



Critical Assessment of Metagenome Interpretation

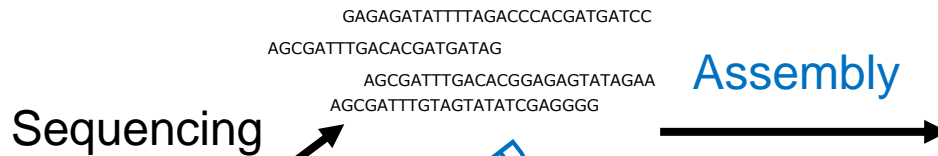
Alice C. McHardy

Computational Biology of Infection Research

Helmholtz Centre for Infection Research

and the CAMI Initiative

Computational Metagenomics



```
[...]sdATGACGATTCCGAAACAGAGCCGCGAGGTTCTTGATCGGGGATCGAGCATACGCGGGCGCTTTCC
CGGCCACTTTGATCAGTGGTTGCGGGGAGTTCAAGTTCGATGATCTGACCGAGCGCTCTGACGCTGCG
CGAGCAGACGCGCCATAGCGGCCAGCACGCTCGCGGCCCGCGCATGCGCTCGCCGCTGCTCTC
GGATCCGTTGAAGTAGACACGAGCGCGTACCTCGCGCCGCTCGC
```

```
CCTCGACGGGGTGCGCTGTTGCCCCCTTCGCTCTCTTGAAGACGCGCTCGCTGGATCTCGGC
```

```
GCTCTTCTCTCGTGAACGGCTCCAGGAAATCGGCGAGGTCGTTGCGCCGAACCGCTCGGCGTCCGG
TAGAAGAAGCCGAGGAGCGCTCTGTCAGCCGCCGCGCTCGGGGAACGATCTCGGGGTTCTTCGCA
GCATCGTGCTCAGCAGCTCGGCGAGCGACCGGGGACGTCGGGGCGCACGAGCGGAGCGCGGATACTC
```

```
GCTCGCTCGACGCCCGCATGATCTCGAACGCCGTCGGCGCCGGAACGGGTTGACCCGGCGAGCATC
GATGGTCTGTTTCGCGGGATGAGCTGAAGCTCAGCCGCCGCGGTTCTGACGCCGAGCGACAGCGGC
```

```
GTACGTCGAGGAGCAGCACTCTGTCACCTGCCCGGAGAGCGCGGCTTGCAGCGCAGCGCGGACCG
CGACGACCTCTCGGGGTTGACGCCCTTGTTCGGCTCCCGGCCGAAGAACTCGGCGACCGCGCTCGAC
CGCGGGCATCGGGTCATGCCCGCAGCAGGACACCGTGTACGCGCGGAGACGGGAGCTTGGGCTCC
CGAGCGTCGCGCAGACAGCTCGATGGTCCGCCGAGGATGAGGCCCTCGCAGAGCATCTCGAGCTGTTGC
GCCGATCTGCTCGCTCGAGGTGAGCGGCCCTCGCCGCGGCCGACGCGGATGAAGGGGATGTTGATCTC
```

```
GGTCTCGAGCGACGACGAGCTCTGCTTCCGCTTCTCGGCCCTCTTGAAGCGCTGACGCGCATG
CGATCCGGCGCAGATCGATCGGCTCTTCCCTCGAACTCTGCGGCGAGCAGGTCGATGATCCGCTGG
CGAAGTCTCGCGCGGAGGTGCTGTGCGCGCGCTCGCTTGAAGCTGAAGACCGCTCGGATCTC
GAGGATCGAGATATCAAGCTGCTCCGCCGAGATCGTAGCCGCGATGCTCGGCCCTTCACTTGTGC [...]
```

Bacteria 0.7

Archaea 0.3

Proteobacteria 0.2

Firmicutes 0.05

...

Genome & Taxonomic 'binning'

```
[...]sdATGACGATTCCGAAACAGAGCCGCGAGGTTCTTGATCGGGGATCGAGCATACGCGGGCGCTTTCC
CGGCCACTTTGATCAGTGGTTGCGGGGAGTTCAAGTTCGATGATCTGACCGAGCGCTCTGACGCTGCG
CGAGCAGACGCGCCATAGCGGCCAGCACGCTCGCGGCCCGCGCATGCGCTCGCCGCTGCTCTC
GGATCCGTTGAAGTAGACACGAGCGCGTACCTCGCGCCGCTCGC
```

```
CCTCGACGGGGTGCGCTGTTGCCCCCTTCGCTCTCTTGAAGACGCGCTCGCTGGATCTCGGC
```

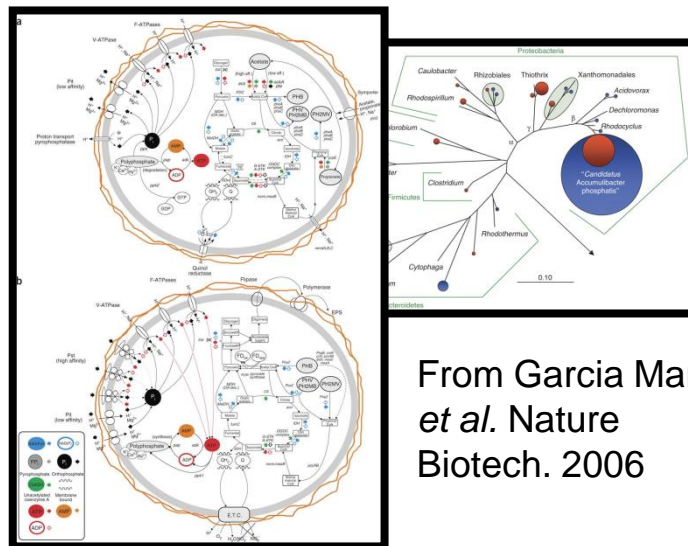
```
GCTCTTCTCTCGTGAACGGCTCCAGGAAATCGGCGAGGTCGTTGCGCCGAACCGCTCGGCGTCCGG
TAGAAGAAGCCGAGGAGCGCTCTGTCAGCCGCCGCGCTCGGGGAACGATCTCGGGGTTCTTCGCA
GCATCGTGCTCAGCAGCTCGGCGAGCGACCGGGGACGTCGGGGCGCACGAGCGGAGCGCGGATACTC
```

```
GCTCGCTCGACGCCCGCATGATCTCGAACGCCGTCGGCGCCGGAACGGGTTGACCCGGCGAGCATC
GATGGTCTGTTTCGCGGGATGAGCTGAAGCTCAGCCGCCGCGGTTCTGACGCCGAGCGACAGCGGC
```

```
GTACGTCGAGGAGCAGCACTCTGTCACCTGCCCGGAGAGCGCGGCTTGCAGCGCAGCGCGGACCG
CGACGACCTCTCGGGGTTGACGCCCTTGTTCGGCTCCCGCGGAAGAACTCGCGCACGCGCGCTCGAC
CGCGGGCATCGGGTCATGCCCGCAGCAGGAGACACCGTGTGACGCGCGAGACGGGAGCTTGGGCTCC
CGAGGCTCGCGCGAGCAGCTCGATGCTCGCGGATGAGGCCCTCGCAGAGCATCTCGAGCTGTTGC
GCCGATCTGCTCGCTCGAGGTGAGCGGCCCTCGCCGCGGCCGACGCGGATGAAGGGGATGTTGATCTC
```

```
GGTCTCGAGCGACGAGAGCTCTGCTTCCGCTTCTCGGCCCTCTTGAAGCGCTGACAGGCCATG
CGATCCGGCGCAGATCGATCGGCTCTTCCCTCGAACTCTGCGGCGAGCAGGTCGATGATCCGCTGG
CGAAGTCTCGCGCGGAGGTGCTGTGCGCGCGCTCGCTTGAAGCTGAAGACCGCTCGGATCTC
GAGGATCGAGATATCAAGCTGCTCCGCCGAGATCGTAGCCGCGATGCTCTCGGCCCTTCACTTGTGC [...]
```

Annotation



From Garcia Martin
et al. Nature
Biotech. 2006

The Critical Assessment of Metagenome Interpretation (CAMI) competition

27 Jun 2014 | 6:36 PM | Posted by Tal Nawy | Category: Bioinformatics, Computational, Guest Post, Metagenomics

Alice McHardy, Alex Sczyrba and Thomas Rattei announce a new initiative for assessing metagenomics methods in this guest post.



Alice McHardy
FOLKER MEYER



Alex Sczyrba
A. SCZYRBA



Thomas Rattei
ANJA VENIER

In just over a decade, metagenomics has developed into a powerful and productive method in microbiology and microbial ecology. The ability to retrieve and organize bits and pieces of genomic DNA from any natural context has opened a window into the vast universe of uncultivated microbes. Tremendous progress has been made in computational approaches to interpret this sequence data but none can completely recover the complex information encoded in metagenomes.

A number of challenges stand in the way. Simplifying

Towards a comprehensive and objective evaluation of computational metagenomics software

CASP 1 (1994)
Critical Assessment of Techniques for
Protein Structure Prediction

PROTEINS: Structure, Function, and Genetics 23:301–317 (1995)

A Critical Assessment of Comparative Molecular Modeling of Tertiary Structures of Proteins*

Steven Mosimann, Ron Meleshko, and Michael N.G. James

Medical Research Council of Canada, Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

ABSTRACT In spite of the tremendous increase in the rate at which protein structures are being determined, there is still an enormous gap between the numbers of known DNA-derived sequences and the numbers of three-dimensional structures. In order to shed light on the biological functions of the molecules, researchers often resort to comparative molecular modeling. Earlier work has shown that when the sequence alignment is in error, then the comparative model is guaranteed to be wrong. In addition, loops, the sites of insertions and deletions in families of homologous proteins, are exceedingly difficult to model. Thus, many of the current problems in comparative molecular modeling are minor versions of the global pro-

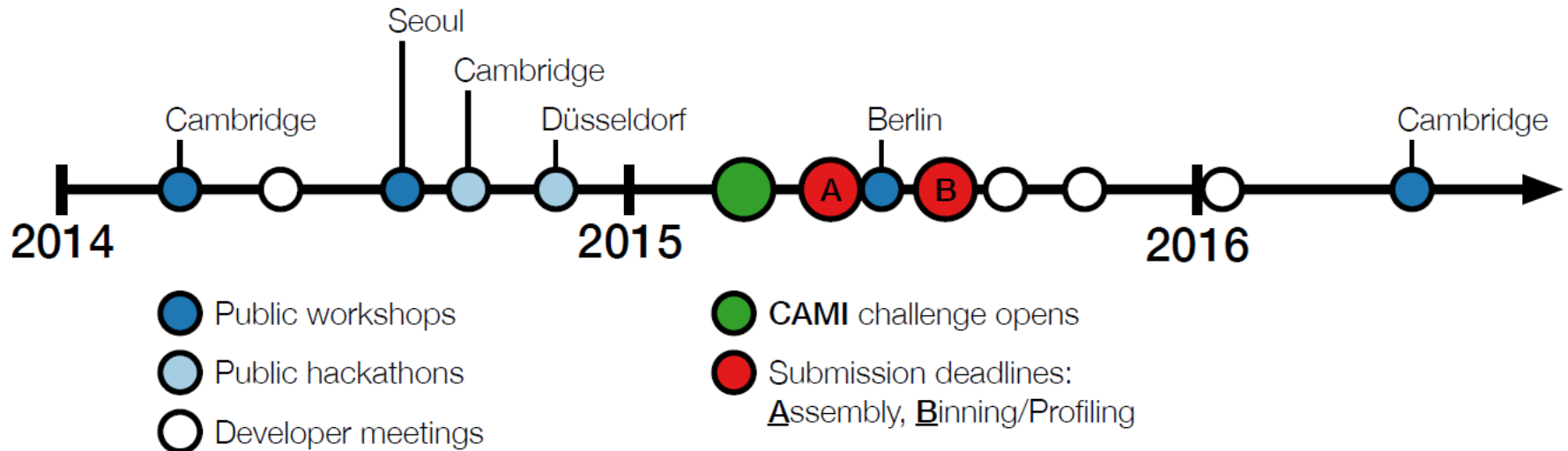
there are several commercial and public domain computer programs that have been developed for modeling; these programs remove much of the tedium from the process. There are numerous reasons for constructing comparative molecular models of proteins. The molecular model may explain the structural basis of existing experimental results and can provide one with structural information on which further experiments can be planned, executed, and evaluated. Site-specific mutations of the gene coding for the specific protein can provide important data regarding the protein's function. Perhaps, some of the most revealing experiments are those designed to predict and to probe the molecular reasons for an enzyme's specificity.³ On a more practical note, a molecular model can sometimes be used successfully



Principles

- Design decisions made by the community (data sets, evaluation measures and principles)
- Extensive, high-quality benchmark datasets from unpublished data
- Evaluation measures: informative to developers and the applied community
- Reproducibility: data generation, programs, evaluation
- Benchmark assembly, (tax.) binning and taxonomic profiling software

Timeline 1st CAMI Challenge

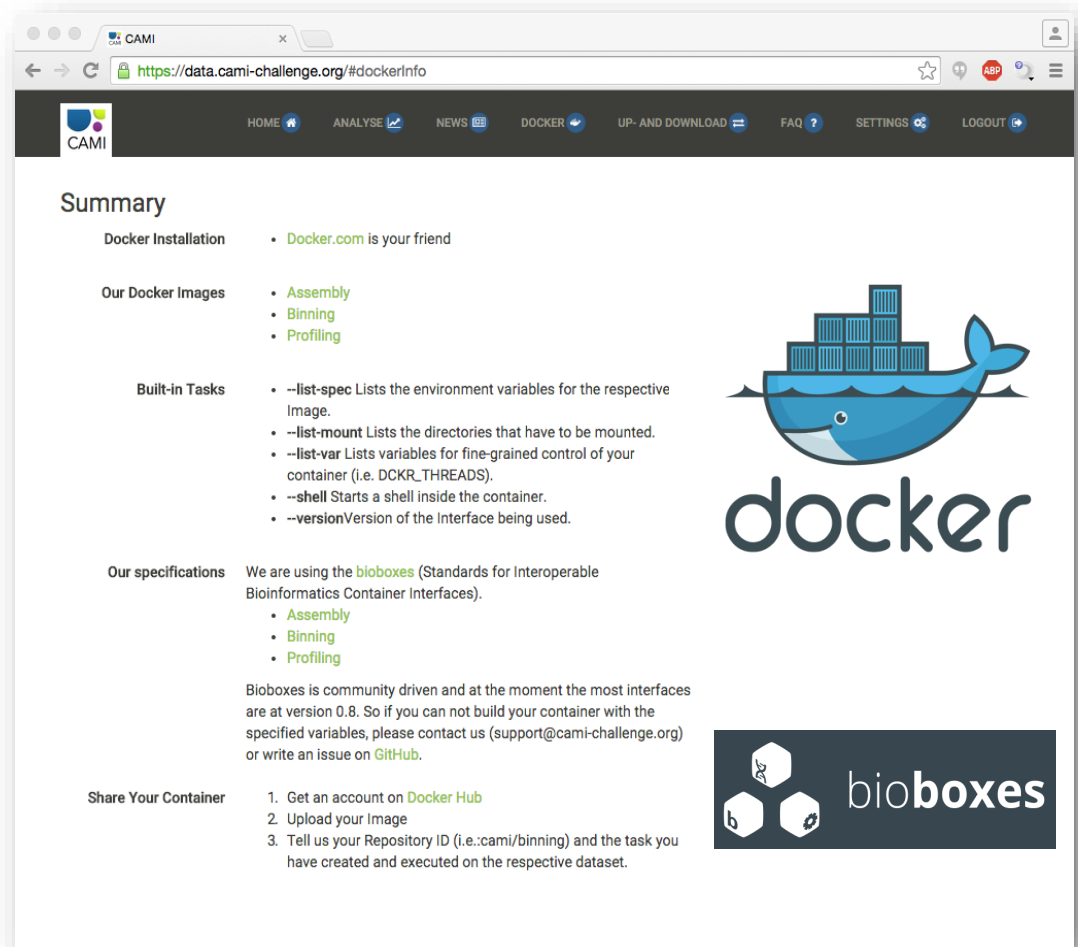


 Follow @CAMI_challenge

Sczyrba *et al.*, Nature Methods 2017

Reproducibility and Standardization

- Output formats for binning and profiling
- Standard interfaces
- Docker-based Bioboxes for programs and metrics
- Semi-automatic benchmarking in future



The screenshot shows a web browser displaying the CAMI DockerInfo page. The page has a dark navigation bar with links: HOME, ANALYSE, NEWS, DOCKER, UP- AND DOWNLOAD, FAQ, SETTINGS, and LOGOUT. The main content area is titled 'Summary' and contains several sections:

- Docker Installation:**
 - Docker.com is your friend
- Our Docker Images:**
 - Assembly
 - Binning
 - Profiling
- Built-in Tasks:**
 - list-spec Lists the environment variables for the respective Image.
 - list-mount Lists the directories that have to be mounted.
 - list-var Lists variables for fine-grained control of your container (i.e. DCKR_THREADS).
 - shell Starts a shell inside the container.
 - version Version of the Interface being used.
- Our specifications:**

We are using the **bioboxes** (Standards for Interoperable Bioinformatics Container Interfaces).

 - Assembly
 - Binning
 - Profiling

Bioboxes is community driven and at the moment the most interfaces are at version 0.8. So if you can not build your container with the specified variables, please contact us (support@cam-challenge.org) or write an issue on [GitHub](#).
- Share Your Container:**
 - Get an account on [Docker Hub](#)
 - Upload your Image
 - Tell us your Repository ID (i.e.:cam/binning) and the task you have created and executed on the respective dataset.

On the right side of the page, there is a large Docker logo (a blue whale with a stack of containers on its back) and a smaller bioboxes logo at the bottom right.

Belmann *et al.*, Gigascience 2015

CAMI Challenge Datasets

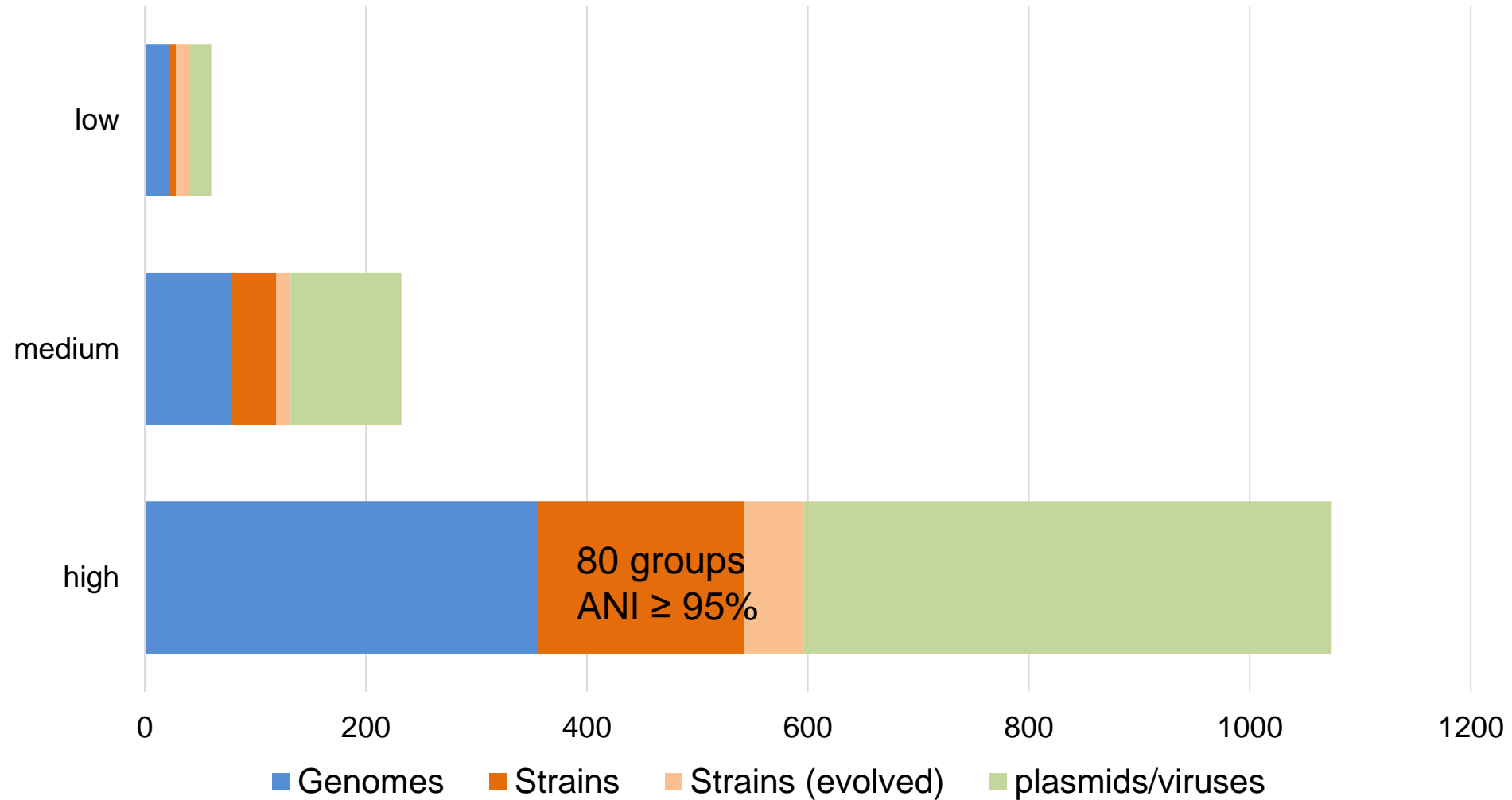
- Common experimental setups and community types
- Strain-level variation
- Different ev. distances to public genomes
- Non-bacterial sequences (archaea, plasmids, viruses)

CAMI_low	CAMI_medium (differential abundance)	CAMI_high (time series)
1 sample	2 samples	5 samples
15 Gb	40 Gb	75 Gb
2 x150 bp	2 x150 bp	2 x150 bp
Insert size: 270 bp	Insert sizes: 270 bp & 5kbp	Insert size: 270 bp

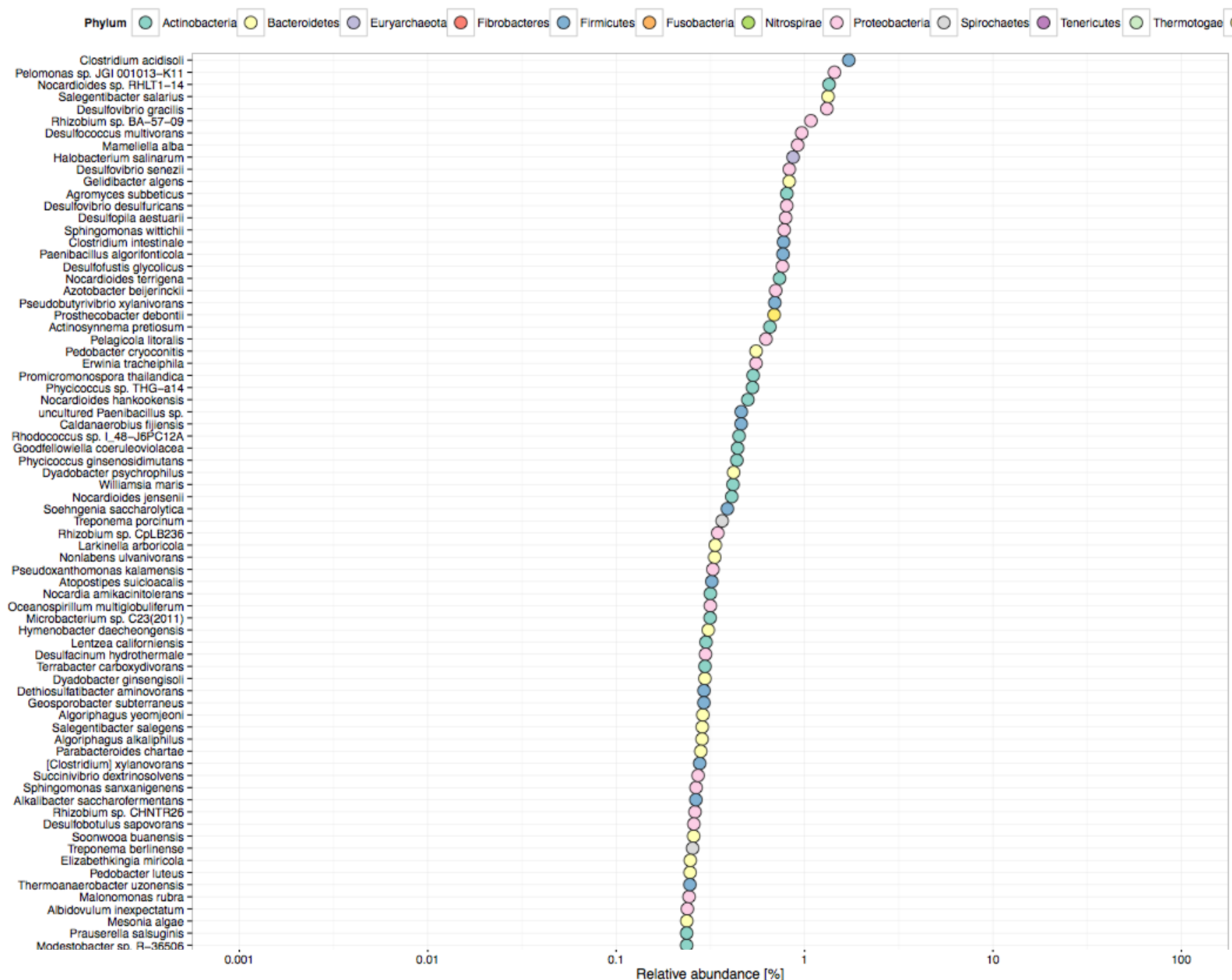
Simulated from ~700 novel microbial genomes, 600 novel viruses, plasmids and other circular elements

<https://github.com/CAMI-challenge/MetagenomeSimulationPipeline>

CAMI Challenge Datasets



High



Challenge Participants



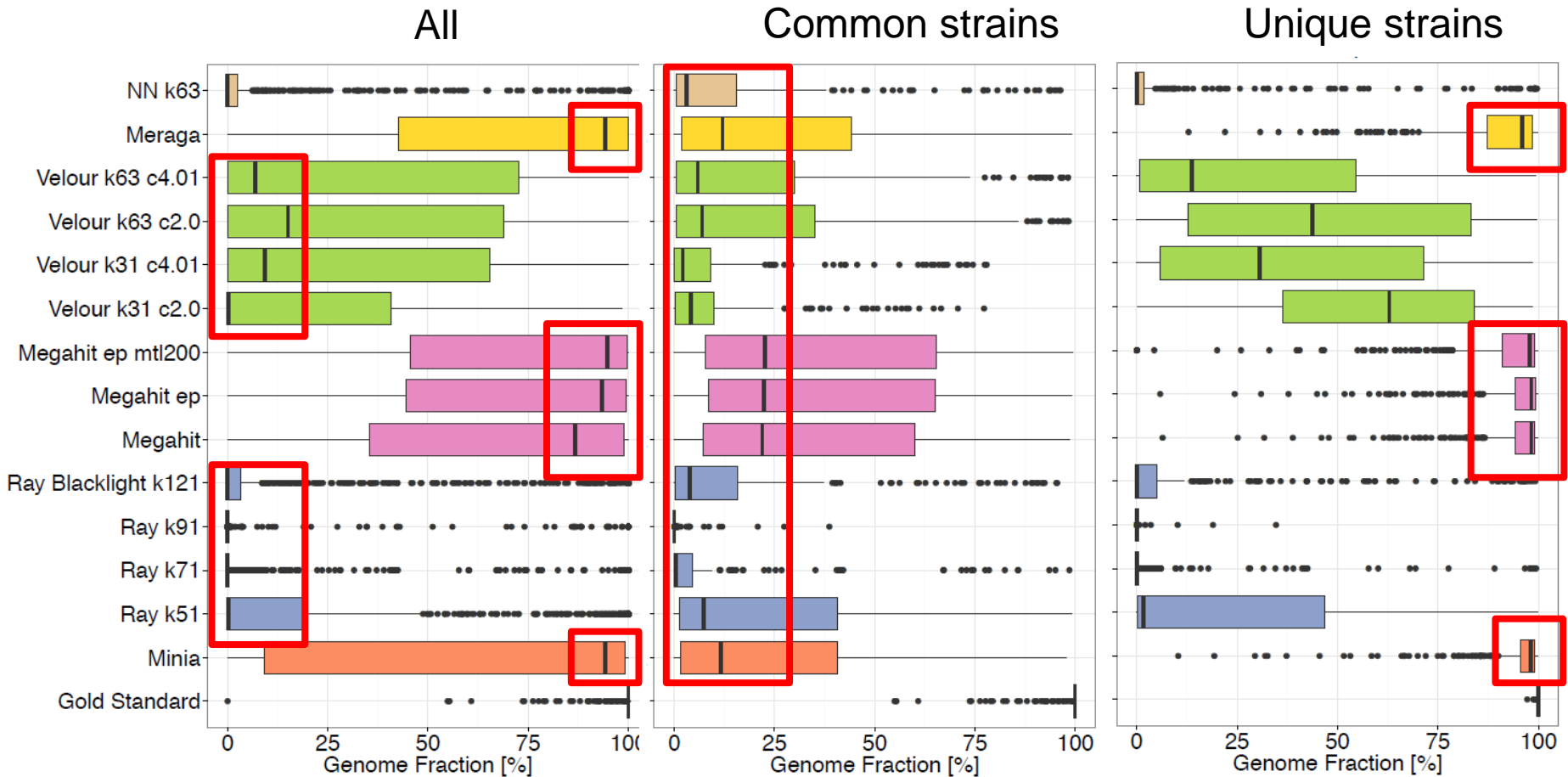
<https://data.cami-challenge.org>

Submissions to the CAMI Challenge

- With consent to publish: 215 submissions; 16 teams; 25 programs; 36 bioboxes

Assemblers		Genome binners and taxonomic binners	
Megahit ⁴	Metagenome assembler using succinct de Bruijn graph	CONCOCT ¹¹	Binner using differential coverage, tetranucleotide frequencies, paired-end linkage
Ray Meta ⁶	Distributed de Bruijn graph	MaxBin 2.0 ⁸	Binner using multi-sample coverage, tetranucleotide frequencies
Meraga	Meraculous ⁴³ + MEGAL		
Minia ⁵	De Bruijn graph	Kraken ¹⁵	Taxonomic binner using long k-mers and Lowest
A* ⁴⁴	MetaPhyler ²⁵	Phylogenetic marker genes	ted assignments
	mOTU ²⁶	Phylogenetic marker genes	ence similarities and
	Quikr/ARK/SEK ²⁷⁻²⁹	k-mer based nonnegative least squares	
Velour ₁	Taxy-Pro ³⁰	Mixture model analysis of protein signatures	verage, tetranucleotide ge
	TIPP ³¹	Marker genes and SATÉ phylogenetic placement	requencies and differential
	CLARK ²⁰	Phylogenetically discriminative k-mers	
	Common Kmers/MetaPalette ²¹	Long k-mer based nonnegative least squares	encies, multi-sample ylogenetic marker genes
			r frequencies (4-6mers),
	DUDes ²²	Read mapping and deepest uncommon ancestor	
	FOCUS ²³	k-mer based nonnegative least squares	ence homology and tax.
	MetaPhlAN 2.0 ²⁴	Clade specific marker genes	

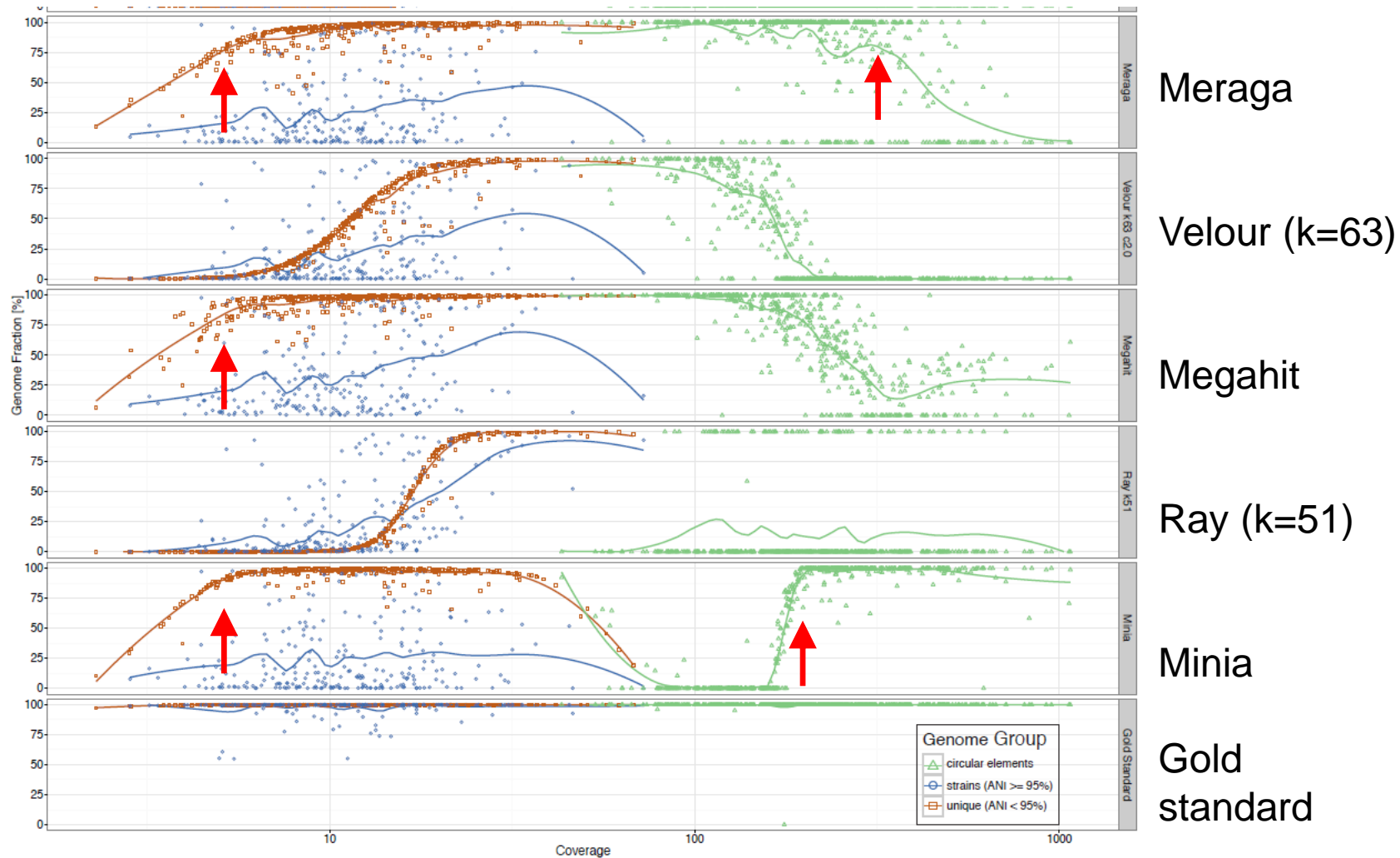
Results of the Assembly Challenge



Good performance in genome assembly for unique strains

Subspecies diversity is a challenge!

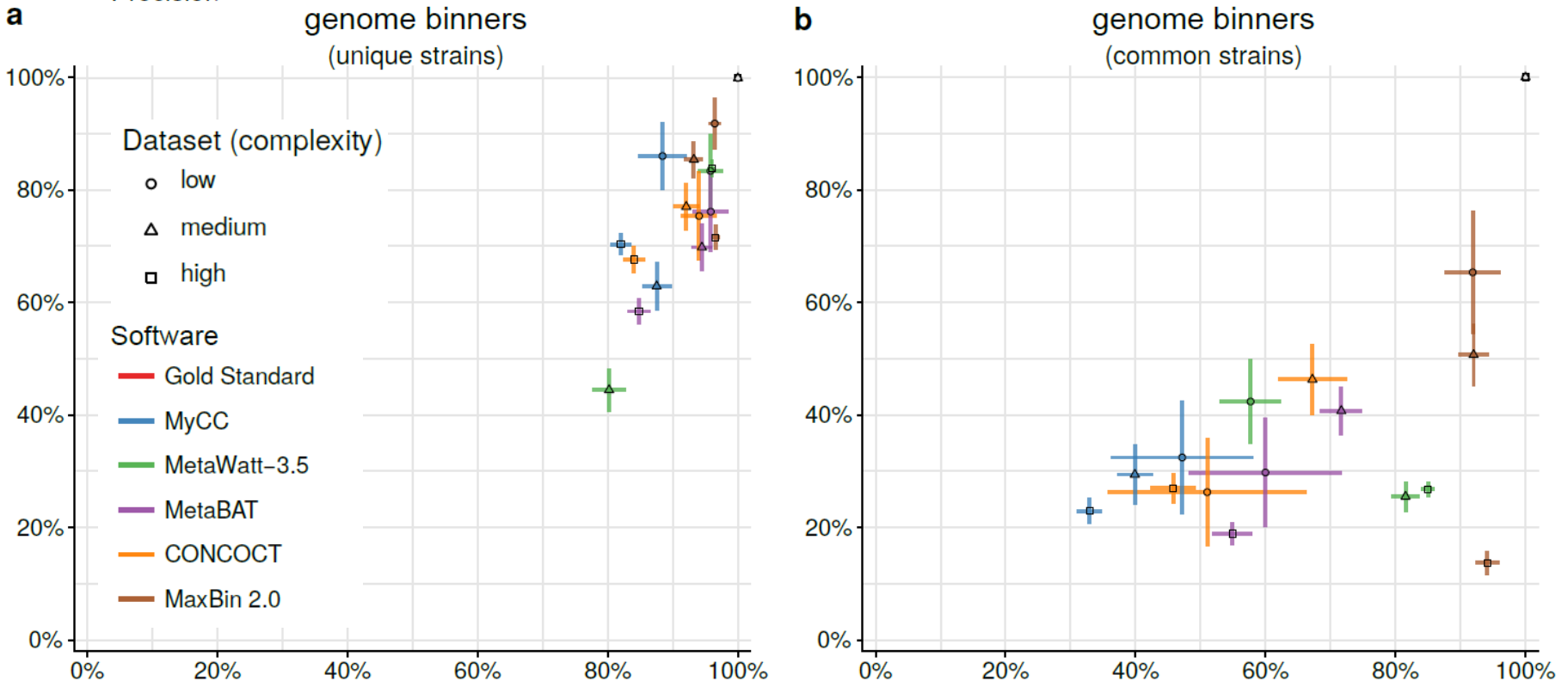
Genome Recovery at different Coverages



- Assemblers using multiple kmers performed better
- Minia (and Meraga) are good in plasmid assembly

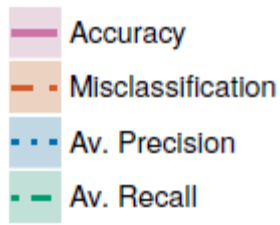
Results for Genome Binner

Recall
Precision



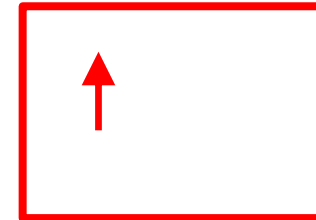
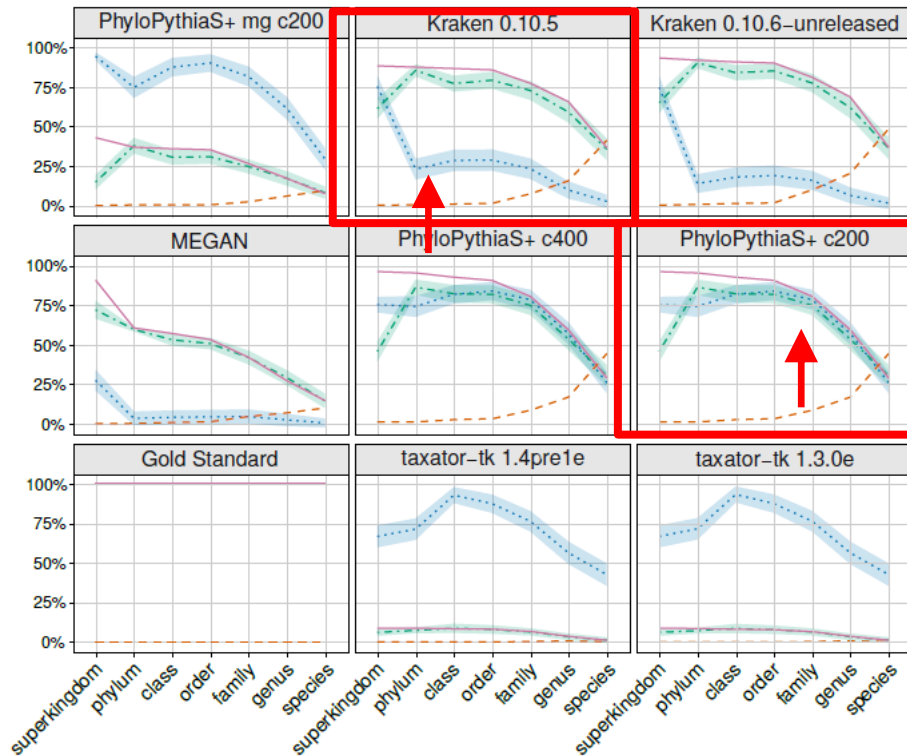
- Good performance in genome reconstruction for unique strains
- Subspecies diversity is a challenge!

Metric



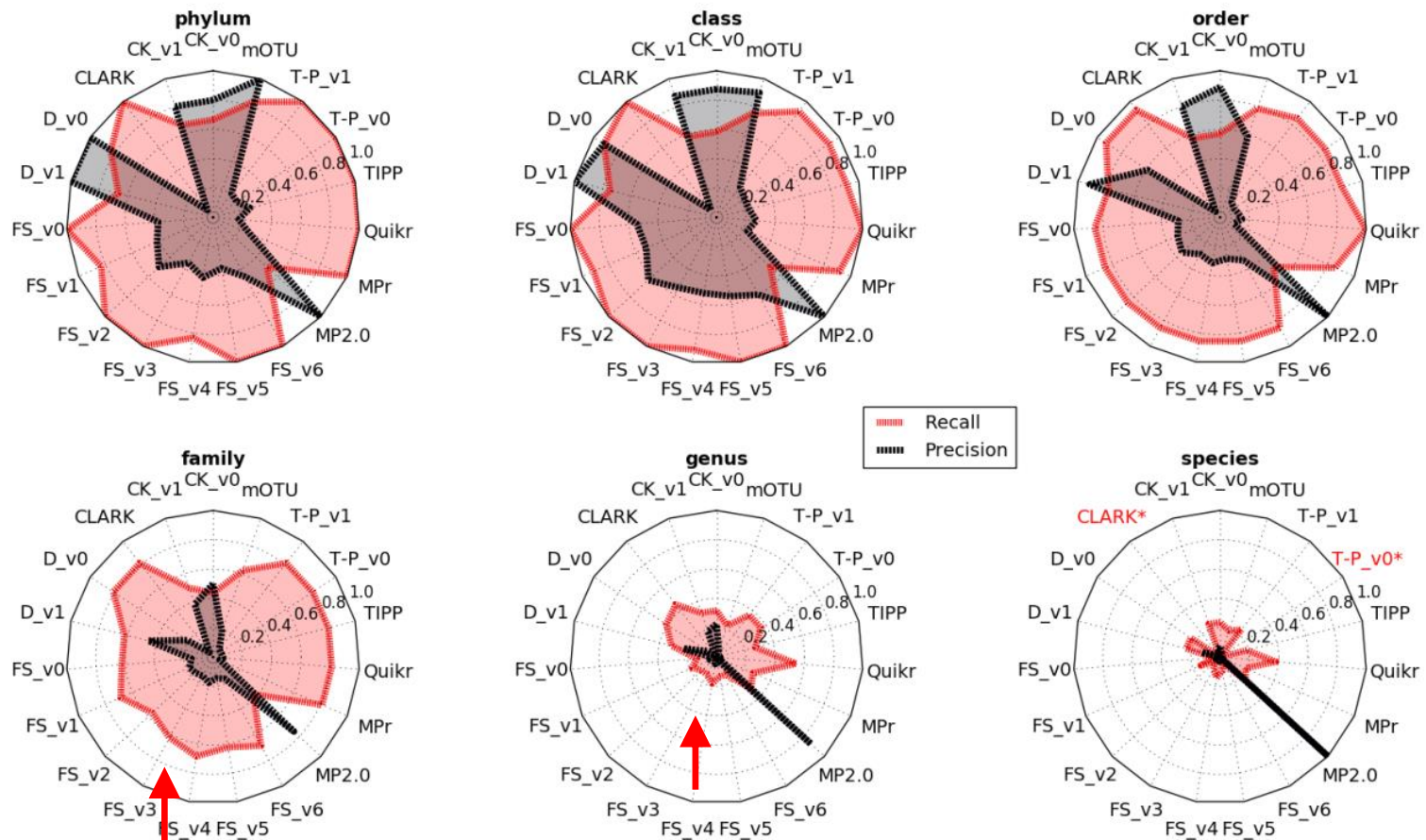
Results for Taxonomic Binner

MC - All data



- Good performance until family rank and substantial decrease below
- Small predicted bins are oftentimes false positives

Results for Taxonomic Profilers



- Performance decreases substantially below family rank
- Virus and plasmid data affect abundance estimates

Main Conclusions

- Good assembly and genome binning of „single strain“ species
- Strain-level diversity is a challenge
- Taxonomic profiling and taxon binning good until family level
- Reproducibility is very important, large variability of program performances with parameter settings

What is next?

- Benchmark programs on CAMI challenge datasets with CAMI benchmarking platform (www.data.cami-challenge.org)
- 2nd CAMI challenge
 - Illumina/PacBio/ONP data sets
 - specific environments
 - strain madness
 - workflows
- **Get in touch:** alice.mchardy@helmholtz-hzi.de; contact@camichallenge.org



Participants of CAMI Workshop at INI in Cambridge May 2016

Contributors 1st CAMI Challenge

P. Hofmann, P. Belmann, D. Koslicki, T. Woyke, N. Shapiro, S. Janssen, M. Barton, P. D. Blood, S. Majda, J. Dröge, I. Gregor, J. Fiedler, E. Dahms, R. Garrido-Oter, A. Bremges, A. Fritz, M. Z. DeMaere, C. Quince, T. Sparholt Jørgensen, L. Hestbjerg Hansen, S. J. Sørensen, Y. Bai, D. Turaev, M. Beckstette, M. Balvociute, F. Meyer, N. Nagarajan, B. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. Cook, R. Chikhi, C. Deltel, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, T. Lingner, H.-H. Lin, Y.-C. Liao, G. Gueiros Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Goeker, M. Balvociute, N. Kyrpides, J. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, A. Sczyrba



46 institutions





Why CAMI?

Tool development for shotgun metagenome data sets is a very active area: Assembly, (tax.) binning, taxonomic profiling

- Method papers present evaluations using many different metrics, simulated data sets (snapshots) and are difficult to compare
- It is unclear to everyone which tools are most suitable for a particular task and for particular data sets
- Comparative benchmarking requires extensive resources and there are pitfalls

Towards a comprehensive, independent and unbiased evaluation of computational metagenome analysis methods

CASP 1 (1994)
Critical Assessment of techniques for
protein Structure Prediction

PROTEINS: Structure, Function, and Genetics 23:301–317 (1995)

A Critical Assessment of Comparative Molecular Modeling of Tertiary Structures of Proteins*

Steven Mosimann, Ron Meleshko, and Michael N.G. James

Medical Research Council of Canada, Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

ABSTRACT In spite of the tremendous increase in the rate at which protein structures are being determined, there is still an enormous gap between the numbers of known DNA-derived sequences and the numbers of three-dimensional structures. In order to shed light on the biological functions of the molecules, researchers often resort to comparative molecular modeling. Earlier work has shown that when the sequence alignment is in error, then the comparative model is guaranteed to be wrong. In addition, loops, the sites of insertions and deletions in families of homologous proteins, are exceedingly difficult to model. Thus, many of the current problems in comparative molecular modeling are minor versions of the global pro-

there are several commercial and public domain computer programs that have been developed for modeling; these programs remove much of the tedium from the process. There are numerous reasons for constructing comparative molecular models of proteins. The molecular model may explain the structural basis of existing experimental results and can provide one with structural information on which further experiments can be planned, executed, and evaluated. Site-specific mutations of the gene coding for the specific protein can provide important data regarding the protein's function. Perhaps, some of the most revealing experiments are those designed to predict and to probe the molecular reasons for an enzyme's specificity.³ On a more practical note, a molecular model can sometimes be used successfully



GASP1 (1999) Genome Annotation Assessment Project

Participating Groups

Introduction

Following groups are participating:

1. Team Gaasterland, program name: MAGPIE

Terry Gaasterland	group leader, Laboratory for Computational Genomics, Rockefeller
Alexander Sczyrba	Laboratory for Computational Genomics, Rockefeller
Elizabeth Thomas	Laboratory for Computational Genomics, Rockefeller
Gulriz Kurban	Laboratory for Computational Genomics, Rockefeller
Paul Gordon	Institute for Marine Biosciences, Canada
Christoph Sensen	Institute for Marine Biosciences, Canada

2. [Computational Genomics Group](#), The Sanger Centre, program name: FGenes/FGenesH

Victor [Solovyev](#), group leader
Asaf [Salamov](#)

3. Genome Annotation Group, The Sanger Centre, program name: [Wise2](#)/GeneWise

Ewan [Birney](#), group leader

4. [Chair for Pattern Recognitions](#), The University of Erlangen/Nuremberg, program name: LME/MCPromoter

Uwe [Ohler](#), group leader
George [Stemmer](#)
Stefan [Harbeck](#)
Heinrich Niemann

5. [Computational Biosciences](#), Oakridge National Laboratory, program name: GRAIL

Richard J. [Mural](#), group leader
Douglas [Hyatt](#)
Frank Larimer
Manesh Shah
Morey Parang

A large-scale evaluation of computational protein function prediction

Predrag Radivojac¹, Wyatt T Clark¹, Tal Ronnen Oron², Artem Sokolov^{4,5}, Kiley Graim⁴, Christopher Funk⁶, Karim Jeffrey M Yunes¹⁰, Ameet S Talwalkar¹¹, Susanna Repo^{8,12}, Rita Casadio¹⁴, Zheng Wang¹⁵, Jianlin Cheng¹⁵, Hai Fang¹⁶, Jussi Nokso-Koivisto¹⁷, Liisa Holm¹⁷, Domenico Cozzetto¹⁸, David T Jones¹⁸, Bhakti Limaye¹⁹, Harshal Inamdar¹⁹, Avik Datta¹⁹, Sunitha Meghana Chitale²⁰, Daisuke Kihara^{20,21}, Andreas M Lisewski²², Serkan Erdin²², Eric Robert Rentzsch²³, Haixuan Yang²⁴, Alfonso E Romero²⁴, Prajwal Bhat²⁴, Alberto Paccanaro²⁵, Rebecca Kaßner²⁵, Stefan Seemayer²⁵, Esmeralda Vicedo²⁵, Christian Schaefer²⁵, Dominik Achten²⁵, Florian Auer²⁵, Ariane Boehm²⁵, Tatjana Braun²⁵, Maximilian Hecht²⁵, Mark Heron²⁵, Peter Hönigsmid²⁵, Thomas A Hopf²⁵, Stefanie Kaufmann²⁵, Michael Kiening²⁵, Denis Krompass²⁵, Cedric Landerer²⁵, Yannick Mahlich²⁵, Manfred Roos²⁵, Jari Björne²⁶, Tapio Salakoski²⁶, Andrew Wong²⁷, Hagit Shatkay^{27,28}, Fanny Gatzmann²⁹, Ingolf Sommer²⁹, Mark N Wass^{30,31}, Michael J E Sternberg³⁰, Nives Škunca³², Fran Supek³², Matko Bošnjak³², Panče Panov³³, Sašo Džeroski³³, Tomislav Šmuc³², Yiannis A I Kourmpetis^{34,35}, Aalt D J van Dijk^{34,36}, Cajo J F ter Braak³⁴, Yuanpeng Zhou³⁷, Qingtian Gong³⁷, Xinran Dong³⁷, Weidong Tian³⁷, Marco Falda³⁸, Paolo Fontana³⁹, Enrico Lavezzo³⁸, Barbara Di Camillo⁴⁰, Stefano Toppo³⁸, Liang Lan⁴¹, Nemanja Djuric⁴¹, Yuhong Guo⁴¹, Slobodan Vucetic⁴¹, Amos Bairoch^{42,43}, Michal Linial⁴⁴, Patricia C Babbitt³, Steven E Brenner⁸, Christine Orengo²³, Burkhard Rost²⁵, Sean D Mooney² & Iddo Friedberg^{45,46}

CAFA (2013)
Critical Assessment of protein
Function Annotation

Automated annotation of protein function is challenging. As the number of sequenced genomes rapidly grows, the overwhelming majority of protein products can only be annotated computationally. If computational predictions are to be relied upon, it is crucial that the accuracy of these methods be high. Here we report the results from the first large-scale community-based critical assessment of protein function annotation (CAFA) experiment. Fifty-four methods representing the state of the art for protein function prediction were evaluated on a target set of 866 proteins from 11 organisms. Two findings stand out: (i) today's best protein function prediction algorithms substantially outperform widely used first-generation methods, with large gains on all types of targets; and (ii) although the top methods perform well enough to guide experiments, there is considerable need for improvement of currently available tools.

available¹. The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology.

Many solutions have been proposed in the last four decades^{2–10}, yet the task of computational functional inference in a laboratory often relies on traditional approaches such as identifying domains or finding Basic Local Alignment Search Tool (BLAST)¹¹ hits among proteins with experimentally determined function. Recently, the availability of genomic-level sequence information for thousands of species, coupled with massive high-throughput experimental data, has created new opportunities for function prediction. A large number of methods have been proposed to exploit these data, including function prediction from amino acid sequence^{12–16}, inferred evolutionary relationships and genomic context^{17–21}, protein-protein interaction networks^{22–25}, protein structure data^{26–28}, microarrays²⁹ or a combination^{30–34}. An unbiased evaluation of these different



First CAMI challenge

- Benchmark assembly, (tax.) binning and taxonomic profiling software
- Extensive, high-quality benchmark data sets from unpublished data
- Publication with participants and data contributors

Aims

- Overview of tools and use cases
- Standards
- Facilitate future benchmarking
- Indicate promising directions for tool development
- Suggestions for experimental design

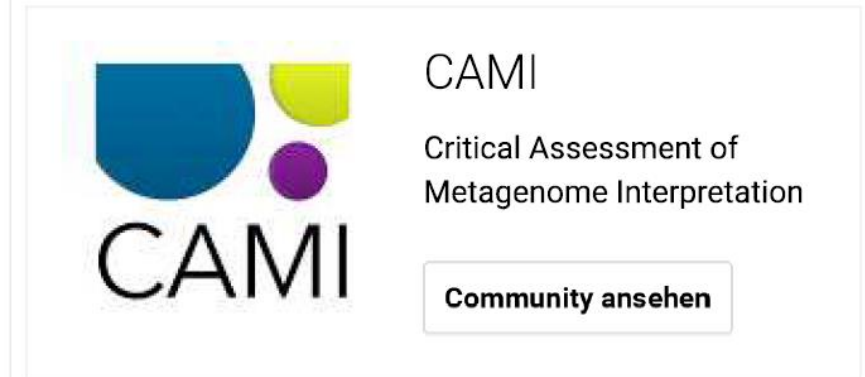
Contest opened in **early 2015**

- All design decisions (data sets, evaluation measures and principles) should involve the community
- Data sets should be as realistic as possible
- Evaluation measures should be informative to developers and understandable also by applied community
- Reproducibility (data generation, tools, evaluation)
- Participants should not see any of the data before

Community Involvement

- ISME Roundtable, Hackathons & workshops
- Announcements in blogs & tweets
- www.cami-challenge.org with newsletter

■ Google+ community



NATURE METHODS | METHAGORA

The Critical Assessment of Metagenome Interpretation (CAMI) competition

27 Jun 2014 | 6:36 PM | Posted by Tal Nawy | Category: Bioinformatics, Computational, Guest Post, Metagenomics

Alice McHardy, Alex Sczyrba and Thomas Rattei announce a new initiative for assessing metagenomics methods in this guest post.



Alice McHardy
FOLKER MEYER



Alex Sczyrba
A. SCZYRBA



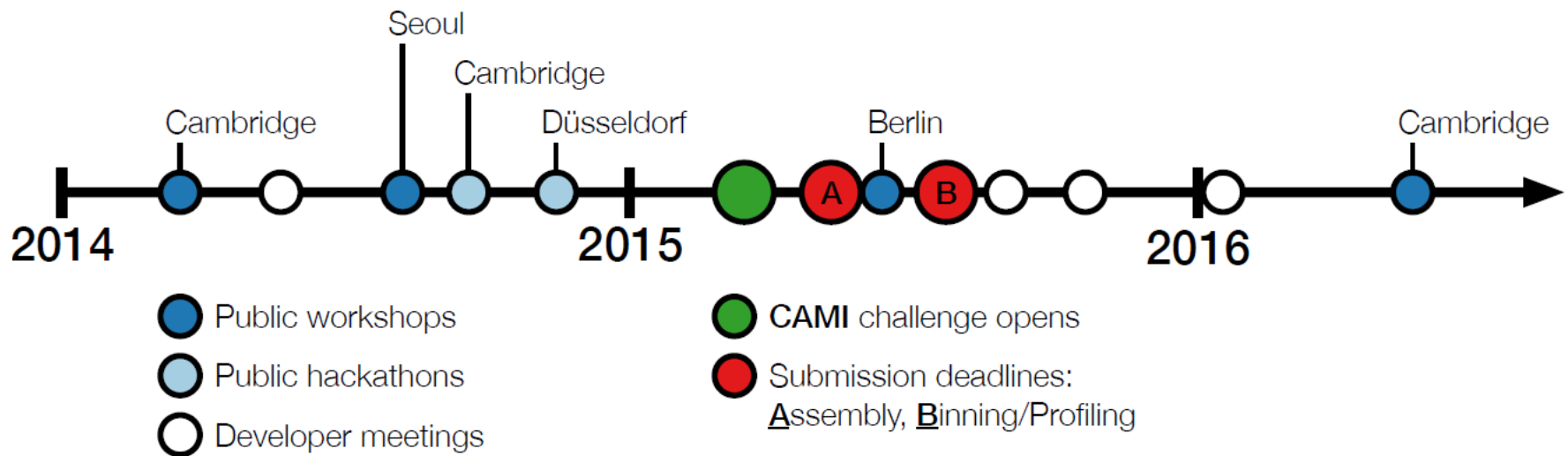
Thomas Rattei
ANJA VENIER

In just over a decade, metagenomics has developed into a powerful and productive method in microbiology and microbial ecology. The ability to retrieve and organize bits and pieces of genomic DNA from any natural context has opened a window into the vast universe of uncultivated microbes. Tremendous progress has been made in computational approaches to interpret this sequence data but none can completely recover the complex information encoded in metagenomes.

A number of challenges stand in the way. Simplifying

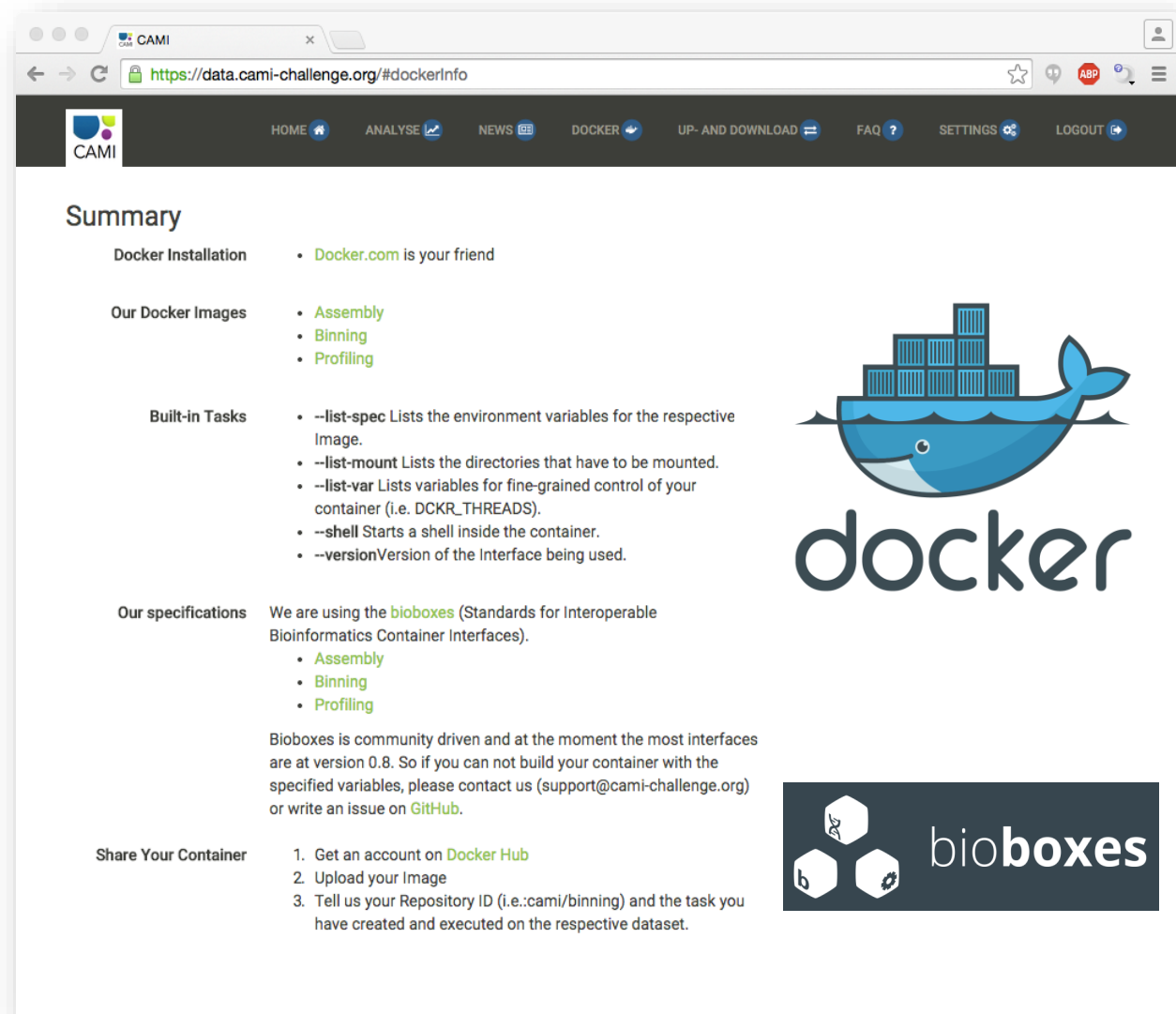
 Follow @CAMI_challenge

Timeline first CAMI Challenge



Reproducibility and Standardization

- Standard formats for binning and profiling
- Standard interfaces for tool execution
- Bioboxes (docker containers) for tools and metrics
- Currently 25 tools in bioboxes – **semi-automatic benchmarking in future challenges**



The screenshot shows a web browser displaying the CAMI Docker page at <https://data.cami-challenge.org/#dockerInfo>. The page features a navigation bar with links to HOME, ANALYSE, NEWS, DOCKER, UP- AND DOWNLOAD, FAQ, SETTINGS, and LOGOUT. The main content is titled "Summary" and includes sections for Docker Installation, Our Docker Images, Built-in Tasks, Our specifications, and Share Your Container. The Docker logo is prominently displayed on the right side of the page.

Summary

Docker Installation

- [Docker.com](#) is your friend

Our Docker Images

- [Assembly](#)
- [Binning](#)
- [Profiling](#)

Built-in Tasks

- `--list-spec` Lists the environment variables for the respective Image.
- `--list-mount` Lists the directories that have to be mounted.
- `--list-var` Lists variables for fine-grained control of your container (i.e. DCKR_THREADS).
- `--shell` Starts a shell inside the container.
- `--version` Version of the Interface being used.

Our specifications

We are using the [bioboxes](#) (Standards for Interoperable Bioinformatics Container Interfaces).

- [Assembly](#)
- [Binning](#)
- [Profiling](#)

Bioboxes is community driven and at the moment the most interfaces are at version 0.8. So if you can not build your container with the specified variables, please contact us (support@cam-challenge.org) or write an issue on [GitHub](#).

Share Your Container

1. Get an account on [Docker Hub](#)
2. Upload your Image
3. Tell us your Repository ID (i.e.: `cam/binning`) and the task you have created and executed on the respective dataset.

docker

bioboxes

Barton *et al.*, Gigascience 2015

Challenge Data sets – Design principles

- As realistic as possible, challenging
- Common experimental setups and community types
- Unpublished data
- Strain-level variation
- Different taxonomic distances to sequenced genomes (deep branchers included)
- State-of-the-art sequencing technologies
- Non-bacterial sequences



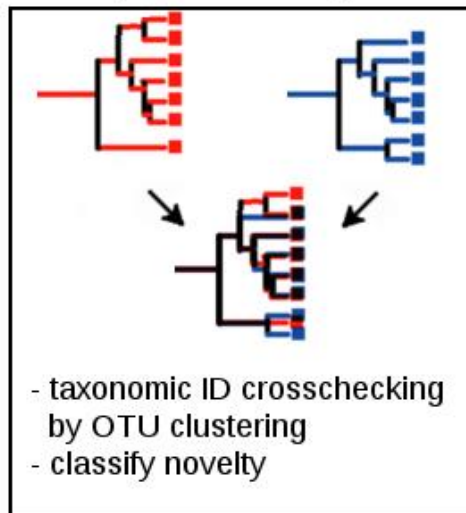
CAMI

Simulated Metagenome Sample Generation

Taxonomic metadata definition

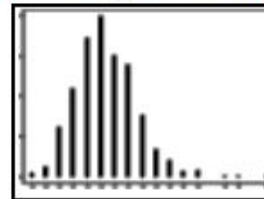
Isolate strain
genome
assemblies

Reference¹
genome
assemblies

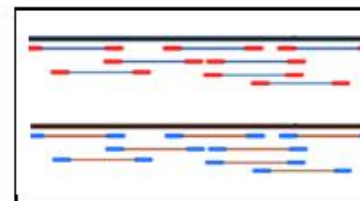


1. Prepared from NCBI, HMP, JGI data

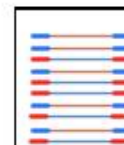
Community & sequence sample



- select specified # strains
per novelty category
- adjustable distribution



- simulate paired-end
reads



- FASTQ of anonymized &
shuffled reads

Datasets simulated from ~700 unpublished microbial genomes and additional sequence material

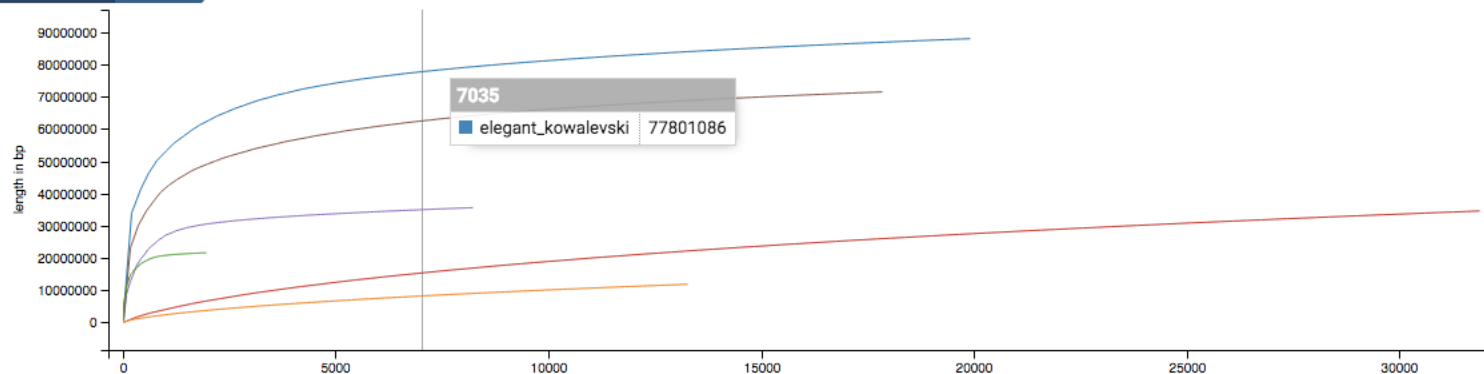
1st Challenge Timeline



CAMI Evaluation Metrics

Anonymous Name	# contigs	Largest contig	Total length	N50	GC (%)	# misassemblies	# relocations	# translocations	# inversions	# misassembled contigs	# indels
pensive_babbage	10.0	6503724.0	8361599.0	6503724.0	46.96	-	-	-	-	-	-
focused_bardeen	8864.0	487875.0	3.680766E7	23804.0	54.17	28.0	5.0	23.0	0.0	27.0	-
sharp_perلمان	17911.0	888870.0	7.155452E7	19216.0	54.92	-	-	-	-	-	-
adoring_jones	16018.0	457213.0	4.7181912E7	13601.0	54.0	-	-	-	-	-	-
goofy_darwin	15795.0	888811.0	4.8721356E7	28403.0	53.75	172.0	33.0	130.0	9.0	125.0	93.0
elegant_kowalevski	20004.0	2780101.0	8.807724E7	24752.0	54.62	0.0	0.0	0.0	0.0	0.0	-
trusting_colden	184.0	8.0	133.0	816.0	0.18	67.0	14.0	4775.0	9.77	-	-
lonely_franklin	32148.0	10239.0	3.4692516E7	1183.0	56.54	45.0	6.0	37.0	2.0	45.0	-
drunk_galileo	8218.0	307410.0	3.558718E7	23934.0	54.41	-	-	-	-	-	-
hungry_jones	13325.0	15618.0	1.1841911E7	889.0	55.27	0.0	0.0	0.0	0.0	0.0	4.0
elated_wright	1957.0	515047.0	2.1586394E7	65409.0	57.52	80.0	24.0	49.0	7.0	65.0	100.0

CONTIG LENGTH NX GC

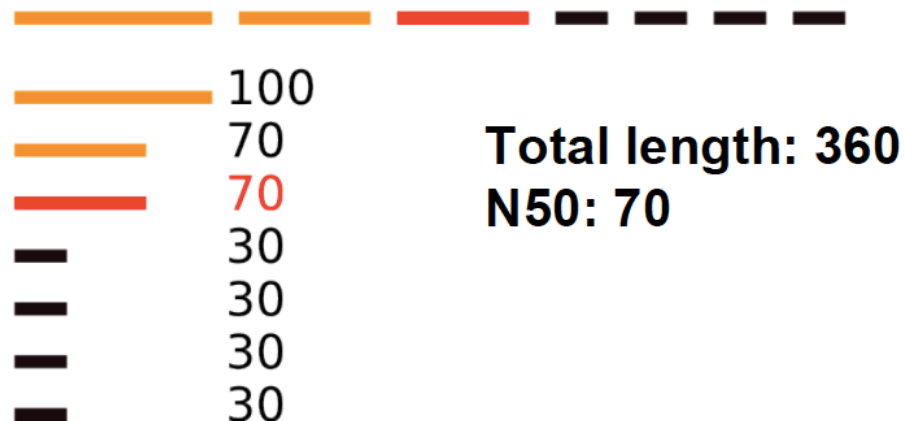


Assembly Evaluation Metrics

Basic Statistics:

- Number of contigs
- Number of large contigs (i.e. > 1000 bp)
- Largest contig length
- Total assembly length
- N50:

The length for which the collection of all contigs of that length or longer covers at least half an assembly (50%)

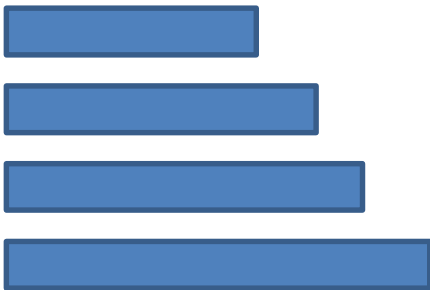


Assembly Evaluation Metrics

Reference-based statistics:

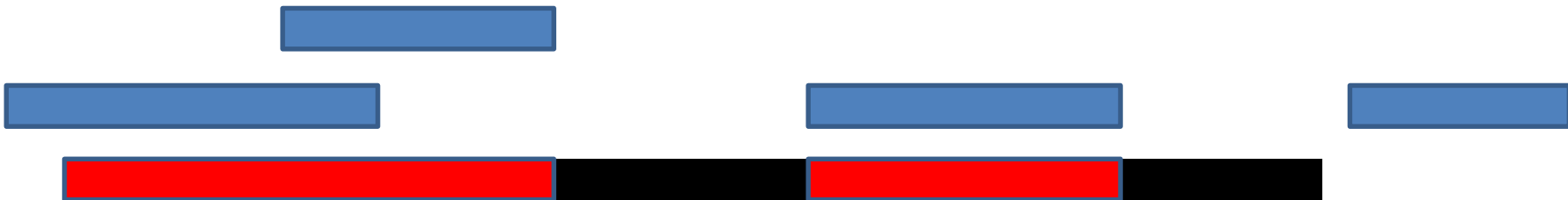
- Reference length
- Reference GC %
- Number of chromosomes
- Number of genes/operons
- NGx, LGx

Alignment Statistics



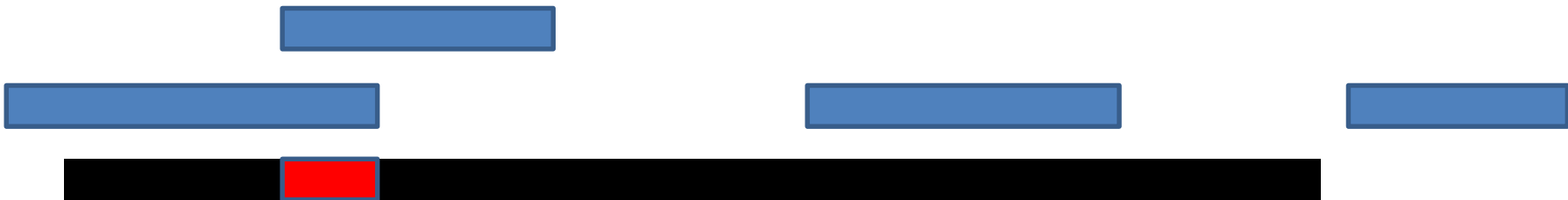
Alignment Statistics

- Genome fraction %



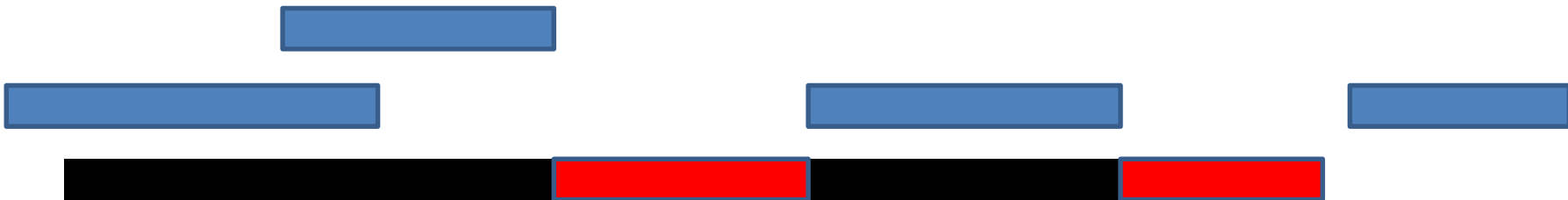
Alignment Statistics

- Genome fraction %
- Duplication ratio



Alignment Statistics

- Genome fraction %
- Duplication ratio
- Number of gaps



Alignment Statistics

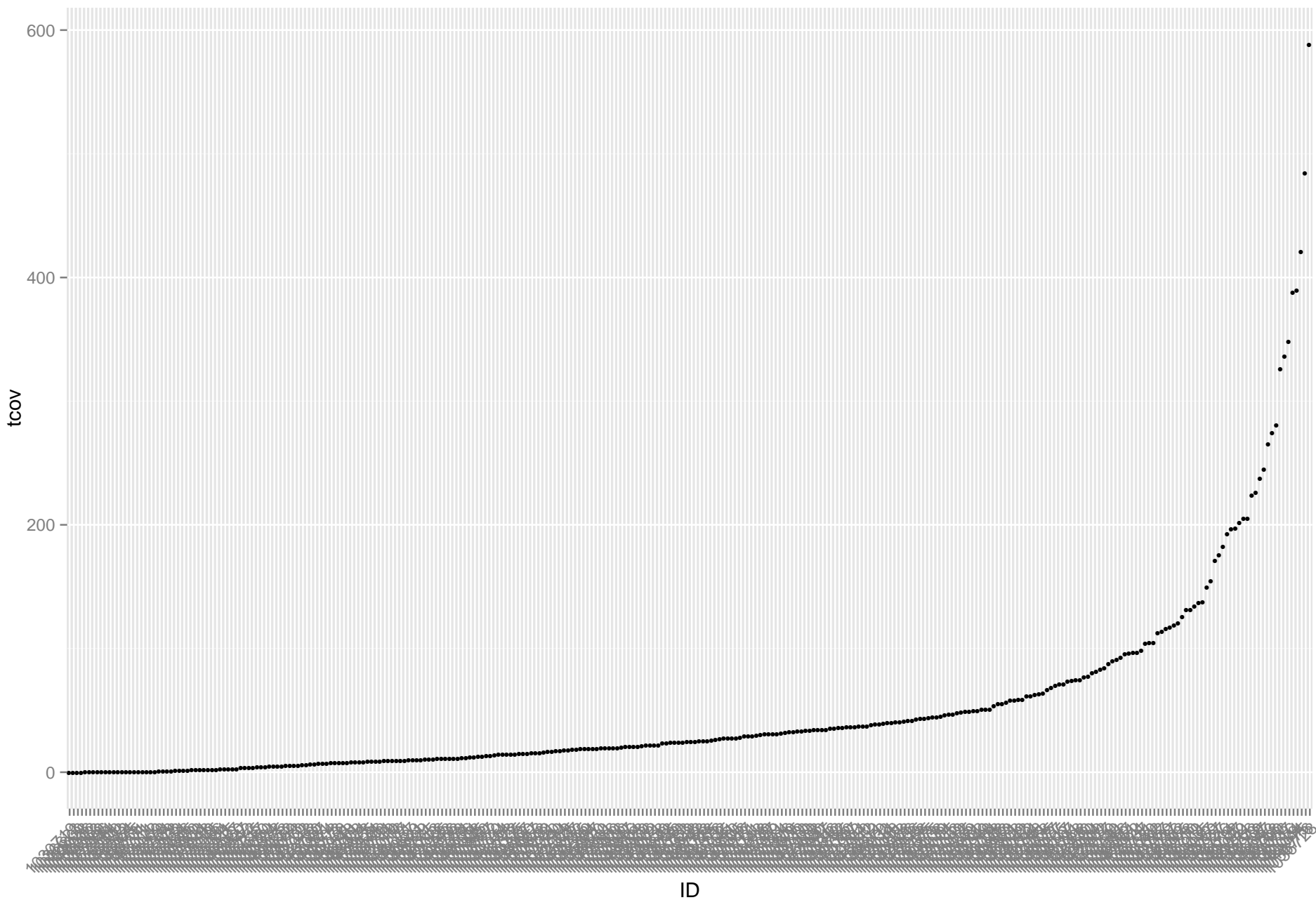
- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length

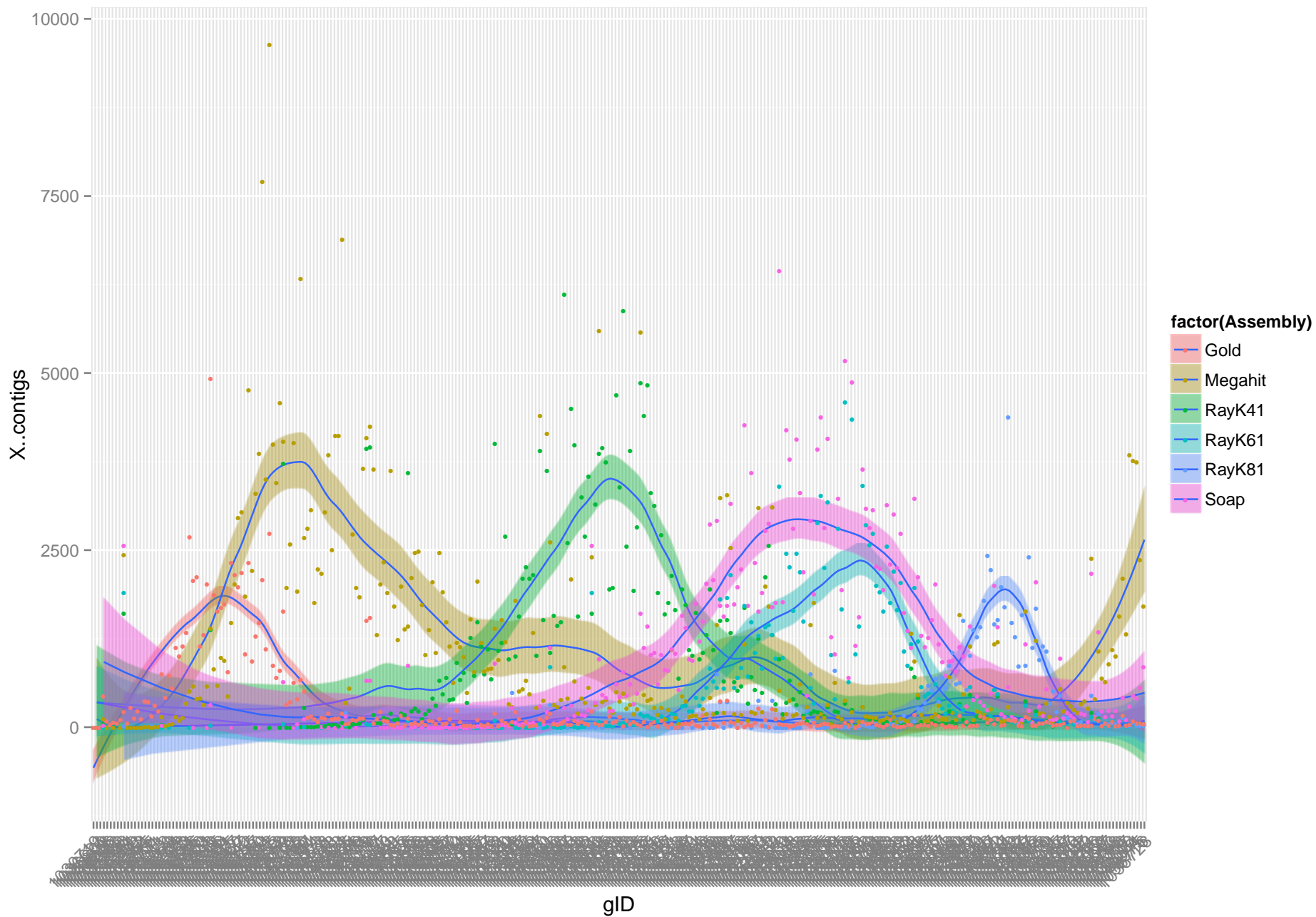


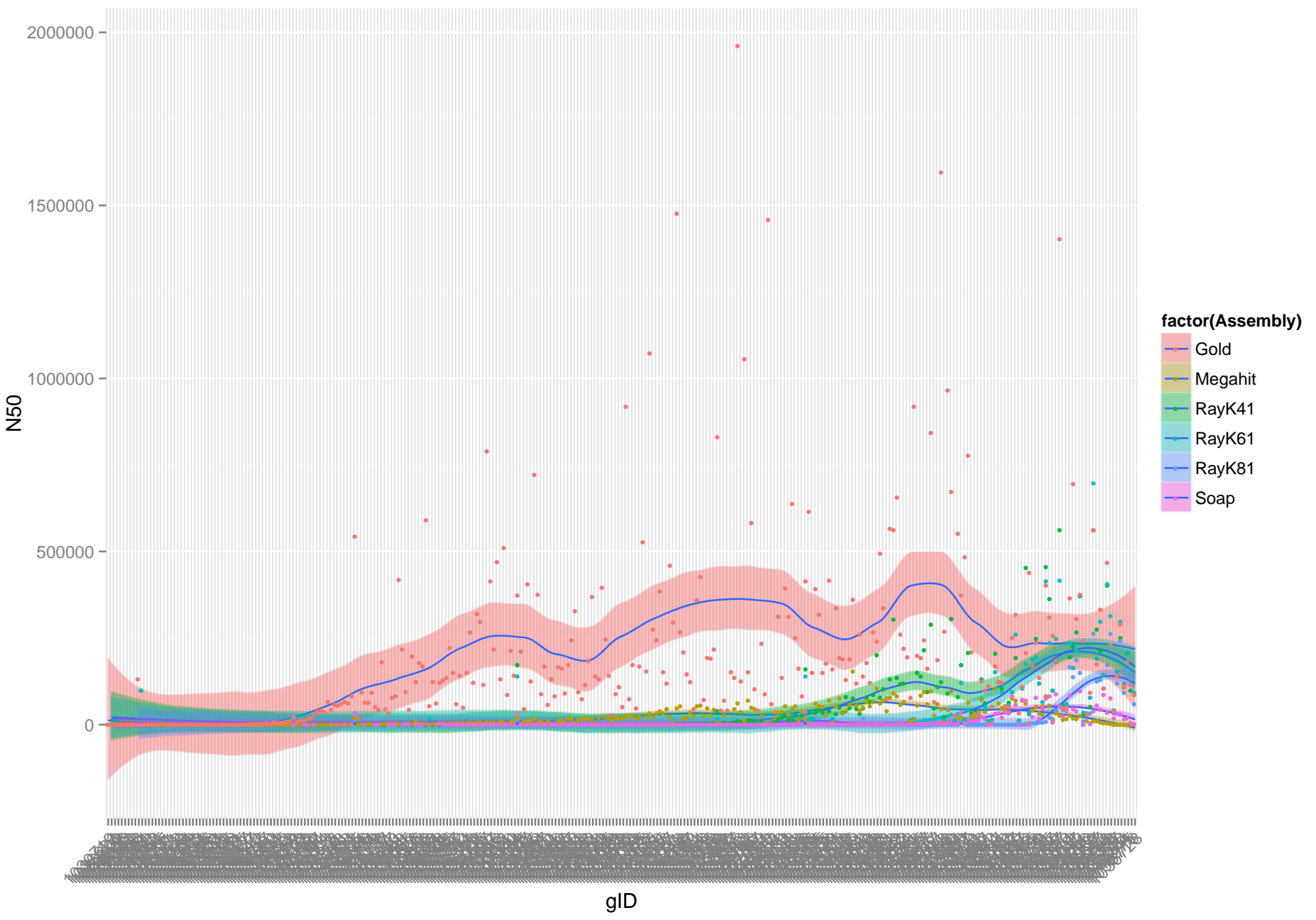
Alignment Statistics

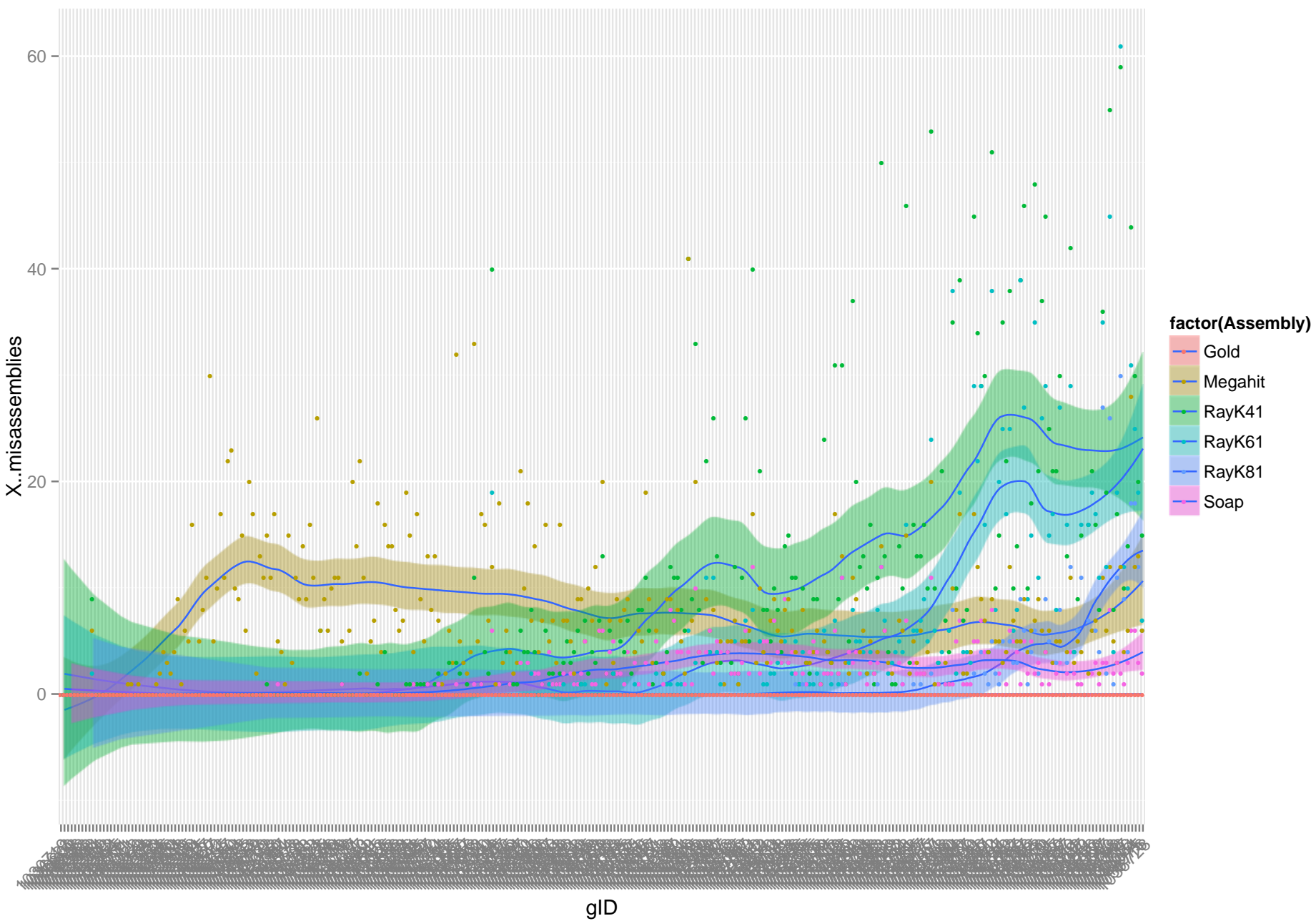
- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length
- Number of unaligned contigs (full & partial)

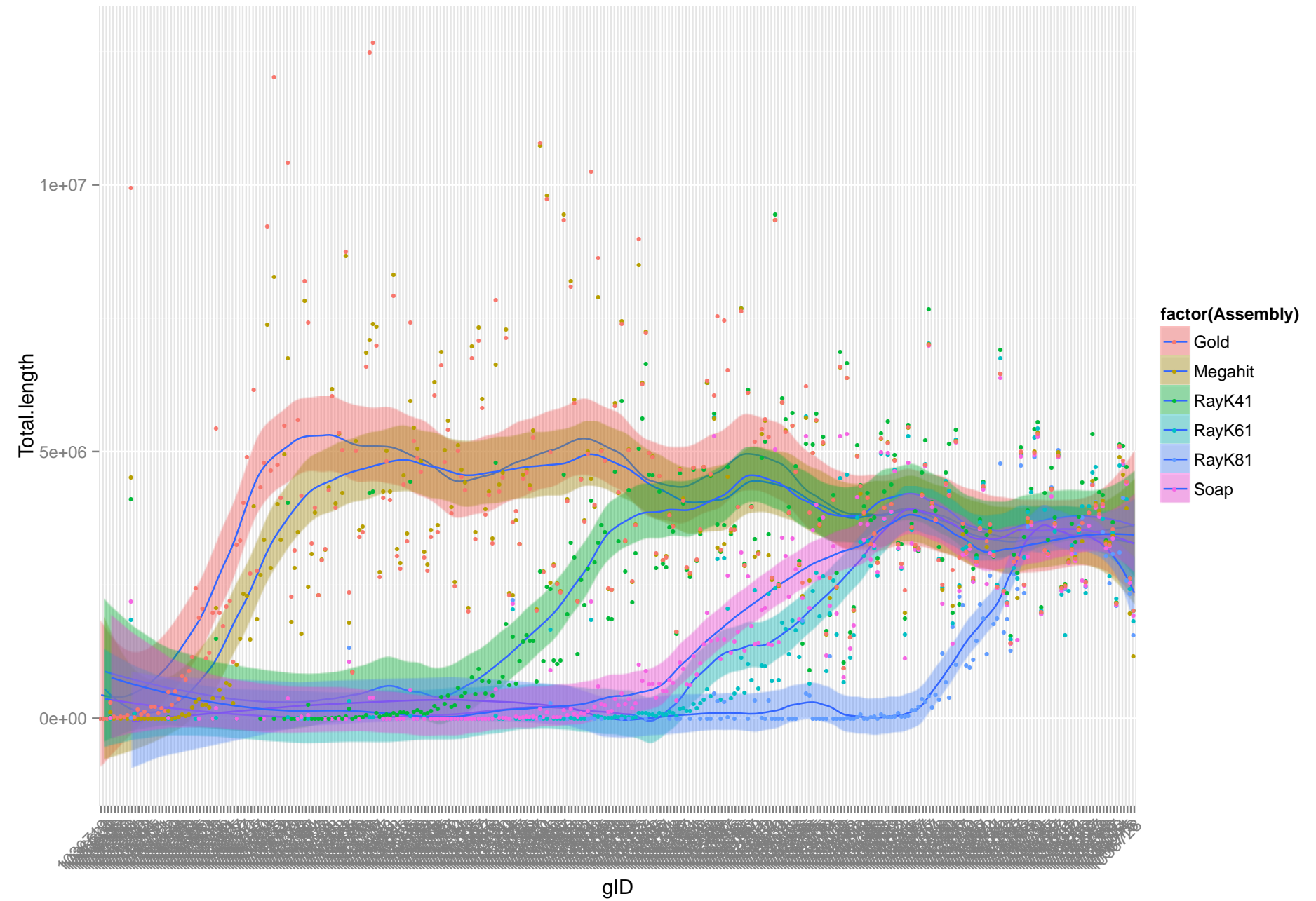










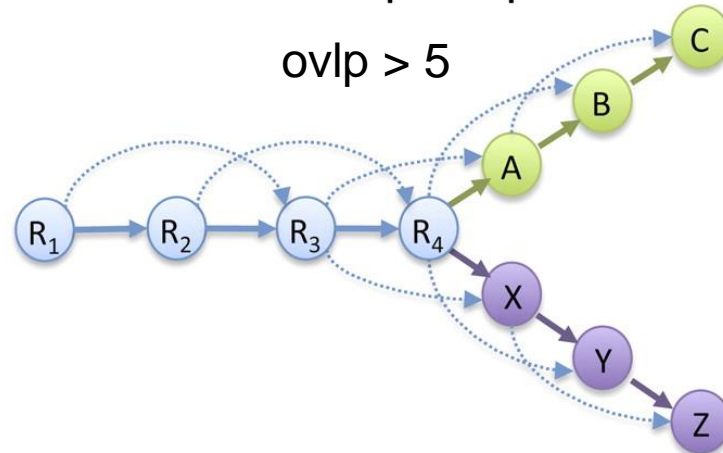


Overlap graph vs de Bruijn graph for assembly.

A Read Layout

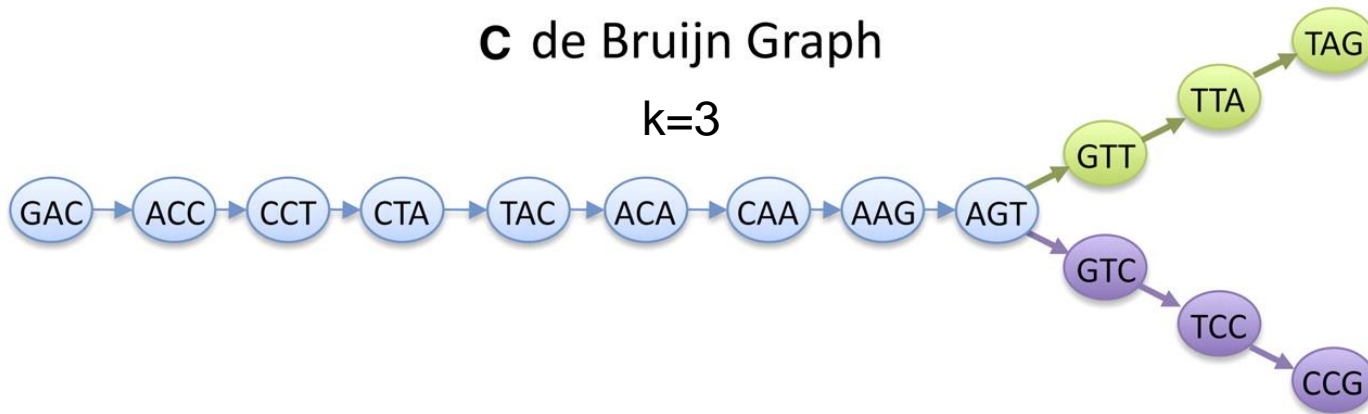
R1: GACCTACA
R2: ACCTACAA
R3: CCTACAAG
R4: CTACAAGT
A: TACAAGTT
B: ACAAGTTA
C: CAAGTTAG
X: TACAAGTC
Y: ACAAGTCC
Z: CAAGTCCG

B Overlap Graph



C de Bruijn Graph

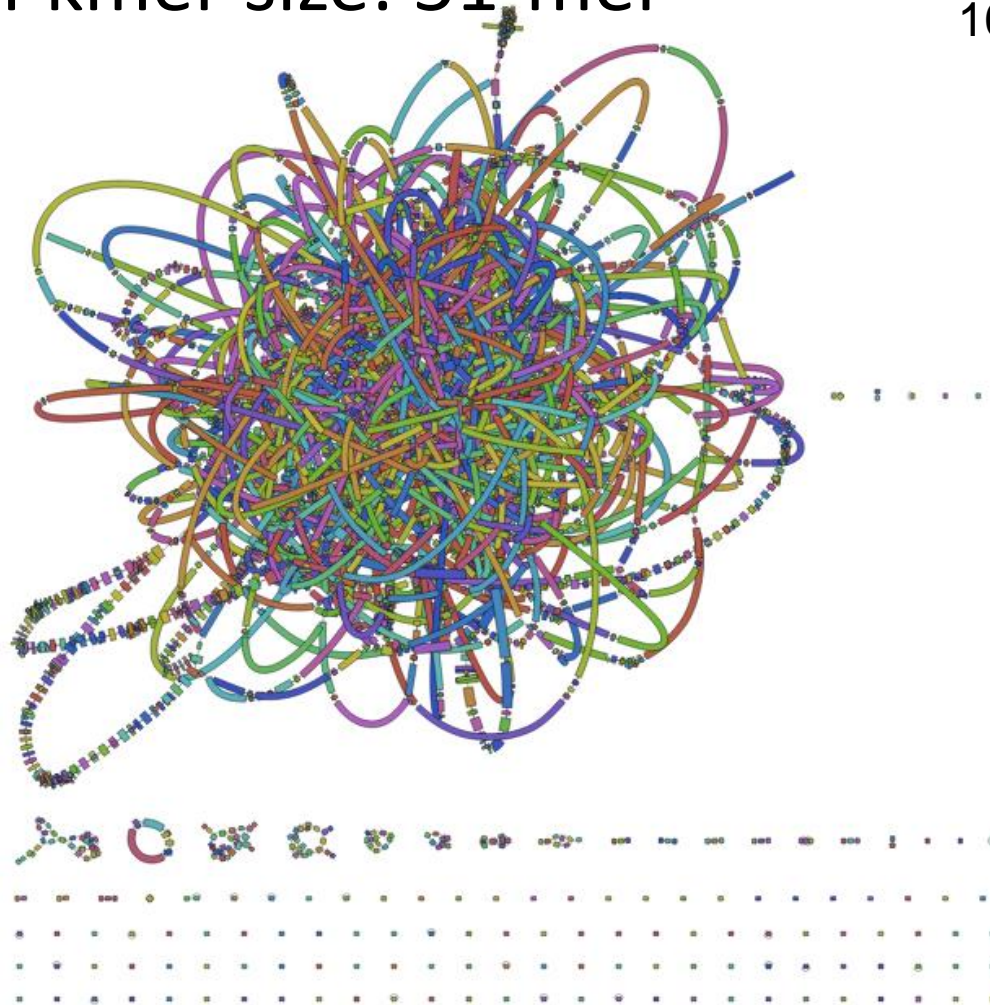
k=3



Schatz M C et al. *Genome Res.* 2010;20:1165-1173

Effect of kmer size: 51-mer

Salmonella genome
100 bp Illumina reads



Bandage

Effect of kmer size: 61-mer

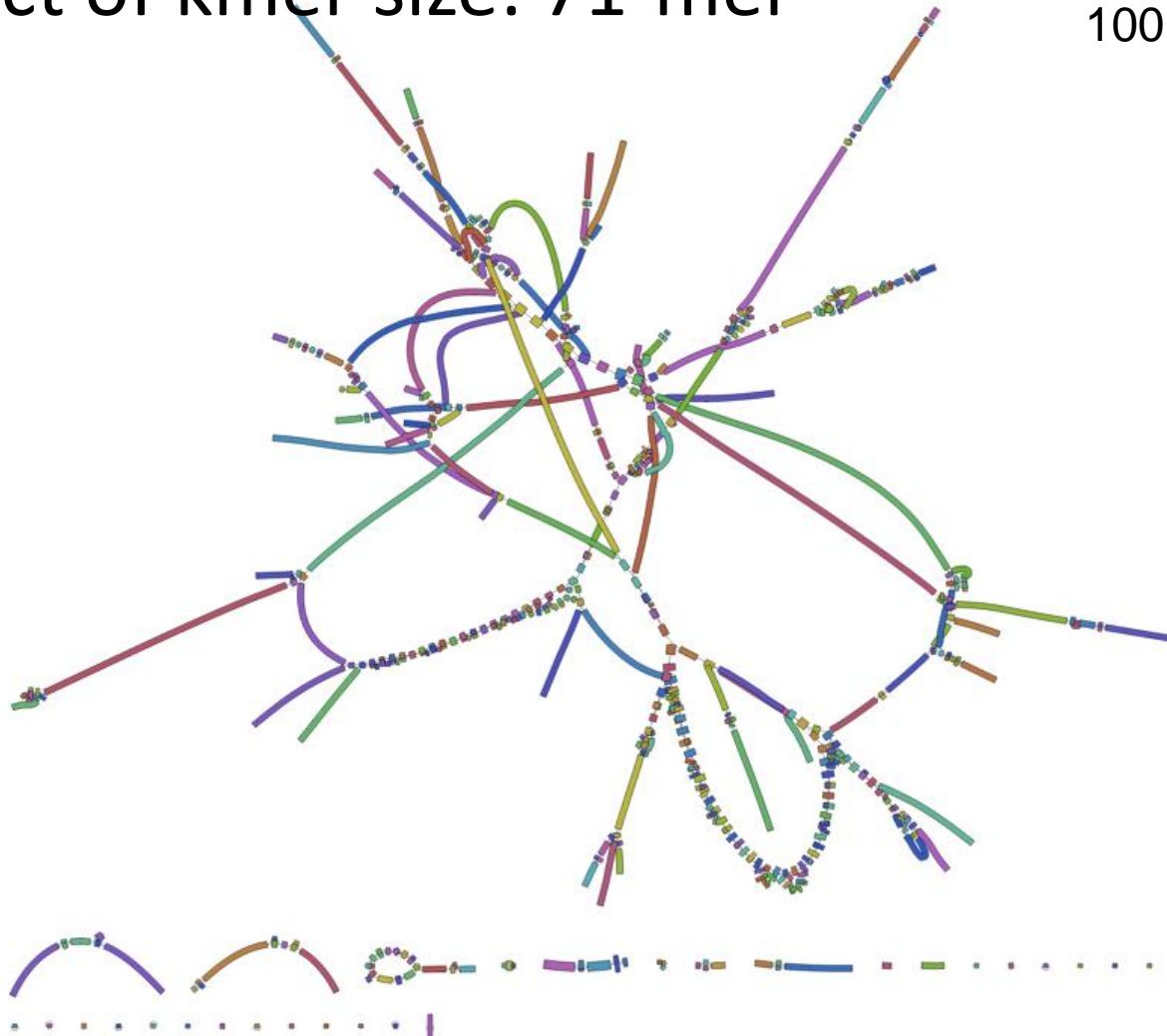
Salmonella genome
100 bp Illumina reads



Bandage

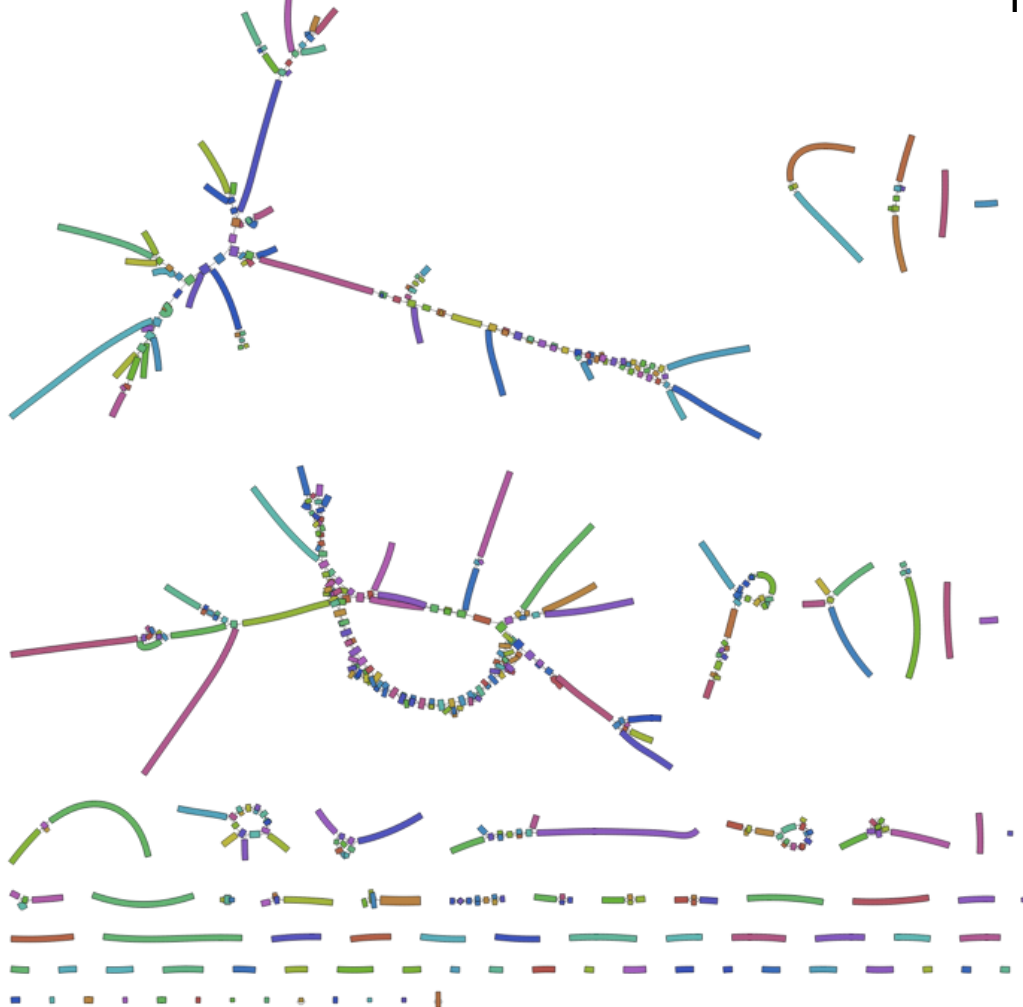
Effect of kmer size: 71-mer

Salmonella genome
100 bp Illumina reads



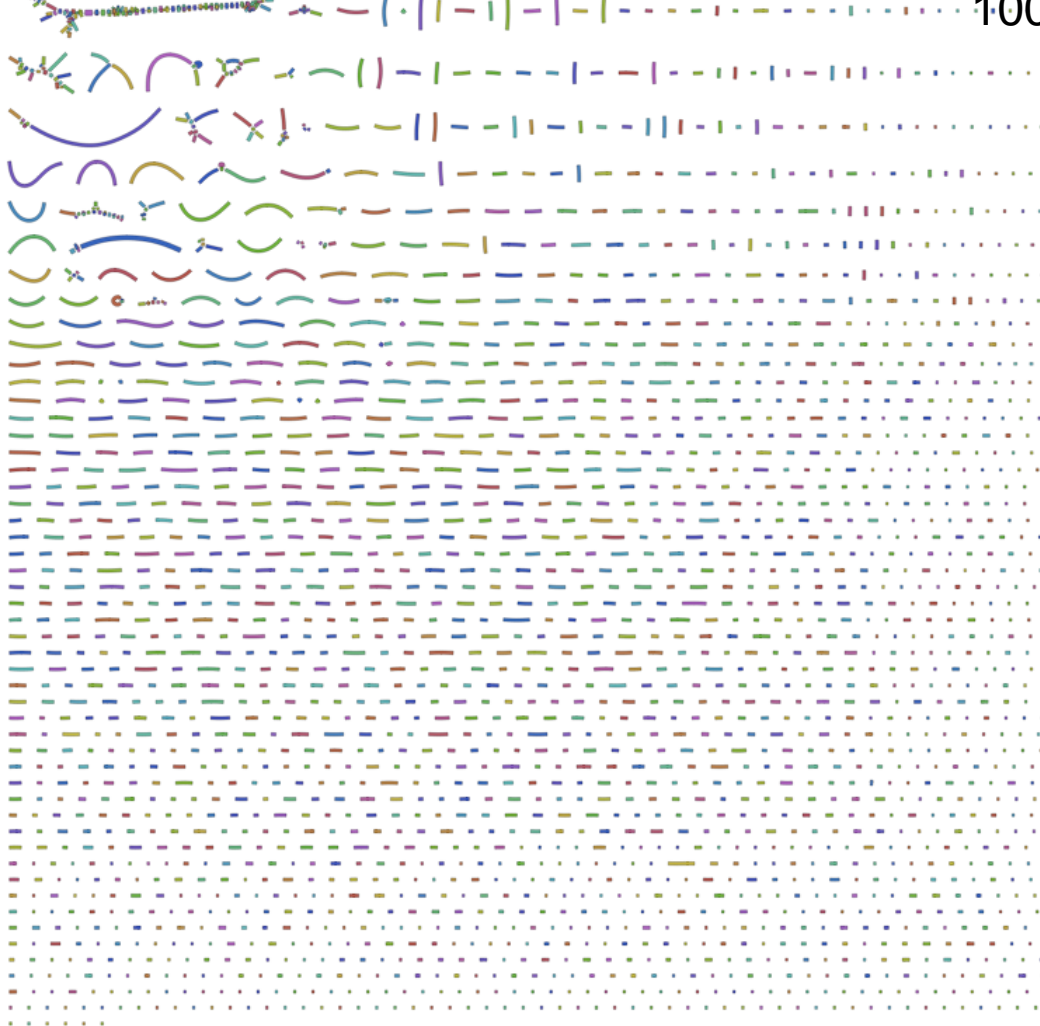
Effect of kmer size: 81-mer

Salmonella genome
100 bp Illumina reads



Effect of kmer size: 91-mer

Salmonella genome
100 bp Illumina reads



Bandage

Be part of it!
<http://www.cami-challenge.org>



Assemblathon 1 (2010)
Assemblathon 2 (2011)

The Assemblathon • Assemblathon 1

← → ↻ assemblathon.org/assemblathon1

⊕ ☆ ⚙ ABP 🔍 ☰

Follow assemblathon


tumblr.

BACKGROUND ASSEMBLATHON 1 ASSEMBLATHON 2 ASSEMBLATHON 3 MAILING LISTS CONTACT US

🕒 📡 🔍 Search

THE ASSEMBLATHON

About



An offshoot of the [Genome 10K](#) project, and primarily organized by the [UC Davis Genome Center](#), Assemblathons are contests to assess state-of-the-art methods in the field of genome assembly.

Assemblathon 1

This page will serve as an archive of all material relating to Assemblathon 1 (2010 – 2011). Existing Assemblathon 1 webpages will be converted to blog posts and linked to from this page. Also note that the UC Santa Cruz Assemblathon team, [have their own webpage](#) with code, documents and data from Assemblathon 1.

- [Data](#)