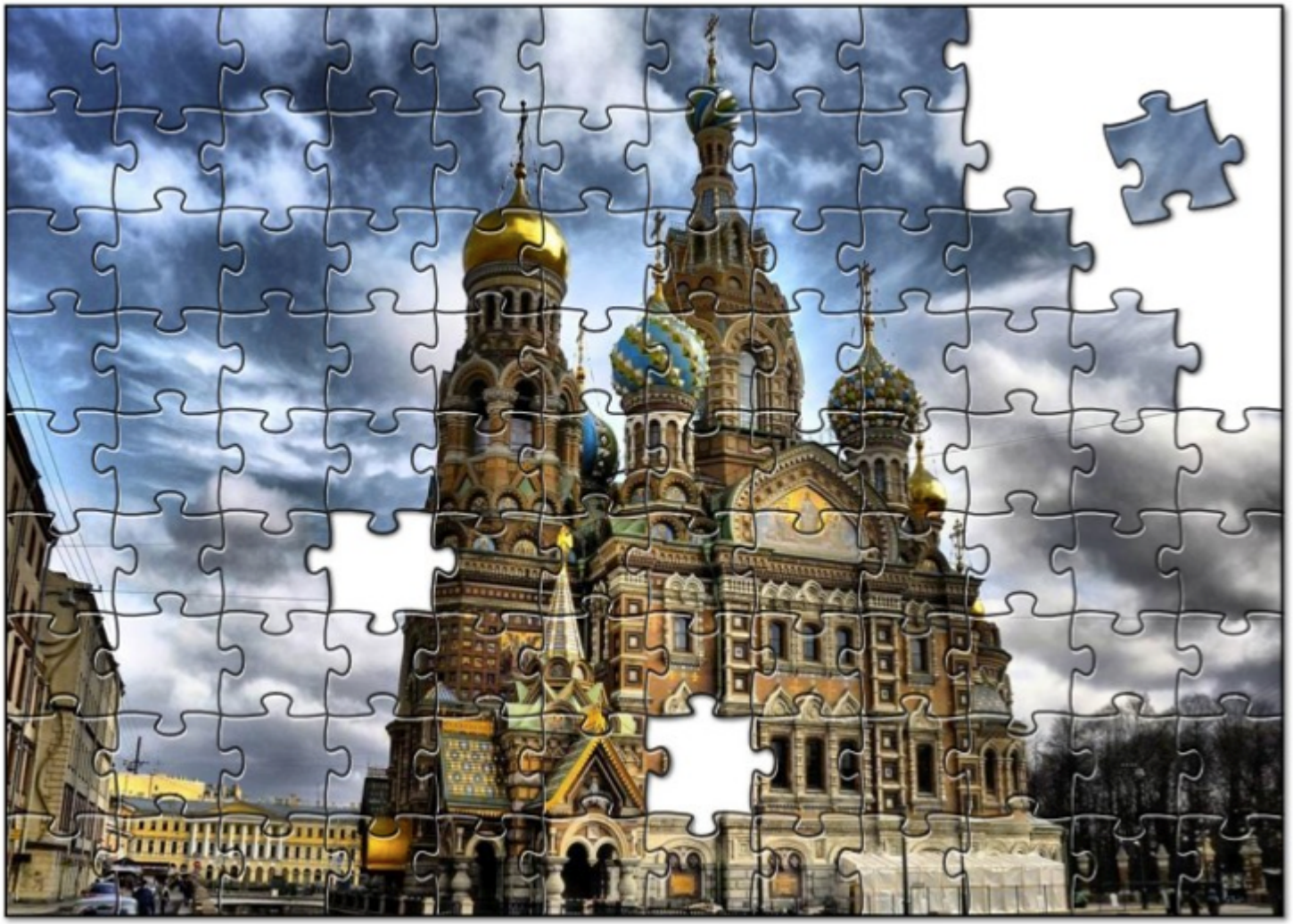
New Frontiers of Genome Assembly with SPAdes 3.1

Andrey Prjibelski

Algorithmic Biology Laboratory, St. Petersburg Academic University,
St. Petersburg, Russia
http://bioinf.spbau.ru/spades



SPAdes (Saint Petersburg Assembler)

SPAdes

- Originally designed as single-cell assembler
- Can deal with big variations of the coverage and MDA-imposed chimeric read connections
- Turned out to work well for multi-cell isolate assemblies
- Among with MaSuRCA are top 2 assemblers in GAGE-B study by Salzberg's lab (Magoc et al, 2013)

ABLAB

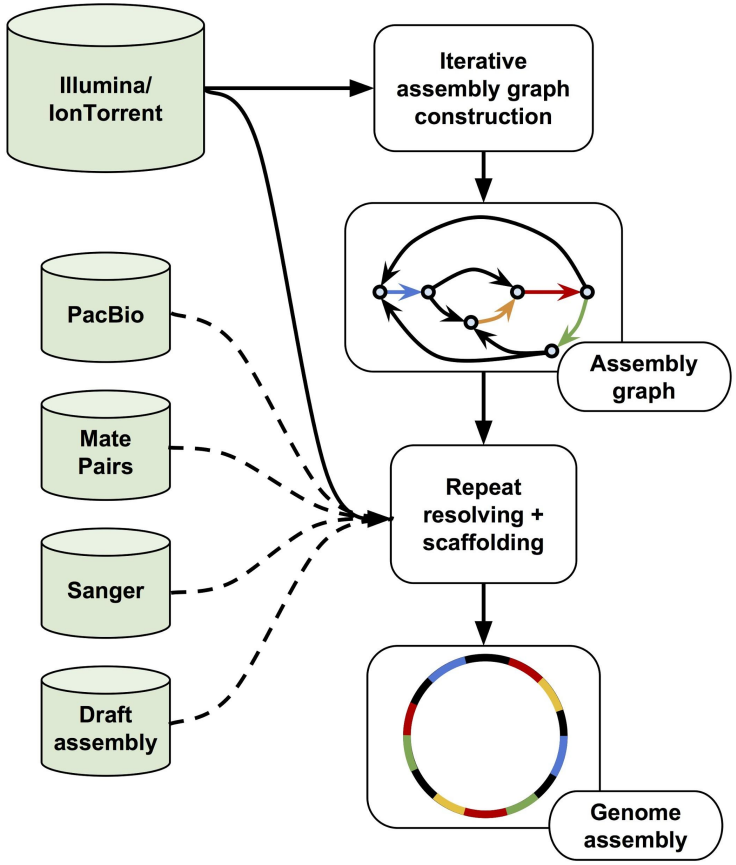
The Past: SPAdes 3.0

Major additions:

- Multiple libraries support: PE, MP, single reads
- IonTorrent read error correction and assembly
- (not only) PacBio hybrid assemblies
- dipSPAdes: assembler for highly polymorphic diploid genomes

ABLAB

Hybrid Assemblies



- New universal repeat resolution algorithm
- Can utilize different sources of genomic distance information

ABLAB

Andrey Prjibelski, Irina Vasilinetc, et al.
ExSPAnder: a Universal Repeat Resolver for DNA Fragment Assembly. ISMB 2014

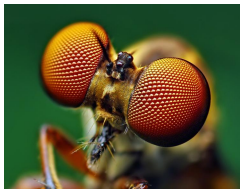
dipSPAdes

The first de Bruijn graph assembler designed for highly polymorphic diploid genomes:

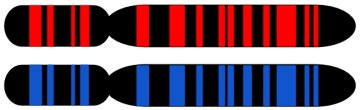
Fungus
heterozygosity up to 20%

Sea squirts
heterozygosity up to 12%

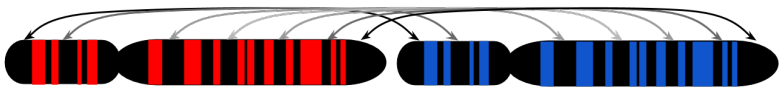
Plants
avg heterozygosity 7%



Insects
avg heterozygosity 9%



diploid genome with high heterozygosity



conventional approaches assemble such genome as two highly repetitive sequences and construct very fragmented assemblies

dipSPAdes takes advantage of the assembly graph structure and constructs longer consensus contigs

ABLAB

Yana Safonova, Anton Bankevich, Pavel A. Pevzner.
dipSPAdes: an assembler for highly polymorphic diploid genomes. RECOMB 2014

IonTorrent Error Correction & Assembly

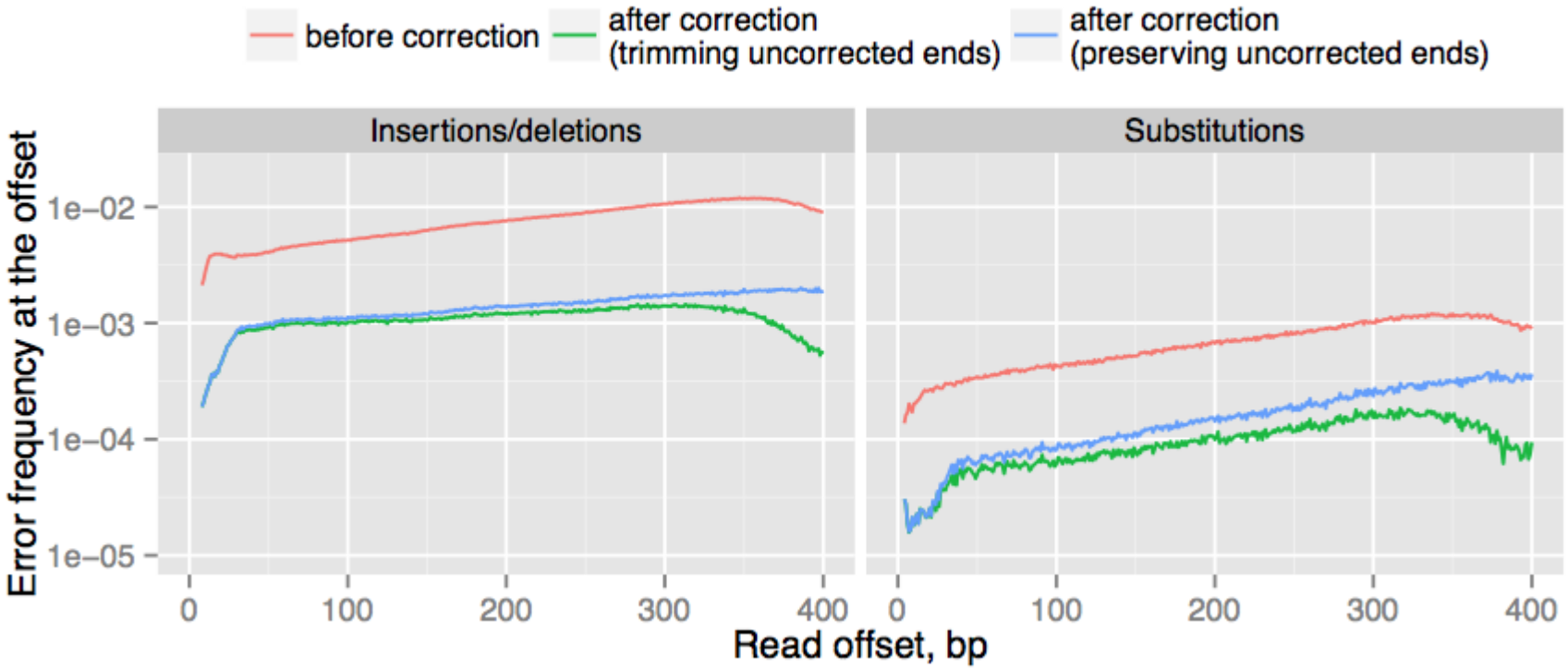
IonHammer: read error correction

- Bases on ideas of BayesHammer
- Works in homopolymer space
- Corrects both indels and mismatches
- Knows about undercall-overcall pairs

ABLAB

This enables accurate assemblies with inaccurate IonTorrent reads

IonTorrent Error Correction & Assembly



E. coli O157:H7 str. Sakai from IonCommunity

ABLAB

5-8 fold reduction of both indel and mismatch rates!

The Present: SPAdes 3.1

- Nextera mate pairs support and mate pair-only assemblies
- Better scaffolding and repeat resolution
- Runtime & RAM consumption improvements
- Further IonTorrent improvements
- Integration with cloud services

ABLAB

The Present: SPAdes 3.1

- Nextera mate pairs support and mate pair-only assemblies
- Better scaffolding and repeat resolution
- Runtime & RAM consumption improvements
- Further IonTorrent improvements
- Integration with cloud services

ABLAB

Available today at http://bioinf.spbau.ru

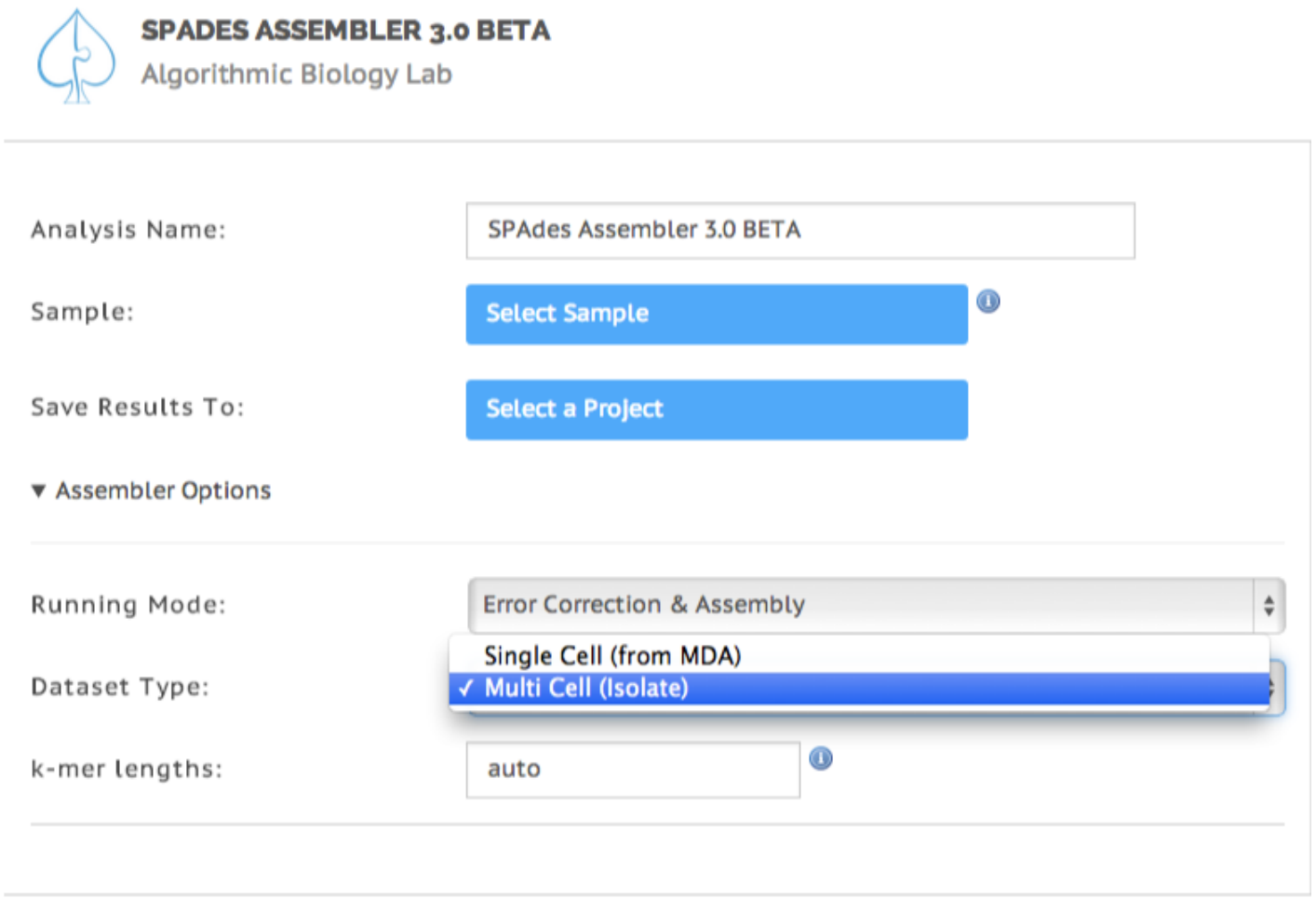
SPAdes Goes Cloud: BaseSpace

- SPAdes tuned to Illumina BaseSpace platform
- Integrates SPAdes and QUAST
- Single push-button de novo assembly solution

Check it out: https://basespace.illumina.com/apps/160160

ABLAB

SPAdes Goes Cloud: BaseSpace



SPAdes for TorrentServer

- AssemblerPlus TS plugin
- Integrates SPAdes and QUAST
- Native support of BAM files as input
- Read error correction additionally uses raw flow information

Available soon!

ABLAB

SPAdes for Cloud Platforms

SPAdes runs on:

- Illumina BaseSpace
- DNAnexus
- TorrentServer
- Galaxy (available from Galaxy Tool Shed)

ABLAB

Runtime Improvements

Improvements to BayesHammer (error correction)

- On 100 Mbp diploid genome:
  - Was: 90 hours
  - Now: 16 hours

Improvements to repeat resolution

- On 60 Mbp repeat-rich genome:
  - Was: 78 hours
  - Now: 2 hours

ABLAB

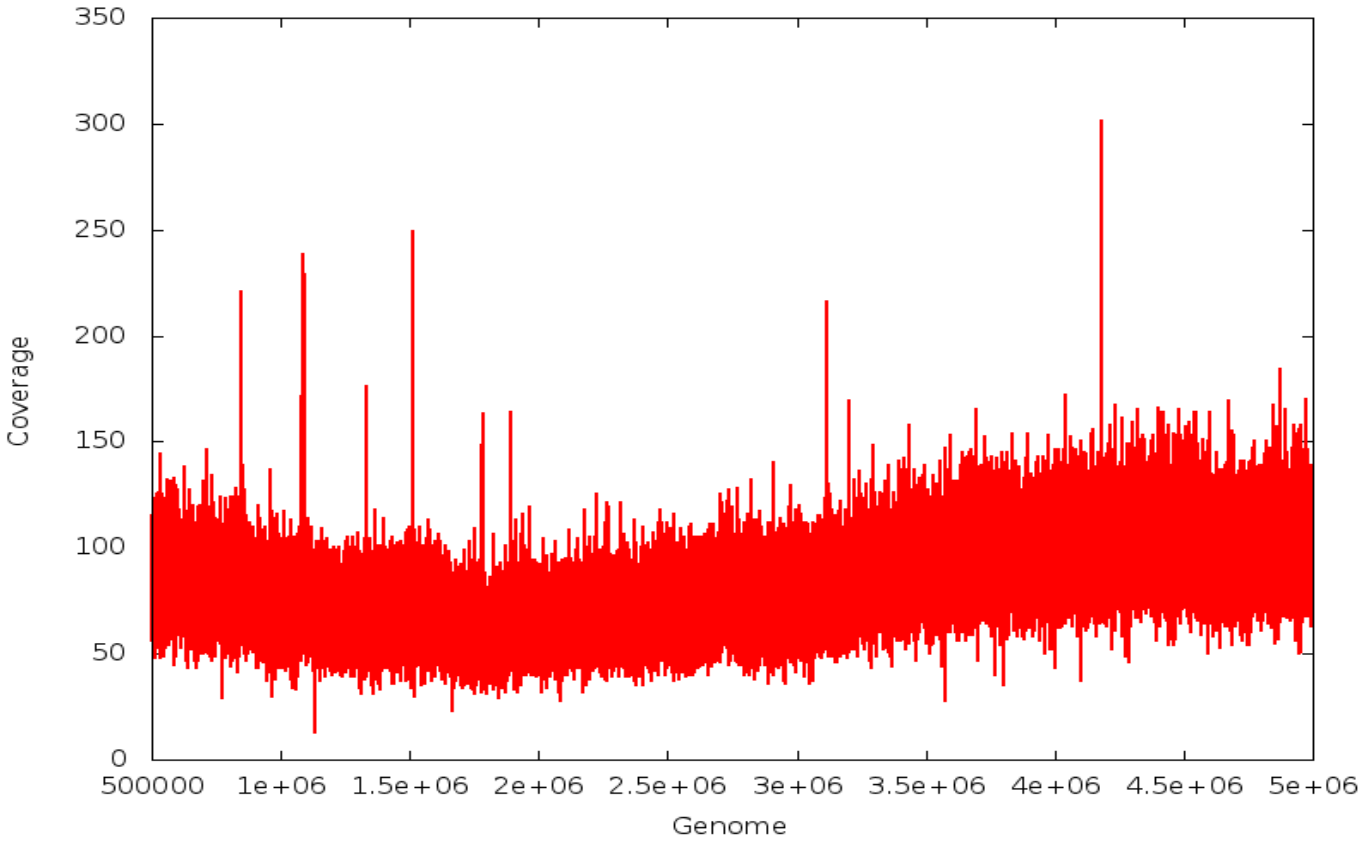
Illumina Nextera Mate Pairs

Reasonably uniform coverage

ABLAB

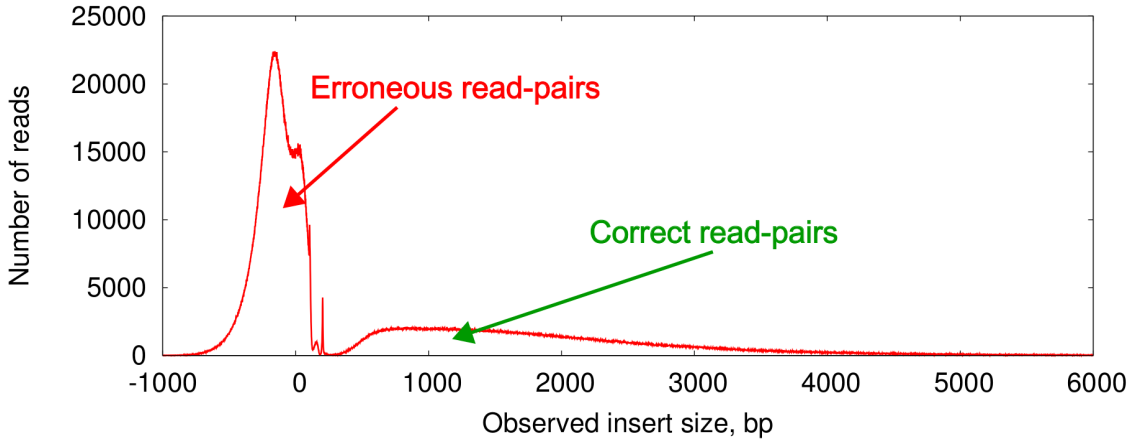
Illumina Nextera Mate Pairs

Reasonably uniform coverage:

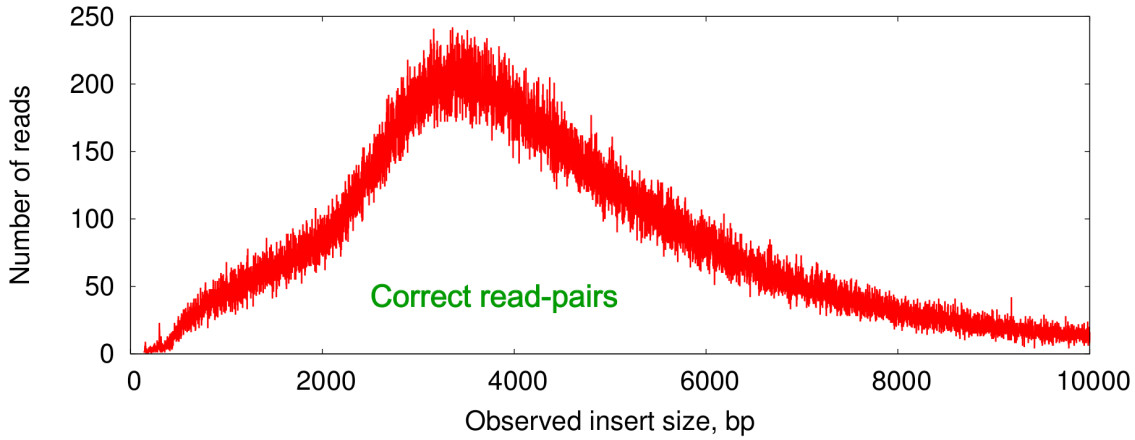


Illumina Nextera Mate Pairs

Conventional mate-pairs:



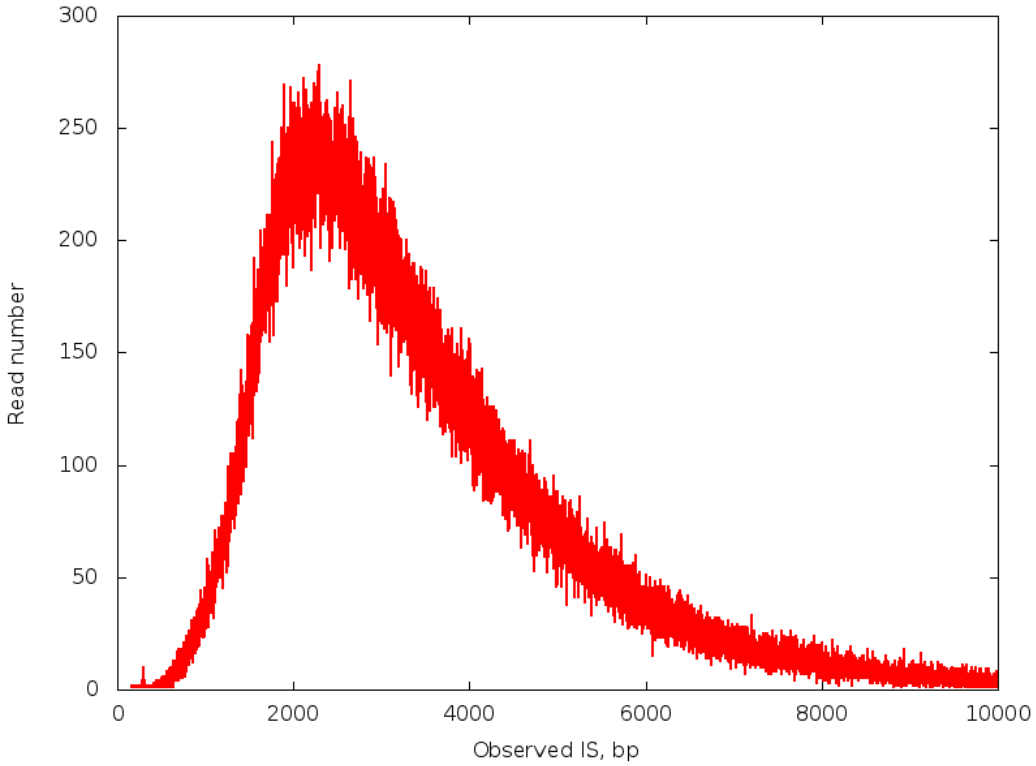
Nextera mate-pairs:



ABLAB

Illumina Nextera Mate Pairs

No “paired-end” fragments:



ABLAB

Everything looks like large insert-size paired-end library!

Mate Pair Only Assemblies

	SPAdes	Velvet, k=67	Velvet, k=99
# scaffolds	11	**5**	6
Largest contig	1496k	**2817k**	2006k
Reference length	2944k	2944k	2944k
N50	1496k	**2817k**	2006k
N75	739k	**2817k**	814k

L. monocytogenes data from Illumina

ABLAB

Why one would need SPAdes then?

Mate Pair Only Assemblies

	SPAdes	Velvet, k=67	Velvet, k=99
# scaffolds	11	5	6
Reference length	2944k	2944k	2944k
# misassemblies	**0**	**6**	**3**
Largest alignment	**1496k**	1493k	1493k
NGA50	**1496k**	1493k	1493k
NGA75	**739k**	505k	509k
Genome fraction	**99.463%**	99.133%	99.391%

L. monocytogenes data from Illumina

ABLAB

Acknowledgement

SPAdes team:

Dmitry Antipov
Anton Bankevich
Sergey Nurk
Alexey Gurevich
Anton Korobeynikov
Yana Safonova
Irina Vasilinetc
Alla Lapidus
Pavel Pevzner

ABLAB

This work was supported by the Government of the Russian Federation (grant 11.G34.31.0018)

Thank you!

SPAdes: bioinf.spbau.ru/spades

QUAST: bioinf.spbau.ru/quast