Finding Biology in the Human Microbiome

George Weinstock

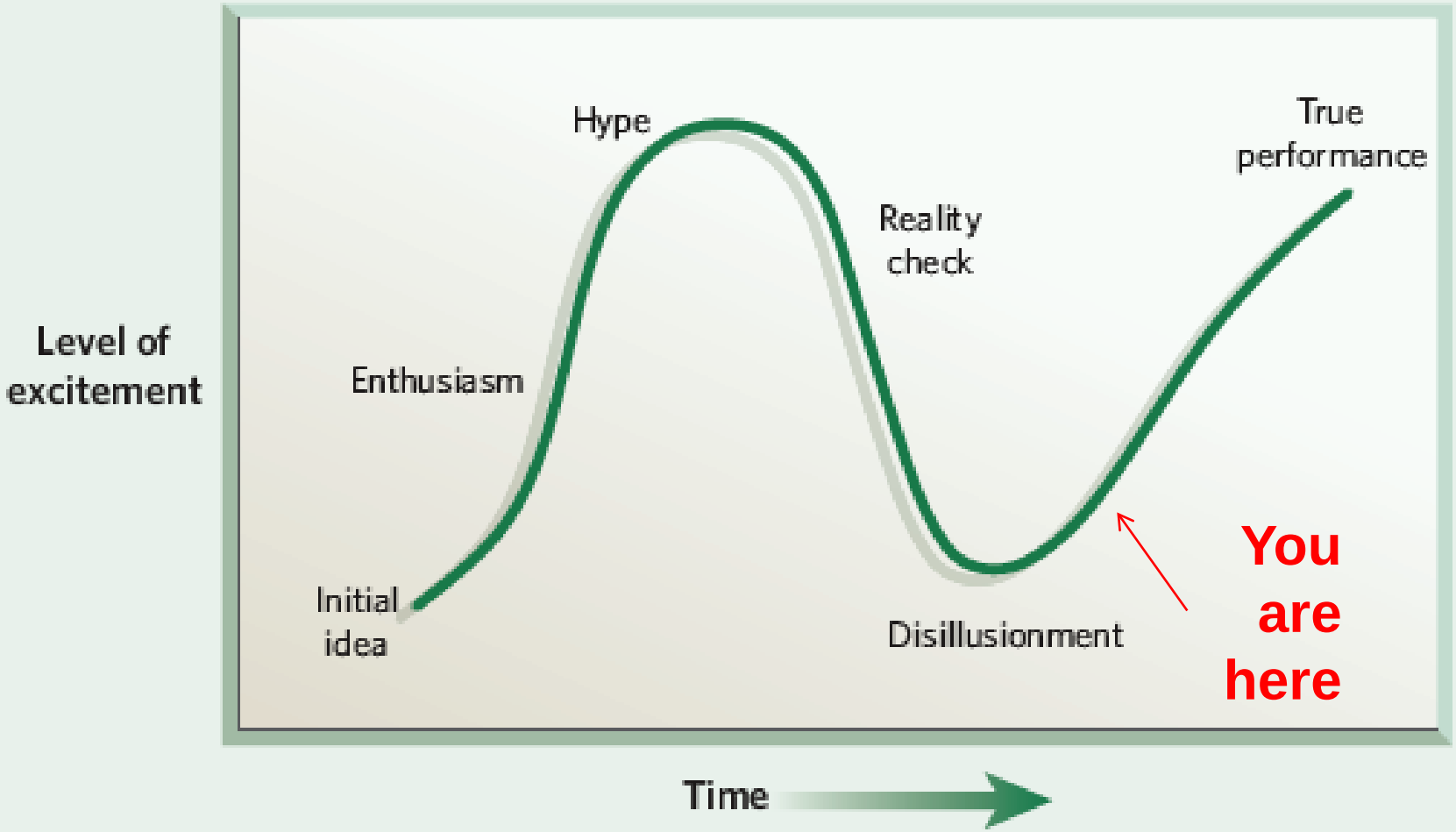
What's next for the Human Microbiome?

George Weinstock

Metagenomics Unfolds

Setting Up → Descriptive Phase → Hypothesis Testing

Metagenomics Unfolds



Setting Up → Descriptive Phase → Hypothesis Testing

Human Microbiome Research Thoughts

- St. Louis (6 months ago) showed a wide-ranging field
- Vancouver shows momentum and maturation continues
- Analogy to Human Genome Project?
  - Probably not a "reference" microbiome
- Distinguish a "healthy" from a "perturbed" microbiome?
  - Genomic, Transcriptomic, Proteomic, Metabolomic features
  - Ecological concepts apply to define disease?
- Involvement of clinical concepts early in the project
  - Close interactions between clinicians and basic scientists
  - Massive amount of clinical samples in play; many clinical studies
  - *Opportunities for new strategic and funding models early on*?

Sequencing Technology

- HMP
  - 3000 reference genomes, euks and viruses too
  - Metagenomic 16S: 50 million reads at first data freeze
  - Metagenomic shotgun: >7 , ~10^{11} reads at first data freeze
- 454 for 16S, 200 samples/run
- Illumina for shotgun
  - Then: GAIIx @ 40 Gb/run
  - Now: HiSeq @ 500 Gb/run and increasing
- Coming: PacBio, 454 Junior, Ion Torrent, MiSeq, SOLiD

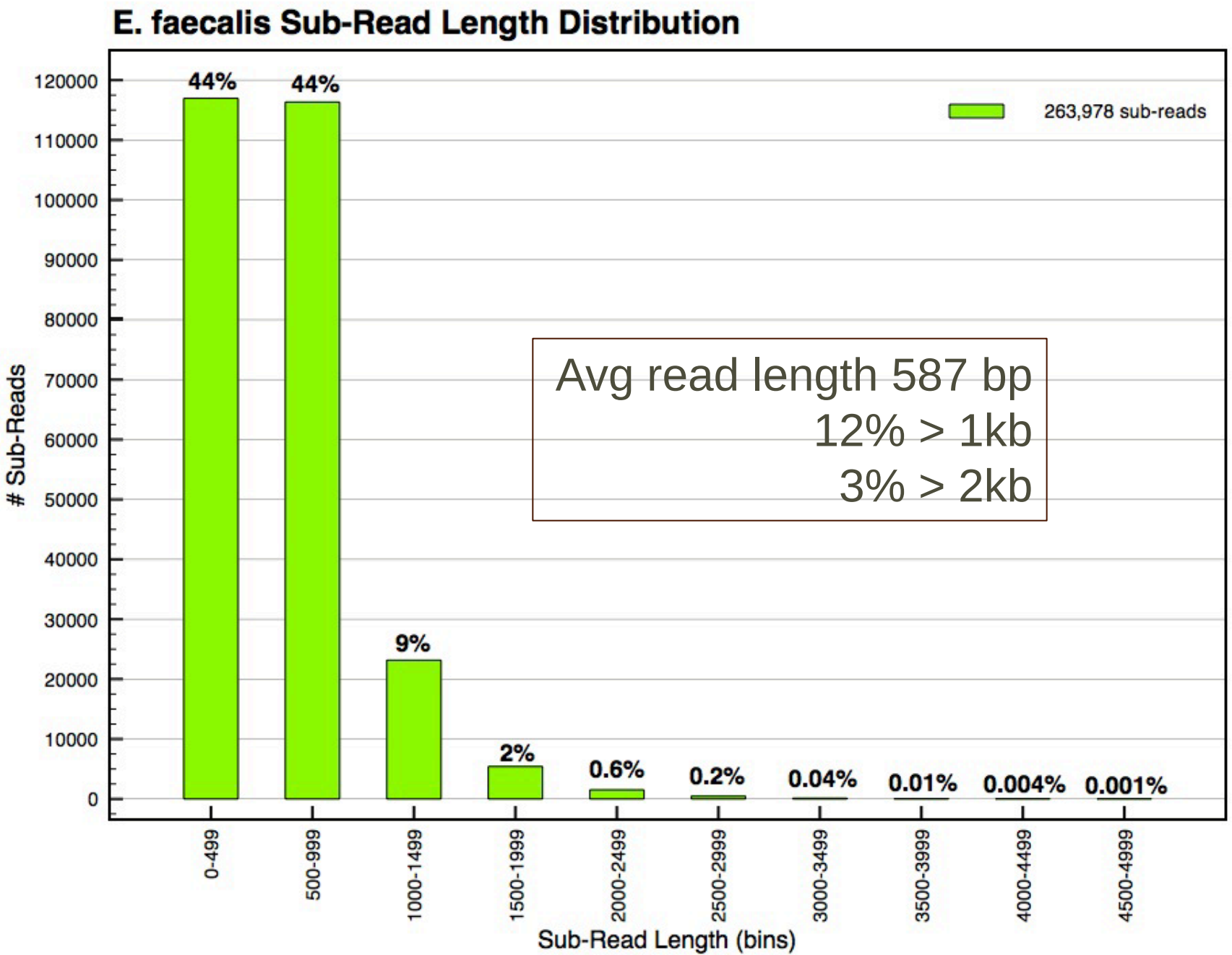
Early Pac Bio Experiences

- What's special?
  - Longer reads: assembly, novel organism/virus detection, 16S
  - Shorter run time: flexible throughput, match experiment better
- Whole genome reference sequencing
  - Enterococcus faecalis example
- 16S sequencing
  - Whole 16S gene

E. faecalis genome sequencing with Pac Bio

Alignment All Reads

Percent of Reference Bases Covered	**93% (Ref = V583)**
Average Coverage Depth	48x
Number of Gaps	50
Average Gap Length	4464


E. faecalis Sub-Read Length Distribution
263,978 sub-reads
Sub-Reads
120000
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
44%
44%
9%
2%
0.6%
0.2%
0.04%
0.01%
0.004%
0.001%
0-499
500-999
1000-1499
1500-1999
2000-2499
2500-2999
3000-3499
3500-3999
4000-4499
4500-4999
Sub-Read Length (bins)
Avg read length 587 bp
12% > 1kb
3% > 2kb

E. faecalis genome sequencing with Pac Bio

Alignment All Reads

Percent of Reference Bases Covered	**93% (Ref = V583)**
Average Coverage Depth	48x
Number of Gaps	50
Average Gap Length	4464

Alignment Reads >2kb

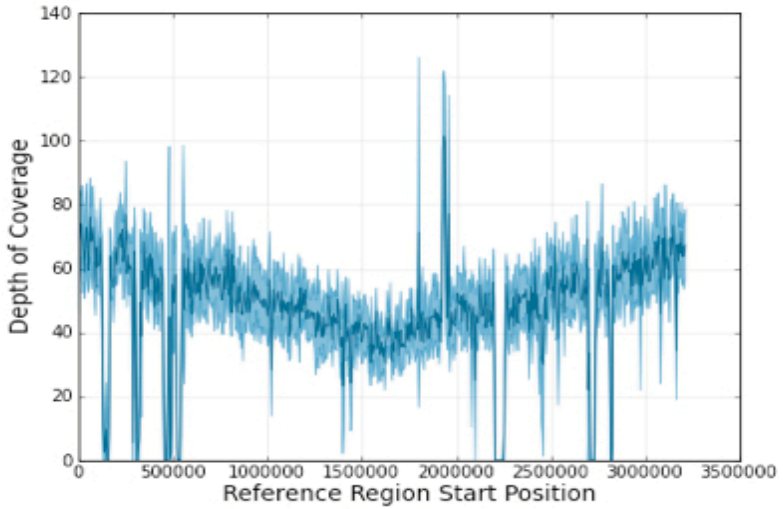
Percent of Reference Bases Covered	**75%**
Average Coverage Depth	1.6x
Number of Gaps	380
Average Gap Length	2144

Coverage Metrics

Report Summary

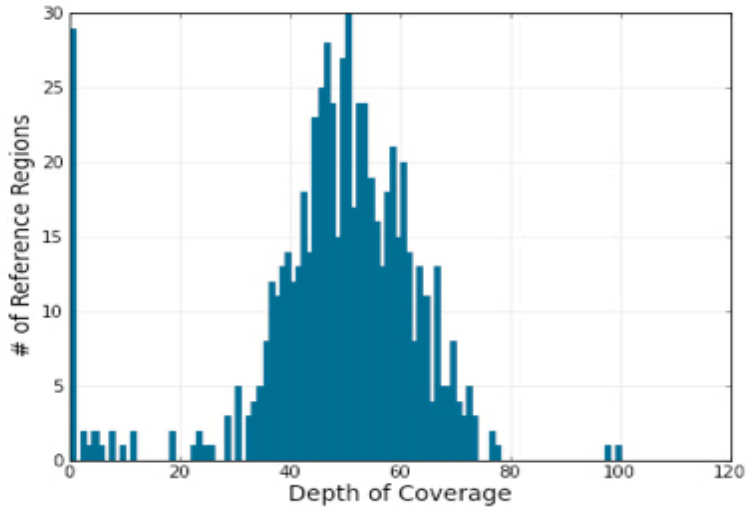
Mean Depth of Coverage	48.15
% Missing Bases	6.94

Depth Of Coverage Across Reference



Observed depth of coverage across E_faecalis_V583_NC_004668_1 (window size = 5004bp).

Depth Of Coverage Histogram



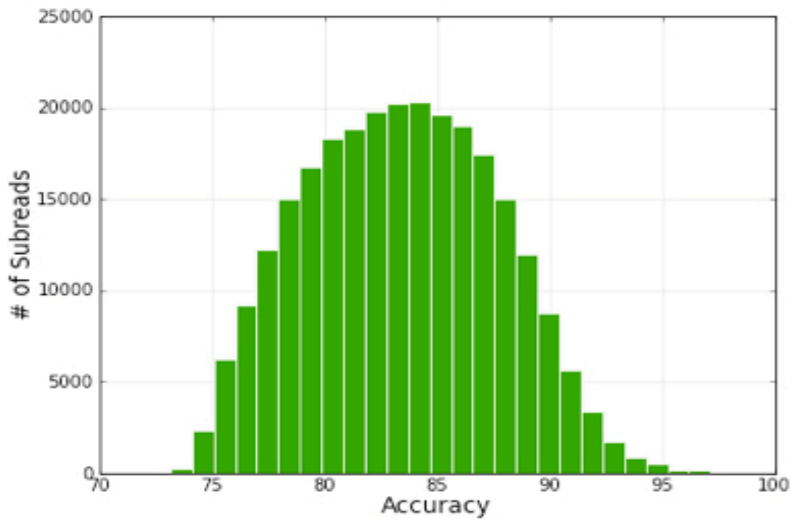
Depth of Coverage Distribution

Read Length & Accuracy

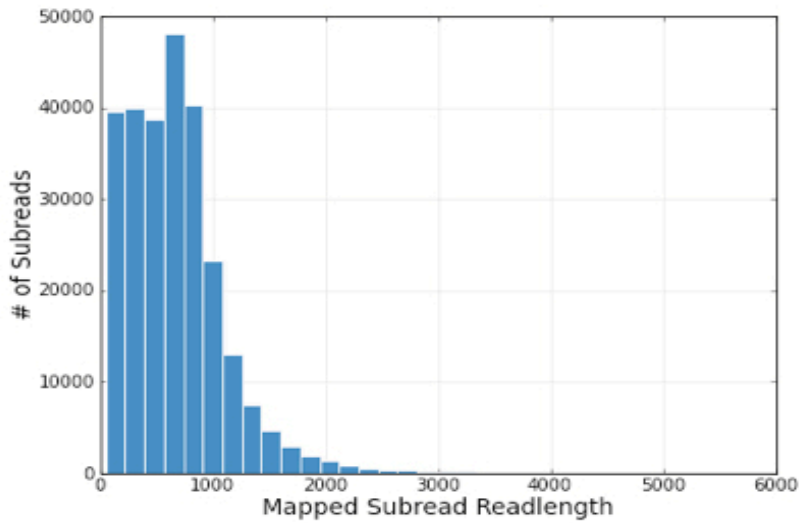
Report Summary

# of Post-Filter Reads	266730	# of Mapped Reads	151107
# of Mapped Bases	213650285 bp	95th Percentile Mapped Readlength	3616 bp
Maximum Mapped Readlength	6503 bp	Mean Mapped Readlength	1414 bp
# of Mapped Subreads	263978	Mean Mapped Subread Readlength	657 bp
Mean Mapped Subread Accuracy	83.36%		

Accuracy Histogram



Mapped Subread Readlength Histogram



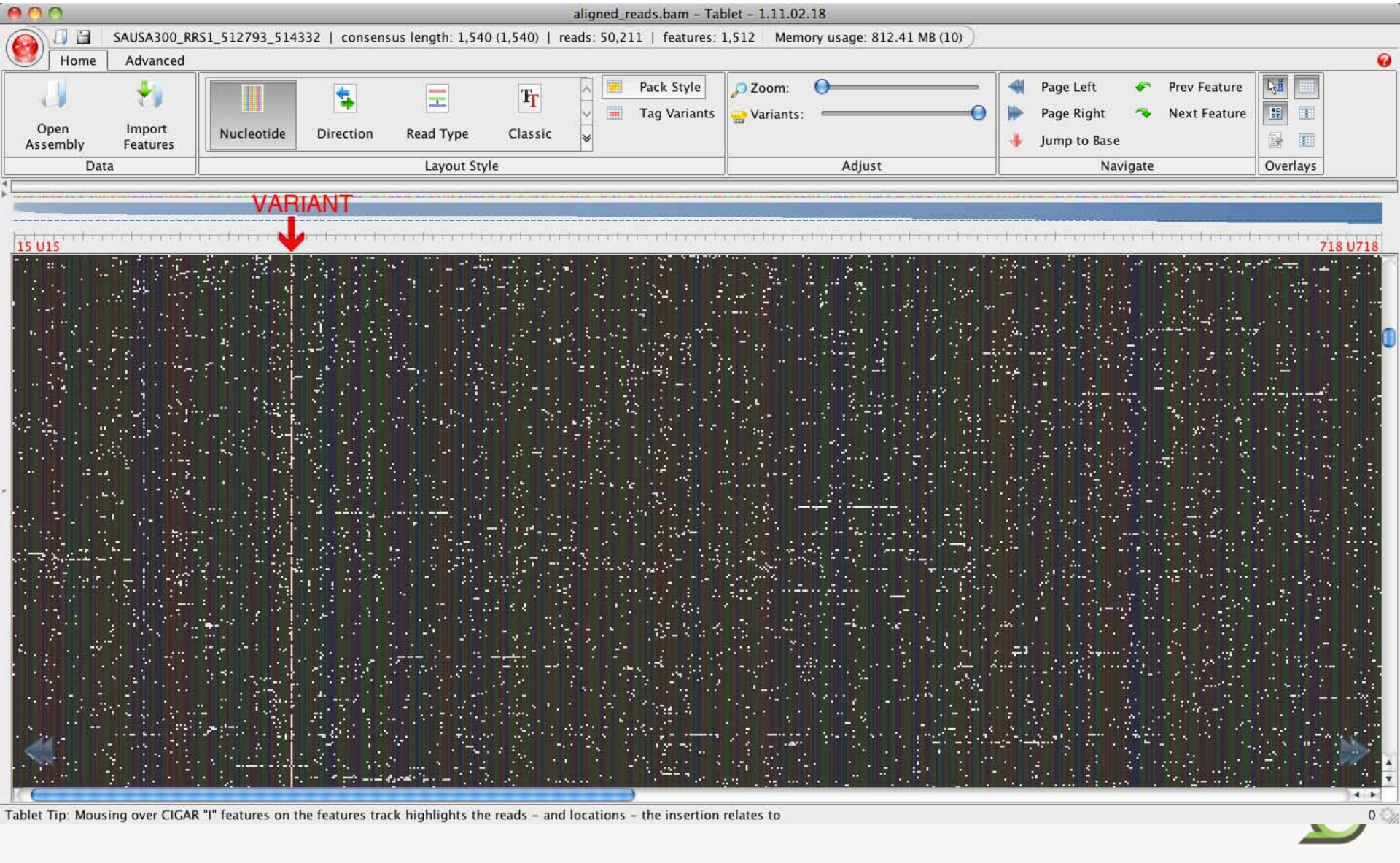
Allora Assembly of *E. faecalis*

- Velvet, Illumina: 162 contigs/scaffolds, N50 19kb
- Allora, Pac Bio: 622 contigs/scaffolds, N50 7.8kb
- Allora, Illumina + Pac Bio: 49 contigs/scaffolds, N50 243kb

16S sequencing on Pac Bio

- Full length sequences possible due to read length
- High error rate (85%) challenging:
  - Different ribotypes vs sequencing errors
  - High coverage needed for each fragment sequenced
- Alignment algorithm critical:
  - Conserved regions low information for clustering
  - Emphasize variable regions

16S genes of *S. aureus* on Pac Bio



Computational Technology Evolving

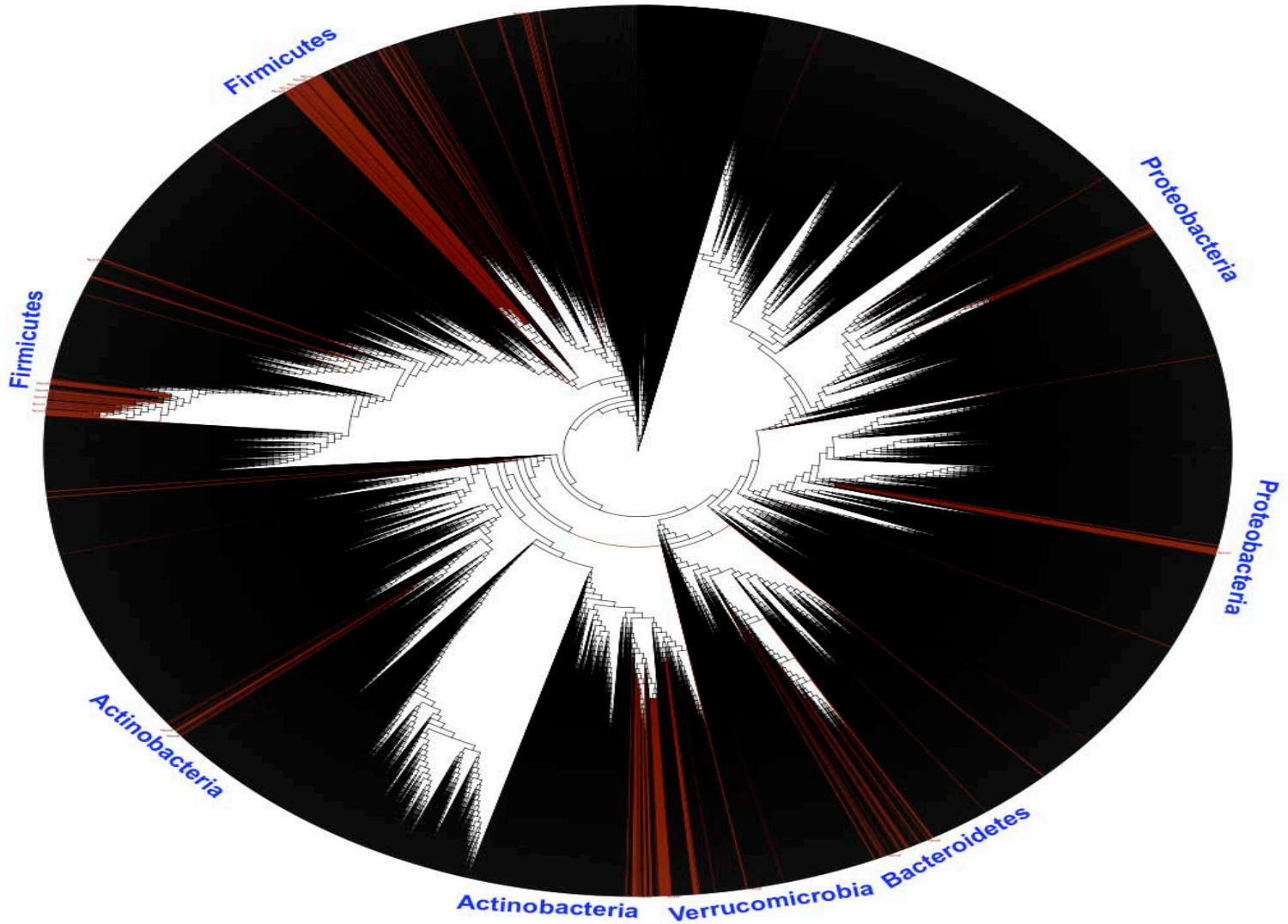
- Continued development of methods for community comparisons (16S)
  - What's a person to do?
  - Qiime (Rob Knight)
- Large-scale shotgun metagenomic data
  - HUMAnN pipeline (Curtis Huttenhower)
  - HMP (Makedonka Mitreva)
    - MulticoreWare, Real Time Genomics accelerated Blast and more
- Power calculations (Bill Shannon)

Novel organism discovery

- Need to know who's there to understand phenotypes
- Phylum level analysis
  - 454 16S reads, Illumina shotgun reads (Erica Sodergren et al)
- OTU/Species level analysis
  - Mothur (Sue Huse et al)
  - PhylOTU (Katie Pollard et al)
- Assembly of shotgun metagenomic data (HMP)
- Technologies for sequencing uncultured organisms

Novel Phyla?

1200 reads in initial HMP 16S set



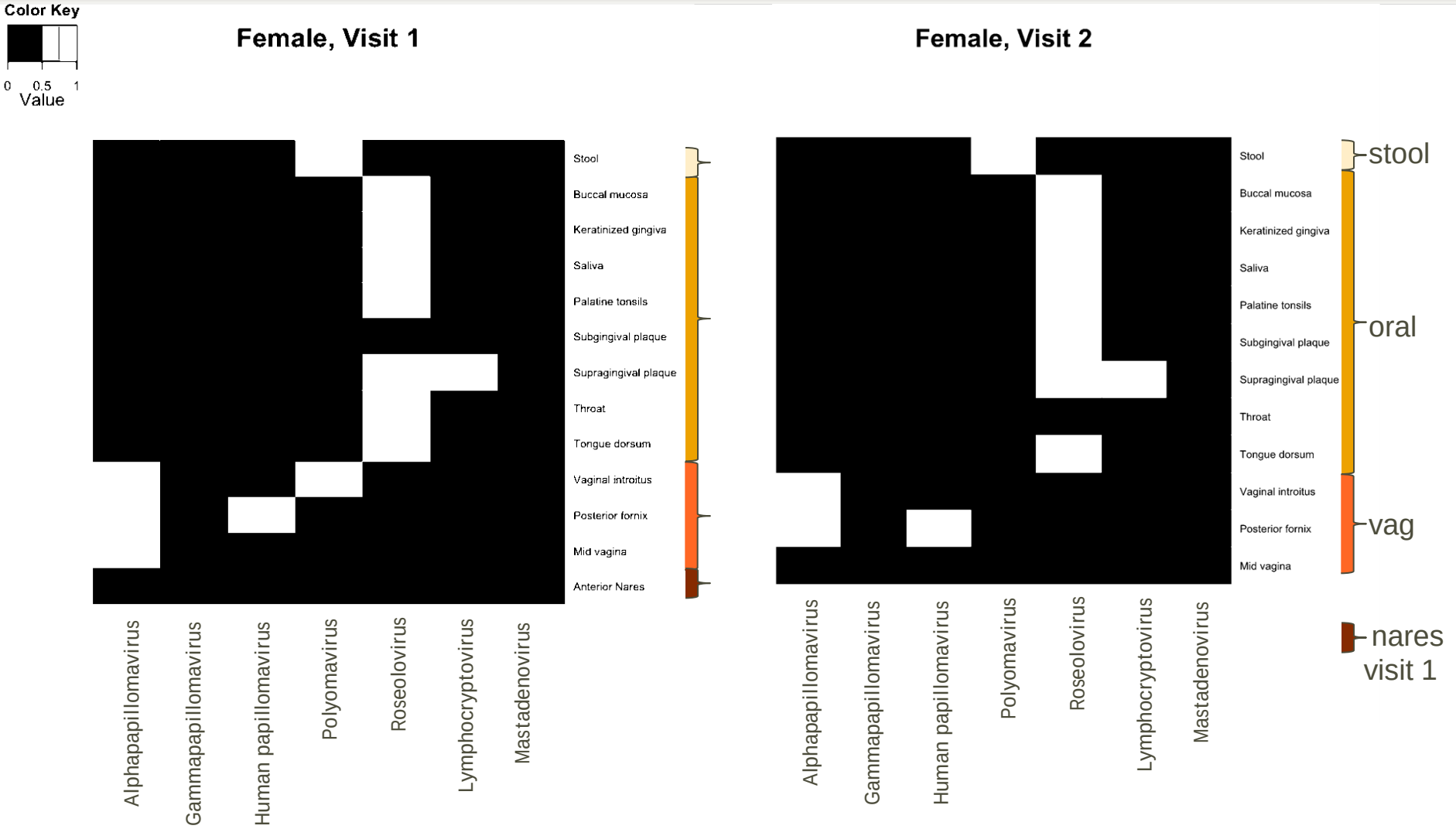
Load of deleterious organisms/genes

- Virulence factors
- Antibiotic resistances
- Found in all body sites
  - Low levels but easily detectable with shotgun sequencing

The Virome

- Growing activity in this area
- Healthy people have many viruses, sick people have more
- Animal viruses and Bacteriophages (and hybrids!)
- Methodology for detection of novel viruses improving
  - Known taxa but highly diverged
  - New taxa
- Relation to prokaryotic microbiome?

Virome of an individual over two visits



Presence/absence
White=virus present

Challenges to the field

- Minor organisms
  - Long tail in abundance distributions
  - Mainly unimportant?
  - Mix of communities?
- New types of data being included
  - Transcriptome
  - Proteome
  - Metabolome
- Host genotype
  - Enough said!!

Acknowledgments

- Washington Univ Genome Center Staff
  - Rick Wilson, Elaine Mardis, Tim Ley, Erica Sodergren, Li Ding, Makedonka Mitreva, Lucinda Fulton, Bob Fulton, David Dooling, Wes Warren, Pat Minx, Asif Chinwalla, Sandy Clifton, Vince Magrini, Kathie Mihindukulasuriya, Yanjiao Zhou, Lei Chen, Darina Cejkova, Otis Hall, Production, Tech Devel, Informatics, …
- Genome Centers
  - Baylor College of Medicine
  - Broad Institute
  - J. Craig Venter Institute
  - Washington University, St. Louis
- Funders
  - NIH, esp. NHGRI, NIAID, NIDCR

gweinsto@genome.wustl.edu

TYVM for Conference Help

- FNIH
  - Richard Scarfo
  - Peggy Diab
  - Bonnie Knight
  - Chianti Seitz
  - Laura Harwood
  - Jenna Mills
  - Joshua Walker
- Genome BC
  - Sally Greenwood
  - Cece Gnocato
- NIH
  - Shaila Chhibba
  - Tsegahiwot Belachew
  - Chris Wellington
  - Lita Proctor
  - Jane Peterson
  - Jean McEwen
  - Maria Giovanni