PATRIC Bioinformatics Resource Center

Genome Assembly in PATRIC

Presented by Neal Conrad

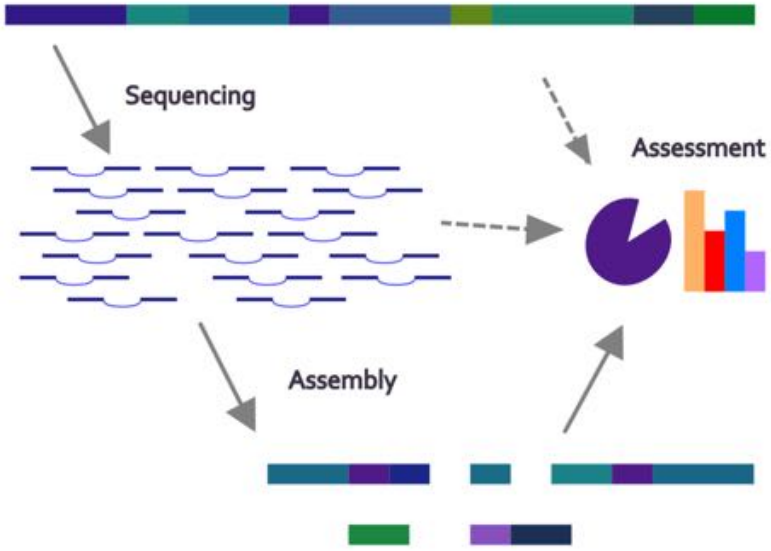
Slides originally by Fangfang Xia

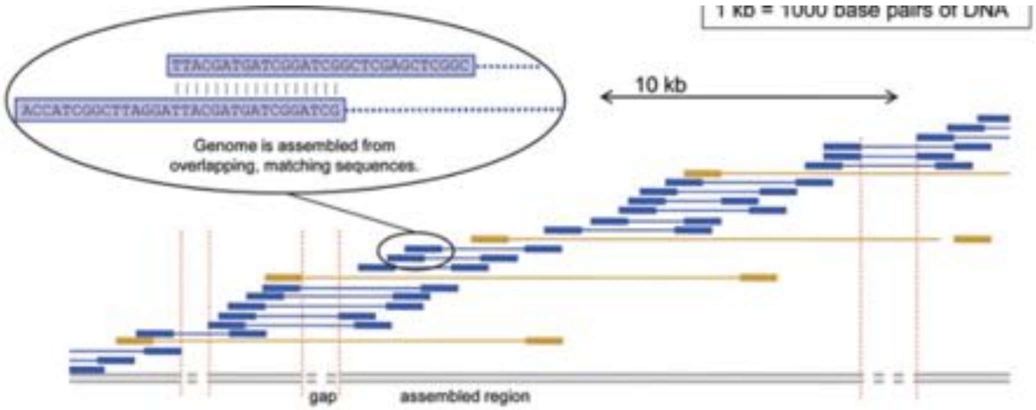
PATRIC

PATHOSYSTEMS RESOURCE INTEGRATION CENTER

The Sequence Assembly problem

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

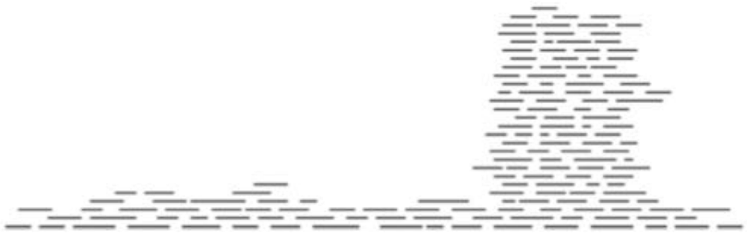


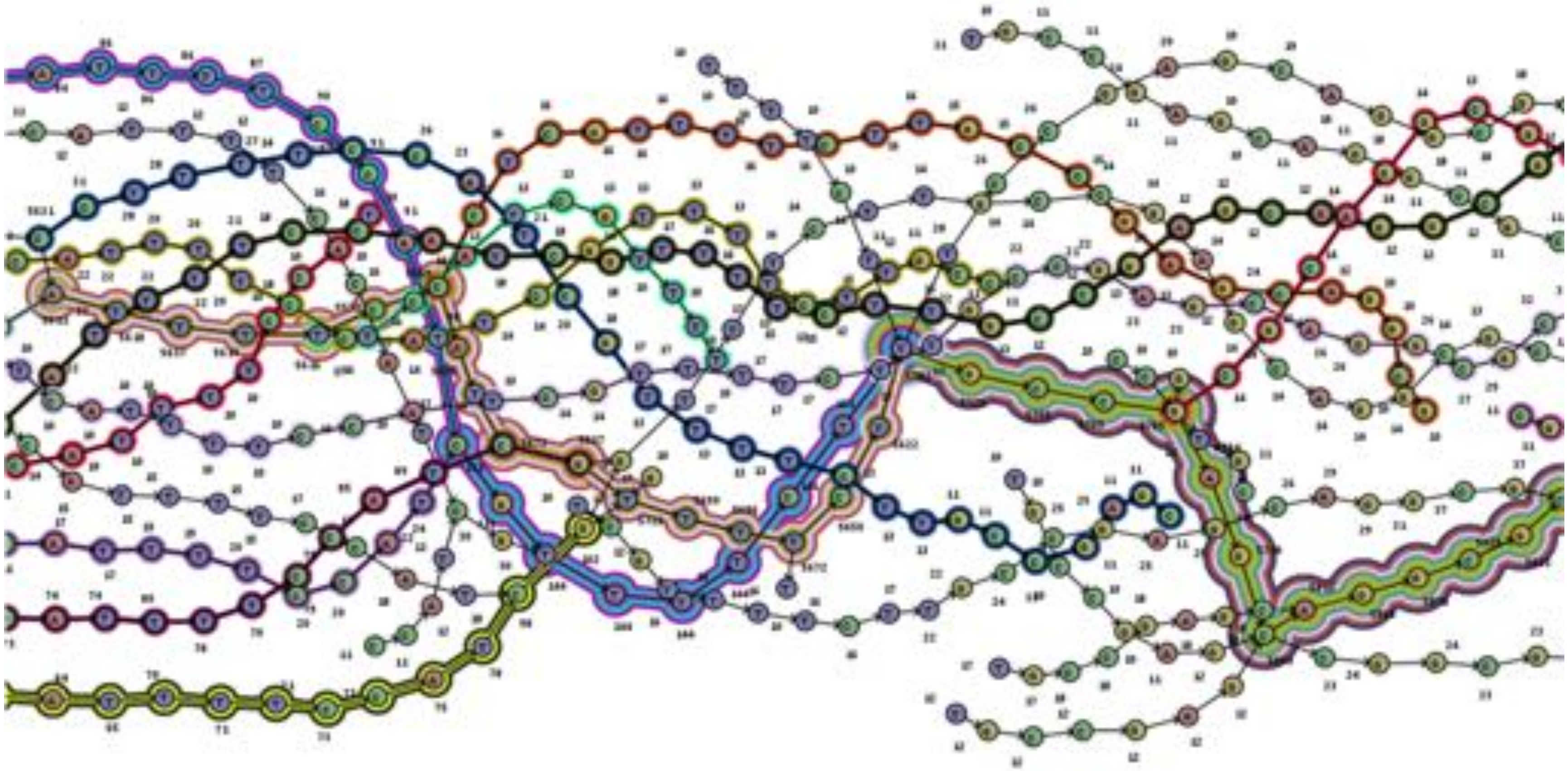

1 kb = 1000 base pairs of DNA
TTACGATGATCGGATCGGCTCGAGCTCGGC
ACCATCGGCTTAGGATTACGATGATCGGATCG
Genome is assembled from overlapping, matching sequences.
10 kb
gap
assembled region

An incomplete list of assemblers

Abyss, AllPaths, AllPaths-LG, AMOS, Arapan-M, Arapan-S, Celera, CLC, Clustgun, Cortex, Discovar, 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



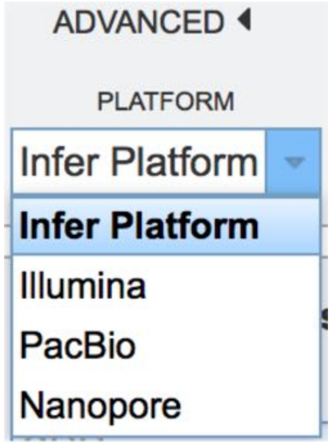


"Assemblers should take in your data and automatically do the best possible job with it."

– A reviewer for Assemblathon

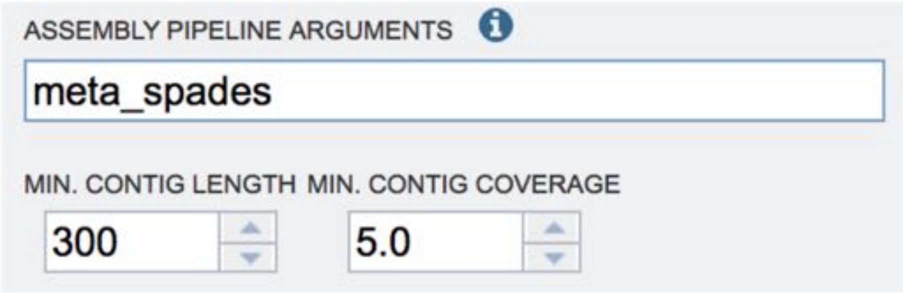
Common Assembly Scenarios

- *Short read assembly* — Illumina sequencers
- *Long read assembly* — PacBio, Oxford Nanopore
- *Hybrid assembly* — Short reads + long reads
  - Short for assembly + long for scaffolding (spades)
  - Long for assembly + short for finetuning (auto; planned)



PLATFORM: Infer Platform

- *Whole genome assembly*
- *Plasmid assembly*
- *Metagenome assembly*

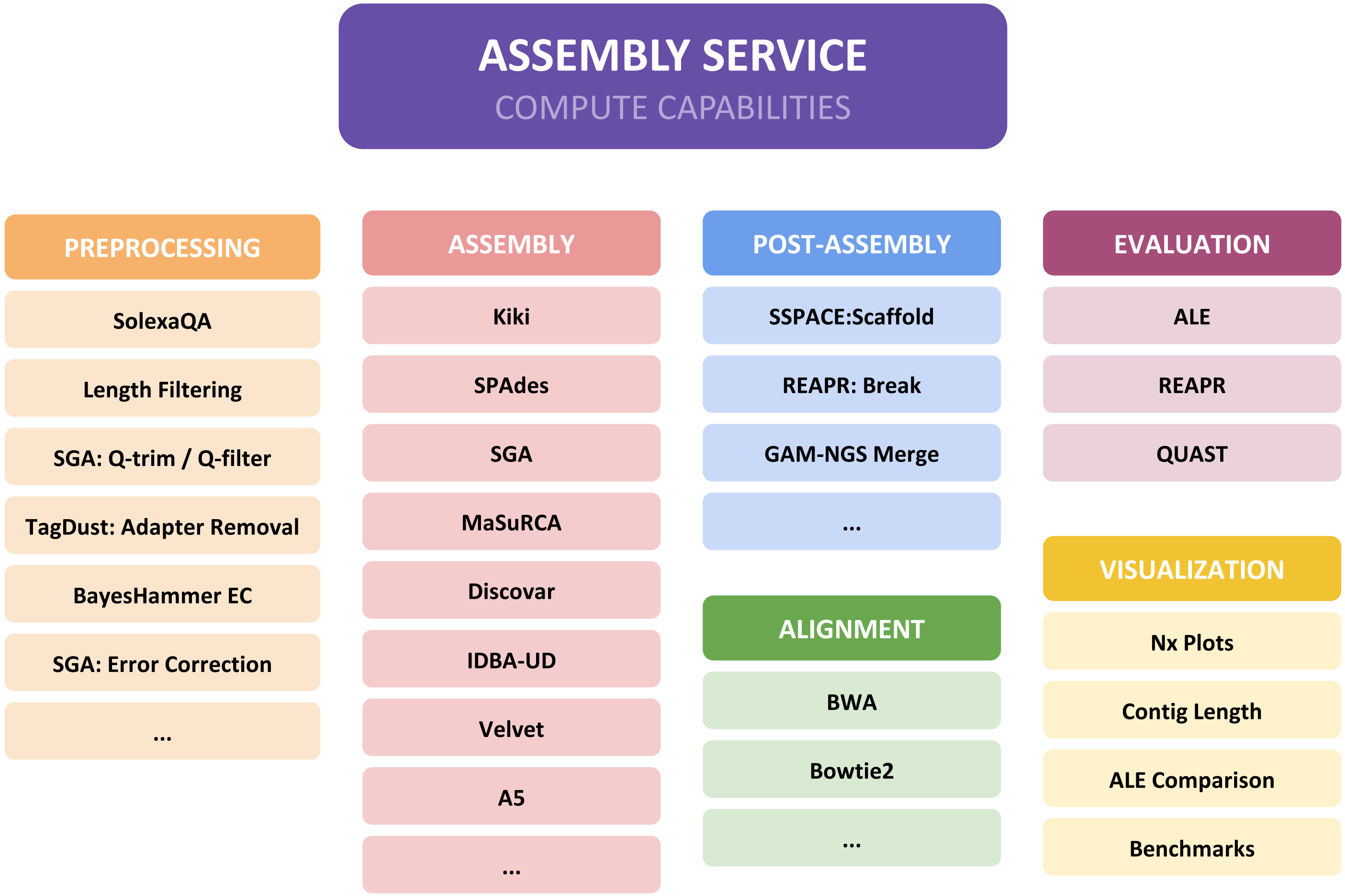


ASSEMBLY PIPELINE ARGUMENTS: meta_spades

MIN. CONTIG LENGTH: 300

MIN. CONTIG COVERAGE: 5.0

- *De novo assemby vs reference-guided assembly*
  - lib.consensus.fa output in variation analysis



ASSEMBLY SERVICE

COMPUTE CAPABILITIES

PREPROCESSING

- SolexaQA
- Length Filtering
- SGA: Q-trim / Q-filter
- TagDust: Adapter Removal
- BayesHammer EC
- SGA: Error Correction
- …

ASSEMBLY

- Kiki
- SPAdes
- SGA
- MaSuRCA
- Discovar
- IDBA-UD
- Velvet
- A5
- …

POST-ASSEMBLY

- SSPACE:Scaffold
- REAPR: Break
- GAM-NGS Merge
- …

ALIGNMENT

- BWA
- Bowtie2
- …

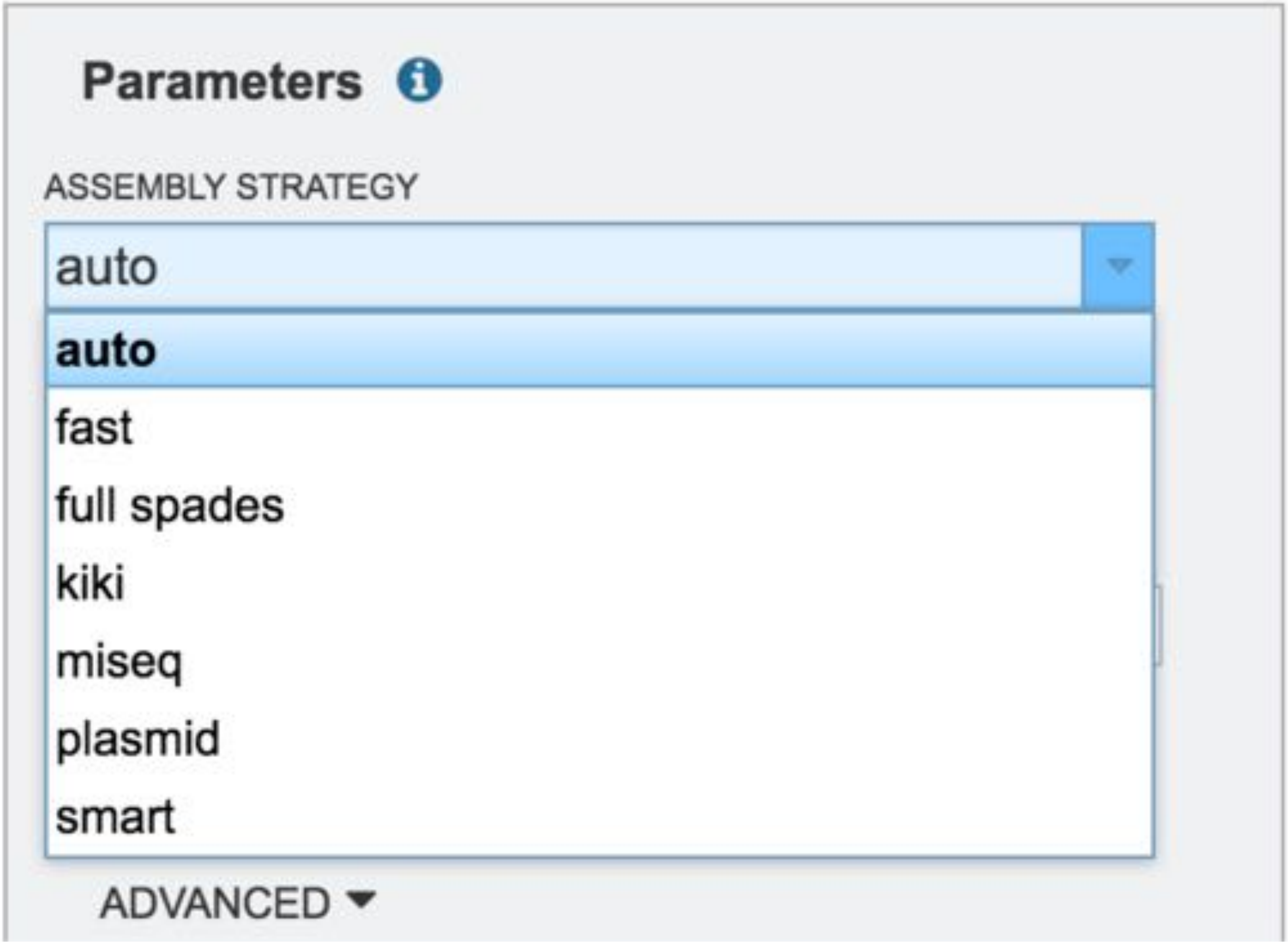
EVALUATION

- ALE
- REAPR
- QUAST

VISUALIZATION

- Nx Plots
- Contig Length
- ALE Comparison
- Benchmarks

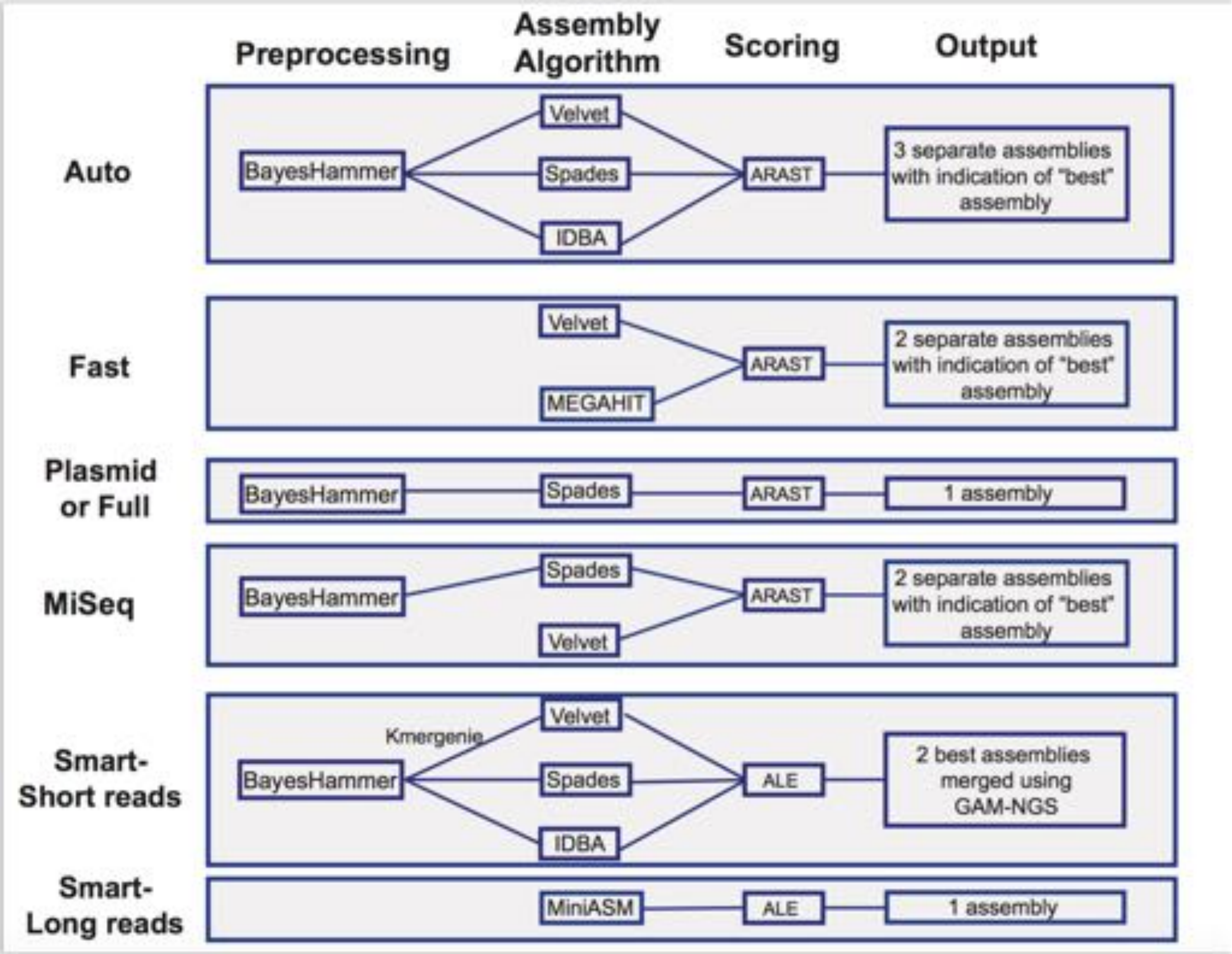
Curated assembly strategies



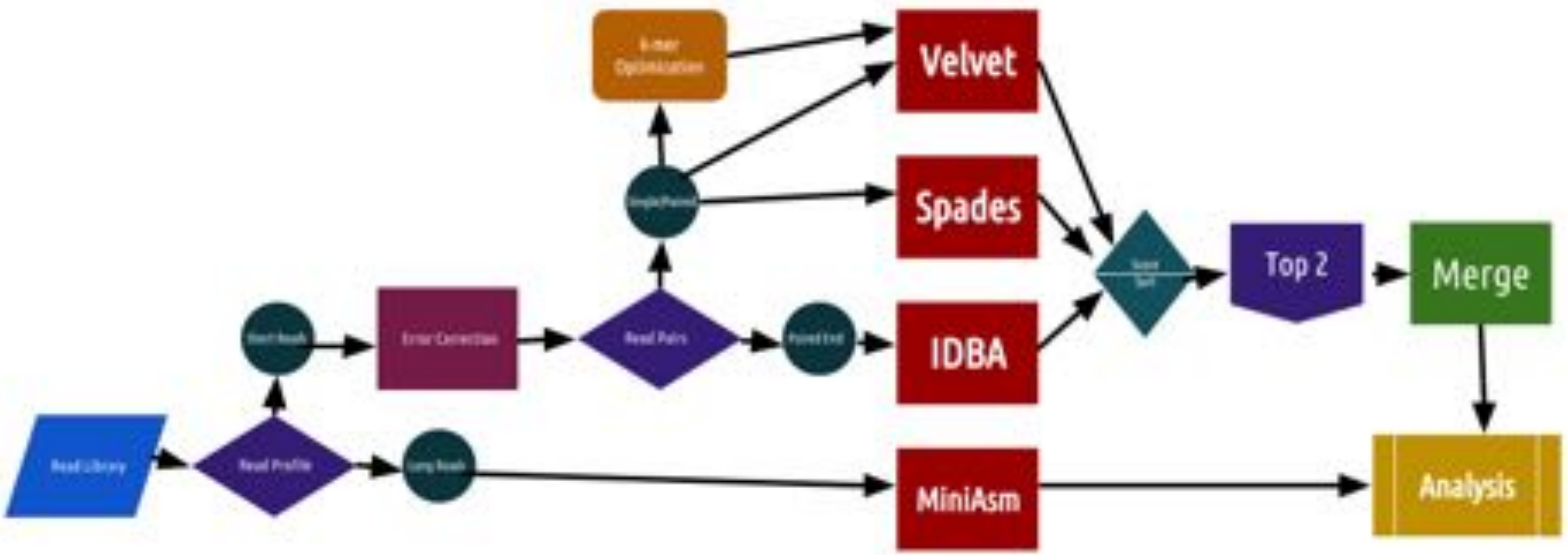
Parameters

ASSEMBLY STRATEGY

auto


Preprocessing
Assembly Algorithm
Scoring
Output
Auto
BayesHammer
Velvet
Spades
IDBA
ARAST
3 separate assemblies with indication of "best" assembly
Fast
Velvet
MEGAHIT
ARAST
2 separate assemblies with indication of "best" assembly
Plasmid or Full
BayesHammer
Spades
ARAST
1 assembly
MiSeq
BayesHammer
Spades
Velvet
ARAST
2 separate assemblies with indication of "best" assembly
Smart-Short reads
Kmergenie
BayesHammer
Velvet
Spades
IDBA
ALE
2 best assemblies merged using GAM-NGS
Smart-Long reads
MiniASM
ALE
1 assembly

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)

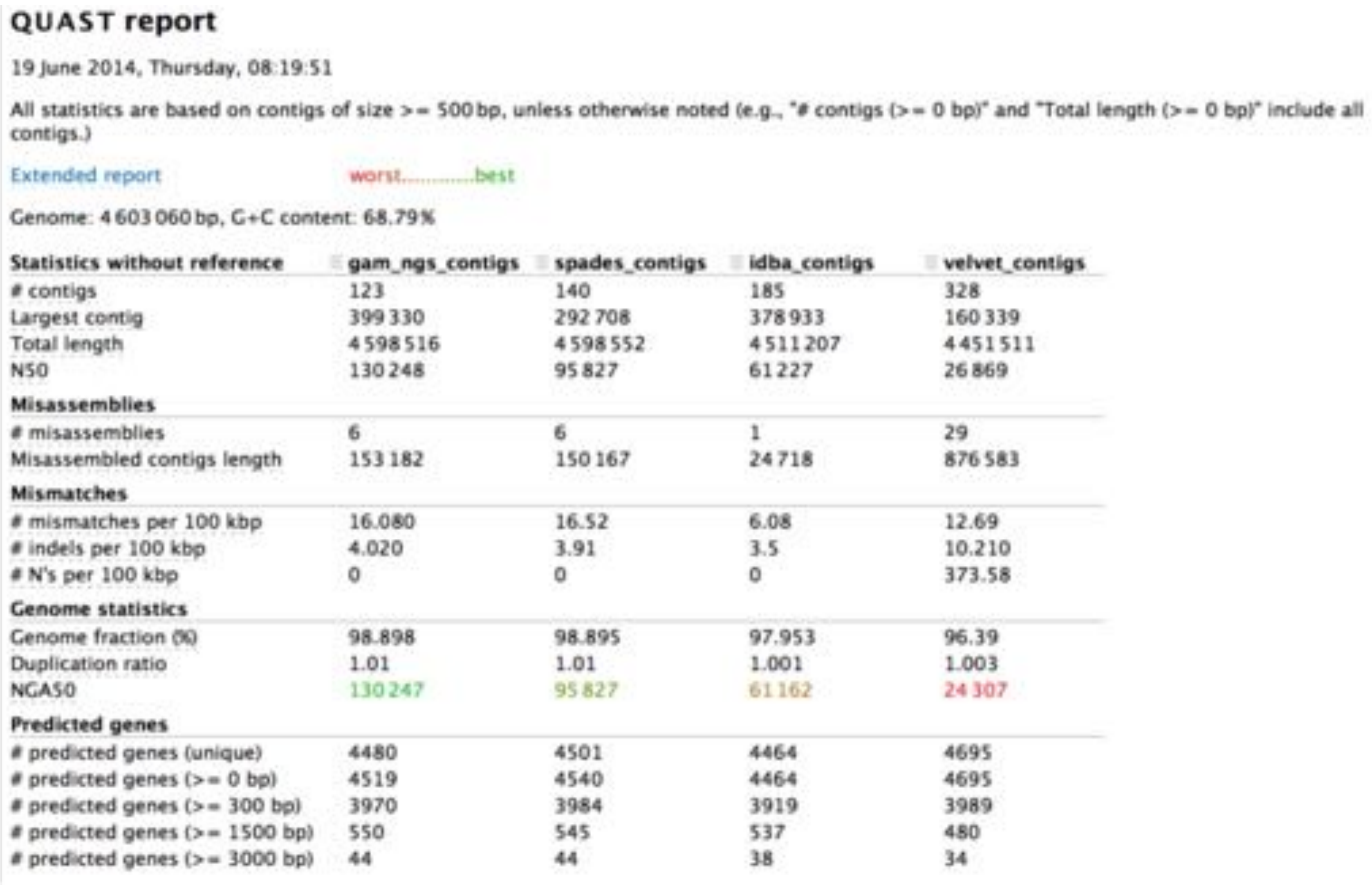
Typical execution times

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.

Evaluate assembled contigs



QUAST report

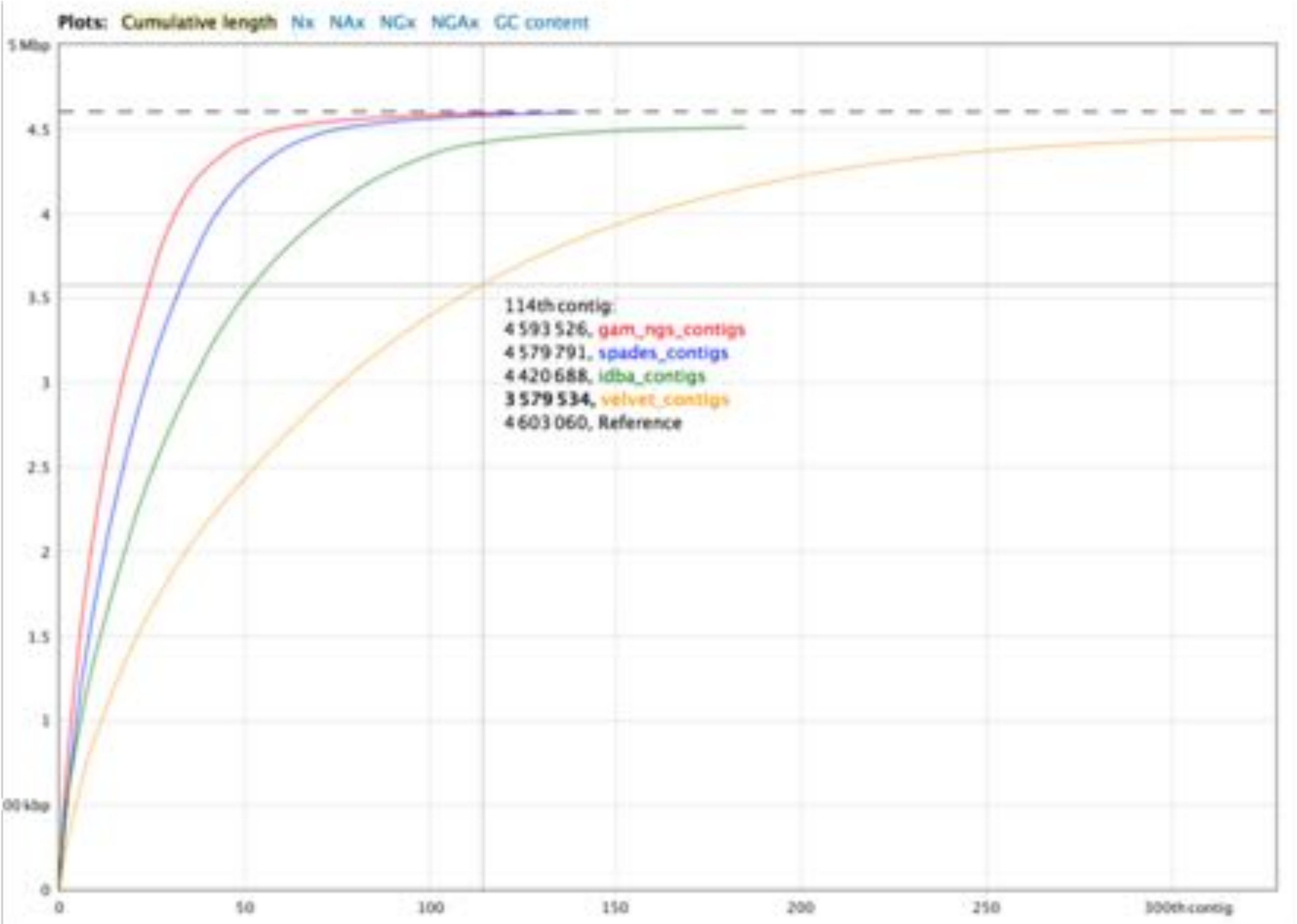
19 June 2014, Thursday, 08:19:51

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

Extended report worst.............best

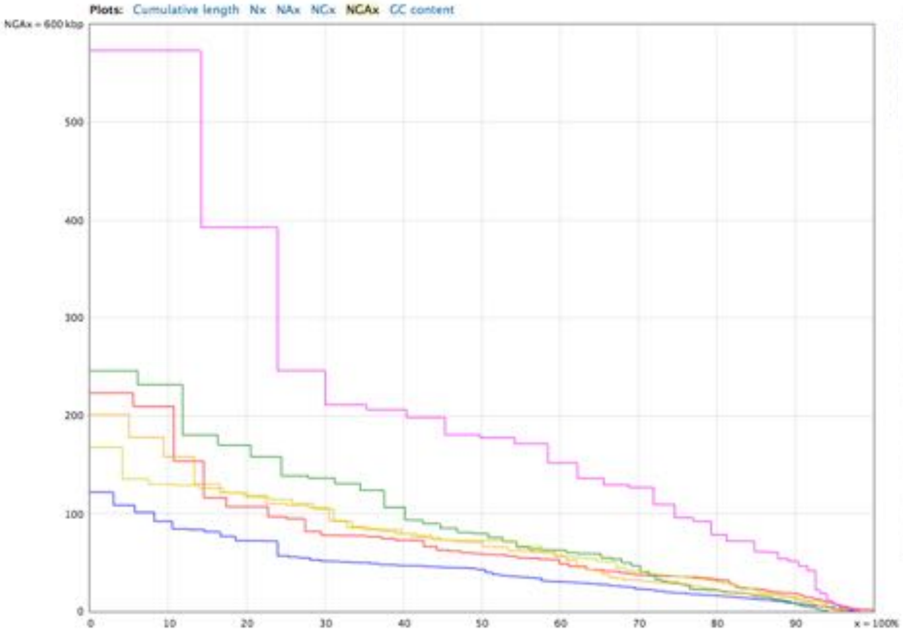
Genome: 4 603 060 bp, G+C content: 68.79%

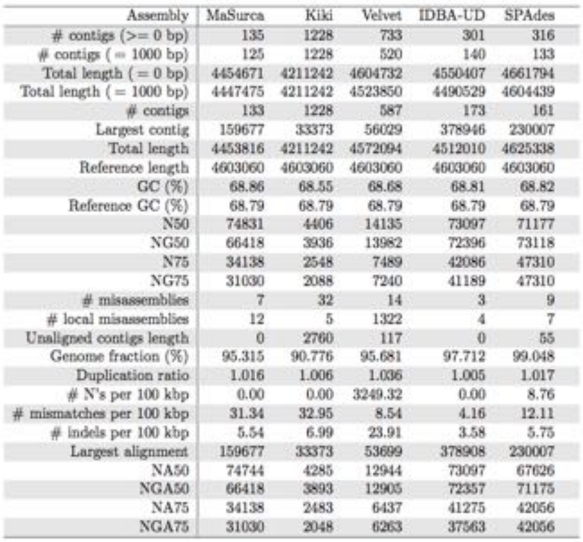
Statistics without reference	gam_ngs_contigs	spades_contigs	idba_contigs	velvet_contigs
# contigs	123	140	185	328
Largest contig	399 330	292 708	378 933	160 339
Total length	4 598 516	4 598 552	4 511 207	4 451 511
N50	130 248	95 827	61 227	26 869
Misassemblies				
# misassemblies	6	6	1	29
Misassembled contigs length	153 182	150 167	24 718	876 583
Mismatches				
# 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
Genome statistics				
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
Predicted genes				
# predicted genes (unique)	4480	4501	4464	4695
# predicted genes (>= 0 bp)	4519	4540	4464	4695
# predicted genes (>= 300 bp)	3970	3984	3919	3989
# predicted genes (>= 1500 bp)	550	545	537	480
# predicted genes (>= 3000 bp)	44	44	38	34


Plots: Cumulative length Nx NAx NGx NGAx GC content
5 Mbp
4.5
4
3.5
3
2.5
2
1.5
1
00 kbp
0
0
50
100
150
200
250
300th contig
114th contig:
4 593 526, gam_ngs_contigs
4 579 791, spades_contigs
4 420 688, idba_contigs
3 579 534, velvet_contigs
4 603 060, Reference

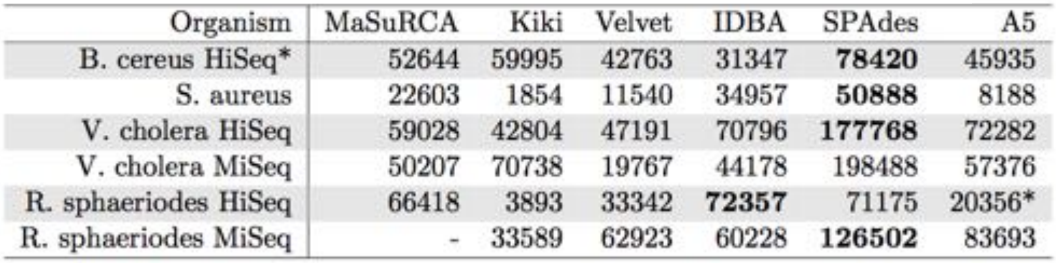
AssemblyRAST: Assembler Comparison

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

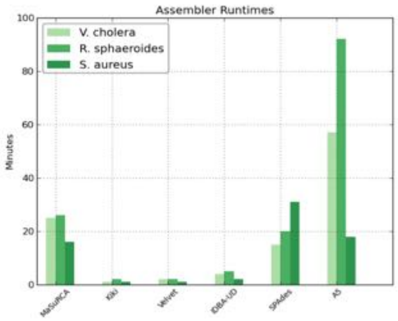




Assembly	MaSurca	Kiki	Velvet	IDBA-UD	SPAdes
# contigs (>= 0 bp)	135	1228	733	301	316
# contigs (= 1000 bp)	125	1228	520	140	133
Total length (= 0 bp)	4454671	4211242	4604732	4550407	4661794
Total length (= 1000 bp)	4447475	4211242	4523850	4490529	4604439
# contigs	133	1228	587	173	161
Largest contig	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
B. cereus HiSeq*	52644	59995	42763	31347	**78420**	45935
S. aureus	22603	1854	11540	34957	**50888**	8188
V. cholera HiSeq	59028	42804	47191	70796	**177768**	72282
V. cholera MiSeq	50207	70738	19767	44178	198488	57376
R. sphaeriodes HiSeq	66418	3893	33342	**72357**	71175	20356*
R. sphaeriodes MiSeq	-	33589	62923	60228	**126502**	83693


Assembler Runtimes

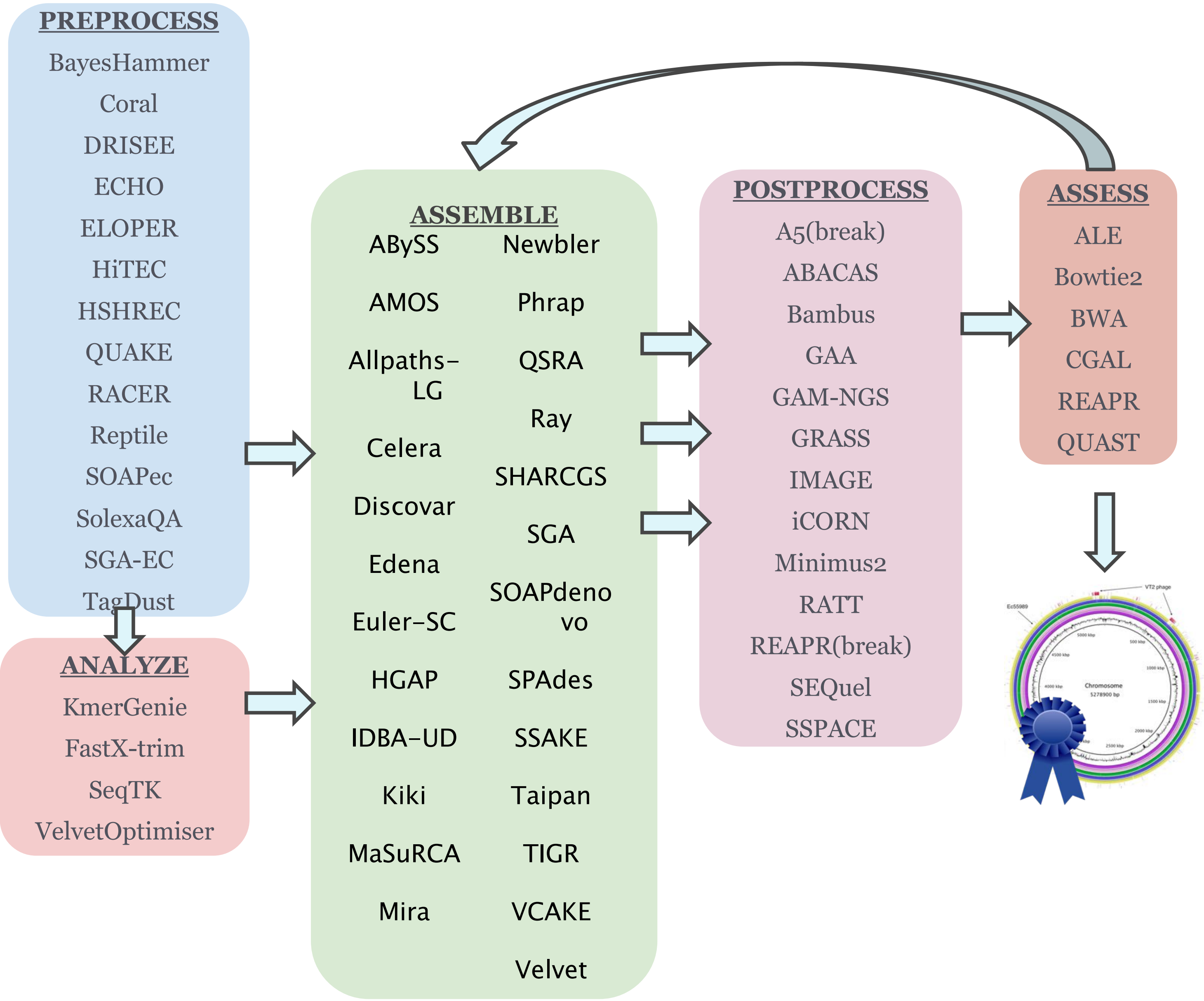
Interactive Demo

Acknowledgements

- PATRIC Team:
  - University of Chicago
    - Ryan Aydelott
    - Tom Brettin
    - Neal 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

Advanced pipelines



PREPROCESS

BayesHammer
Coral
DRISEE
ECHO
ELOPER
HiTEC
HSHREC
QUAKE
RACER
Reptile
SOAPec
SolexaQA
SGA-EC
TagDust

ANALYZE

KmerGenie
FastX-trim
SeqTK
VelvetOptimiser

ASSEMBLE

ABySS
AMOS
Allpaths-LG
Celera
Discovar
Edena
Euler-SC
HGAP
IDBA-UD
Kiki
MaSuRCA
Mira

Newbler
Phrap
QSRA
Ray
SHARCGS
SGA
SOAPdenovo
SPAdes
SSAKE
Taipan
TIGR
VCAKE
Velvet

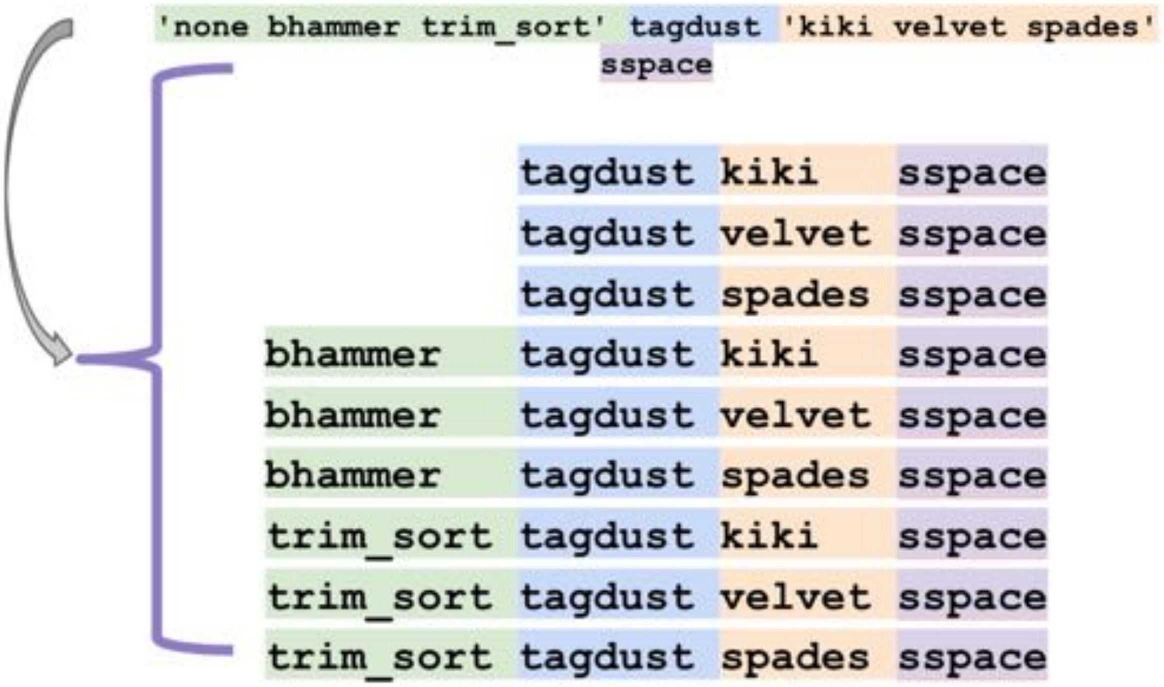
POSTPROCESS

A5(break)
ABACAS
Bambus
GAA
GAM-NGS
GRASS
IMAGE
iCORN
Minimus2
RATT
REAPR(break)
SEQuel
SSPACE

ASSESS

ALE
Bowtie2
BWA
CGAL
REAPR
QUAST

Pipeline design

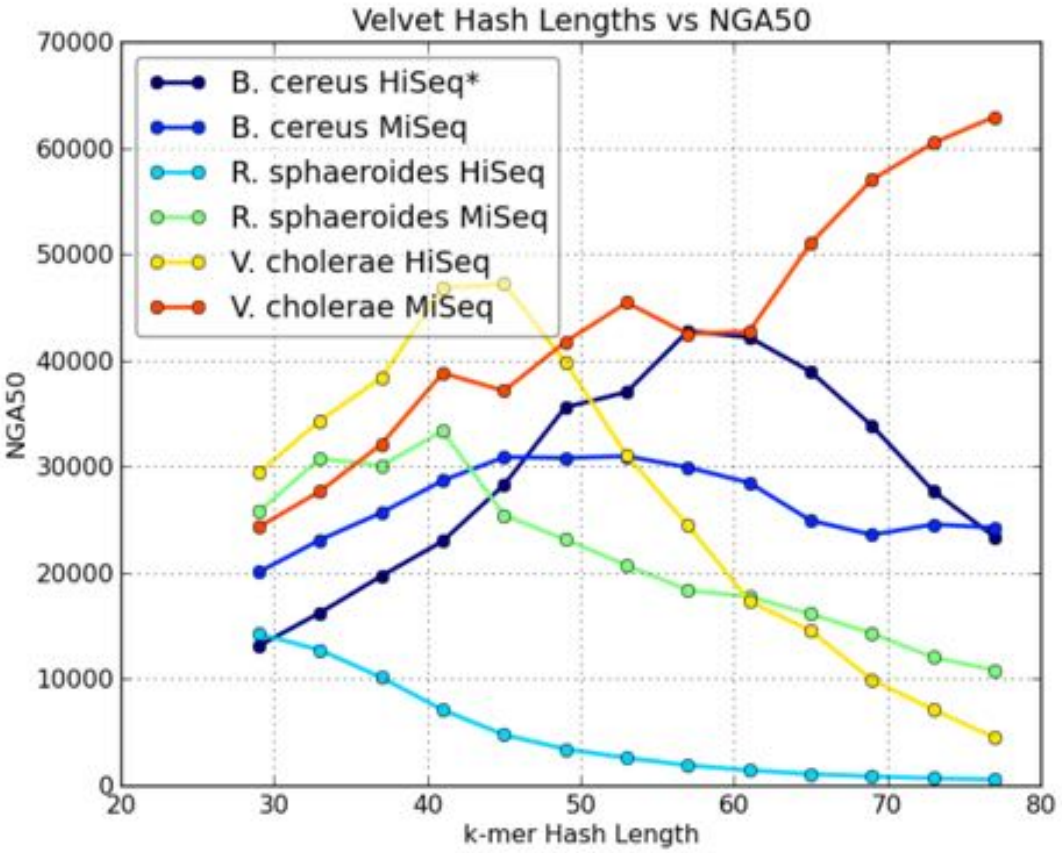


'none bhammer trim_sort' tagdust 'kiki velvet spades'
sspace

	tagdust	kiki	sspace
	tagdust	velvet	sspace
	tagdust	spades	sspace
bhammer	tagdust	kiki	sspace
bhammer	tagdust	velvet	sspace
bhammer	tagdust	spades	sspace
trim_sort	tagdust	kiki	sspace
trim_sort	tagdust	velvet	sspace
trim_sort	tagdust	spades	sspace

Parameter scan: k-mer Optimization

```
velvet ?hash_length=29-77:4
```



Upcoming improvements

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