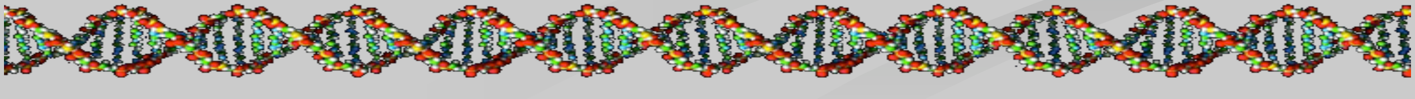
Metagenomic Assembly

Successes, Validation,
And
Challenges

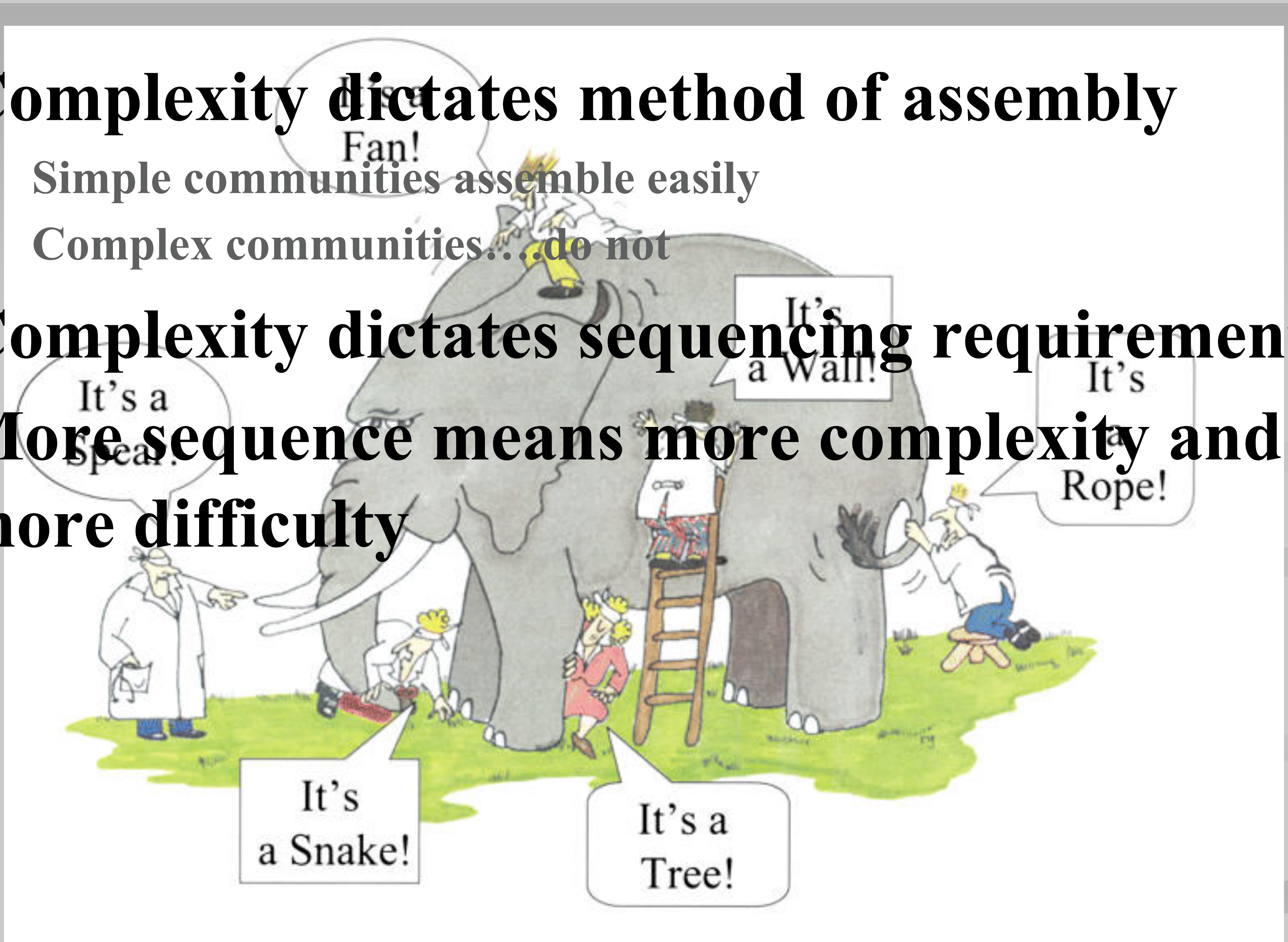
Matthew Scholz
Los Alamos National Laboratory
Genome Sciences Division



LA-UR-11-10900

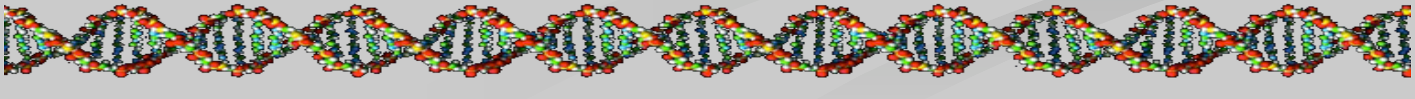
Metagenomic Assembly is Complex

- **Complexity dictates method of assembly**
  - **Simple communities assemble easily**
  - **Complex communities….do not**
- **Complexity dictates sequencing requirement**
- **More sequence means more complexity and more difficulty**



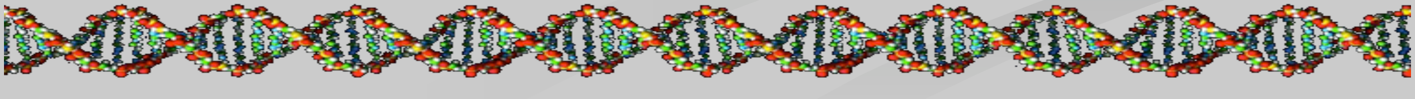
Assembly Successes

- High throughput pipeline
- Improved assemblies
  - JGI/LANL has successfully assembled 123 metagenomes in last year
  - Average time ~ 1 week/sample
- HMP metagenome assemblies
  - LANL assembled 223 metagenomes from whole genome shotgun sequencing of HMP
  - Assembly of 10 site specific samples (multiple samples from same site)
- Validation of several metagenome assemblies



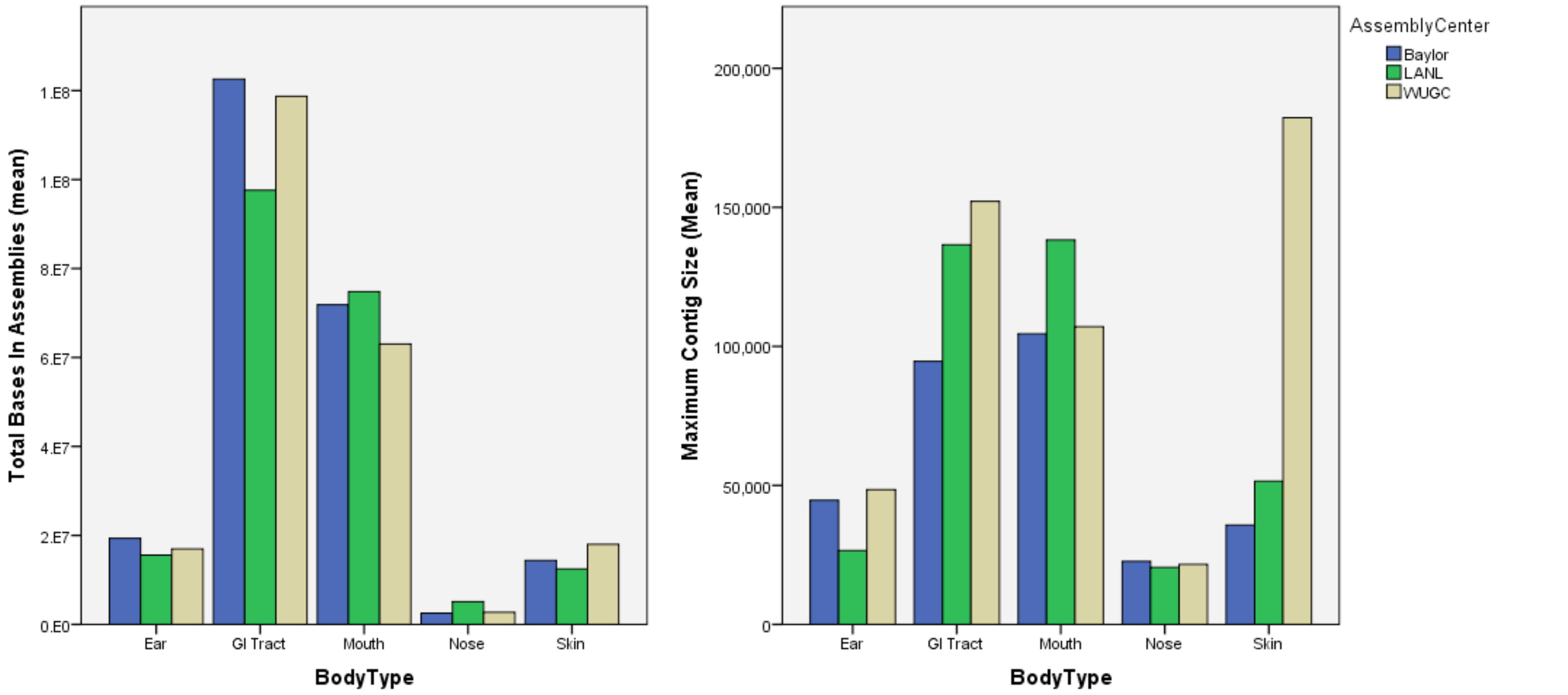
Assembly differences

- **HMP shotgun metagenome assembly**
  - **Optimization**
    - Tool Selection
    - Kmer Selection
    - Selection of # Cores
  - **Volume production**
    - 1 Kmer
    - Metrics
- **JGI Metagenome Assembly**
  - **Multiple tool selection**
  - **Range of Kmers utilized**
  - **Many different sample types**

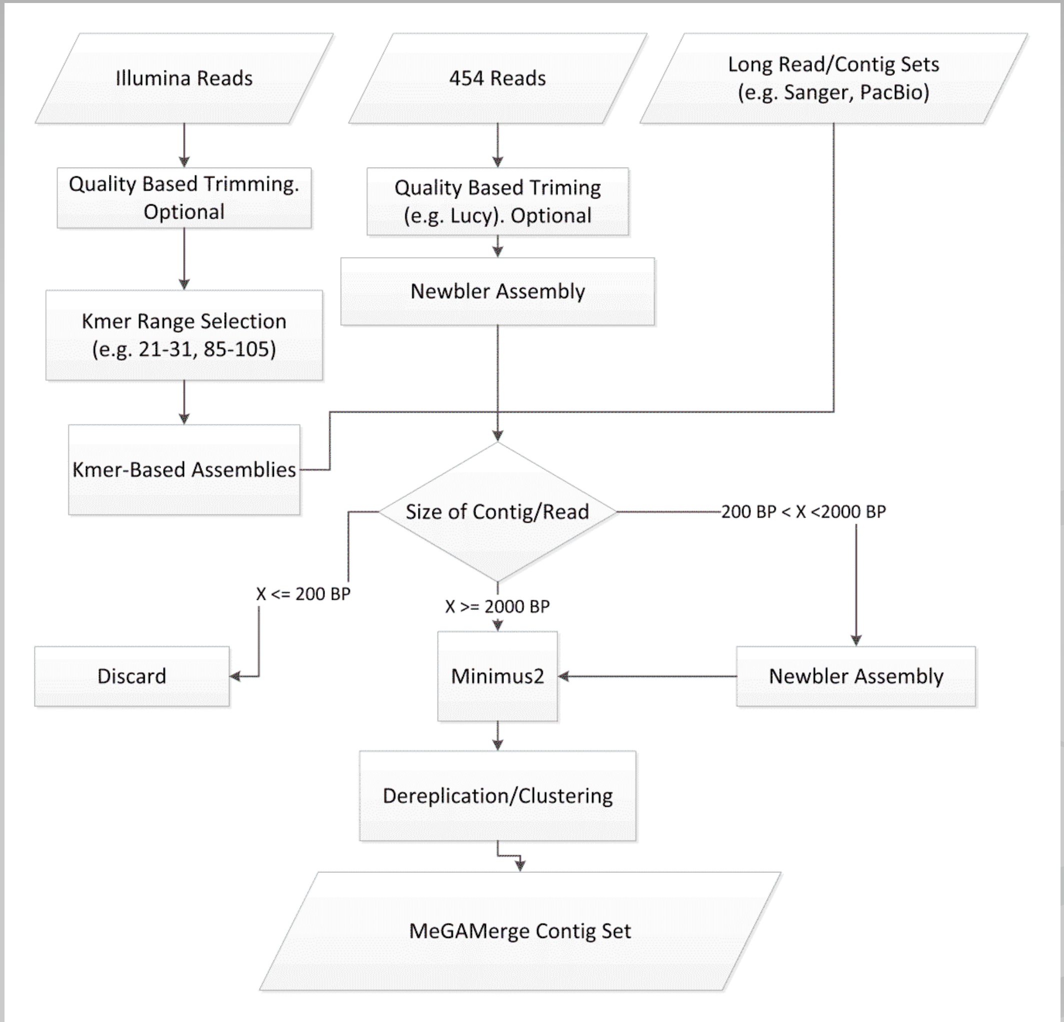


HMP assemblies

- Draft
- Many Samples
- Many metrics



JGI/LANL Metagenome Assembly Pipeline



Illumina Reads

Quality Based Trimming. Optional

Kmer Range Selection (e.g. 21-31, 85-105)

Kmer-Based Assemblies

454 Reads

Quality Based Triming (e.g. Lucy). Optional

Newbler Assembly

Long Read/Contig Sets (e.g. Sanger, PacBio)

Size of Contig/Read

X <= 200 BP

Discard

X >= 2000 BP

Minimus2

200 BP < X <2000 BP

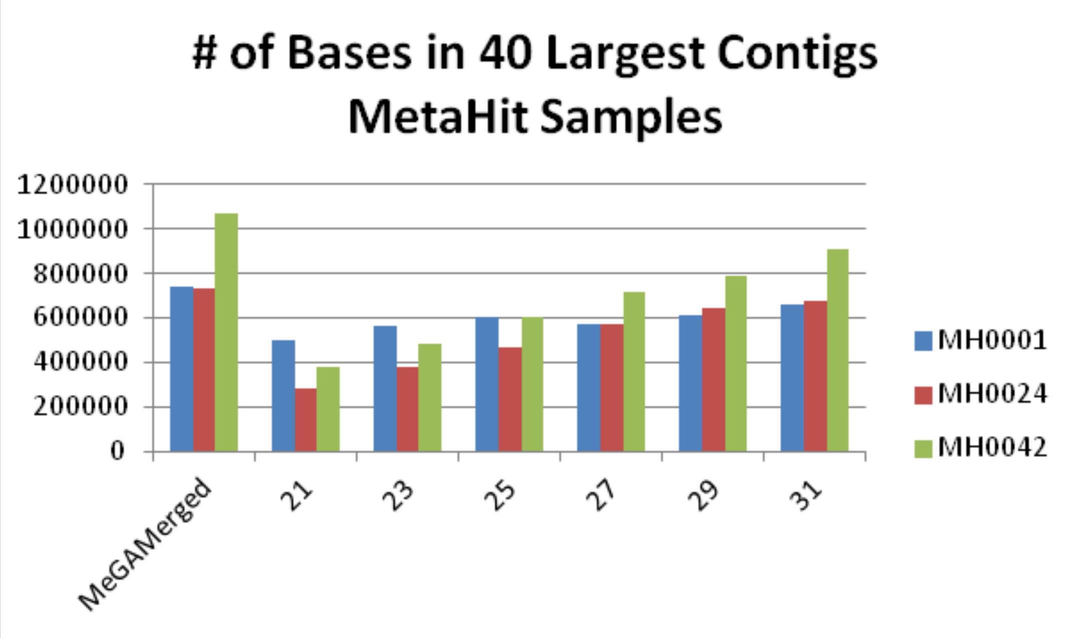
Newbler Assembly

Dereplication/Clustering

MeGAMerge Contig Set

JGI/LANL Assembly Process

- “Improved” merging of Multiple assemblies
- Statistical Metrics are better
- Validation supports these improved assemblies

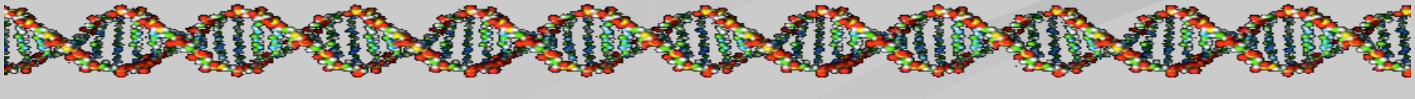


Validation

- **How do you validate metagenomes?**
  - Contig Statistics
  - Read Mapping
  - Annotation
  - Similarity Searches
  - Phylogenetic distribution
  - Reference genomes

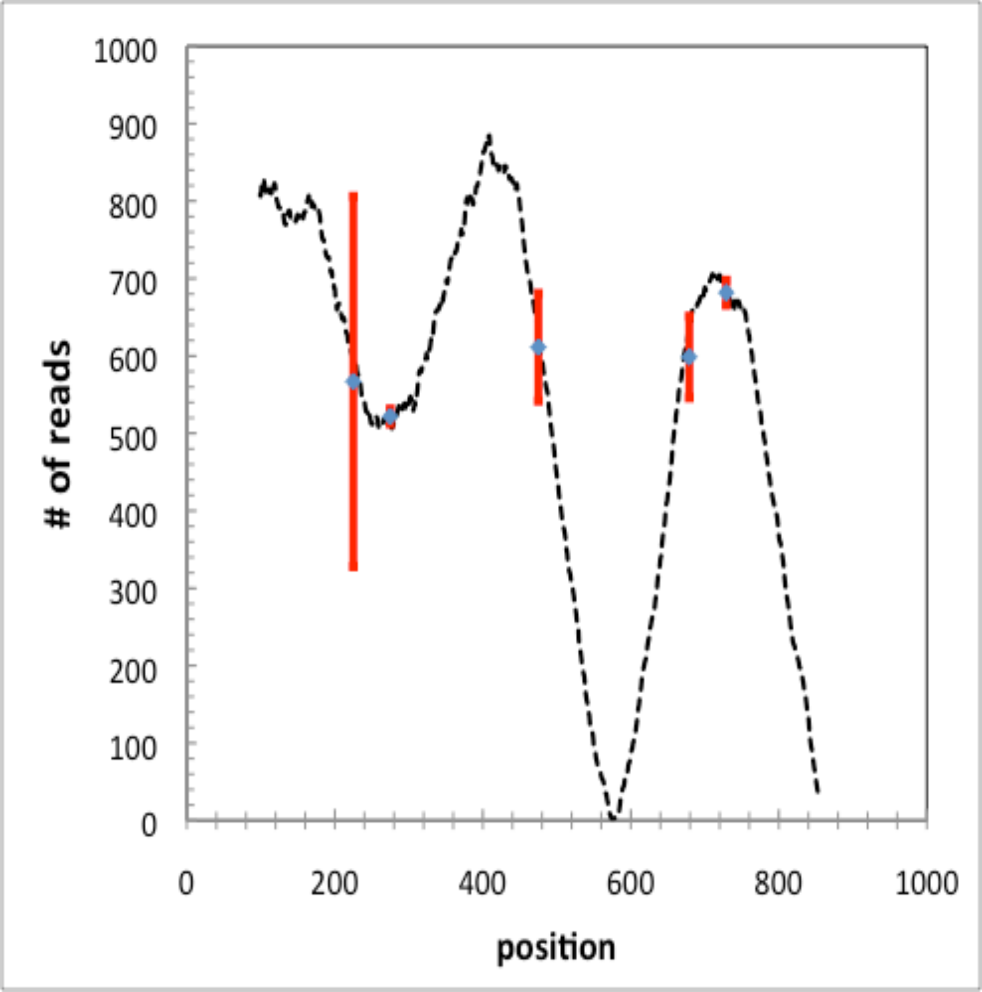
Read Mapping

- **Are contigs correctly assembled?**
  - Do read data confirm contigs?
  - Edge effects
- **Generate information about coverage**
  - Average Fold Coverage
  - Percent of Contig Covered
- **SNP/INDEL information**
  - Indication of diversity within sample

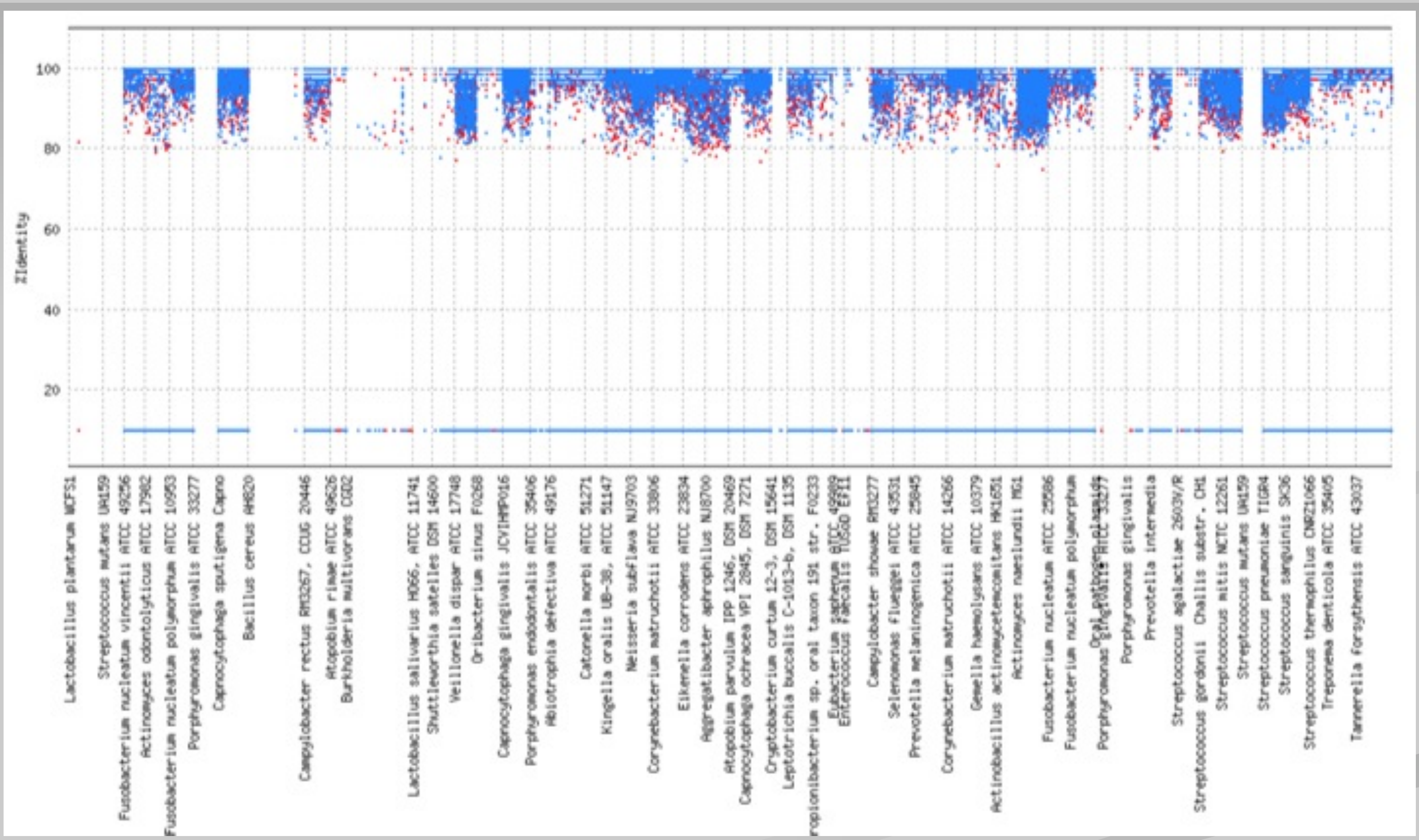


Finding Potentially Erroneous Contigs with Read mapping

- Read coverage of each base from pipeline.
- Lines delineate regions where mean coverage deviates past thresholds
- Is this a good contig?
  - Where did contig come from?
- Can use to automatically break or discard contigs that fail read-mapping



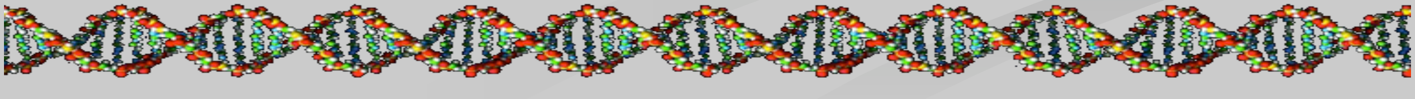
Reference Genomes



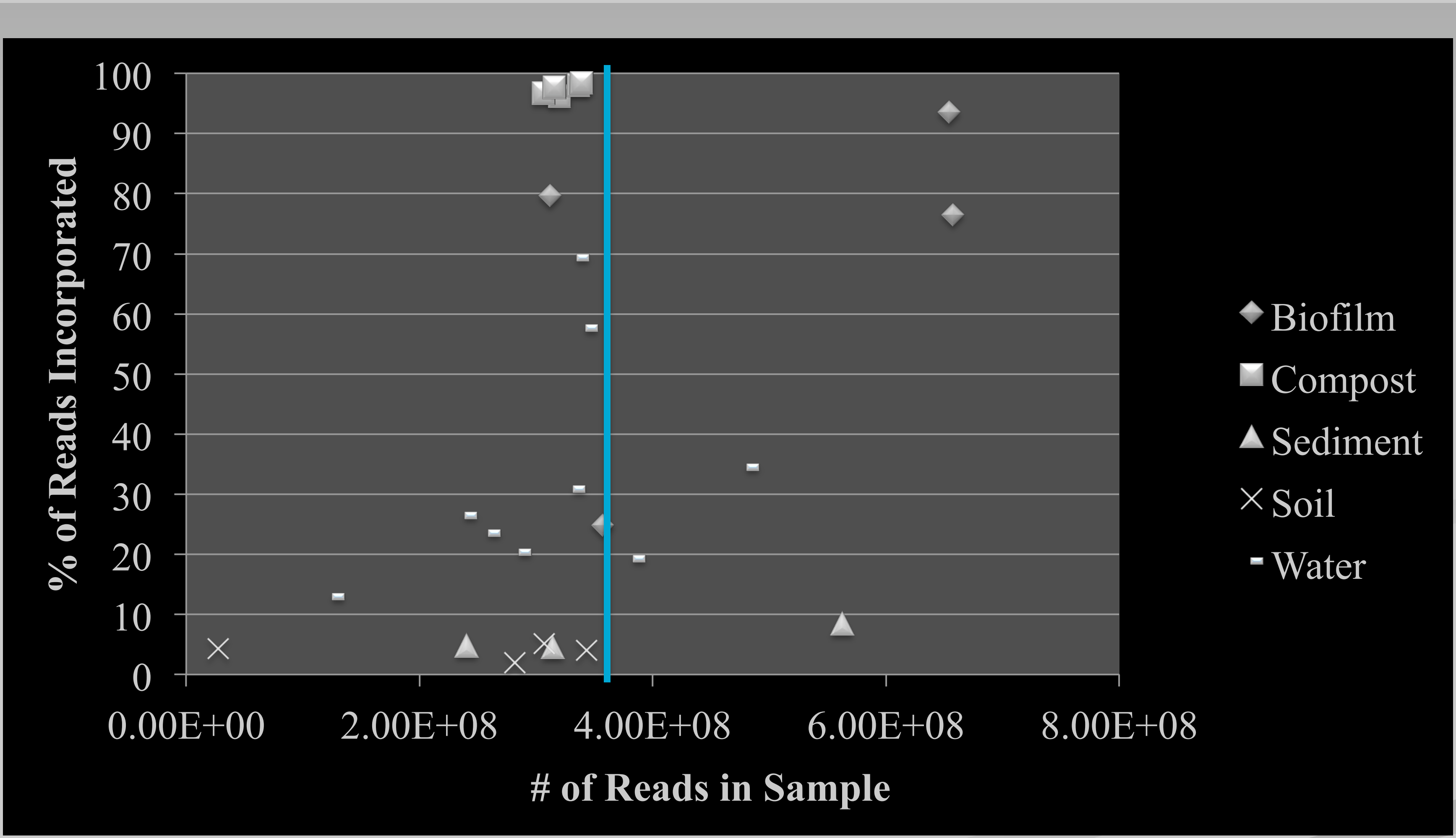
LA-UR-11-10900

Ongoing Challenges

- Hardware/Software
  - Kmer assembly Speed/RAM tradeoff
  - Algorithms for metagenomes
  - Too much software
- How to Determine “Correct” assembly?
- Metadata
- Too much data
- Not enough coverage

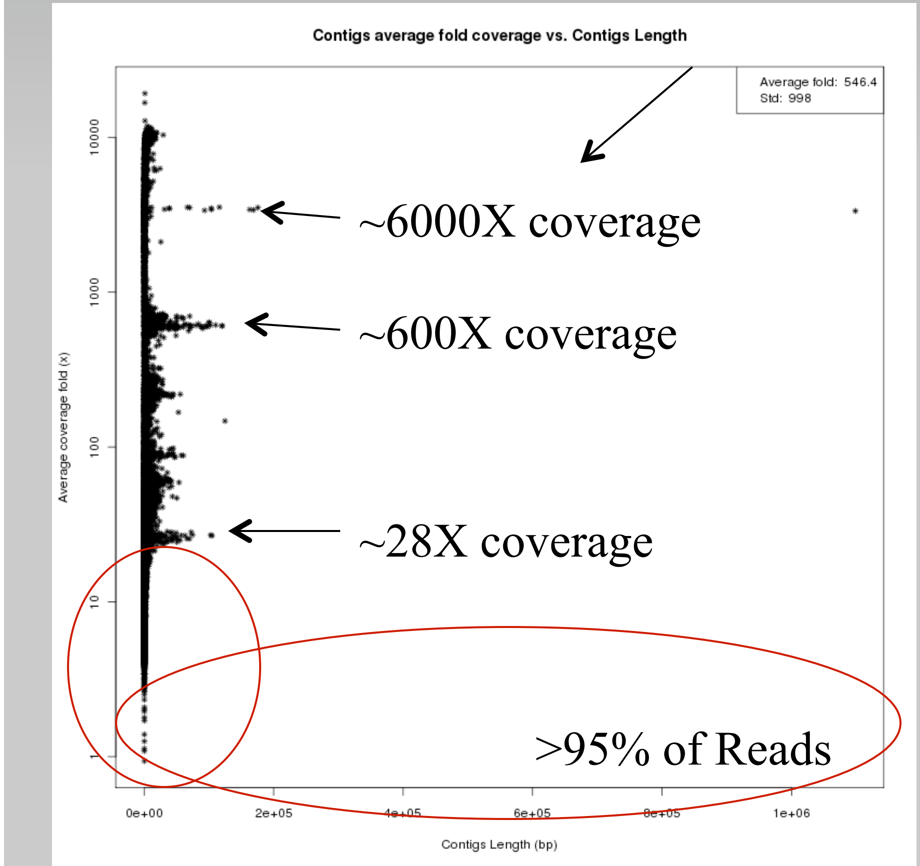


Read Incorporation



Not Enough Coverage

- Coverage varies by sample
  - 1 Lane HiSeq is maximum for current hardware/assemblers (complex samples)



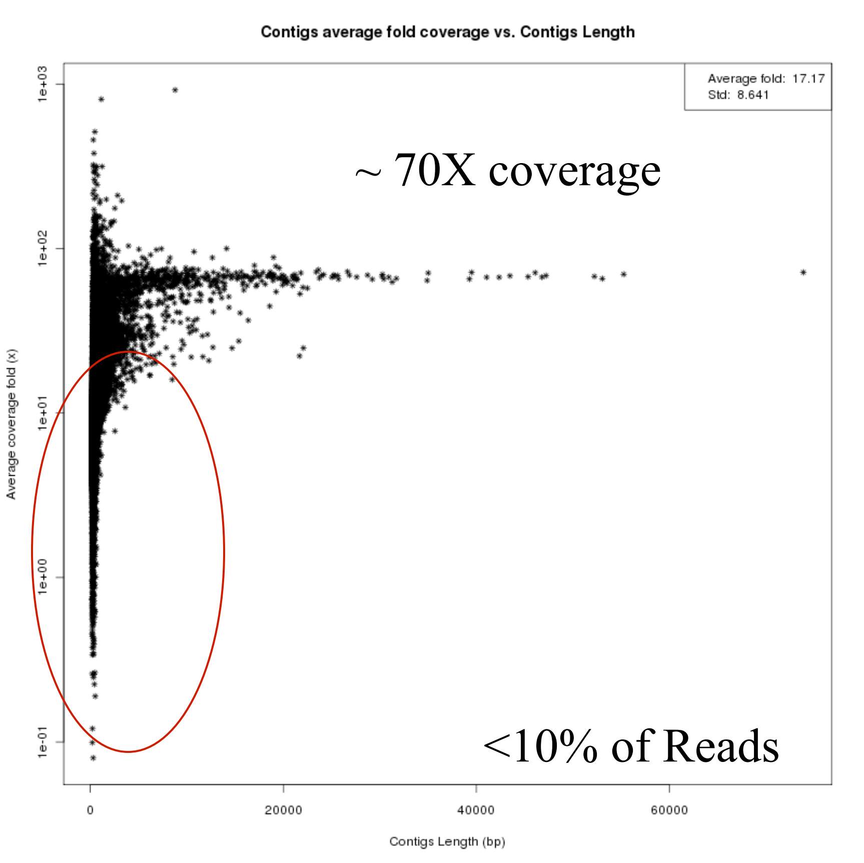
Contigs average fold coverage vs. Contigs Length

~6000X coverage

~600X coverage

~28X coverage

>95% of Reads



Contigs average fold coverage vs. Contigs Length

~ 70X coverage

<10% of Reads

Concluding Thoughts

- We need metadata (standards)
- Tiers
  1. Assembly Statistics
  2. Assembly Level
  3. Contig analysis
- Scores

Classification of Metagenomes

1. **Assembly Statistics**
   - % Read Incorporation
   - Total Assembled Bases (post filtering?)
   - G+C Content
   - Coverage histograms
2. **Assembly Level**
   - Draft
     - 1 assembler, 1 set of parameters
   - Improved Draft
     - Current JGI/LANL assembly/merge method
     - Read based validation
   - High Quality (Theoretical)
     - Binning strategies pre-assembly
     - Read based correction/trimming
     - Single cell genomes from site as references

Classification of Metagenomes

1. **Assembly Statistics**
2. **Assembly Level**
3. **Contig analysis**
   - **Read mapping based validation**
   - **Clustering**
   - **Gene analysis**

Many Thanks:

Metagenomics and Data Analysis Team

- Patrick Chain
- Tracey Freitas
- Ron Croonenberg
- Bin Hu
- Chien-Chi Lo
- Shawn Starkenburg
- Gary Xie
- Shannon Steinfadt
- Others…

Metagenome work

- Jim Tiedje
- Titus Brown
- Adina Howe
- HMP consortium
- Mihai Pop
- Joe Zhou
- Kostas Konstantinidis

Informatics Team

- Ben Allen
- Andy Seirp
- Criag Blackhart
- Yan Xu
- Todd Yilk

Single cell work

- Roger Lasken
- Ramunas Stepanaskus
- Steve Hallam

Management Team

- Chris Detter
- David Bruce
- Tracy Erkkila
- Lance Green
- Shunsheng Han

And many others…

Wet-lab Team

- Cheryl Gleasner
- Kim McMurry
- Krista Reitenga
- Xiaohong Shen
- Others…

Project Management

- Shannon Johnson
- Lynne Goodwin
- Others…

Kmer team

- Joel Berendzen
- Nick Hengartner
- Ben McMahon
- Judith Cohn

Finishing and SCG

- Olga Chertov
- Karen Davenport
- Armand Dichosa
- Michael Fitzsimons
- Ahmet Zeytun
- Others…

Metagenomic Assembly Strategies

- **Bigger computers**
  - **1 Lane HiSeq PE reads = 400 M reads**
    - 1TB RAM (complex communities)
- **Better Assemblers**
  - **MetaIDBA**
  - **MetaVelvet**
  - **Ray**
  - **ABySS**
  - **AllPaths**
- **Binning Reads**
  - **Unsupervised/Heuristic**
  - **Machine Learning**
  - **Statistical**
  - **Reference Based**
- **To Infinity and Beyond…**