

# PATRIC Bioinformatics Resource Center

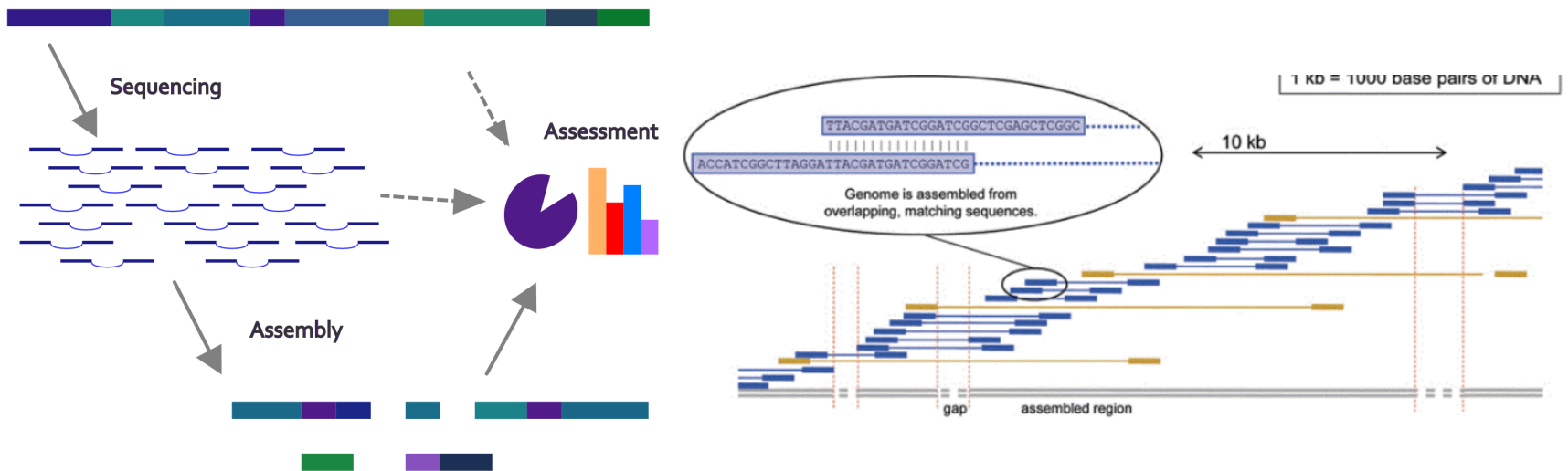
## Genome Assembly in PATRIC

Presented by Fangfang Xia



# The Sequence Assembly problem

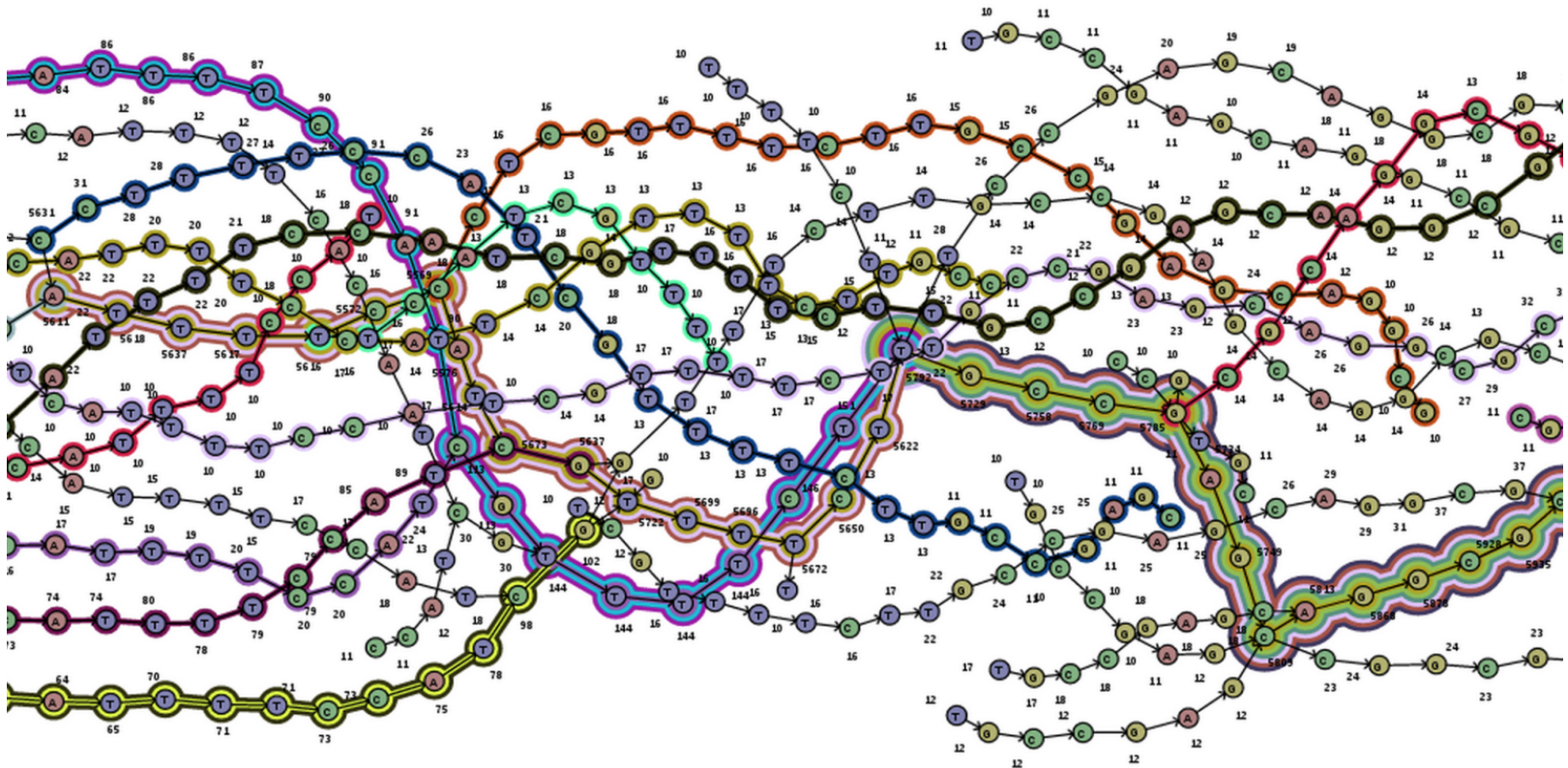
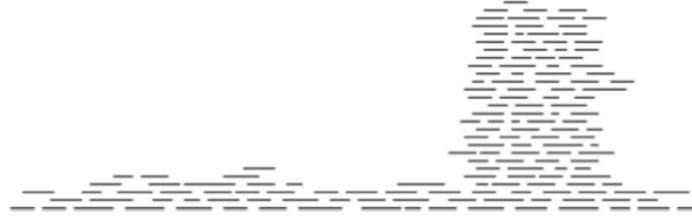
- ▶ Reconstructing contiguous DNA regions (contigs) from a set of short sequences (reads)



# An incomplete list of assemblers

Abyss, AllPaths, AllPaths-LG, AMOS, Arapan-M, Arapan-S, Celera, CLC, Clustgun, Cortex, Discover, DNA Baser, Dragon, Edena, Euler, Euler-sr, FERMI, Forge, Geneious, Graph Constructor, HGAP, IDBA, IDBA-UD, Kiki, Meta-velvet, Minia, MIRA, NextGENe, Newbler, PADENA, PASHA, Phrap, Ray, Ray-meta, REAPR, Sequencher, SeqMan, SGA, SHARCGS, SOPRA, SSAKE, SOAPdenovo, SPAdes, Staden, Taipan, TIGR, VCAKE, Phusion, QSRA, Velvet, YAGA

# Assembly graph



*"Assemblers should take in your data and automatically do the best possible job with it."*

*– A reviewer for Assemblathon*

# Sample Data: Demo FTP Server

The screenshot shows the PATRIC website interface. At the top, there is a navigation bar with the PATRIC logo (version 3.3.15) and menu items: ORGANISMS, DATA, WORKSPACES, SERVICES, and HELP. A search bar contains 'All Data Types' and 'All terms'. There are 'Register' and 'Login' buttons. The main content area features a large banner for 'asm microbe 2017 JUNE 1-5'. Below the banner, there is a 'REGISTER' button. A dropdown menu is open under the 'DATA' tab, listing various data types and specialty data collections. The 'Download Data' sub-menu is also open, with 'FTP Server' highlighted by a red circle. Below the main content, there are four green boxes representing different workshops: ASM MICROBE 2017 NEW ORLEANS, ICSB 2017 BLACKSBURG, VA, INVESTIGATING ANTIBIOTIC RESISTANCE, and SUPPORTING THE TB COMMUNITY. At the bottom, there are sections for 'BROWSE DATA' and 'TOP 10 GENERA'.

## PATRIC WORKSHOP MICROBE 2017 PATRIC WORKSHOP

PATRIC will be hosting a 1-day workshop on "Analyzing Researchers' Private Data: Own Genome using PATRIC, the All Bacteria and Archaea Reference Genomes" (number 001-WS) at the ASM Microbes 2017 conference. The workshop will cover genome assembly, sequencing and transcriptomic analysis, and variation pipeline. The workshop will focus on analyzing researchers' private data compared to the available public data. Seating is limited, so please register soon if you are interested. The main ASM Microbes 2017 registration page is here. Note that you must register for the conference in order to sign up for workshops.

REGISTER

- Data Types**
  - Antibiotic Resistance
  - Genomes
  - Genomic Features
  - Pathways
  - Protein Families
  - Specialty Genes
  - Transcriptomics
- Specialty Data Collections**
  - PATRIC Collaborations
  - PATRIC DBPs
  - NIAID Clinical Proteomics
  - NIAID Genome Sequencing
  - NIAID Structural Genomics
  - NIAID Systems Biology
  - NIAID Functional Genomics
- Download Data**
  - FTP Server**

ASM MICROBE 2017  
NEW ORLEANS

ICSB 2017  
BLACKSBURG, VA

INVESTIGATING  
ANTIBIOTIC RESISTANCE

SUPPORTING  
THE TB COMMUNITY

BROWSE DATA

TOP 10 GENERA



GENOMES



GENOMIC  
FEATURES



SPECIALTY  
GENES



PROTEIN  
FAMILIES



PATHWAYS



TRANSCRIPTOMICS

Name	Genomes
Mycobacterium	11644
Streptococcus	11380
Staphylococcus	10622

Or [ftp://ftp.patricbrc.org/patric2/workshops/workshop\\_data/](ftp://ftp.patricbrc.org/patric2/workshops/workshop_data/)

# Log on to the PATRIC website: patricbrc.org



GENOMES



GENOMIC  
FEATURES



SPECIALTY  
GENES



PROTEIN  
FAMILIES



PATHWAYS



TRANSCRIPTOMICS

Name	Genomes
Mycobacterium	11644
Streptococcus	11380
Staphylococcus	10832

# Create an Account and Log In

### PATRIC User Registration

USERNAME

FIRST NAME LAST NAME

EMAIL ADDRESS

ORGANIZATION

ORGANISMS

INTERESTS



# PATRIC Services

PATRIC 3.3.15 ORGANISMS DATA WORKSPACES SERVICES HELP All Data Types All terms Register Login

## PATRIC WORKSHOP @ MICROBE 2017

### PATRIC WORKSHOP

PATRIC will be hosting a 1-day workshop entitled "Assemble, Annotate & Analyze Your Own Genome using PATRIC, the All Bacterial Bioinformatics Resource Center" on **number 001-WS** at the ASM Microbes 2017 on June 1, 2017 in New Orleans, LA. The workshop will cover genome assembly and annotation, comparative genomics, RNA-Seq and transcriptomic analysis, and calling SNPs, MNPs, and indels using the Variation pipeline. The workshop will focus on analyzing researcher's private data compared to the available public data. Seating is limited, so please register soon if you are interested. The main ASM Microbes 2017 registration page is [here](#). Note that you must register for the conference in order to sign up for workshops.

**REGISTER**

- Genomics**
  - Assembly
  - Annotation
  - BLAST
  - Similar Genome Finder
  - Variation Analysis
  - Tn-Seq Analysis
- Protein Tools**
  - Protein Family Sorter
  - Proteome Comparison
- Metabolomics**
  - Comparative Pathway
  - Model Reconstruction
- Data**
  - ID Mapper
- Transcriptomics**
  - Expression Import
  - RNA-Seq Analysis

ASM MICROBE 2017 NEW ORLEANS ICSB 2017 BLACKSBURG, VA INVESTIGATING ANTIBIOTIC RESISTANCE SUPPORTING THE TB COMMUNITY

BROWSE DATA

TOP 10 GENERA



Name	Genomes
Mycobacterium	11644
Streptococcus	11380
Staphylococcus	10633

# Navigate to the Workspace

The screenshot shows the PATRIC 3.3.15 web interface. The top navigation bar includes links for ORGANISMS, DATA, WORKSPACES, SERVICES, and HELP. A dropdown menu is open under 'WORKSPACES', with 'Your Workspaces' circled in red. Other options in the dropdown include 'home', 'Genome Groups', 'Feature Groups', and 'Experiment Groups'. To the right of the dropdown are links for 'Your Jobs' and 'Your Genomes'. The main content area features a banner for 'PATRIC WORKSPACES MICROBE 2017 PATRIC WORKSHOP' with a 'REGISTER' button. Below the banner is a row of four green boxes representing different workshops: 'ASM MICROBE 2017 NEW ORLEANS', 'ICSB 2017 BLACKSBURG, VA', 'INVESTIGATING ANTIBIOTIC RESISTANCE', and 'SUPPORTING THE TB COMMUNITY'. At the bottom, there are sections for 'BROWSE DATA' and 'TOP 10 GENERA'. The 'BROWSE DATA' section includes icons for GENOMES, GENOMIC FEATURES, SPECIALTY GENES, PROTEIN FAMILIES, PATHWAYS, and TRANSCRIPTOMICS. The 'TOP 10 GENERA' section contains a table with the following data:

Name	Genomes
Mycobacterium	11644
Streptococcus	11380
Staphylococcus	10633

# Bring your own data

Upload

Upload file to: /fangfang@patricbrc.org/home

Upload type: Unspecified

- Unspecified
- Contigs
- Reads
- Diff. Expression Input Data
- Diff. Expression Input Metadata
- feature\_protein\_fasta

File Selected: None



Size









Cancel Upload Files

# Import data from the PATRIC public workspace

Choose or Upload a Workspace Object

Public PATRIC Workspace ▾  
My Workspaces prc.org/home  
**Public PATRIC Workspace**

Name	Owner	Created
 Escherichia coli strain MRSN388634	PATRIC	5/27/16, 1:49 AM
 Experiment Groups	PATRIC	11/25/15, 5:07 PM
 Experiments	PATRIC	11/25/15, 5:07 PM
 Feature Groups	PATRIC	11/25/15, 5:07 PM
 Genome Groups	PATRIC	11/25/15, 5:07 PM
 Reference Data	PATRIC	1/6/16, 7:42 PM
 Special Collections	PATRIC	1/6/16, 7:43 PM
 Workshop	PATRIC	5/18/17, 1:09 PM

Show files with an unspecified type

Cancel OK

# Launch an assembly job

### Paired read library ⓘ

➔

READ FILE 1  
bau\_sim\_R1.fq

READ FILE 2  
bau\_sim\_R2.fq

ADVANCED ▾

### Single read library

➔

READ FILE

### Parameters ⓘ

ASSEMBLY STRATEGY  
auto

OUTPUT FOLDER  
Assemblies

OUTPUT NAME  
bau\_sim.auto

BENCHMARK CONTIGS  
Optional

ADVANCED ◀

### Selected libraries ⓘ

Place read files here using the arrow buttons.

P(bau\_..1.fq, bau\_..2.fq) ✕

Type an informative name

# Selecting public shared datasets

## Genome Assembly

Assemble contigs from sequencing reads.

Paired read library ⓘ ➔ Selected libraries ⓘ

Choose or Upload a Workspace Object ✕

Public PATRIC Workspace ▼  
My Workspaces prc.org/home  
**Public PATRIC Workspace** + ⬆

Name	Owner	Created	+
Escherichia coli strain MRSN388634	PATRIC	5/27/16, 8:49 AM	
Experiment Groups	PATRIC	11/26/15, 12:07 AM	
Experiments	PATRIC	11/26/15, 12:07 AM	
Feature Groups	PATRIC	11/26/15, 12:07 AM	
Genome Groups	PATRIC	11/26/15, 12:07 AM	
PATRIC Workshop	PATRIC	7/12/17, 8:43 AM	
Reference Data	PATRIC	1/7/16, 2:42 AM	
Special Collections	PATRIC	1/7/16, 2:43 AM	

Show files with an unspecified type

Cancel OK

# Submitting SRA datasets for assembly

Services

## Genome Assembly

Assemble contigs from sequencing reads.


The screenshot shows a web interface for submitting SRA datasets for genome assembly. It is divided into several sections:

- Paired read library**: Includes a 'READ FILE 1' input field.
- Single read library**: Includes a 'READ FILE' input field and an 'ADVANCED' dropdown menu.
- SRA run accession**: A field containing the accession number 'SRR5121079', which is circled in red. A tooltip is visible over this field with the text: 'SRA run accession Title: Illumina MiSeq paired end sequencing; ZC192: PCR-amplicon NGS of Zika virus from human serum'.
- Parameters**: Includes 'ASSEMBLY STRATEGY' (set to 'auto'), 'OUTPUT FOLDER', 'OUTPUT NAME' (set to 'Output Name'), and 'BENCHMARK CONTIGS' (set to 'Optional').
- Selected libraries**: A list of libraries to be assembled, containing 'SRR5121079', which is also circled in red. A tooltip is visible over this entry with the text: 'SRA run accession Title: Illumina MiSeq paired end sequencing; ZC192: PCR-amplicon NGS of Zika virus from human serum'.

# Data upload and job status

bau\_sim\_R2.fq 12%

bau\_sim\_R1.fq 16%


Uploads
0·2·14%

Completed · In progress · Queued · Suspended

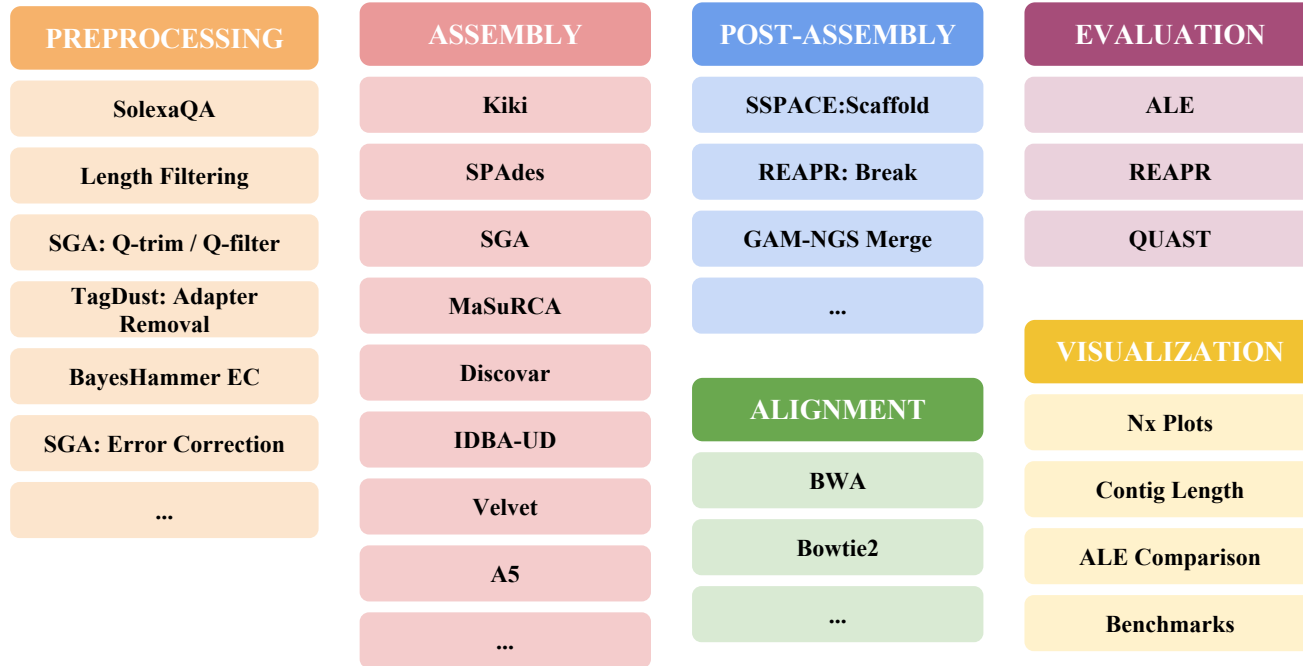
Jobs
6 · 1 · 0 · 4

Status	Submit	App	Output Name	Start	Completed
● completed	6/14/15, 6:58 PM	GenomeAssembly	bau.auto	6/14/15, 6:58 PM	6/14/15, 7:05 PM
● completed	6/14/15, 6:35 PM	GenomeAssembly	bau_demo	6/14/15, 6:35 PM	6/14/15, 6:42 PM
● completed	6/14/15, 10:28 AM	DifferentialExpression	baumannii_test	6/14/15, 10:28 AM	6/14/15, 10:30 AM
● completed	6/14/15, 8:15 AM	GenomeAssembly	rhodo	6/14/15, 8:15 AM	6/14/15, 8:32 AM
● completed	6/14/15, 8:07 AM	GenomeAssembly	test1	6/14/15, 8:07 AM	6/14/15, 8:08 AM
● completed	6/13/15, 1:58 PM	DifferentialExpression	test_exp	6/13/15, 1:59 PM	6/13/15, 1:59 PM
● completed	6/11/15, 5:25 PM	RNASeq	testg37p	6/11/15, 5:25 PM	6/11/15, 5:25 PM



# ASSEMBLY SERVICE

## COMPUTE CAPABILITIES



# Curated assembly strategies

**Parameters** ⓘ

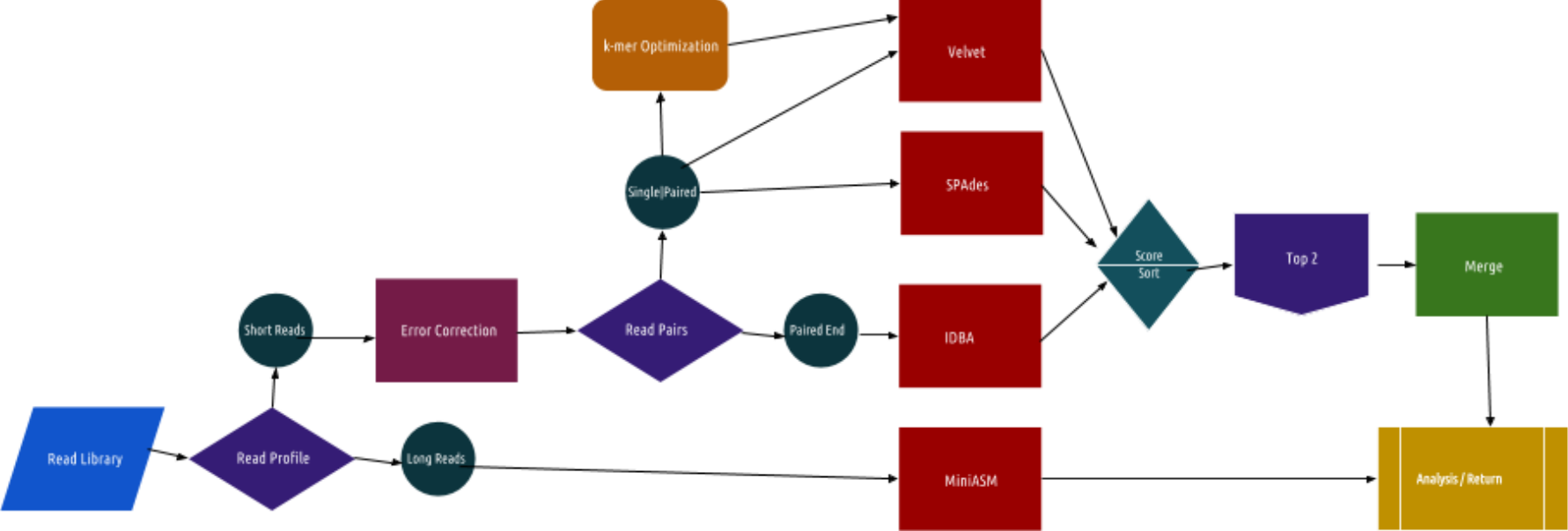
ASSEMBLY STRATEGY

auto ▼

- auto**
- fast
- full spades
- kiki
- miseq
- plasmid
- smart

ADVANCED ▼

# The “smart” assembly recipe



# Which assembly recipe to use

- ▶ ***auto*** — the evolving default strategy recommended for most data
- ▶ ***full spades*** – runs the full SPAdes pipeline, one of the best assemblers for microbial genomes
- ▶ ***fast*** — ~2X faster than *auto*; suited for large genomes or simple microbial communities (velvet + megahit)
- ▶ ***kiki*** — very fast but does not use paired end information; good for metagenome assembly
- ▶ ***miseq*** — good for Illumina MiSeq reads that are 250–350 bp long (Spades with more k-mer iterations)
- ▶ ***smart*** — the slowest and sometimes the most accurate
- ▶ ***plasmid*** — plasmid assembly (plasmidSPAdes)

# Typical execution times

for a typical microbial genome

Recipe	Hours
smart	3 ~ 100
auto / miseq	2 ~ 80
fast	1 ~ 12
kiki	1 ~ 6

Depends on read depth, sequencing errors, genome size, repeat structure, etc.

# You can download output from the workspace



ORGANISMS

DATA

SERVICES

TOOLS

ABOUT



PATRIC\_Workshop / home / Assemblies / A baumannii 1000160



UPLOAD



ADD FOLDER







Name	Size	Owner	Created	
↑ Parent Folder				
📁 SRR1033693.auto	4.4 kB	PATRIC_W	6/11/15, 8:31 PM	
📁 SRR1033693.smart	4.4 kB	PATRIC_W	6/12/15, 9:49 AM	

**PATRIC\_Workshop / home / A baumannii genome sequencing / Acinetobacter baumannii 1207552 / A\_baumannii\_1207552.auto**

Genome Assembly Job Result

<b>Start time</b>	6/16/15, 4:47 PM
<b>End time</b>	6/16/15, 6:16 PM
<b>Run time</b>	1h29m3630s
<b>Parameters</b>	<pre>{   "output_file": "A_baumannii_1207552.auto",   "output_path": "/PATRIC_Workshop@patricbrc.org/home/A_baumannii_genome_sequencing/Acinetobacter_baumannii_1207552",   "paired_end_libs": [     {       "read2": "/PATRIC_Workshop@patricbrc.org/home/A_baumannii_genome_sequencing/Acinetobacter_baumannii_1207552/SRR1030405_2.fastq.gz",       "read1": "/PATRIC_Workshop@patricbrc.org/home/A_baumannii_genome_sequencing/Acinetobacter_baumannii_1207552/SRR1030405_1.fastq.gz",       "interleaved": "false"     }   ],   "recipe": "auto",   "reference_assembly": "" }</pre>

Result Files

Filename	Type	File size
 1_analysis.zip	zip	242.7 kB
 report.txt	txt	151.6 kB
 11_2.idba_contigs.fa	contigs	4.0 MB
 11_3.velvet_contigs.fa	contigs	4.1 MB
 11_1.spades_contigs.fasta	contigs	4.1 MB
 contigs.fa	contigs	4.1 MB

11\_analysis

Name	Date Modified	Size	Kind
▶ predicted_genes	Jun 16, 2015, 6:16 PM	--	Folder
quast.log	Jun 16, 2015, 6:16 PM	18 KB	Log File
quast.out	Jun 16, 2015, 6:16 PM	4 KB	TextW...ument
quast.out.quast	Jun 16, 2015, 6:16 PM	Zero bytes	Document
▶ report_html_aux	Jun 16, 2015, 6:16 PM	--	Folder
report.html	Jun 16, 2015, 6:16 PM	27 KB	HTML
report.tex	Jun 16, 2015, 6:16 PM	1 KB	TextW...ument
report.tsv	Jun 16, 2015, 6:16 PM	653 bytes	Plain Text
report.txt	Jun 16, 2015, 6:16 PM	2 KB	Plain Text
transposed_report.tex	Jun 16, 2015, 6:16 PM	1 KB	TextW...ument
transposed_report.tsv	Jun 16, 2015, 6:16 PM	653 bytes	Plain Text
transposed_report.txt	Jun 16, 2015, 6:16 PM	2 KB	Plain Text



# Evaluate assembled contigs

## QUAST report

19 June 2014, Thursday, 08:19:51

All statistics are based on contigs of size  $\geq 500$  bp, unless otherwise noted (e.g., "# contigs ( $\geq 0$  bp)" and "Total length ( $\geq 0$  bp)" include all contigs.)

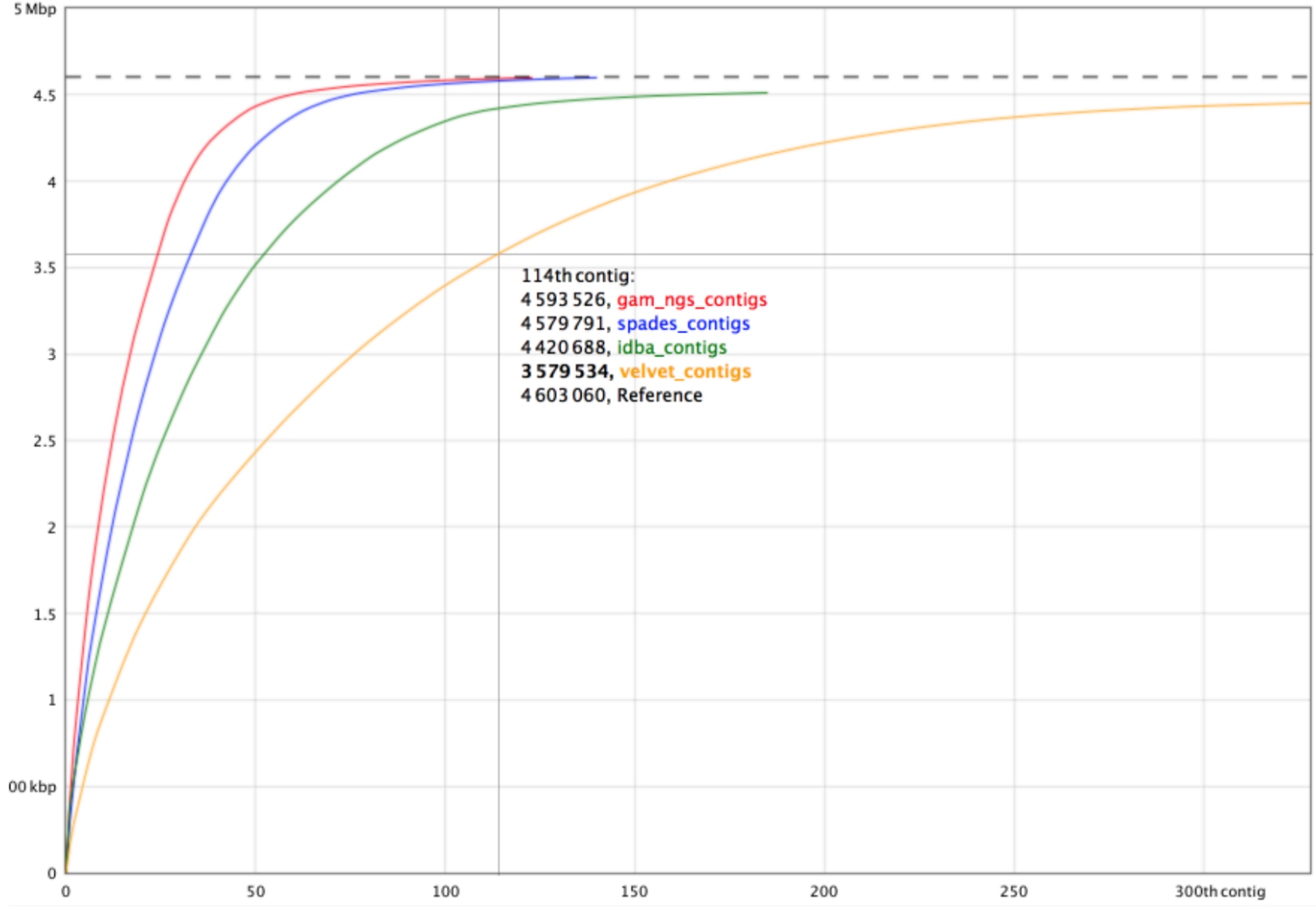
[Extended report](#)

worst.....best

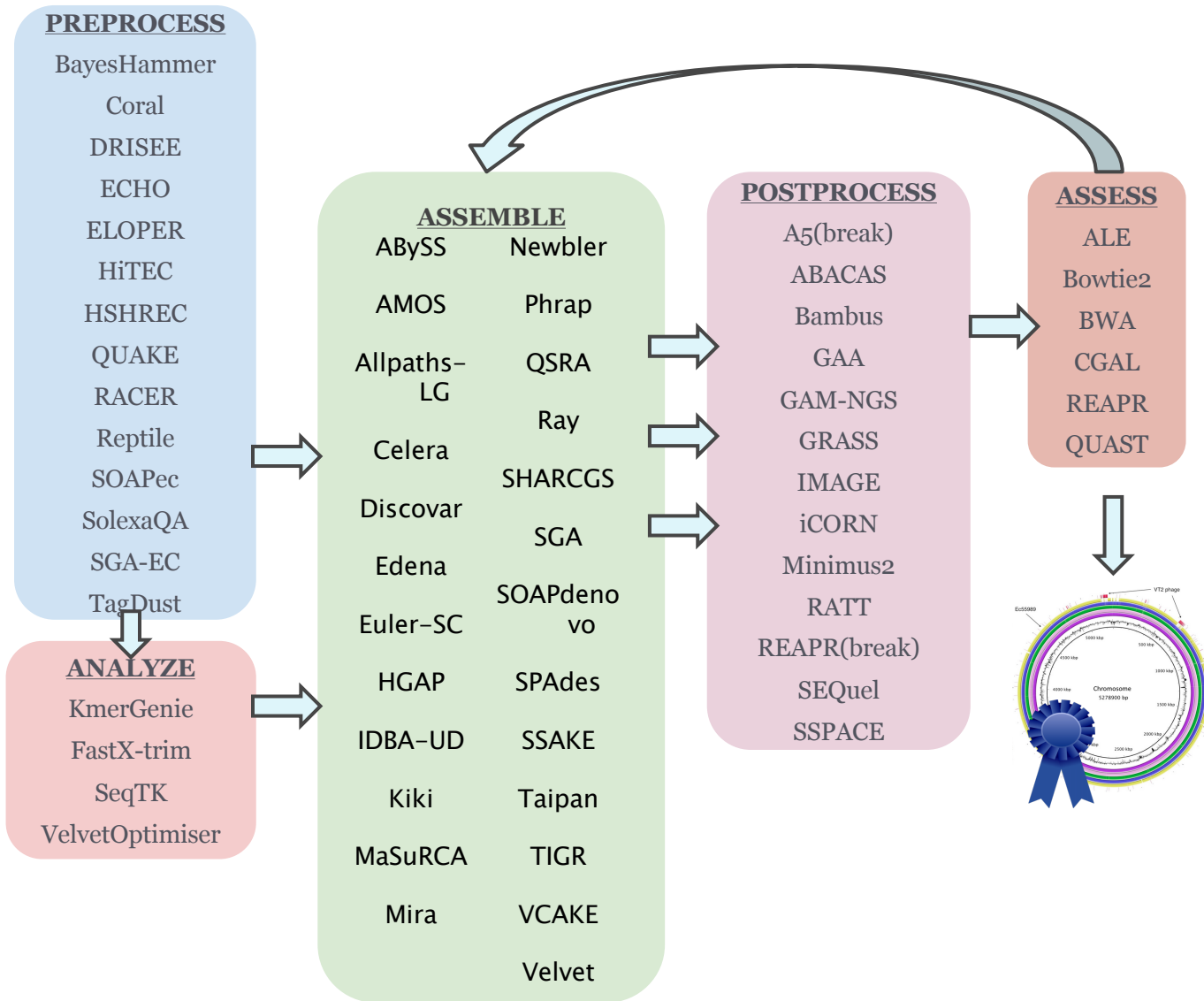
Genome: 4 603 060 bp, G+C content: 68.79%

Statistics without reference	gam_nginx_contigs	spades_contigs	idba_contigs	velvet_contigs
# contigs	123	140	185	328
Largest contig	399 330	292 708	378 933	160 339
Total length	4 598 516	4 598 552	4 511 207	4 451 511
N50	130 248	95 827	61 227	26 869
<b>Misassemblies</b>				
# misassemblies	6	6	1	29
Misassembled contigs length	153 182	150 167	24 718	876 583
<b>Mismatches</b>				
# mismatches per 100 kbp	16.080	16.52	6.08	12.69
# indels per 100 kbp	4.020	3.91	3.5	10.210
# N's per 100 kbp	0	0	0	373.58
<b>Genome statistics</b>				
Genome fraction (%)	98.898	98.895	97.953	96.39
Duplication ratio	1.01	1.01	1.001	1.003
NGA50	130 247	95 827	61 162	24 307
<b>Predicted genes</b>				
# predicted genes (unique)	4480	4501	4464	4695
# predicted genes ( $\geq 0$ bp)	4519	4540	4464	4695
# predicted genes ( $\geq 300$ bp)	3970	3984	3919	3989
# predicted genes ( $\geq 1500$ bp)	550	545	537	480
# predicted genes ( $\geq 3000$ bp)	44	44	38	34

Plots: Cumulative length Nx NAx NGx NGAx GC content



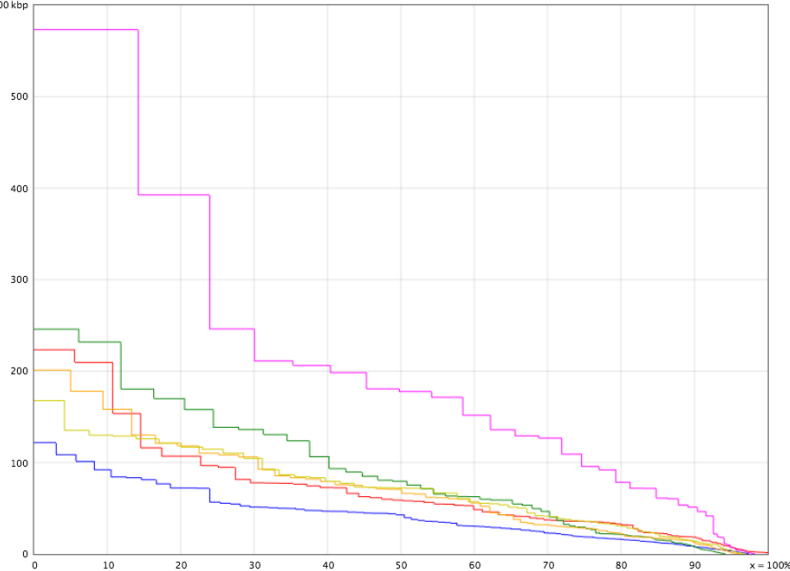
# Advanced pipelines



# AssemblyRAST: Assembler Comparison

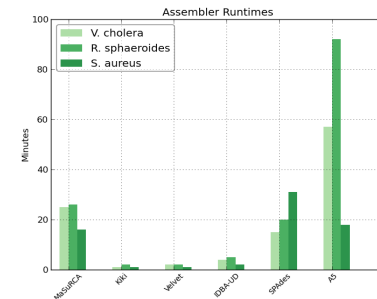
```
for LIB in $(ls)
do
  ar_run -f $LIB/rd*.fq -a masurca kiki velvet spades idba a5
done
```

NGAx = 600 kbp  
Plots: Cumulative length Nx NAX NGx NGAx GC content



Assembly	MaSuRCA	Kiki	Velvet	IDBA-UD	SPAdes
# contigs (>= 0 bp)	135	1228	733	301	316
# contigs (= 1000 bp)	125	1228	520	140	133
Total length (= 0 bp)	4454671	4211242	4604732	4550407	4661794
Total length (= 1000 bp)	4447475	4211242	4523850	4490529	4604439
# contigs	133	1228	587	173	161
Largest contig	159677	33373	56029	378946	230007
Total length	4453816	4211242	4572094	4512010	4625338
Reference length	4603060	4603060	4603060	4603060	4603060
GC (%)	68.86	68.55	68.68	68.81	68.82
Reference GC (%)	68.79	68.79	68.79	68.79	68.79
N50	74831	4406	14135	73097	71177
NG50	66418	3936	13982	72396	73118
N75	34138	2548	7489	42086	47310
NG75	31030	2088	7240	41189	47310
# misassemblies	7	32	14	3	9
# local misassemblies	12	5	1322	4	7
Unaligned contigs length	0	2760	117	0	55
Genome fraction (%)	95.315	90.776	95.681	97.712	99.048
Duplication ratio	1.016	1.006	1.036	1.005	1.017
# N's per 100 kbp	0.00	0.00	3249.32	0.00	8.76
# mismatches per 100 kbp	31.34	32.95	8.54	4.16	12.11
# indels per 100 kbp	5.54	6.99	23.91	3.58	5.75
Largest alignment	159677	33373	53699	378908	230007
NA50	74744	4285	12944	73097	67626
NGA50	66418	3893	12905	72357	71175
NA75	34138	2483	6437	41275	42056
NGA75	31030	2048	6263	37563	42056

Organism	MaSuRCA	Kiki	Velvet	IDBA	SPAdes	A5
<i>B. cereus</i> HiSeq*	52644	59995	42763	31347	<b>78420</b>	45935
<i>S. aureus</i>	22603	1854	11540	34957	<b>50888</b>	8188
<i>V. cholera</i> HiSeq	59028	42804	47191	70796	<b>177768</b>	72282
<i>V. cholera</i> MiSeq	50207	70738	19767	44178	198488	57376
<i>R. sphaeroides</i> HiSeq	66418	3893	33342	<b>72357</b>	71175	20356*
<i>R. sphaeroides</i> MiSeq	-	33589	62923	60228	<b>126502</b>	83693



# Acknowledgements

## ▶ PATRIC Team:

- University of Chicago
  - Ryan Aydelott
  - Tom Brettin
  - Neil Conrad
  - Jim Davis
  - Emily Dietrich
  - Chris Henry
  - Dan Murphy–Olson
  - Bob Olson
  - Bruce Parrello
  - Maulik Shukla
  - Rick Stevens
  - Fangfang Xia

## ◦ FIG

- Terry Disz
- Ross Overbeek
- Gordon Pusch
- Veronika Vonstein

## ◦ VBI

- Joseph Gabbard
- Ron Kenyon
- Dustin Machi
- Chunhong Mao
- Bruno Sobral
- Rebecca Wattam
- Andrew Warren
- Rebecca Will
- Harry Yoo

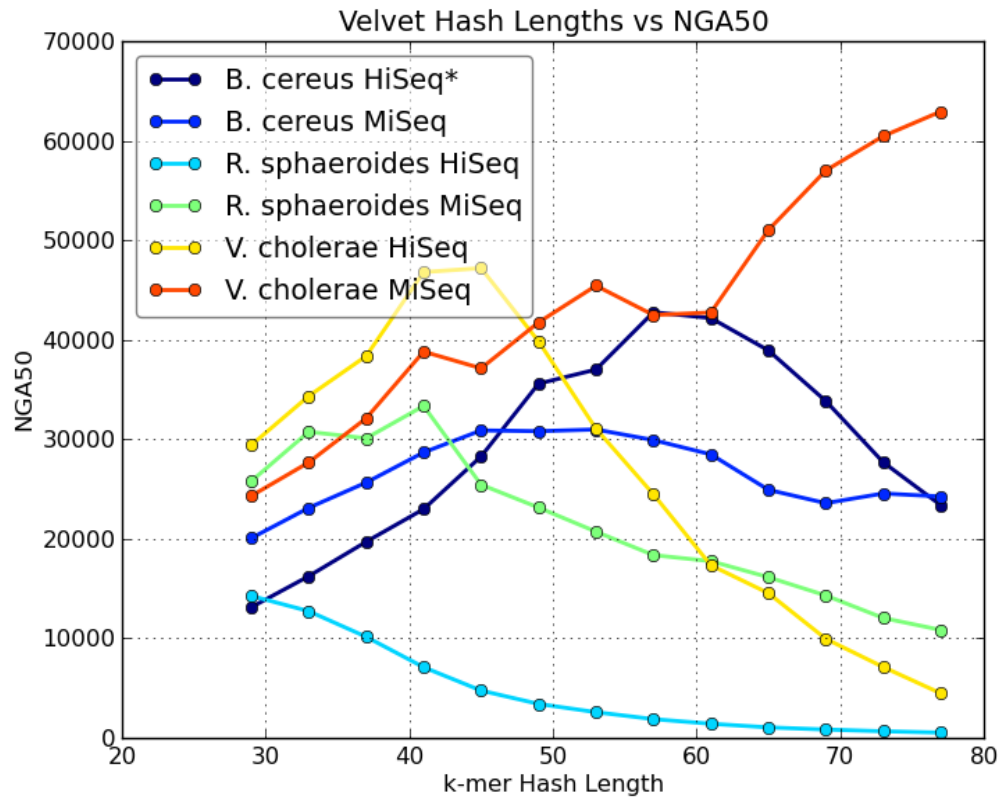
## ◦ Stevens Group

- Chris Bun
- Sebastien Boisvert

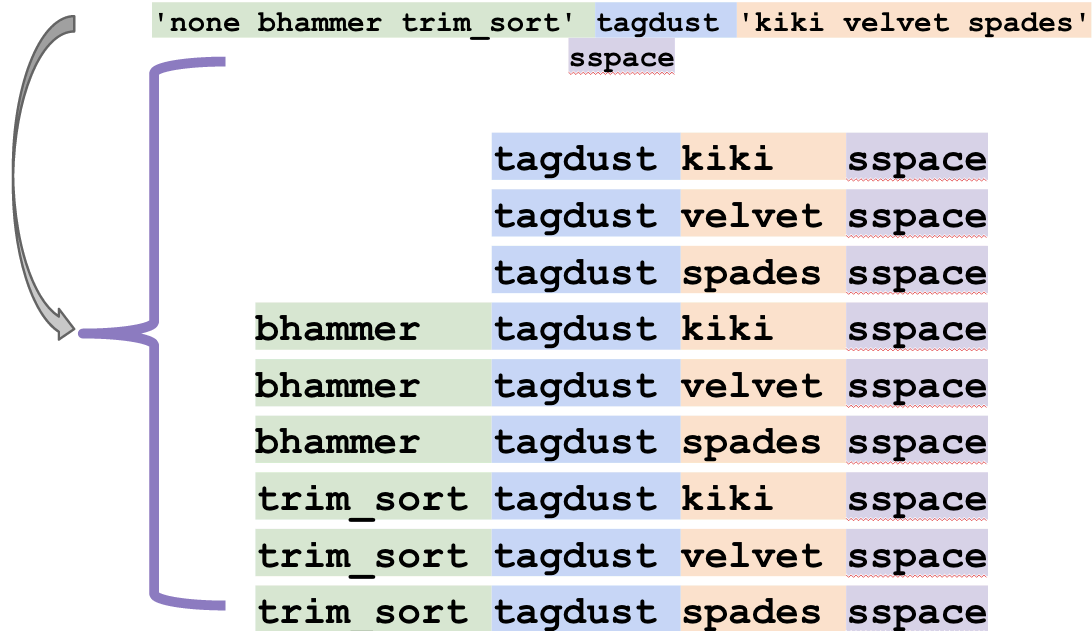
National Institute of Allergy and Infectious Diseases  
Contract No. HHSN272201400027C

# Parameter scan: k-mer Optimization

`velvet ?hash_length=29-77:4`



# Pipeline design



# Upcoming improvements

- Improved error detection and classification using supervised learning
- Workflows for new sequencing technology (MinION), Hybrid assembly