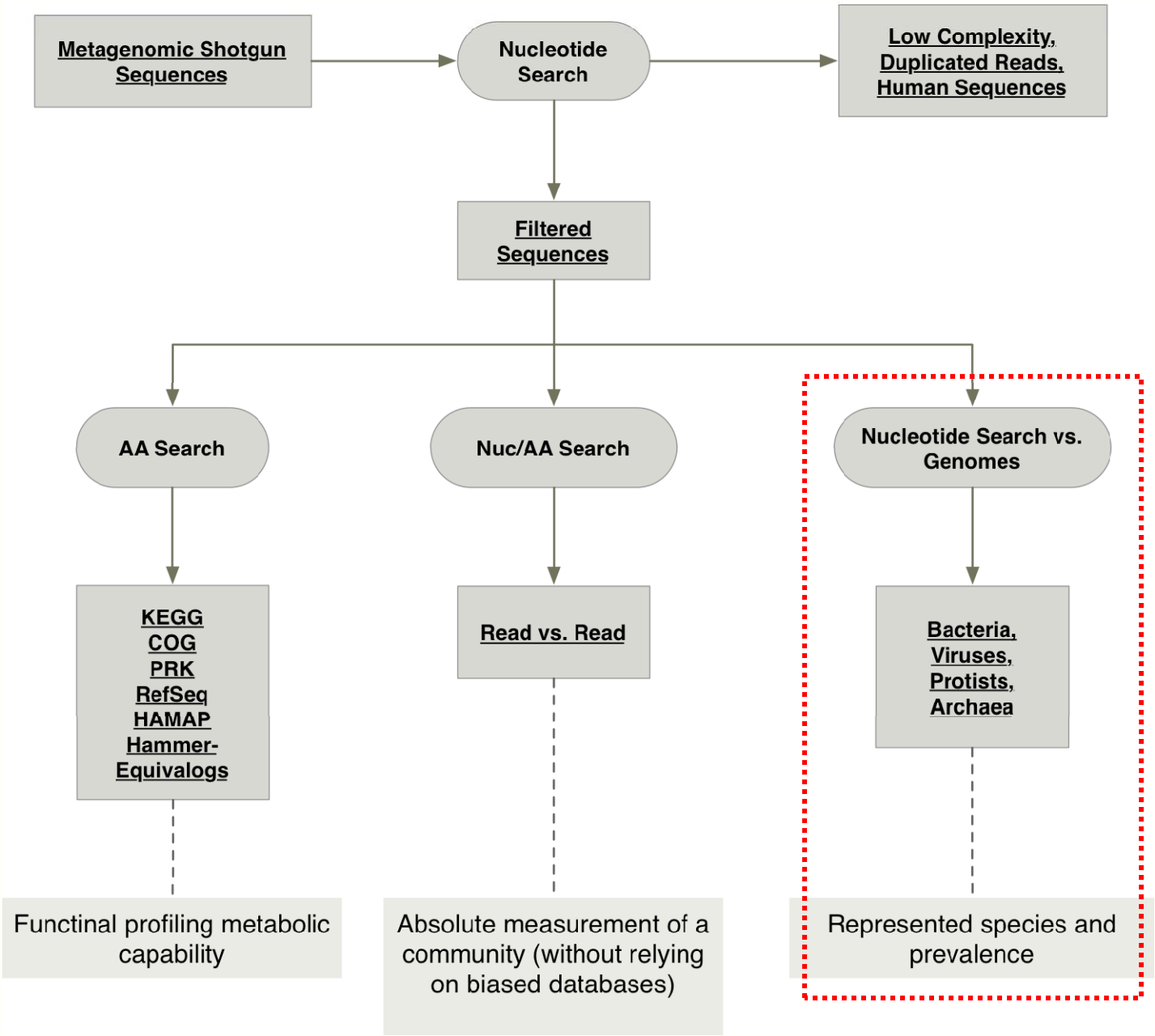
Task	Tool	Reference	Throughput	Processing Rate
High Sensitivity Contaminant Filtering	Mapf	Human genome	100M-200M reads/run	3M rds/hr/core
Read Mapping	Map	3,000 microbial genomes	100M- 200M reads/run	10M rds/hr/core
Translated Protein Search	Mapx	NR protein or KEGG	20M-40M reads/run	>300X vs. blastx

Ability to perform analyses in a timely & cost-effective manner

REALTIME GENOMICS

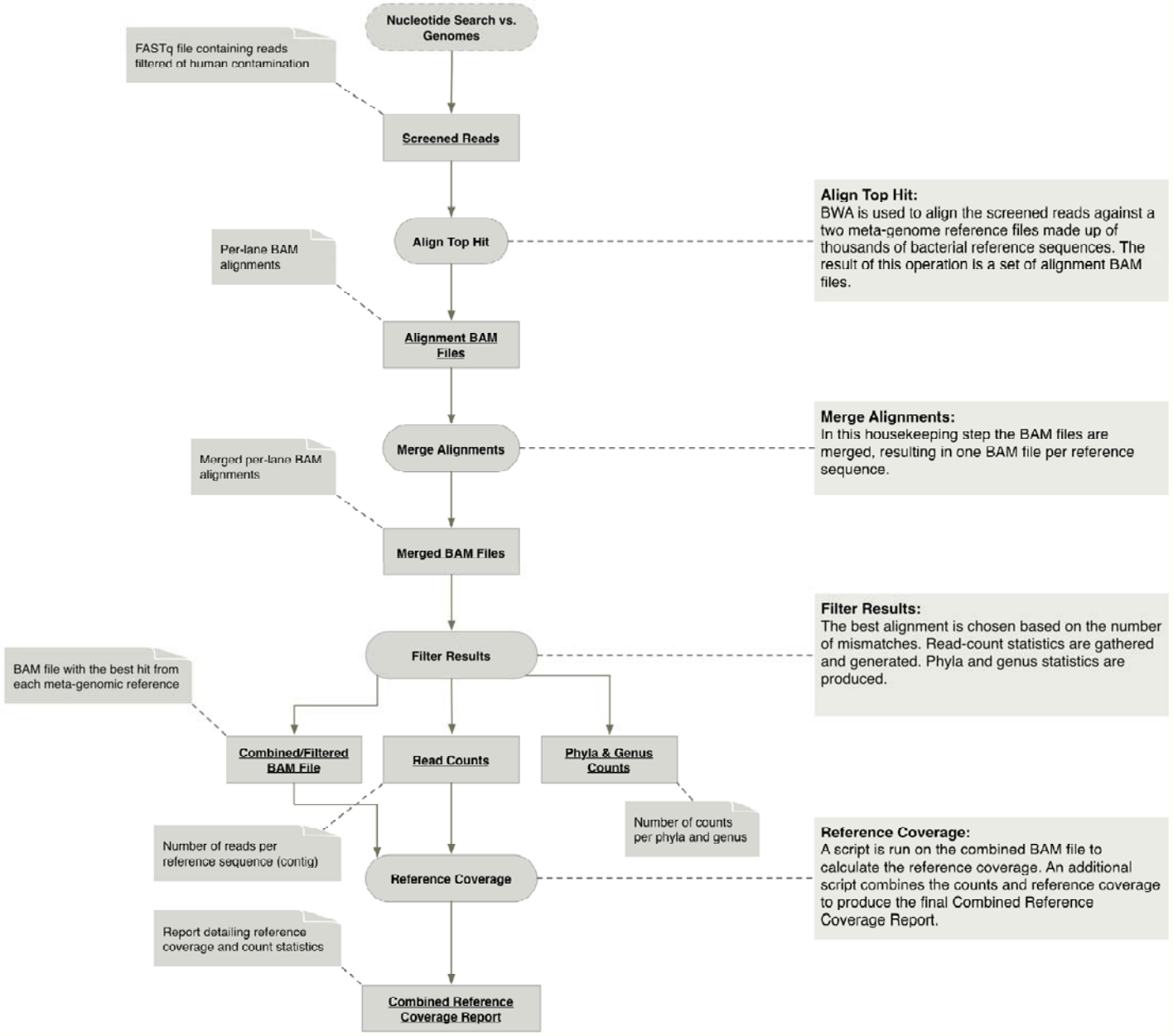
THE Genome CENTER AT WASHINGTON UNIVERSITY

Analyses workflow



THE Genome CENTER AT WASHINGTON UNIVERSITY

Detailed workflow: nucleotide search vs. reference genomes



THE Genome CENTER AT WASHINGTON UNIVERSITY

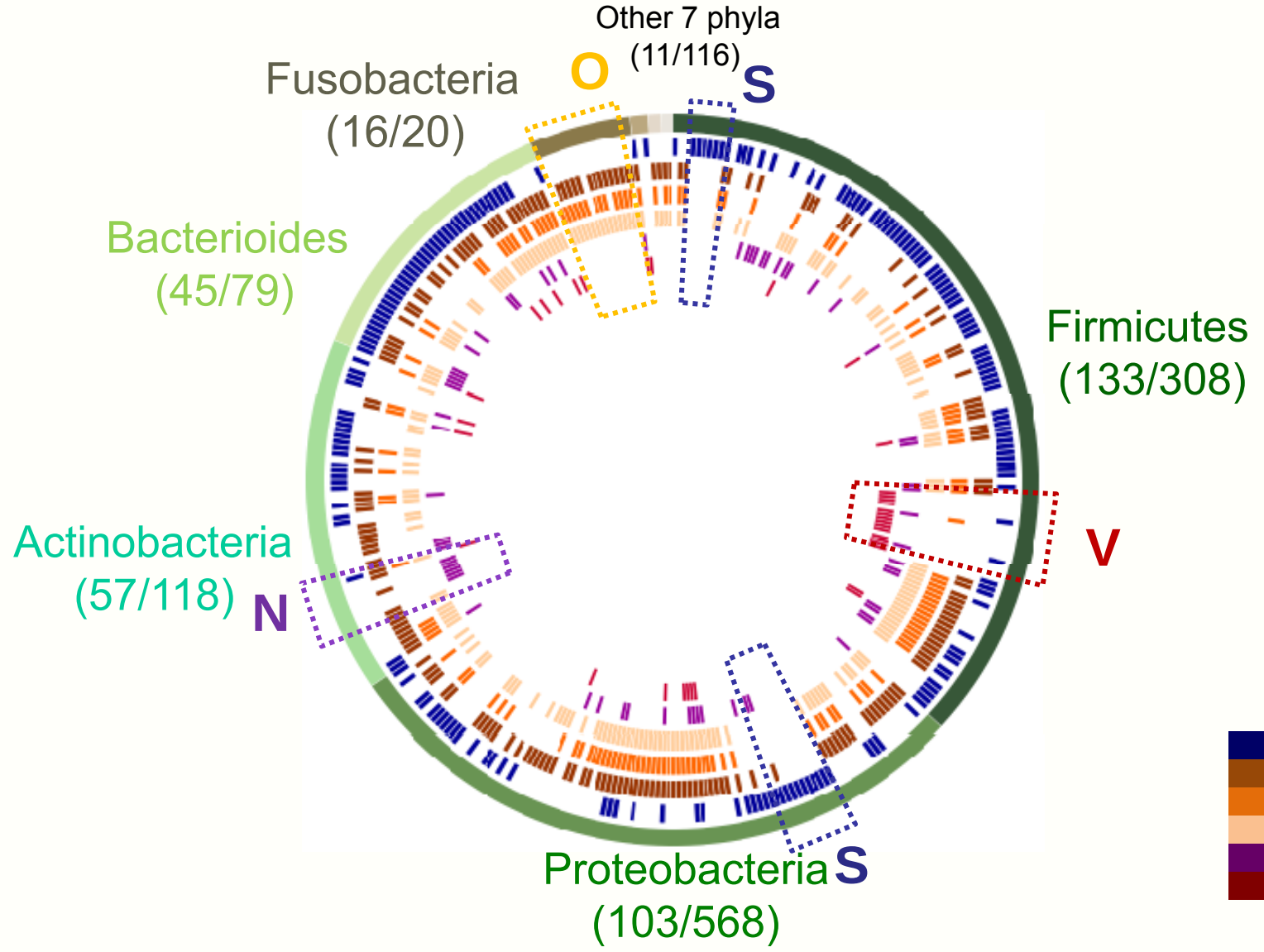
Comparison of taxonomic variations in the human microbiome

- BWA mapping of all the microbial reads to the ~1035 reference species (1,209 strains)

		Samples	Microbial Only		Aligned
Site	**Body habitat**	**#**	**Total reads**	**Reads/sample**	**%**
Oral cavity	**Buccal Mucosa**	16	243,450,640	15,215,665	16
	Supragingival Plaque	16	863,119,159	53,944,947	26
	Tongue Dorsum	16	1,438,867,394	89,929,212	25
Nasal Cavity	**Anterior Nares**	16	84,018,993	5,251,187	11
Vagina	**Posterior Fornix**	8	79,036,696	9,879,587	17
Gut	**Stool**	16	1,606,463,248	100,403,953	50
Total / Avg	**6**	**88**	**4,314,956,130**	**45,770,759**	**24.167**

THE Genome CENTER AT WASHINGTON UNIVERSITY

Comparison of taxonomic variations in the human microbiome



Across all habitats we detected members of 12 bacterial phyla, however 97% of the members were related to 5 phyla (365 species @ 1% b & 0.01X depth of coverage)

THE Genome CENTER AT WASHINGTON UNIVERSITY

Common and marker species

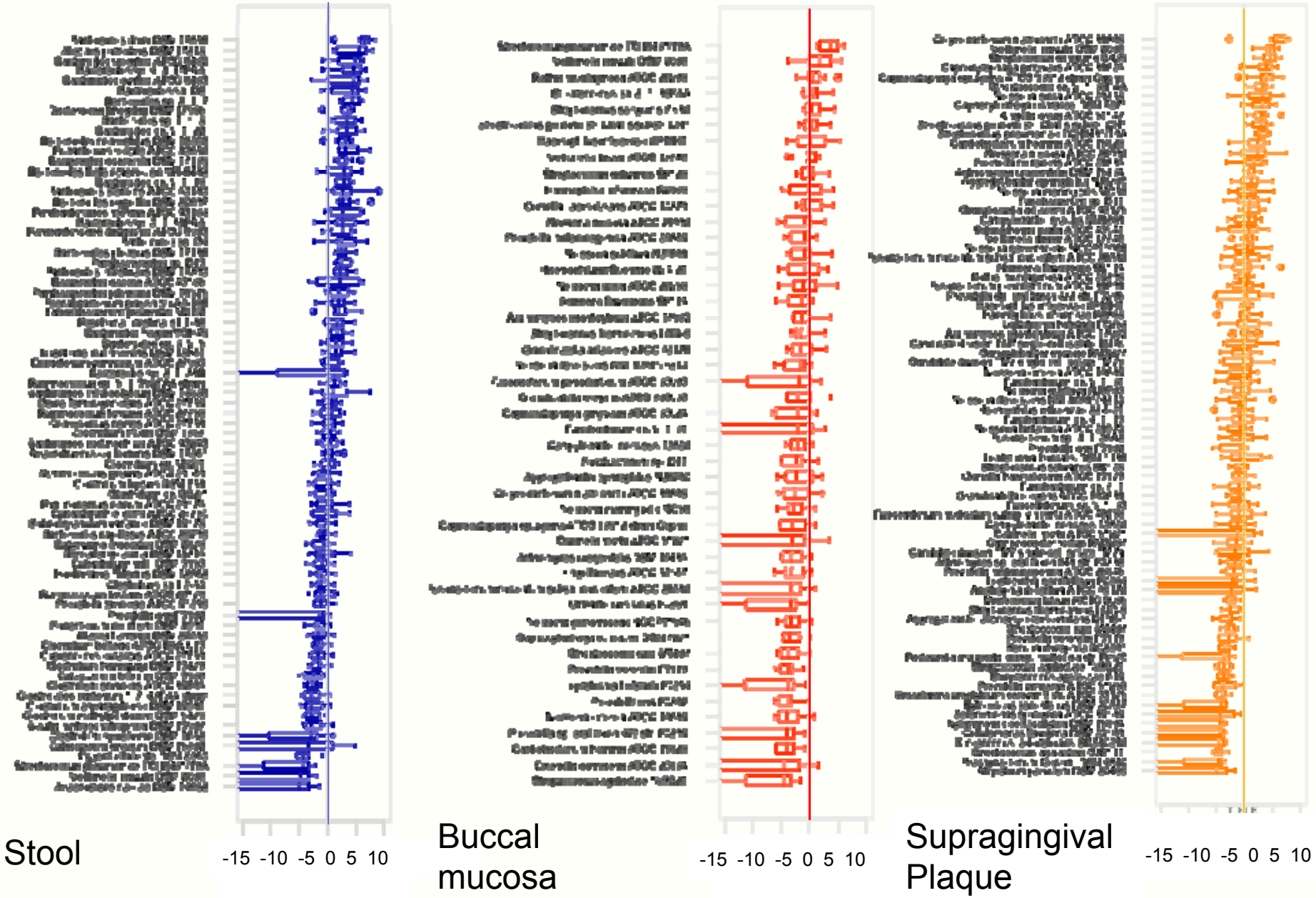
		Common Species			Marker Species			
		100%*	75%	50%	100%	75%	50%	Total Species
Gut	**Stool**	0	71	83	0	69	74	204
Oral	**Buccal Mucosa**	0	42	63	0	0	0	139
	Supragingival Plaque	5	70	97	3	7	12	201
	Tongue Dorsum	2	79	94	0	16	8	195
Nasal	**Anterior Nares**	0	4	6	0	4	5	68
Vaginal	**Posterior Fornix**	0	4	10	0	4	10	37

* % individuals in w hich the species is observed.

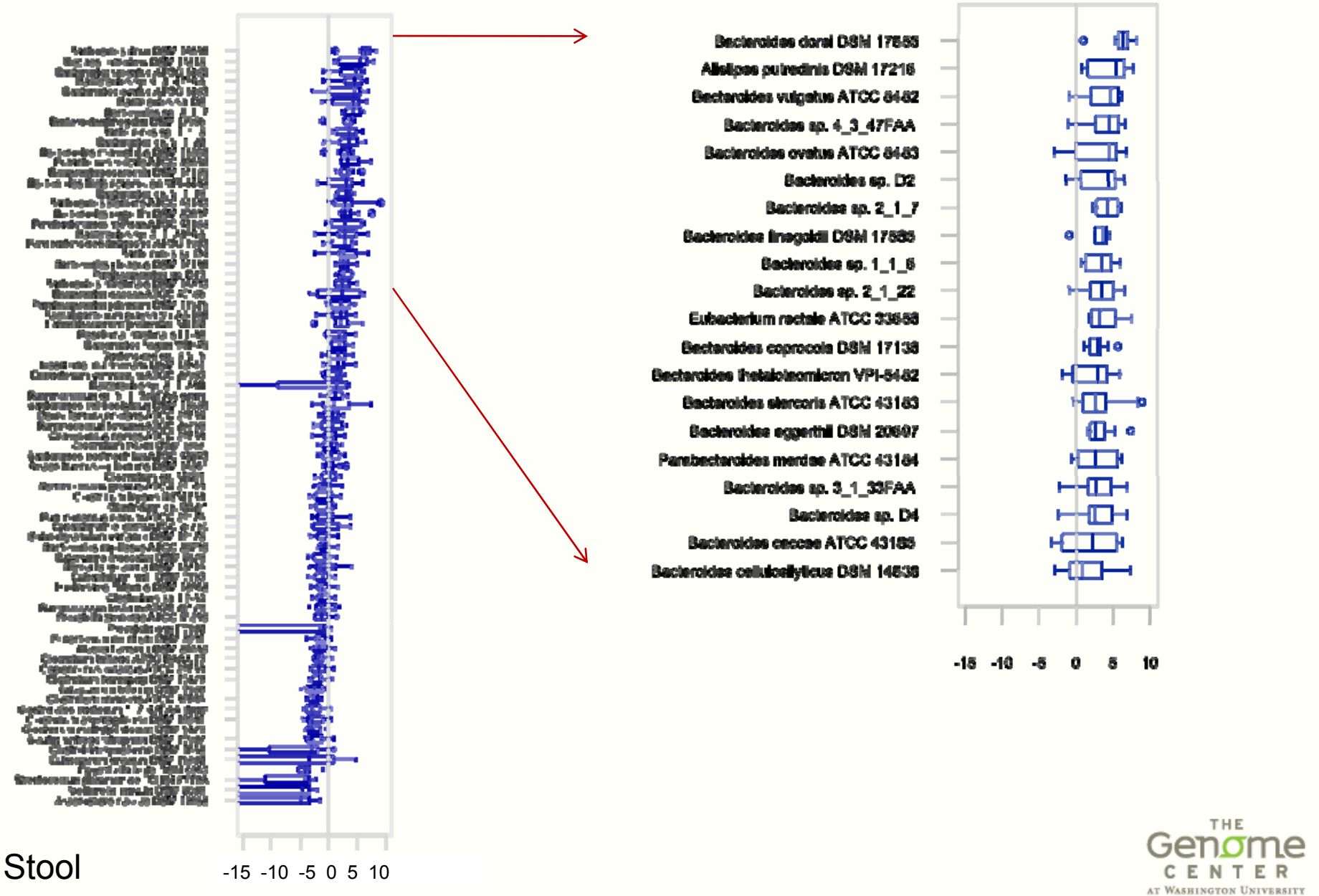
Gut	Oral cavity			Nasal Cavity	Vagina
Stool	**B. mucosa**	**Supragingival Plaque**	**Tongue Dorsum**	**Anterior Nares**	**Posterior Fornix**
Alistipes (1)	-	Actinomyces (1)	Abiotrophia (1)	Corynebacterium (3)	Lactobacillus (4)
Anaerotruncus (1)		Aggregatibacter (1)	Anaerococcus (1)	Staphylococcus (1)	
Bacteroides (26)		Brachybacterium (1)	Atopobium (2)		
Blautia (2)		Kocuria (1)	Parvimonas (1)		
Butyrivibrio (1)		Kytococcus (1)	Prevotella (1)		
Clostridiales (1)		Micrococcus (1)	Selenomonas (1)		
Clostridium (11)		Pasteurella (1)	Shuttleworthia (1)		
Coprococcus (2)			Streptococcus (8)		
Dorea (2)					
Eggerthella (1)					
Eubacterium (6)					
Faecalibacterium (2)					
Holdemania (1)					
Parabacteroides (4)					
Roseburia (2)					
Ruminococcus (5)					
Subdoligranulum (1)					

THE Genome CENTER AT WASHINGTON UNIVERSITY

Relative abundance (log_{10}) of frequent microbial genomes



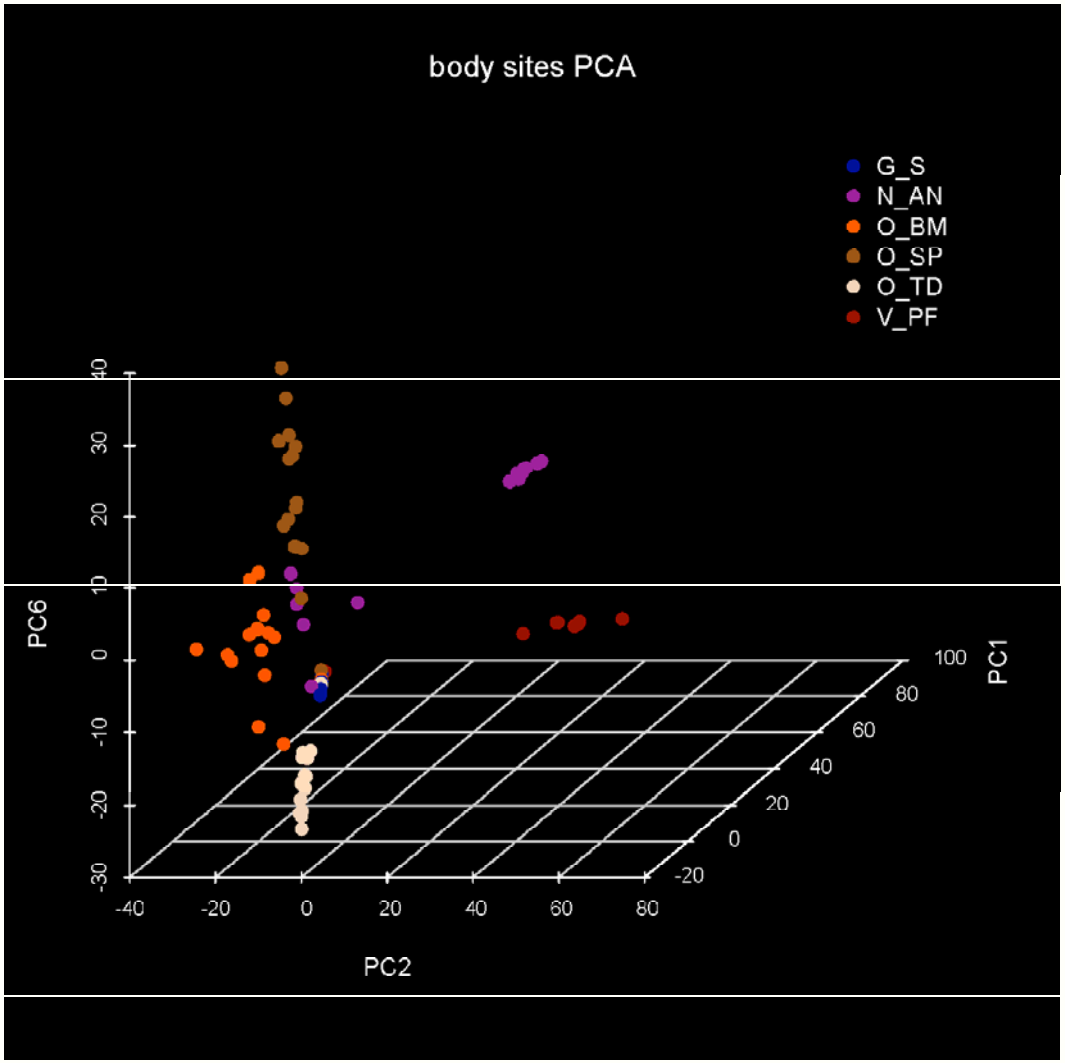
Relative abundance (log_{10}) of frequent microbial genomes

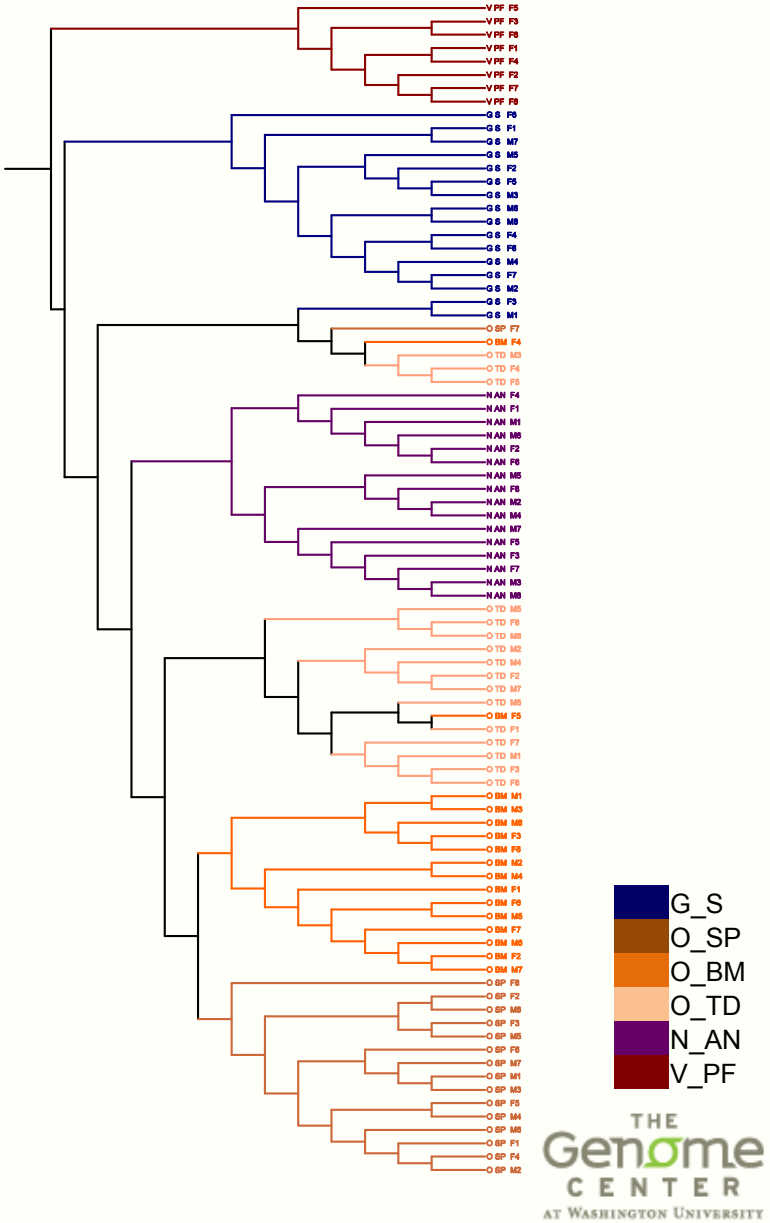


Stool

THE Genome CENTER AT WASHINGTON UNIVERSITY

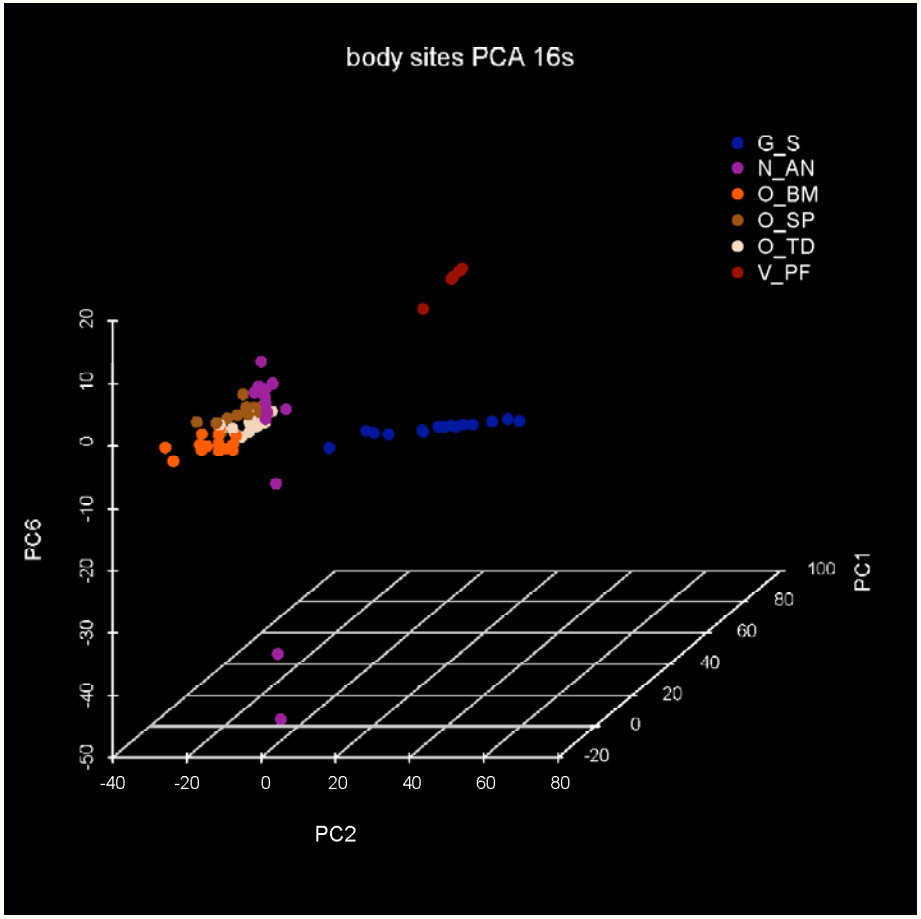
Clustering of communities (based on breadth x depth)

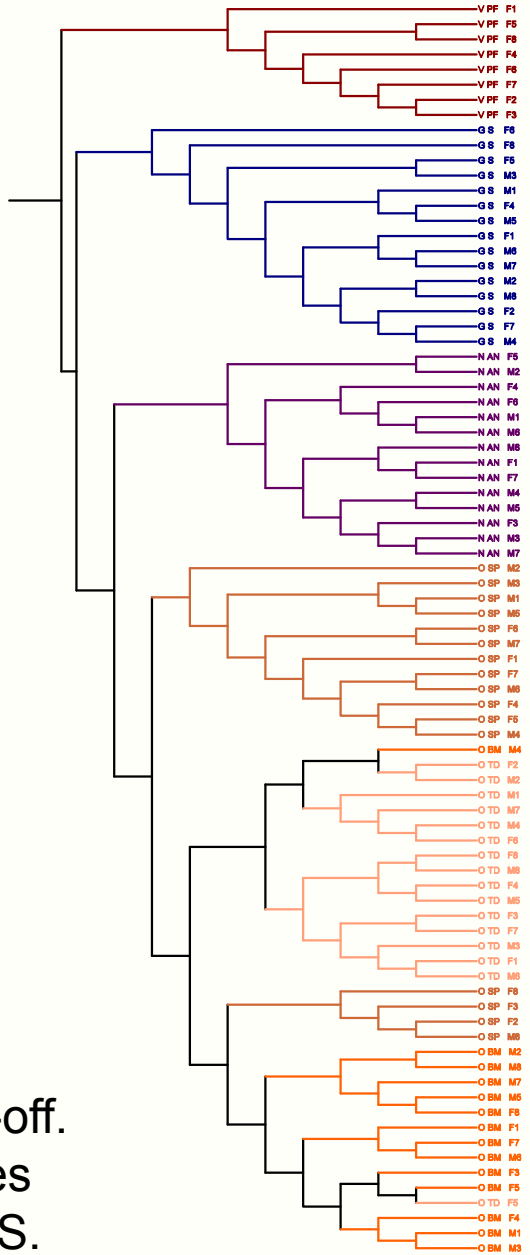




THE Genome CENTER AT WASHINGTON UNIVERSITY

Clustering of communities (based on 16S)

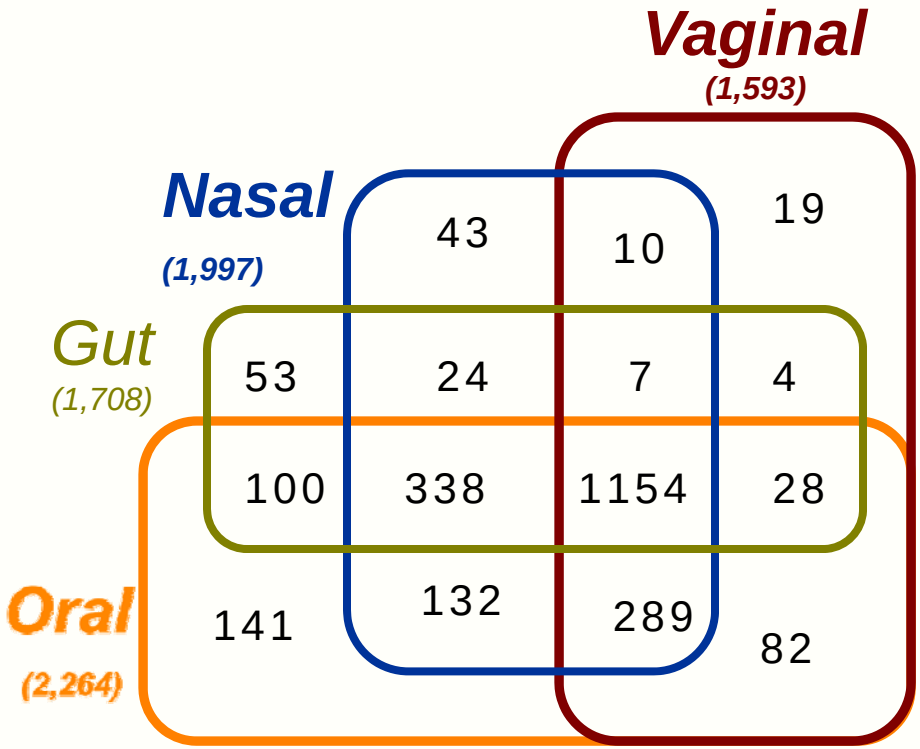


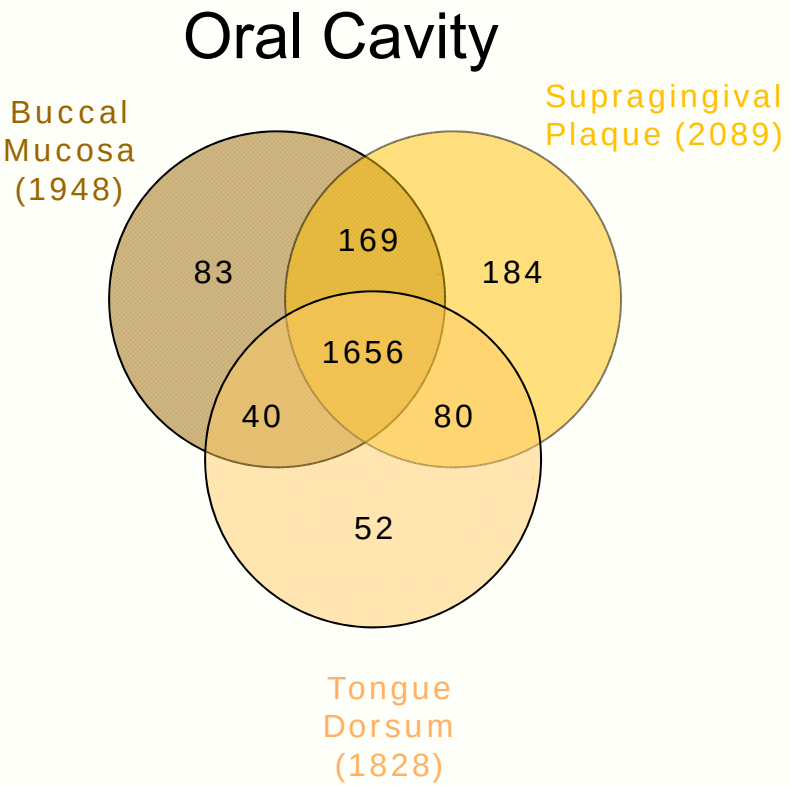


- 16S classification to genus level at 0.8 cut-off.
- It can not separate the three orals as it does not go to species level compared to the WGS.

THE Genome CENTER AT WASHINGTON UNIVERSITY

Metabolic profiling - unique and shared KOs



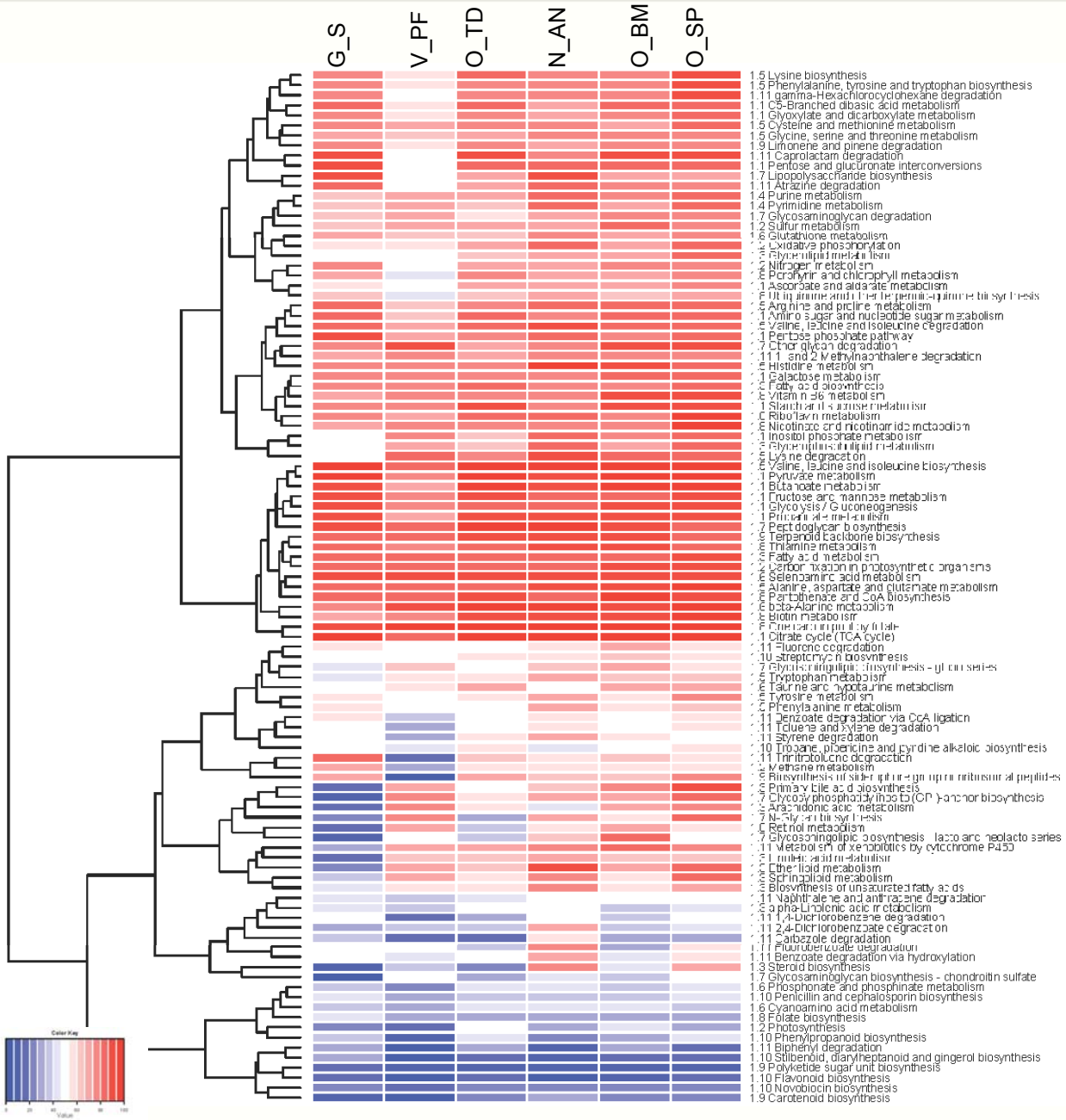


Identified 2,423 KOs

		Species
Gut	**Stool**	**204**
Oral	**Buccal Mucosa**	139
	Supragingival Plaque	201
	Tongue Dorsum	195
Nasal	**Anterior Nares**	68
Vaginal	**Posterior Fornix**	37

RTG/Mapx, W8, a3, P72, E0.01

Pathway coverage (the 66 samples from 6 body sites)



- 153 pathways were detected*
- 103 pathways are active**

*Mapx, W8, a3, P72, E0.01

**Pathway is considered to be active when it has at least 50% of all the Kos represented in the 800 bacterial genomes.

THE Genome CENTER AT WASHINGTON UNIVERSITY

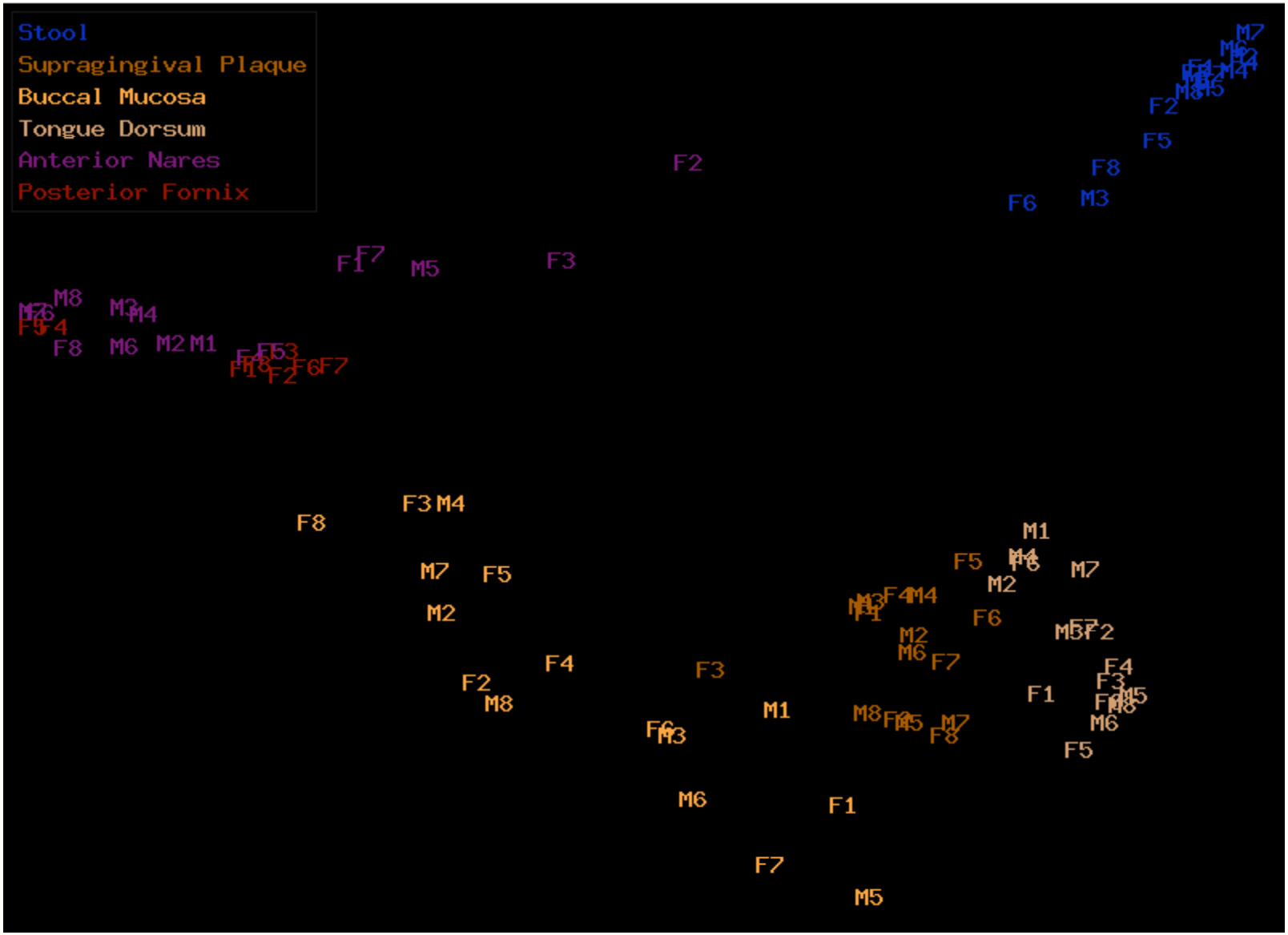
Sequence Phylogeny – taxonomy free comparisons

• Phylogenetic approach solves problems with fragmentary, non-overlapping reads;

• Absolute view of community complexity without biases from existing databases

Approach:
- based on sequence composition, w15-32, s11-30 (RTG phylogeny module)
- metagenomics read trees, similarity distance and hierarchical clustering
-generate k-mer profile, exclude k-mers that ocure more than once,
- compare unique k-mers per sample among samples
Changing parameters the clustering retains per body site

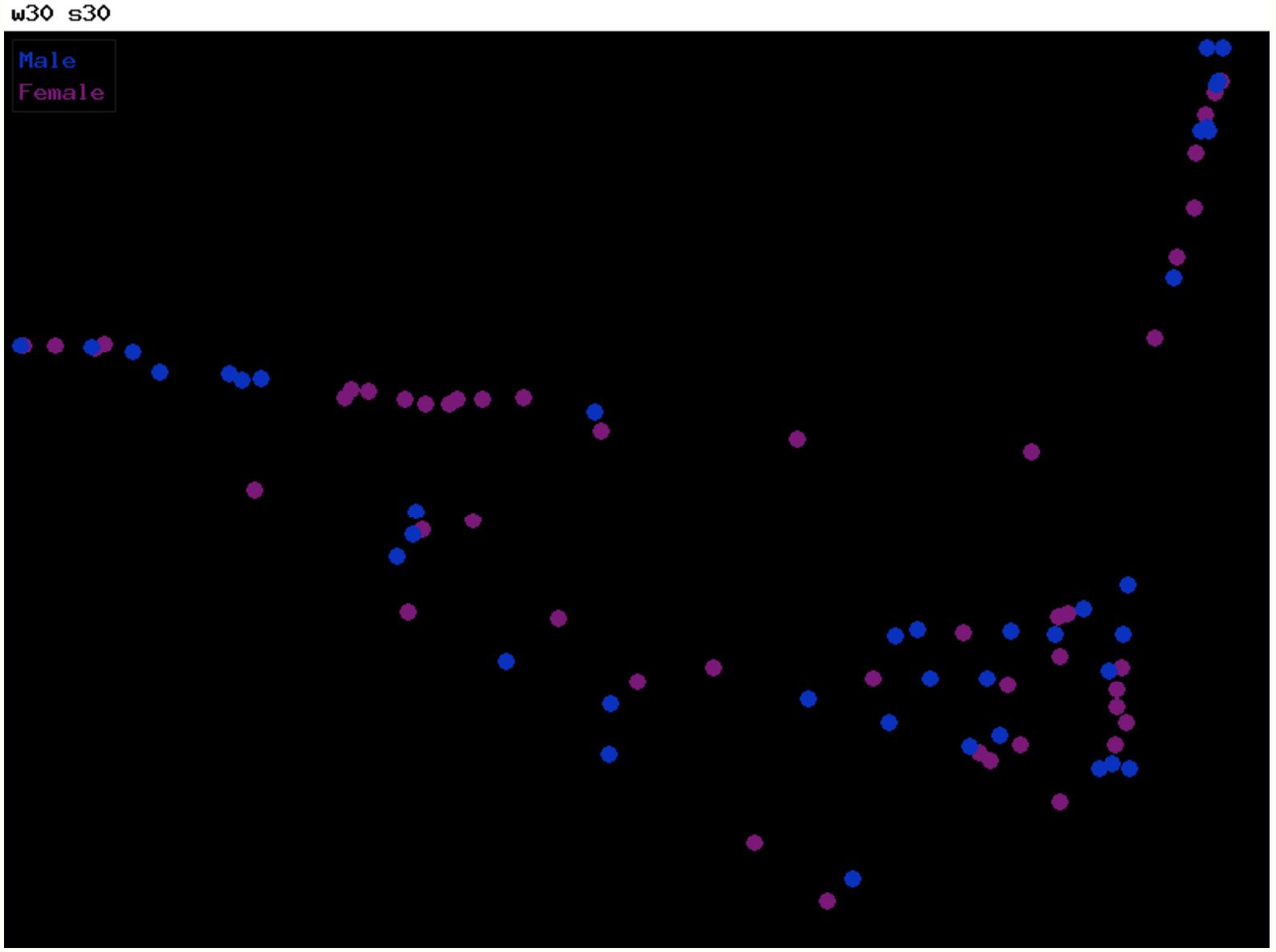
Phylogenetic clustering per body site



W15-30
S11-30

THE Genome CENTER AT WASHINGTON UNIVERSITY

Phylogenetic clustering per gender

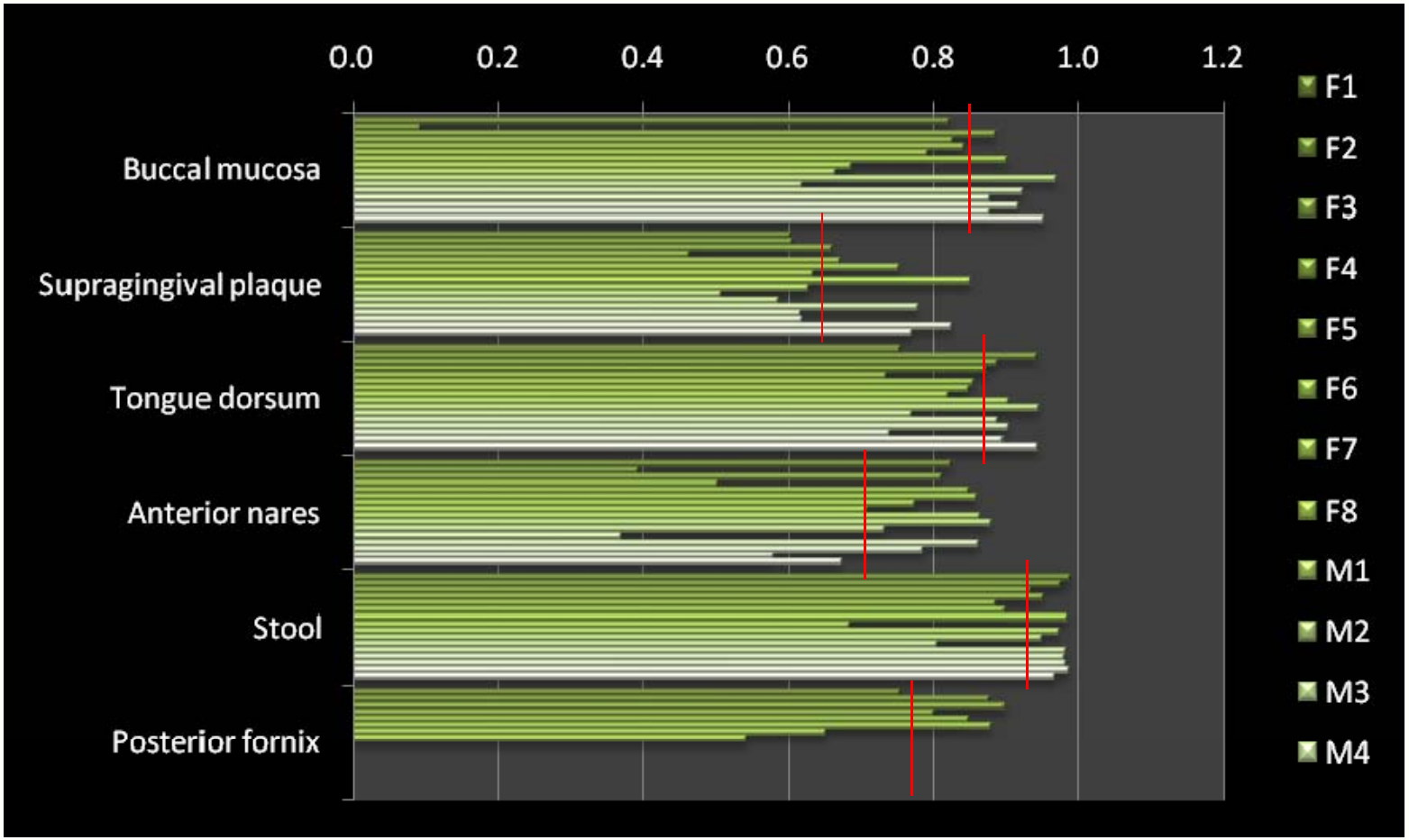


THE Genome CENTER AT WASHINGTON UNIVERSITY

Concordance in community composition
Targeted 16S vs. WGS/16S

Is there a positive correlation? √

How large is the difference? 6.6 - 9.3



Pearson's correlation: degree of linear relationship between 16S and WGS/16S

THE Genome CENTER AT WASHINGTON UNIVERSITY

Validation of novel 16S genes (New Species)

		16S		Supported
		Unclassified	Supported by WGS	(%)
Oral	BM	174	119	68
	SP	1370	1281	94
	TD	2126	2110	99
Nasal	AN	23	5	22
Gut	S	1342	1312	98
Vagina	PF	13	6	46

- Shotgun reads confirmed 4,833 16S out of 5,048 novel 16S genes based on 100% identity over 100% length of the WGS read.
- Rate of new species discovery, 70%, very good due to good shotgun coverage of rare organisms.
- Follow up analysis to increase confidence: Sequence composition, secondary structure prediction, region evolution (loops evolve faster than stems), etc.

THE Genome CENTER AT WASHINGTON UNIVERSITY

Ongoing challenges: experimental work and analysis

- Good progress establishing uniformity, standards

- Increase confidence in presence of rare organisms important and distinguishing them from “noise”

- GCWU scale-up brings issues into focus
  - oInstrument, software, and reagent weaknesses
  - oData and analysis pipelines for human/day scale

- Can’t use old methods for analyses of metagenomic, human, etc data
  - oVarious approaches to speed up

- Scale-up in computational infrastructure still a learning process for both vendors and GC

THE Genome CENTER AT WASHINGTON UNIVERSITY

Acknowledgements – WashU Genome Center Staff

Groups: Resource Bank & Production, Analysis, LIMS/Systems/Automation, Management

Otis Hall, Alex Akerberg, Brandi Herter, Lucinda Fulton, Kim Delehaunty, John Martin, Karthik Kota, Kathie Mihindukulasuriya, Kristine Wylie, Sahar Abubucker, Yanjiao Zhou, Zhengyuan Wang, Guohui Yao, David Dooling, Creig Pohl, Gary Steihr, Adam Dukes, Jim Eldred, Nathan Nutter, Scott Smith, Richard Wilson, Erica Sodergren, George Weinstock and …… . tremendous Production & Info Team

THE Genome CENTER AT WASHINGTON UNIVERSITY