Putting The Pieces Together: From Assembly To Analysis

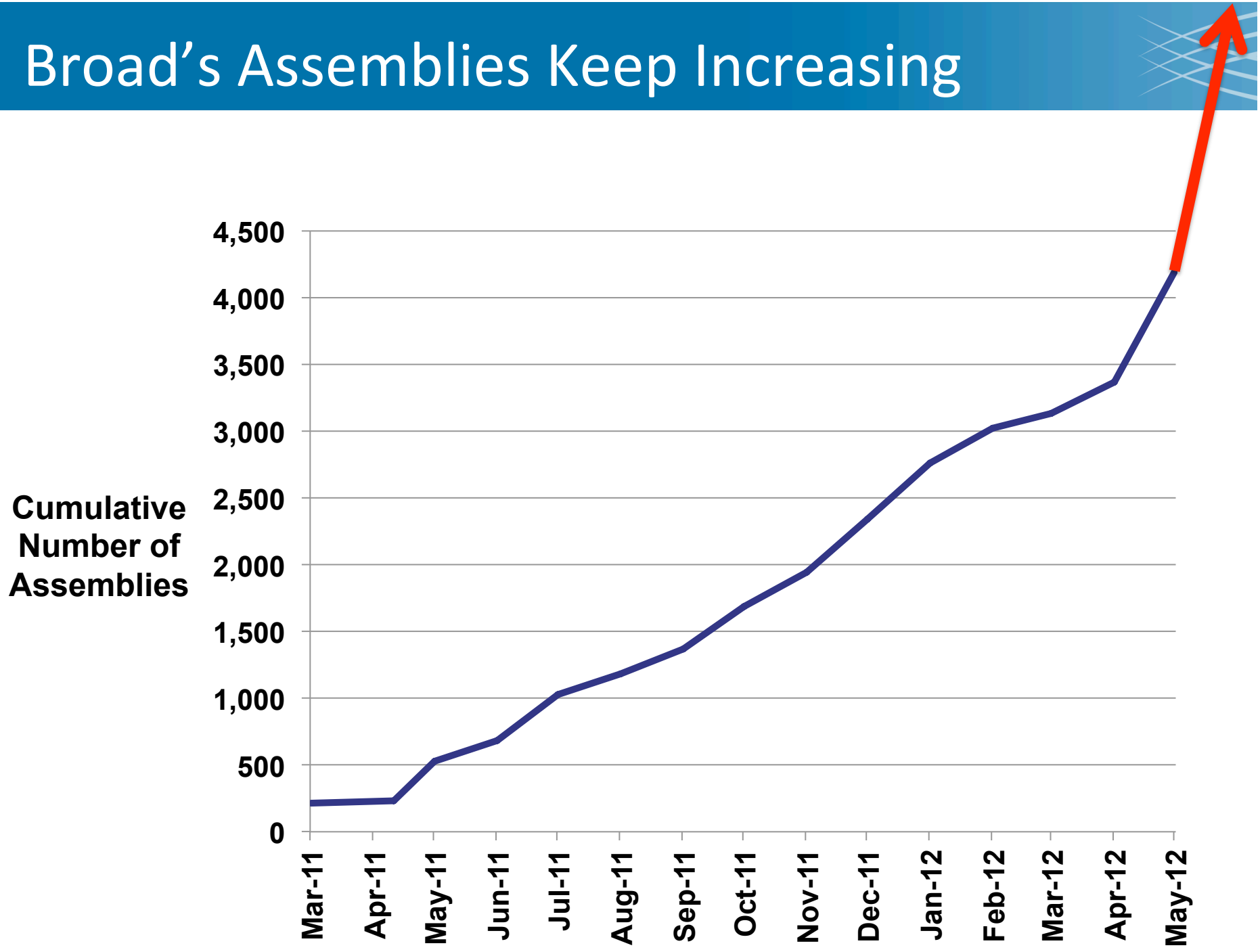
Sean Sykes

Sr. Bioinformatics Analyst, Genome Assembly and Analysis
SFAF
June 6, 2012

ssykes@broadinstitute.org

BROAD INSTITUTE

Broad's Assemblies Keep Increasing



Challenges To Assembly Analysis At Scale

- Shortened analysis time
  - Rapidly increasing sample numbers
- Large selection of bioinformatics tools
  - Wide knowledge-base required
- Numerous input/output file types
  - Challenging to compare results
- Many assembly error modes
  - Need comprehensive analysis tools

Tackling The Challenges Head-on

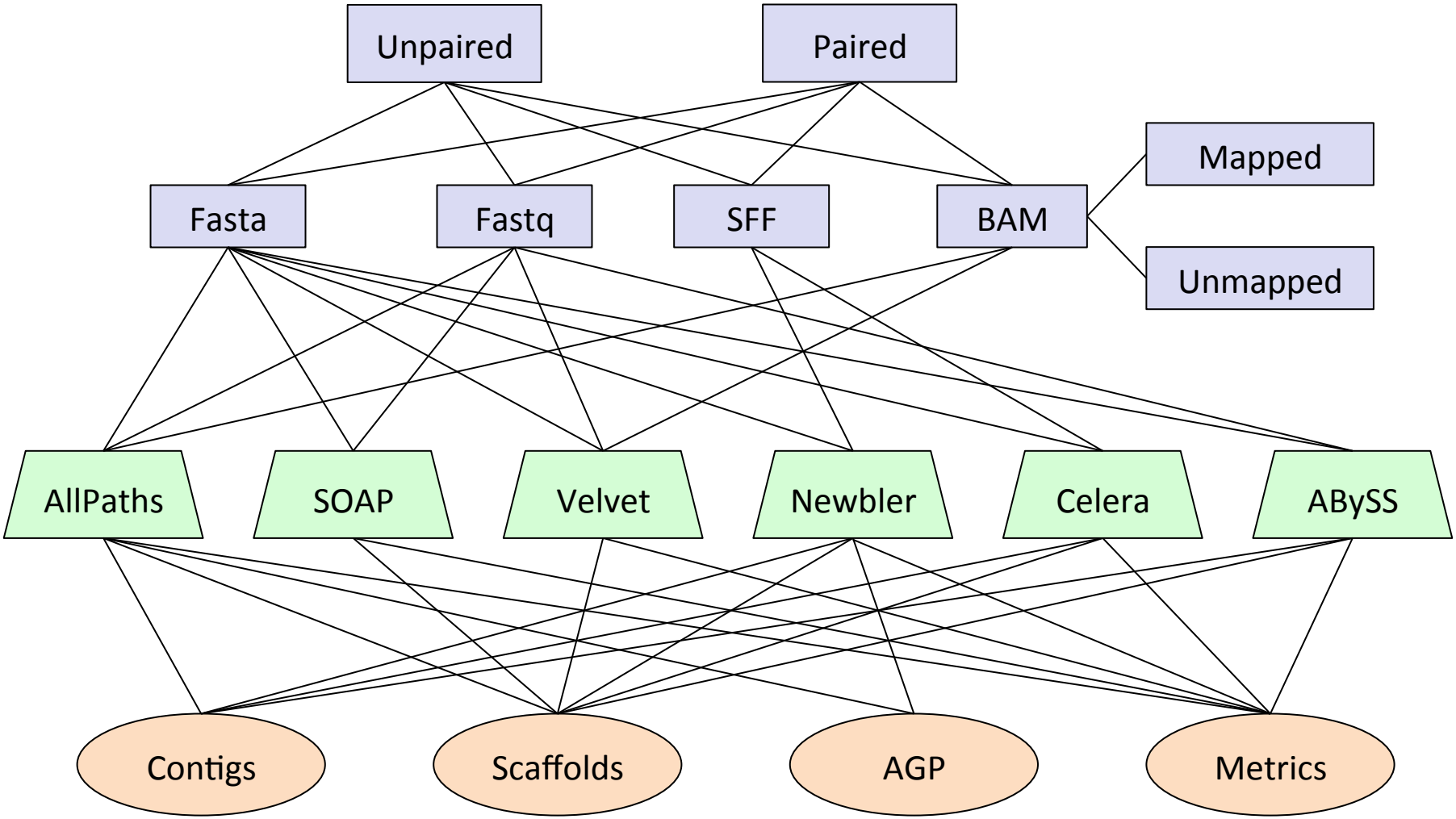
- "Approaches for *de novo* assembly of difficult genomes"
  - Sarah Young, Poster #FF0184
- "Genome Mis-assemblies – A Challenge to Characterize"
  - Sakina Saif, Poster #FF0187
- "Pilon Assembly Improvement Software"
  - Bruce Walker
- Genome Assembly Evaluation Metrics and Reporting (GAEMR)

Goal: Standardized Assembly Analysis

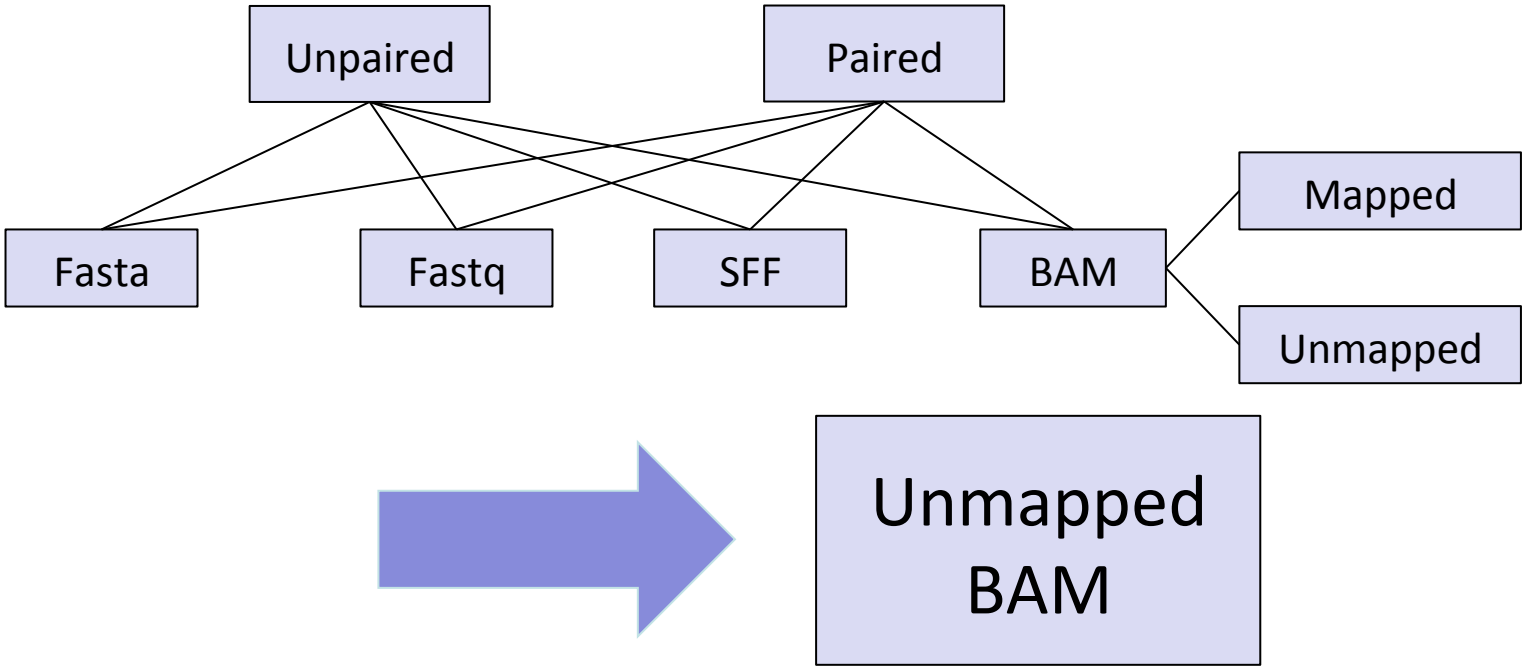
<u>Specifications:</u>

- Fully modular and extensible
- Comprehensive analysis computed upfront
- Assembler neutral
- Rapidly digestible output
- Works with reference and *a priori*

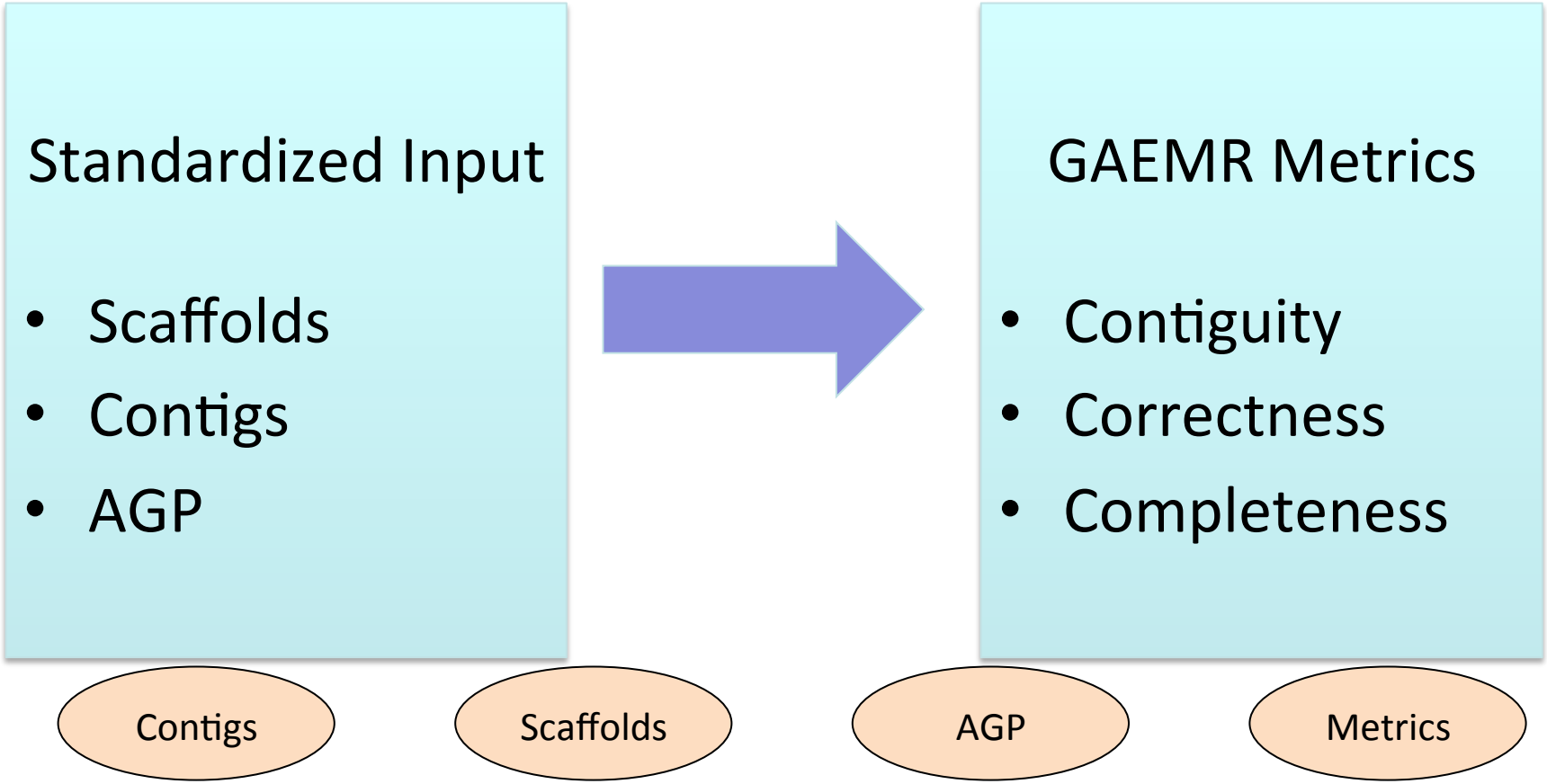
Challenges To Standardizing Analysis



Standardize Inputs For Analysis



Define Controlled Set of Metrics



How We Evaluate An Assembly

- Contiguity
  - How many pieces
- Correctness
  - How many errors
- Completeness
  - How much is missing

How We Evaluate An Assembly

- Contiguity
- Correctness
- Completeness

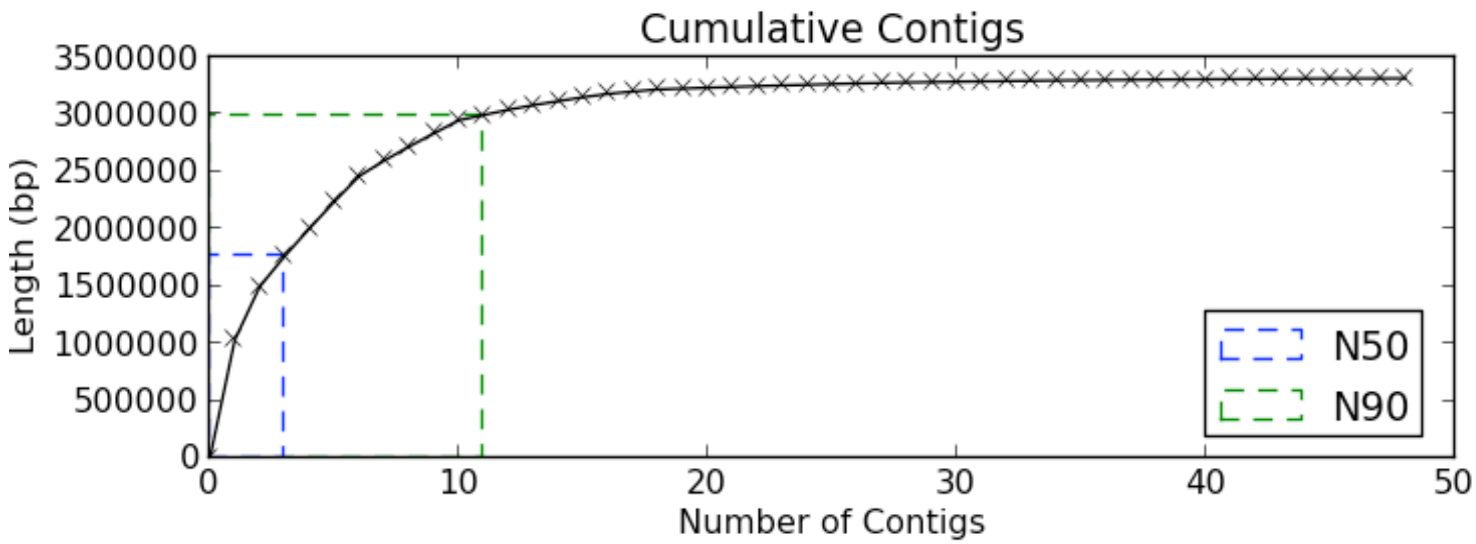
Contiguity

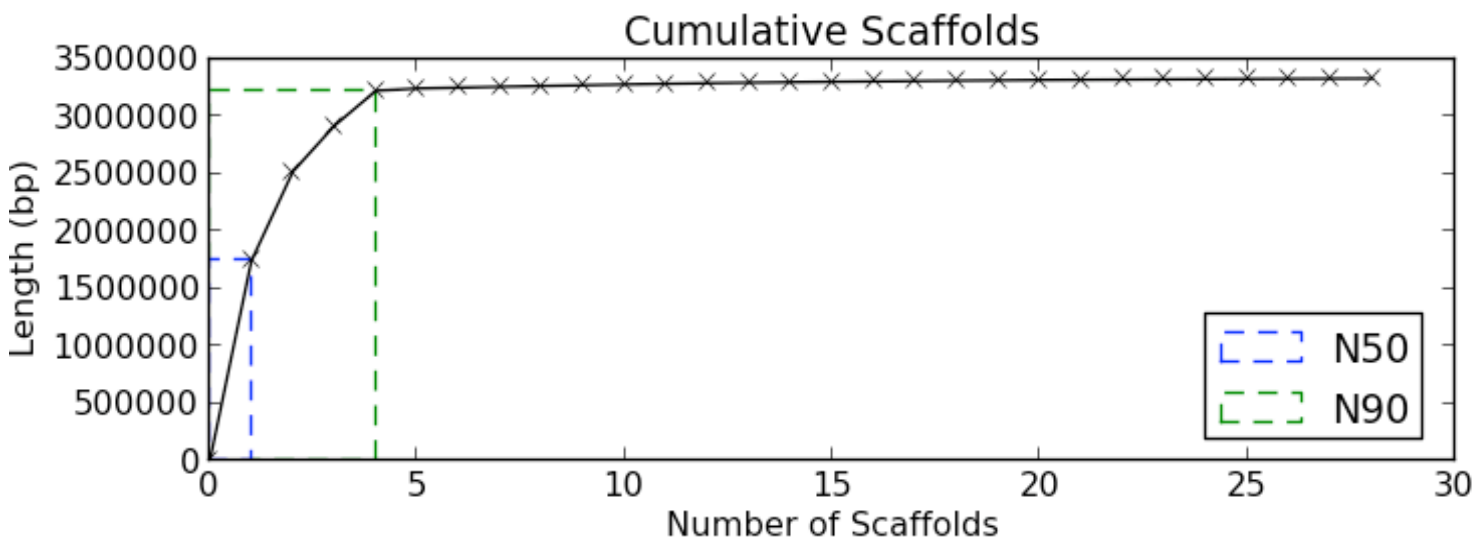
- Does the assembly meet our standard for a "high quality draft"?
- How does this compare to other assemblies?

Standardized Contiguity Statistics

#TITLE: Basic Assembly Stats	
#Name	**G15640_allpaths**
Assembler	allpaths
Contigs	48
Max Contig	1032234
Mean Contig	69084
Contig N50	265874
Contig N90	46178
Total Contig Length	3316026
Assembly GC	37.41
Scaffolds	28
Max Scaffold	1743744
Mean Scaffold	119148
Scaffold N50	1743744
Scaffold N90	305558
Total Scaffold Length	3336135
Gaps	20
Max Gap	5060
Mean Gap	1005
Gap N50	3470
Total Gap Length	20109

Cumulative Charts Spotlight Contiguity





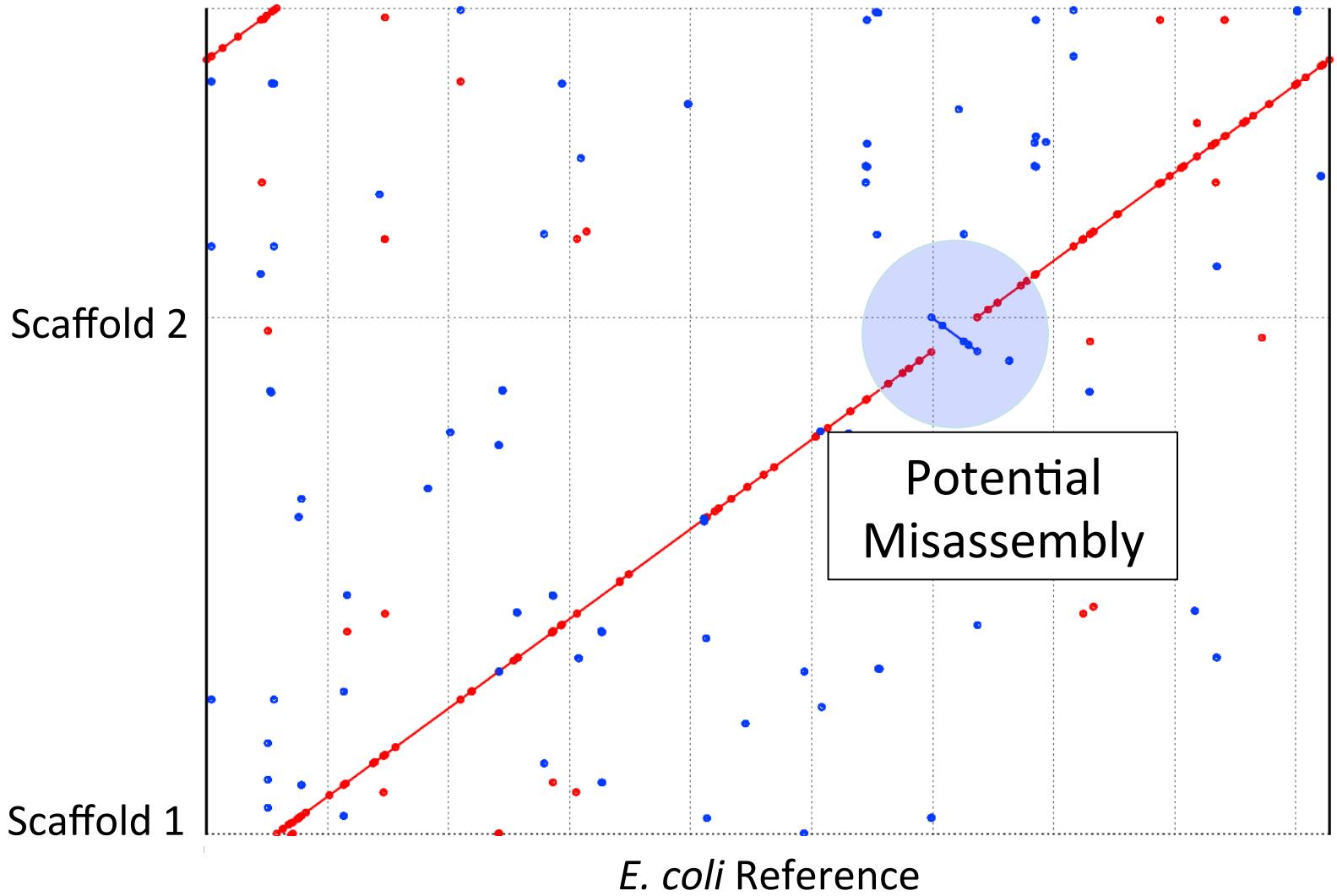
How We Evaluate An Assembly

- Contiguity
- Correctness
- Completeness

Correctness

- Does the assembly represent the genome?
  - Any misassemblies?
  - Any contaminants?

Reference Alignment Highlight Concerns



http://mummer.sourceforge.net/

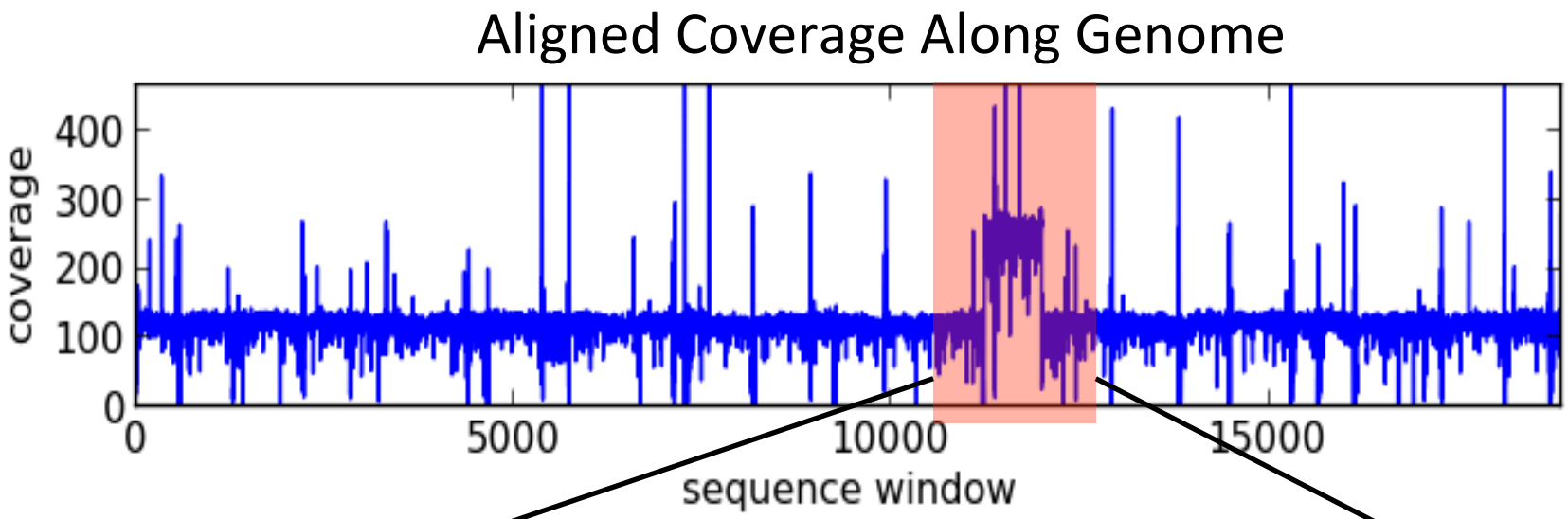
Scaffold Accuracy Assesses Overall "Goodness"

#Title: Scaffold Accuracy Report	
#Name	**G10779_allpaths**
Scaffold Accuracy	0.979
Total Pairs to Assess	8521
Total Valid Pairs	8345
Valid	8251
Valid Circular	94
Total Invalid Pairs	
Invalid Orientation	90
Invalid Length	86
Cross Chromosome	0
Potential Mis-assembled Regions	3

<0.99 = potential misassembly

2.10%

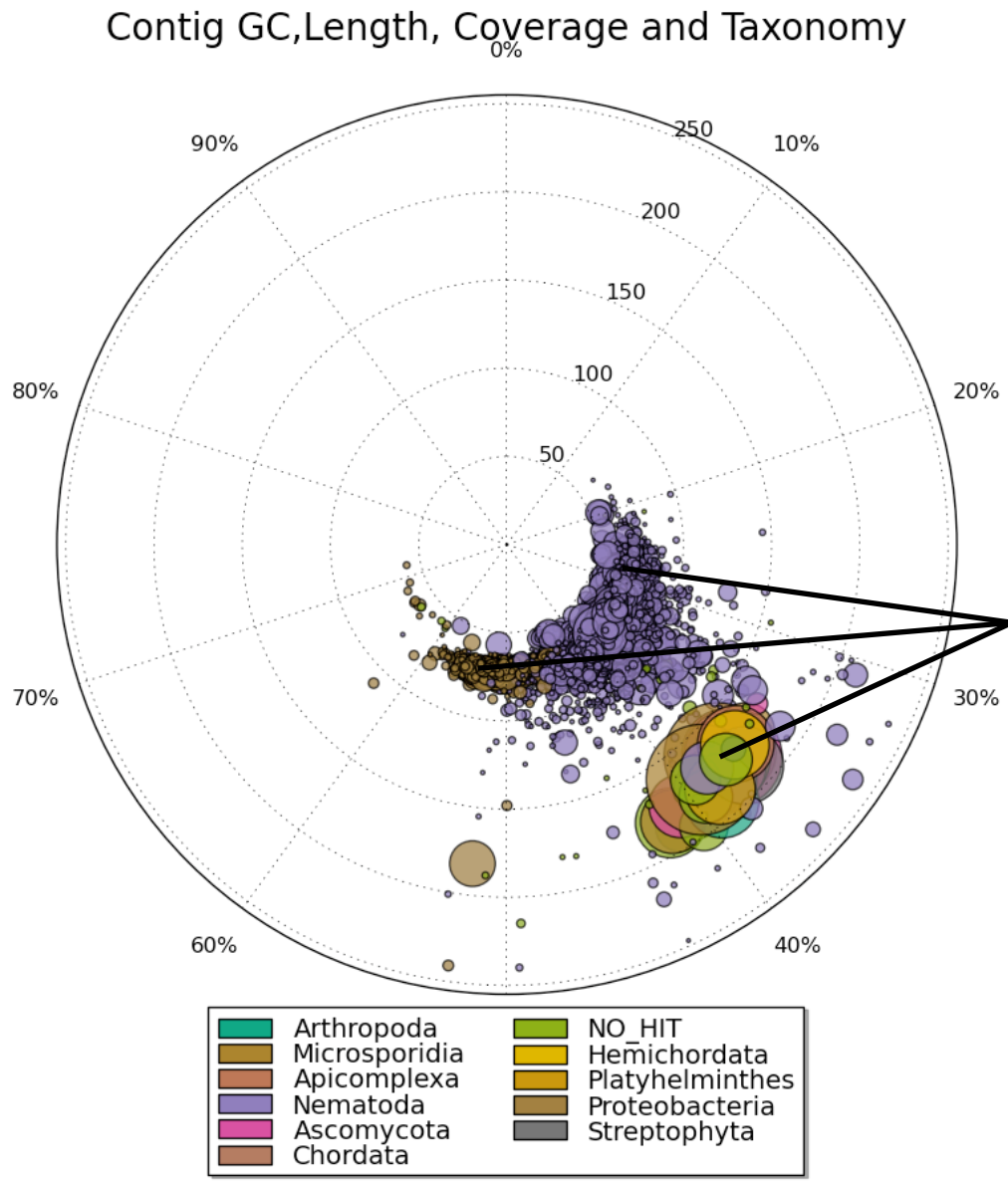
Assembly Read Alignments Show Misassemblies



Long stretch of 2x mean coverage:

- Collapsed repeat
- Genome duplication
- Contaminant

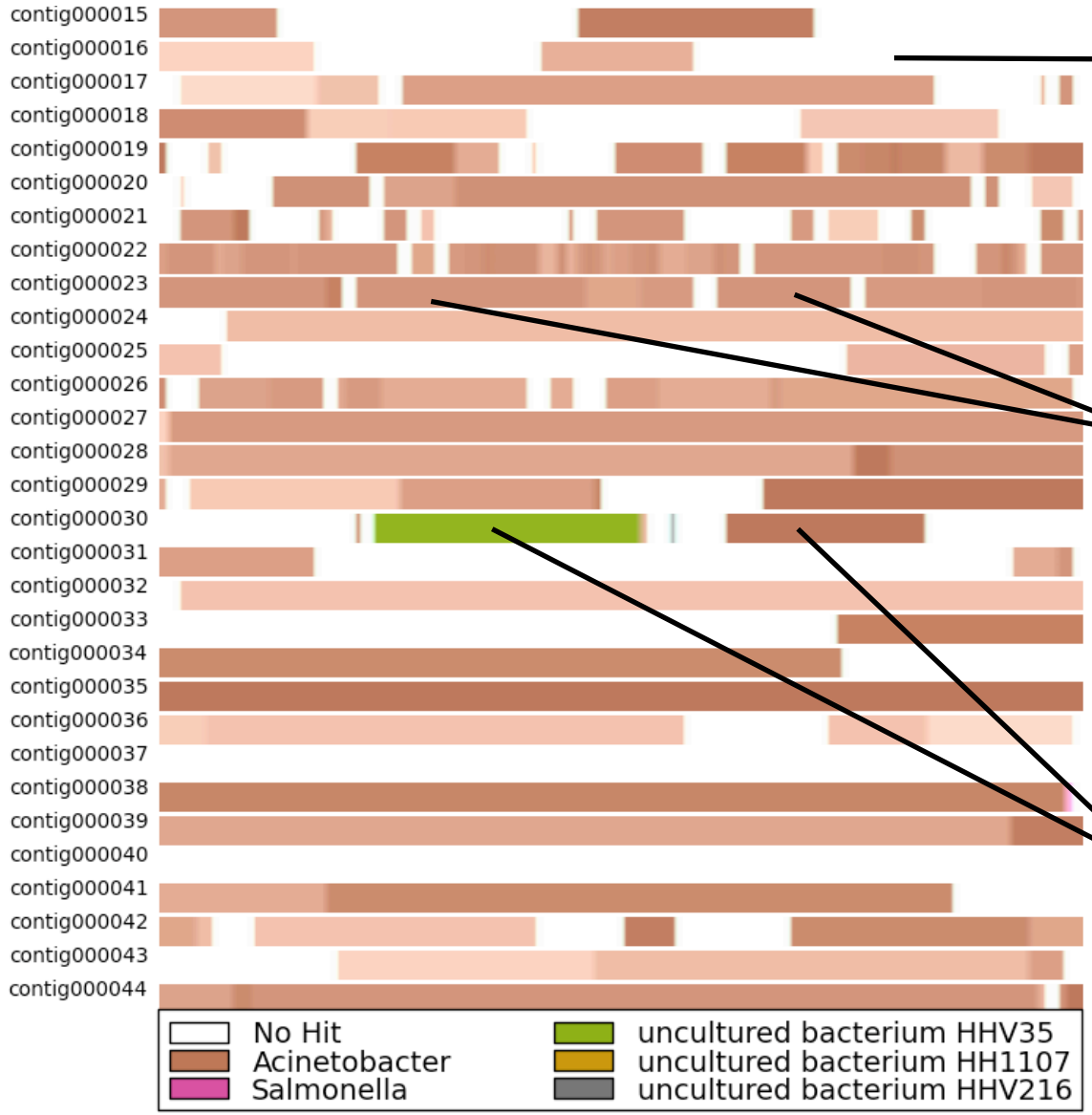
Using BLAST To Pinpoint Contamination



- Point = Contig
- Size = Contig Size
- Radial location = %GC
- Distance from center = aligned read coverage
- Color = Taxonomy

➢ Quickly identify contamination!

Using BLAST To Pinpoint Contamination



No color = “No Hit”

Shading = % Identity

2 different colors =
2 different taxonomy

16S Analysis Quickly Identifies Contaminants

#TITLE: rRNA Analysis Summary

#Gene	Total Copies	Lineage	# Organisms Found	Organism IDs
16S	11	genus	4	*Actinomyces* *Campylobacter* *Peptoniphilus* *Porphyromonas*

16S gene classifies organism

- >1 organism identified usually = contamination
- Detailed output for quick troubleshooting

RNAmmer: http://www.cbs.dtu.dk/services/RNAmmer/
RDP Classifier: http://sourceforge.net/projects/rdp-classifier/

How We Evaluate An Assembly

- Contiguity
- Correctness
- Completeness

Completeness

- Does our assembly capture the whole genome?
  - How does it compare to the reference?
  - Have we used all the data effectively?

K-mer Coverage: % Reference In Assembly

#TITLE: Kmer Coverage

#Category	**Count**	**RefCount**	**Pct**
Covered	3198669	3314682	96.50%
DistinctCovered	3162236	3265649	96.83%
OverCovered	8209	3314682	0.25%
Novel	11125	3314682	0.34%
NovelDistinct	8364	3265649	0.26%

- Pcts => number of *k*-mers in reference
- Assembly covers majority of reference
- Missing ~3% of the reference
  - Real diversity?

Fraction of reads used: Indicator of completeness

#TITLE: Simple Bam Stats	
#Stat	Fragment Pairs (All Reads)
Total Reads	44315334
Paired Reads	44315334
Pct Paired	100
Duplicates	0
Pct Duplicates	0
Total Read 1	22157667
Total Read 2	22157667
Mapped	42545923
Pct Mapped	96.01
Singletons	448165
Pct Singletons	1.05
Mapped w/ Mate	42097758
Pct Mapped w/ Mate	98.95
Properly Paired	41996822
Pct Properly Paired	98.71
Cross-chromosome	80926
Cross-chromosome (MQ >= 5)	61427

High % reads aligned = few missing bases

GAEMR

- Goal: assess contiguity, correctness, and completeness
  - on 1000s of genomes
  - sequenced with different technologies
  - assembled with different algorithms
  - in 10 minutes or less human time??

GAEMR Does Heavy Lifting For You

- Processes analysis in parallel
- Standardizes inputs and outputs
- Tracks process IDs and commands
- Reports successful/failed processes
- Produces HTML ready tables/charts

GAEMR: HTML Reporting

Index

Section 1. Assembly Stats
- Table 1.1. Contig Stats
- Table 1.2. Scaffold Stats
- Table 1.3. Gap Stats
- Table 1.4. Ambiguous Base Stats
- Table 1.5. rRNA Analysis
- Table 1.6. Completeness and Correctness
- Figure 1.7. Cumulative Scaffolds and Contigs

Section 2. Read Stats
- Table 2.1. Read Stats
- Table 2.2. Read Alignment to Assembly
- Table 2.3. Kmer Frequency Stats
- Table 2.4. Read Alignment to Reference
- Table 2.5. Insert Size Stats
- Figure 2.6. Insert Size Histograms

Hyperlinks for quick access to specific assembly analysis output

GAEMR Dependencies

- **Programming Language**:
  - Python-2.7, Biopython, matplotlib, optparse, numpy, pysam
- **Alignment Software:**
  - BWA/Bowtie2
  - BLAST+
  - Nucmer
- **SAM/BAM Utilities**:
  - Picard package
  - Samtools
- **Classification**:
  - RNAmmer/RDP-classifier-2.4
  - NCBI taxonomy dump (optimal)

Conclusions

- Assembly analysis is complicated
- Analysis is complicated further by increased workload
- GAEMR rapidly analyzes assembly regardless of its origin
- GAEMR takes guesswork out of running tools individually
- GAEMR streamlines numerous bioinformatics tools

"GAEMR works hard, so you don't have to."™

GAEMR Availability

- Open source
- Another summer blockbuster!

http://www.broadinstitute.org/scientific-community/science/programs/genome-sequencing-and-analysis/gaag/gaemr

Acknowledgments

Genome Assembly & Analysis

Amr Abouelleil

Harindra Arachchi

Aaron Berlin

Sante Gnerre

Alma Imamovic

Jerome Naylor

Sakina Saif

Terry Shea

Bruce Walker

Sarah Young

Broad Sequencing Platform

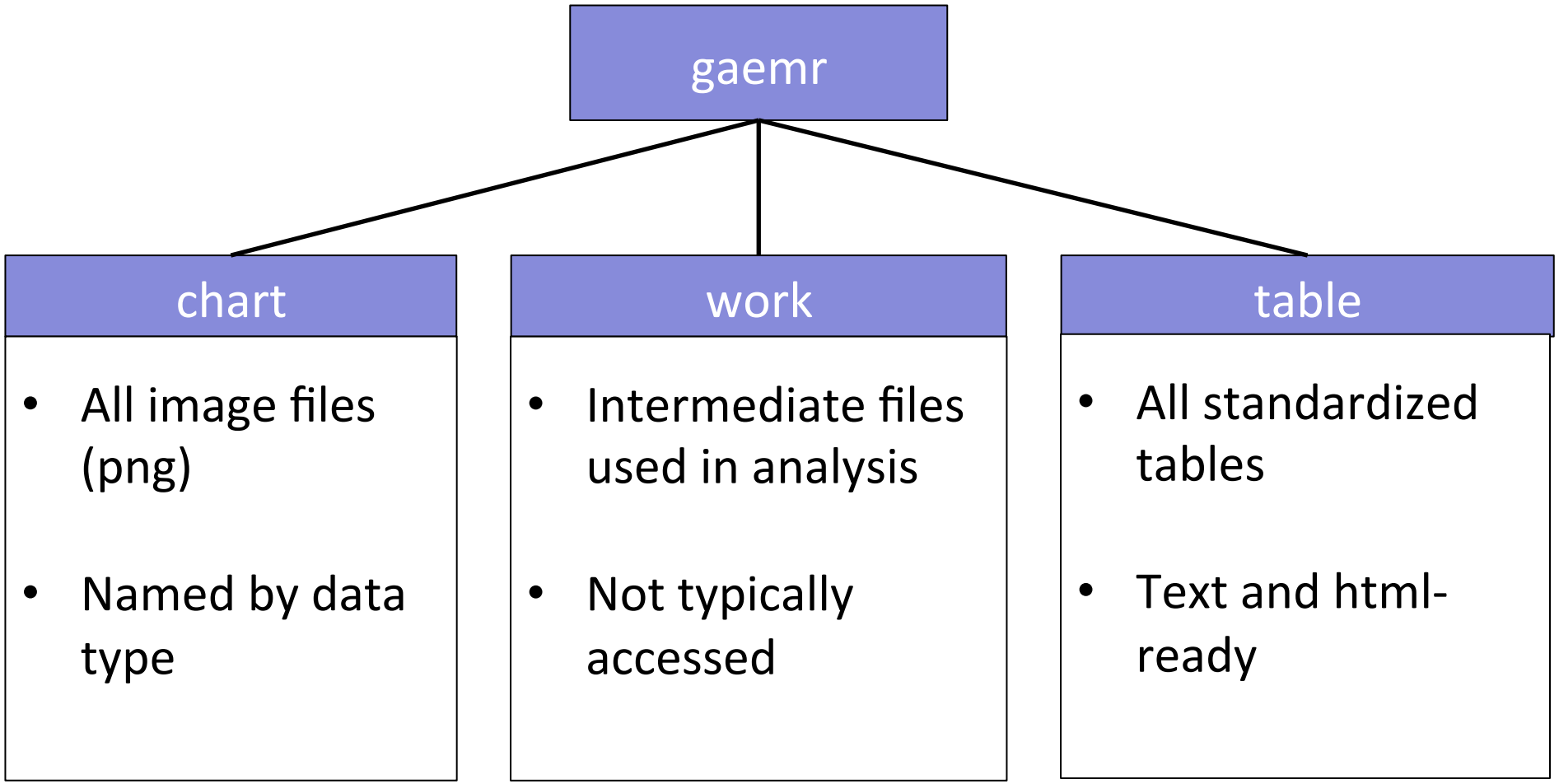
Genome Sequencing & Analysis Program

Chad Nusbaum

Carsten Russ

National Human Genome Research Institute

National Institute of Allergy and Infectious Diseases

BONUS SLIDES

Standardized Directory Structure