SPAdes

Myeongkyu Park (2021-20471)

Illustration from Center for Algorithmic Biotechnology (CAB SPbU)

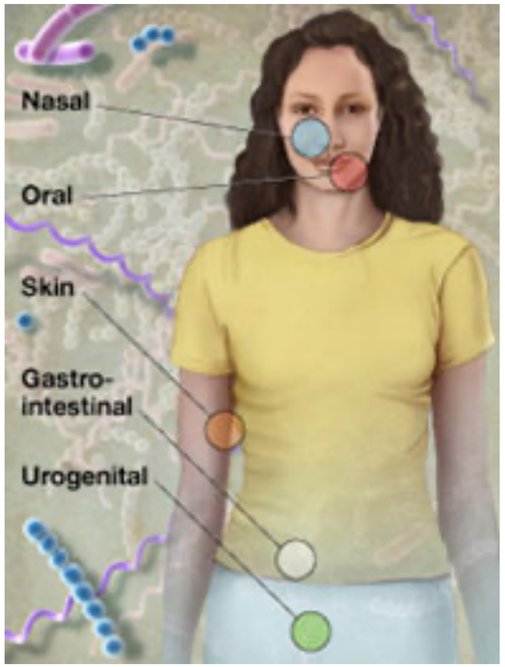
Motivation

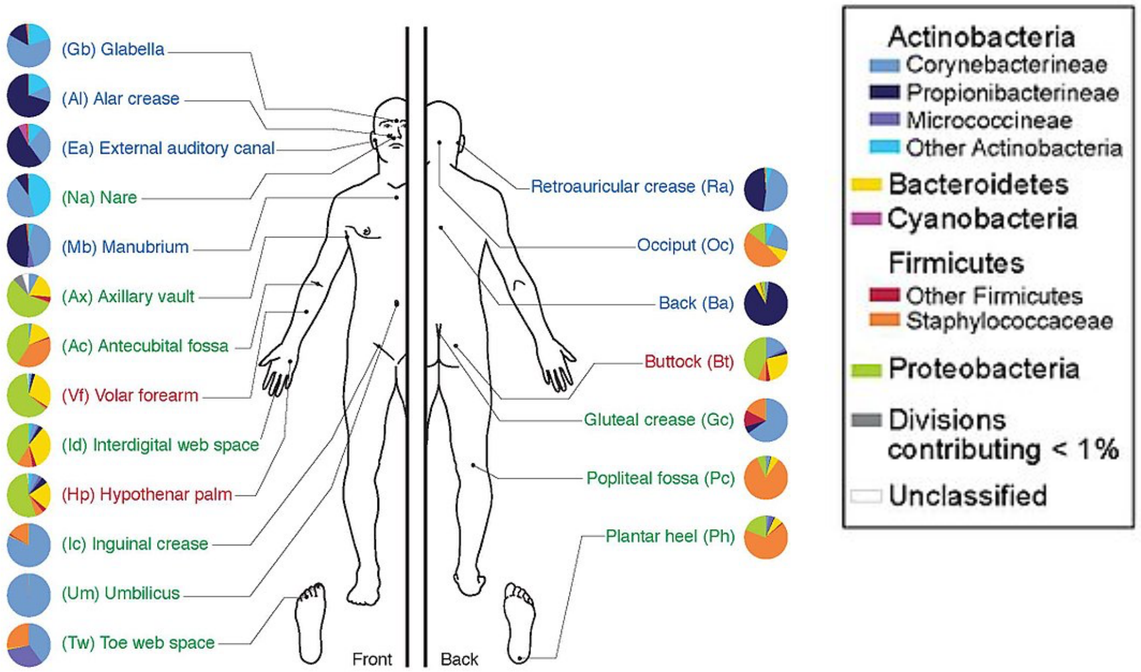
- Human Microbiome Project
- Single-Cell Sequencing
- Challenges

Human Microbiome Project

Motivation

- Genome-wide association study (GWAS)





- Sequencing bottleneck from difficulty of cloning bacteria from various env.

Single-Cell Sequencing (SCS)

Motivation

Metagenomics could not light up the dark matter of microbial life

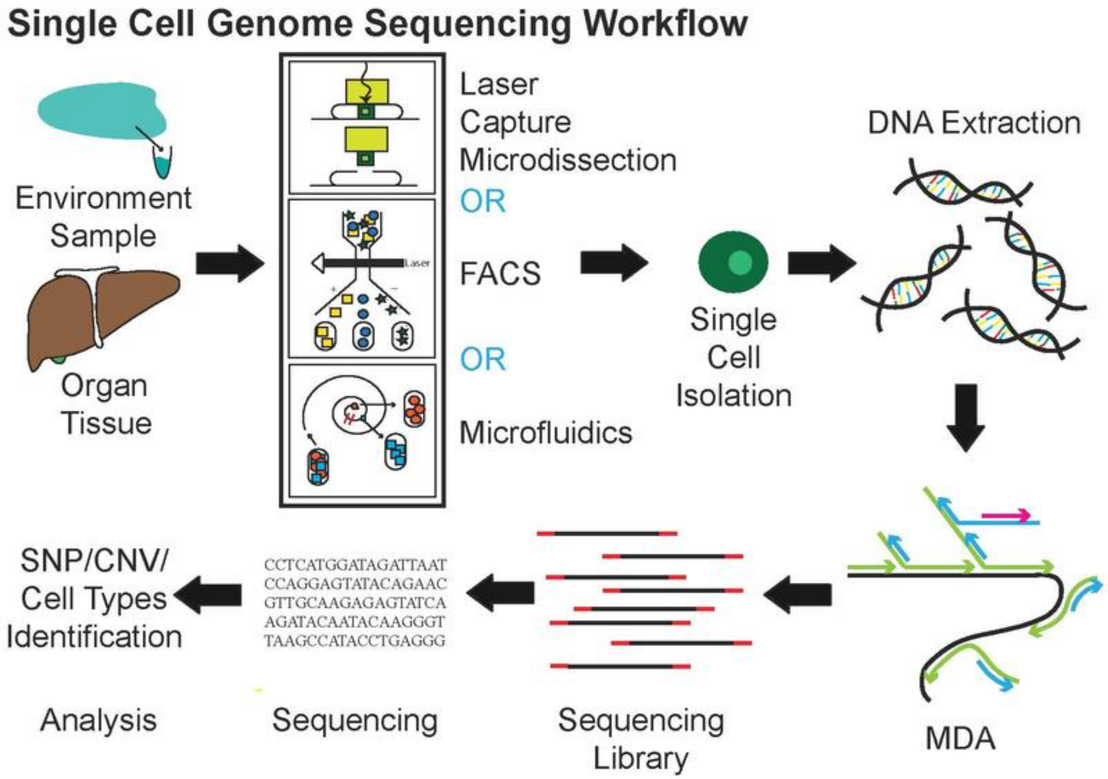


Image from Spits *et al., Nat. Protoc.* 2006

- Provide a higher resolution of cellular differences
- Enable sequencing bacterial genomes from single cells

Multiple Displacement Amplification

Motivation

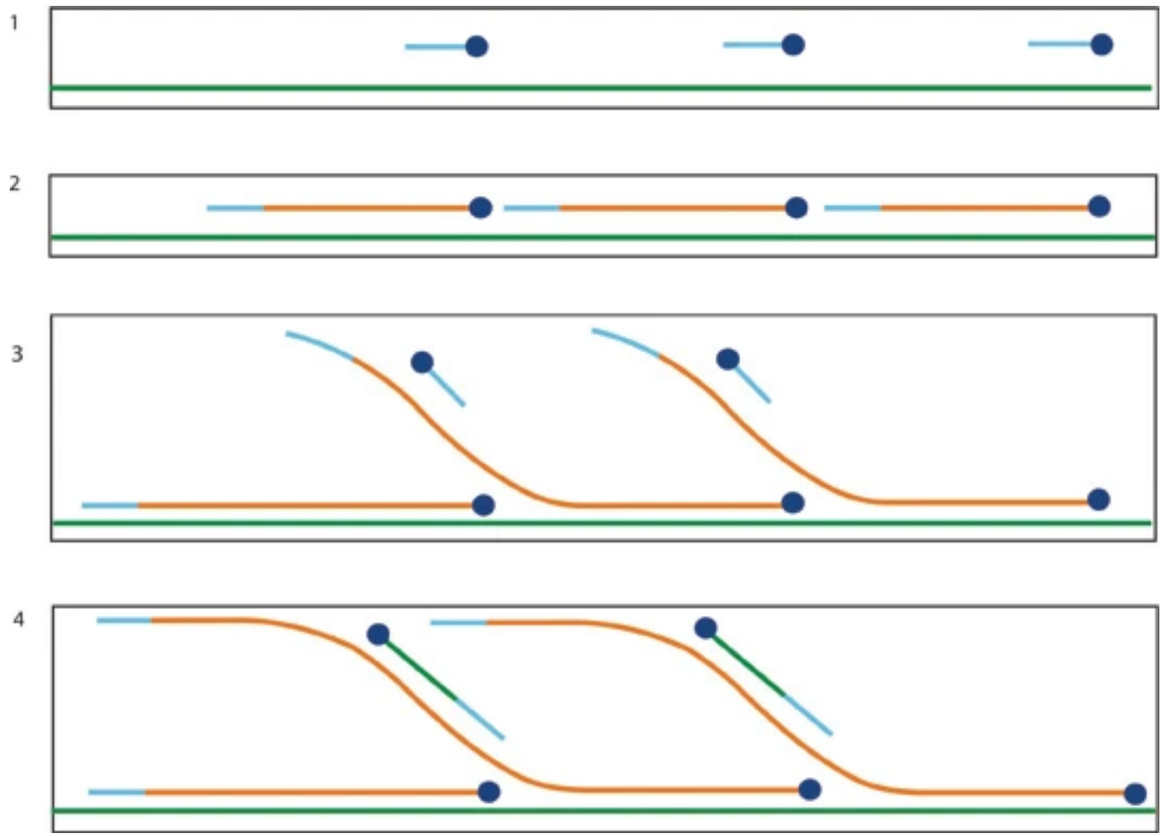


Image from Spits *et al., Nat. Protoc.* 2006

- Dominant technology for SC amplification
- Extreme amplification bias & chimeric reads and read-pairs
- Existing assemblers could not manage the complications

Challenges

Motivation

1. Non-uniform coverage : a coverage threshold would throw a lot of the genome away

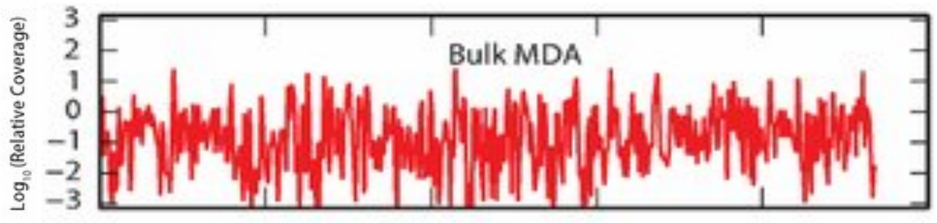
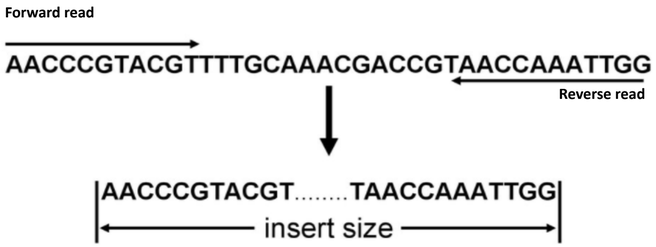
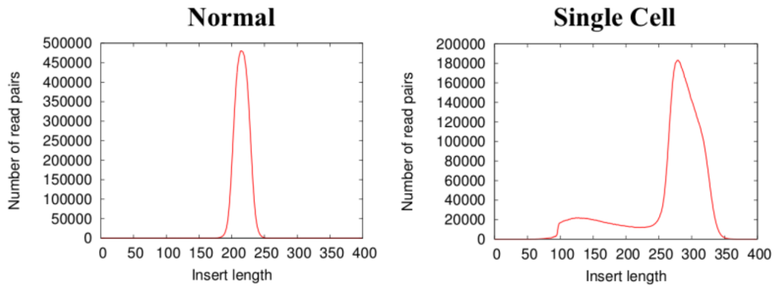


Image modified from Sidore *et al., Nucleic Acids Res.* 2015

2. Insert size variation : no normal distribution in SCS





Images from CAB (Center for Algorithmic Biotechnology)

3. Finally, chimeric connections + also more frequent errors !

E+V-SC assembler

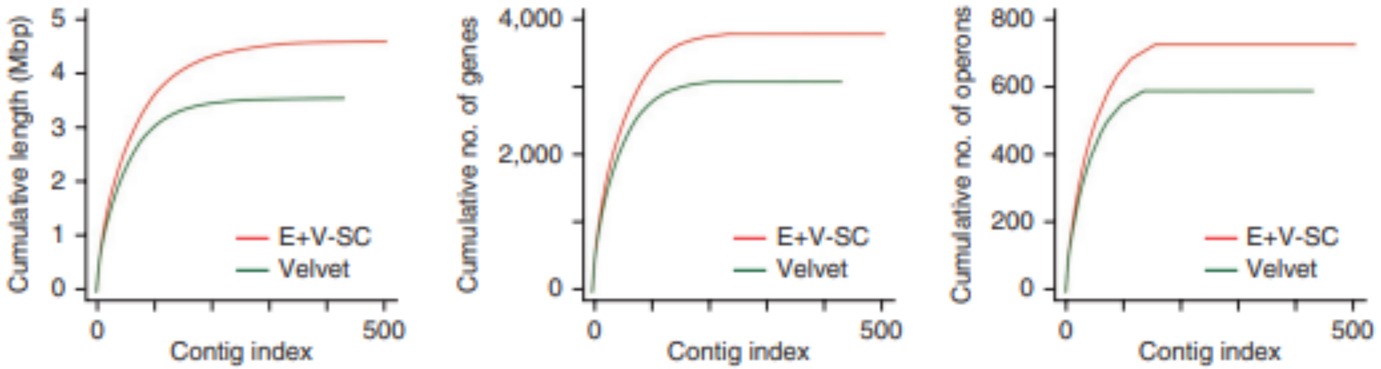
Motivation

- E+V-SC (Euler + Velvet - Single cell) assembler

  For SCS data, 25% more genes detected than the existing assembler

Assembler	No. of contigs	N_{50} (bp)	Largest (bp)	Total (bp)	No. ORFs (MetaGene)	No. ORFs (APIS)	No. COGs	No. conserved single-copy genes
Velvet	1,856	11,531	100,589	3,921,396	4,575	2,462	2,160	55/111 (46%)
Velvet-SC	933	23,230	113,282	4,284,882	4,234	2,627	2,307	75/111 (67%)
E +V-SC	823	30,293	113,282	4,282,110	4,154	2,604	2,281	75/111 (67%)

Total number of contigs; assembly N_{50} (for contigs >110 bp); length of the largest contig (for contigs >110 bp); total nucleotides in the assembly (for contigs >110 bp); number of ORFs > 20 bp predicted by MetaGene; number of ORFs with phylogenetic assignments by APIS; number of ORFs with COGs identified by BLAST; and number of 111 conserved single-copy genes present.



Nevertheless, efforts are needed for better performance

Introduction

- SPAdes
- de Bruijn Graphs

SPAdes

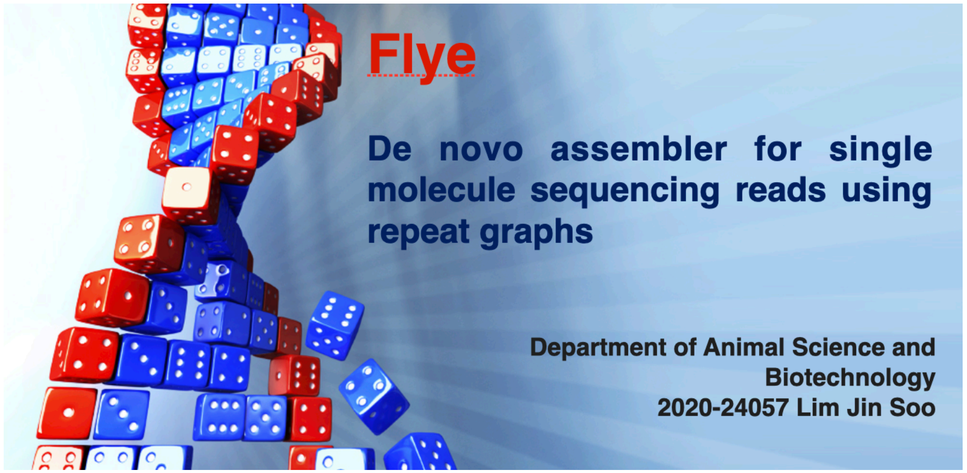
Introduction

- Challenge success !
- A number of new algorithms !
- Better performance !

- Due to sequencing errors and repeats, we cannot just find a Eulerian cycle
- In the presence of errors, …

Standard de Bruijn graph

Introduction



EADS as substrings. We define the *de Bruijn graph* DB(READS, k) as follows (Fig. 2):

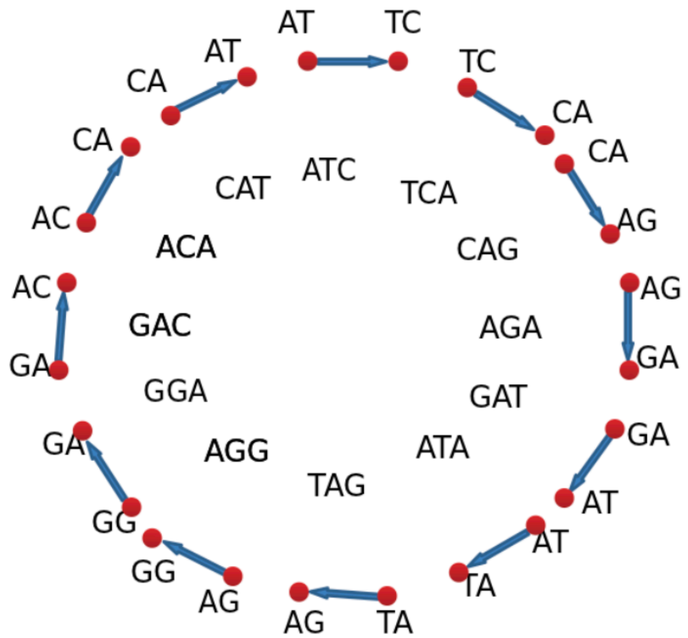
D1. Define an initial graph G_0 on $2N$ vertices. For each k-mer a that occurs in strings in READS as a substring, introduce two new vertices u, v and form an edge $u \to v$. Label the new edge by a, u by PREFIX(a), and v by SUFFIX(a). Note that we label edges by k-mers and vertices by $(k-1)$-mers.

D2. Glue vertices of G_0 together if they have the same label.

Standard de Bruijn graph (example)

Introduction

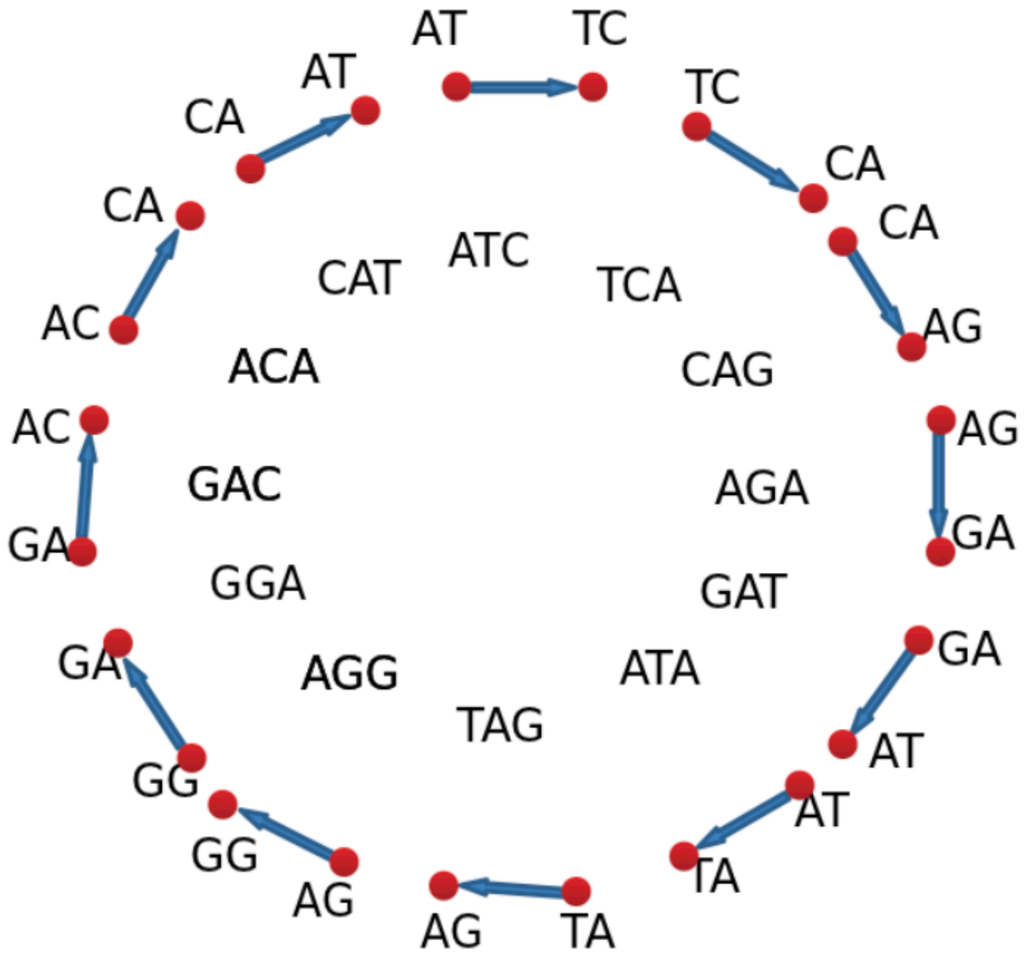
- Given string: ATCAGATAGGACAT
- Let's construct circular de Bruijn graph with $k = 3$!



Images from CAB (Center for Algorithmic Biotechnology)

Standard de Bruijn graph (example)

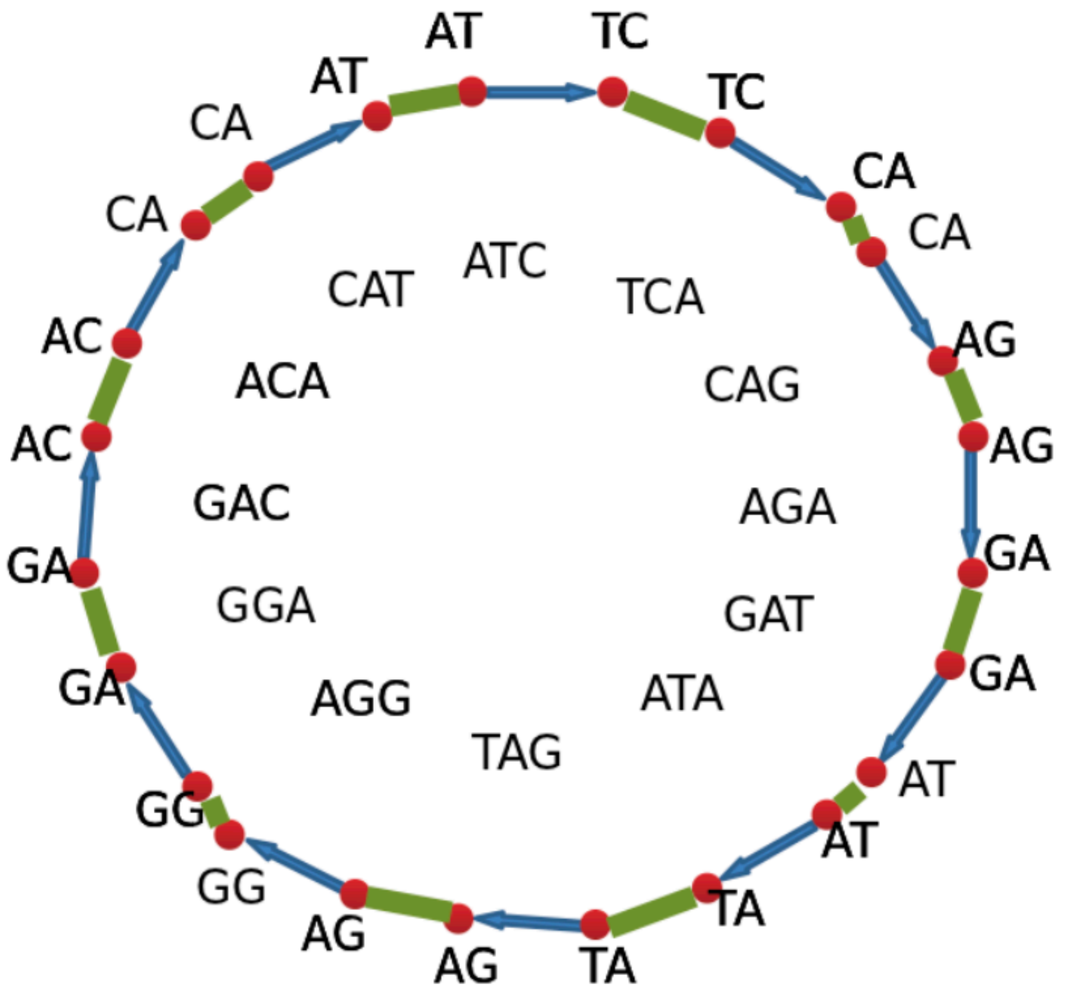
Introduction



Standard de Bruijn graph (example)

Introduction

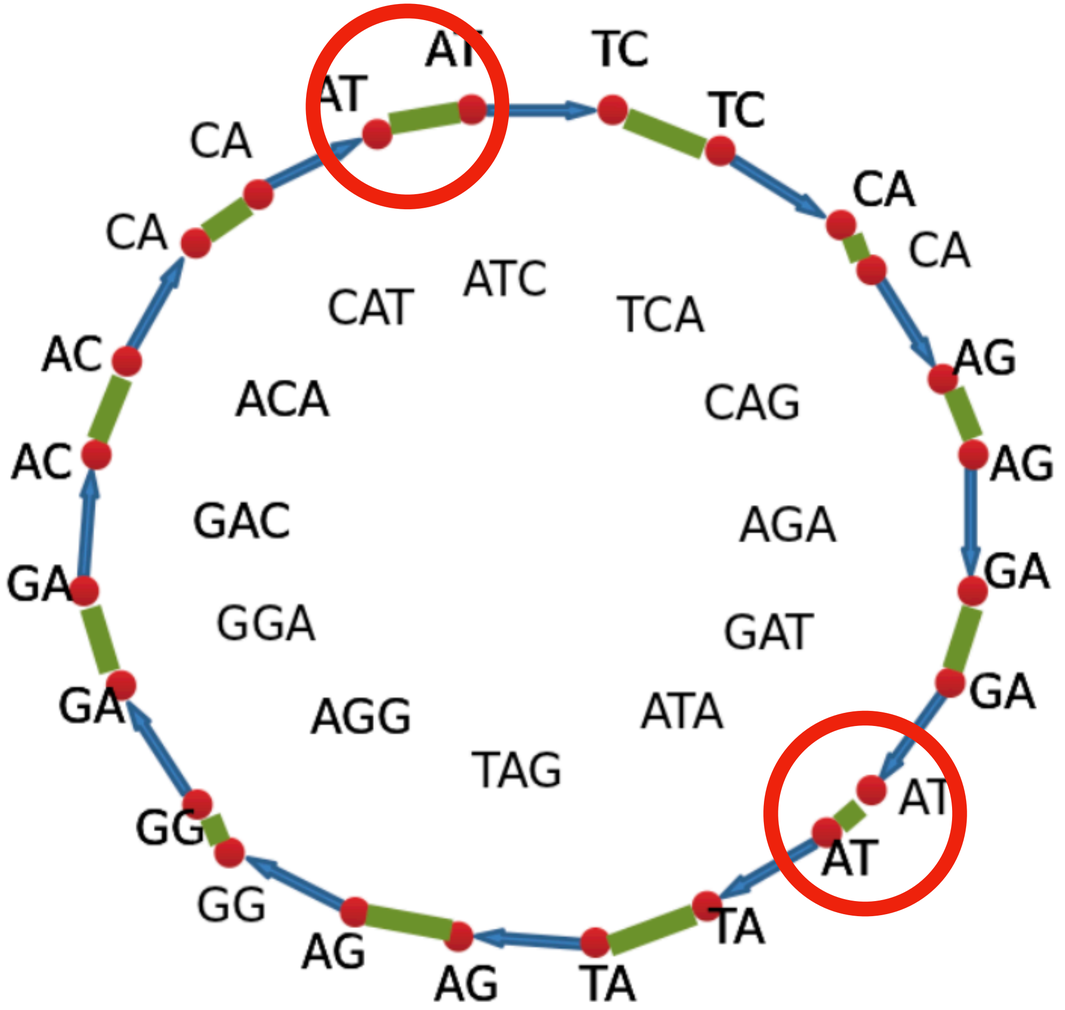
- Glue vertices together if they have same label



Standard de Bruijn graph (example)

Introduction

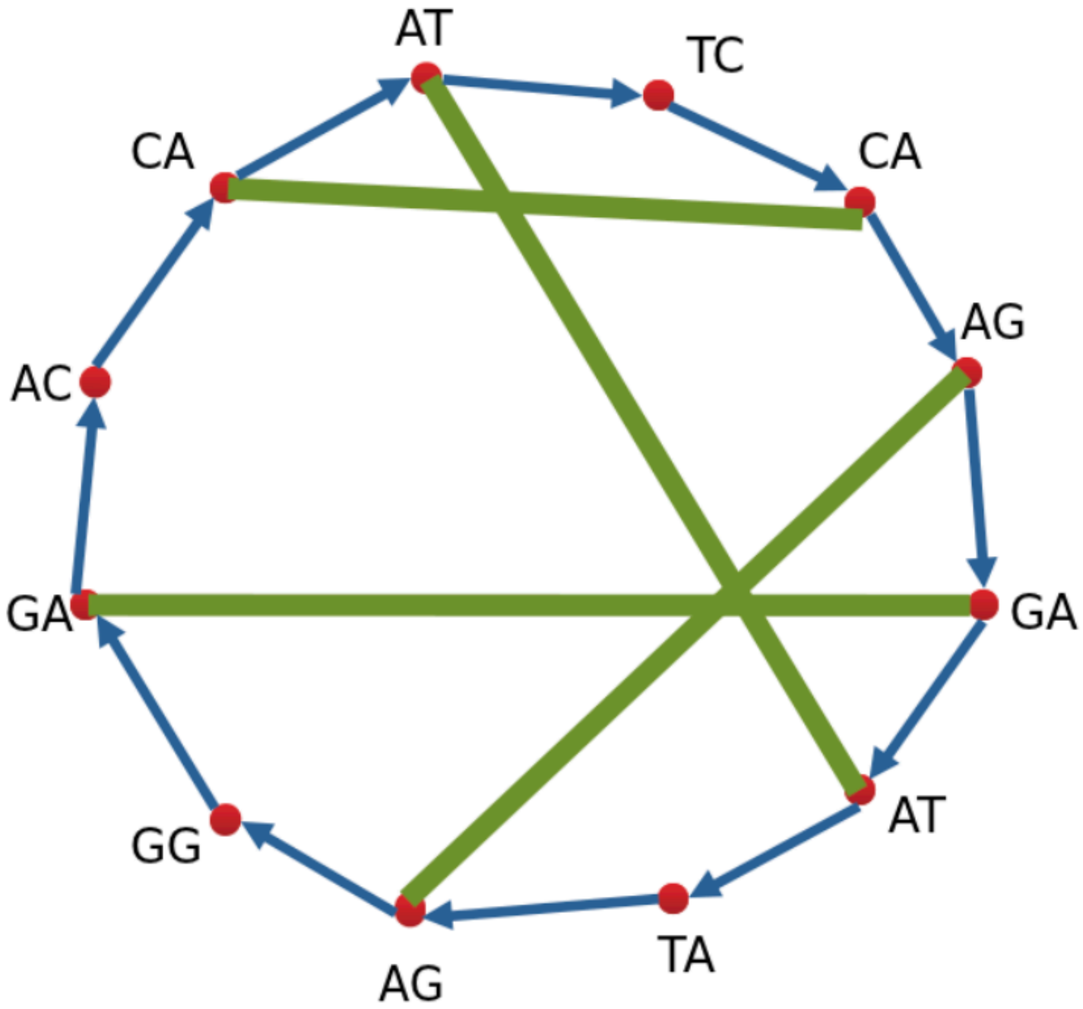
- Glue vertices together if they have same label



Standard de Bruijn graph (example)

Introduction

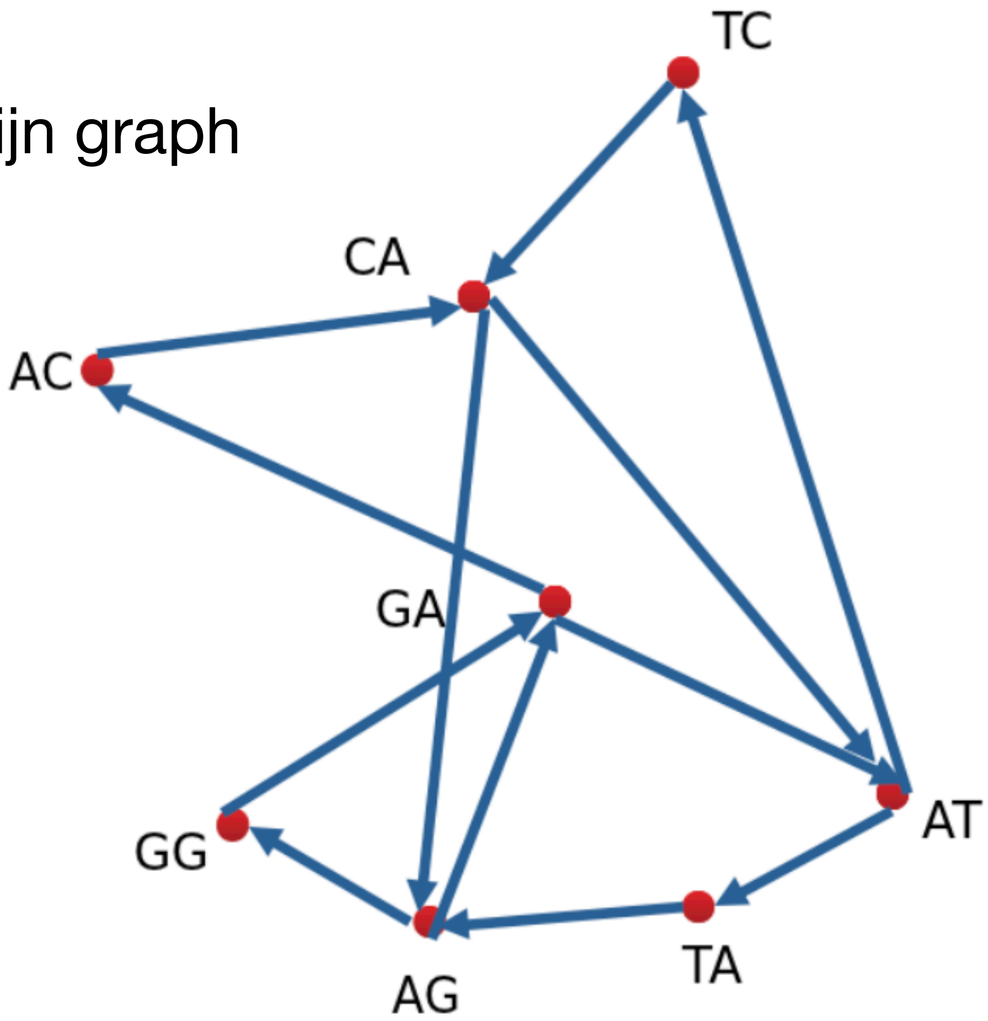
- Ha !



Standard de Bruijn graph (example)

Introduction

- Construction of de Bruijn graph



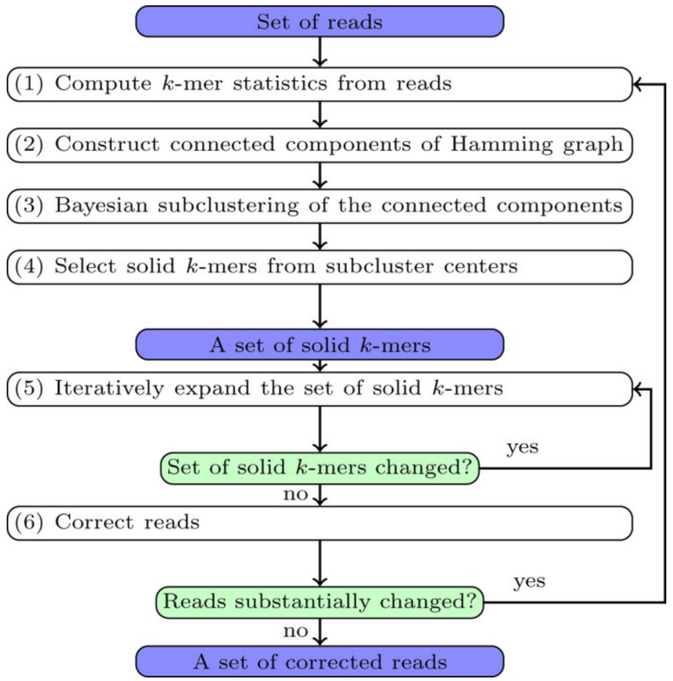
Assembly in SPAdes

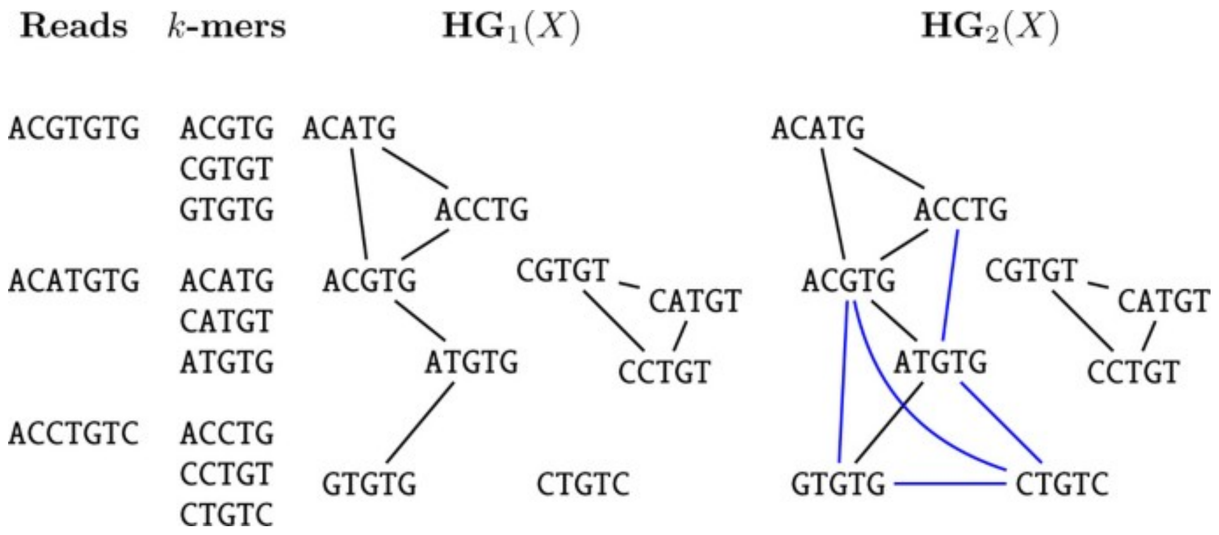
- Error correction
- Four stages in SPAdes

BayesHammer

Assembly in SPAdes

- Difficult error correction due to coverage evenness issue
- **BayesHammer**: Hamming graph + sub-clustering





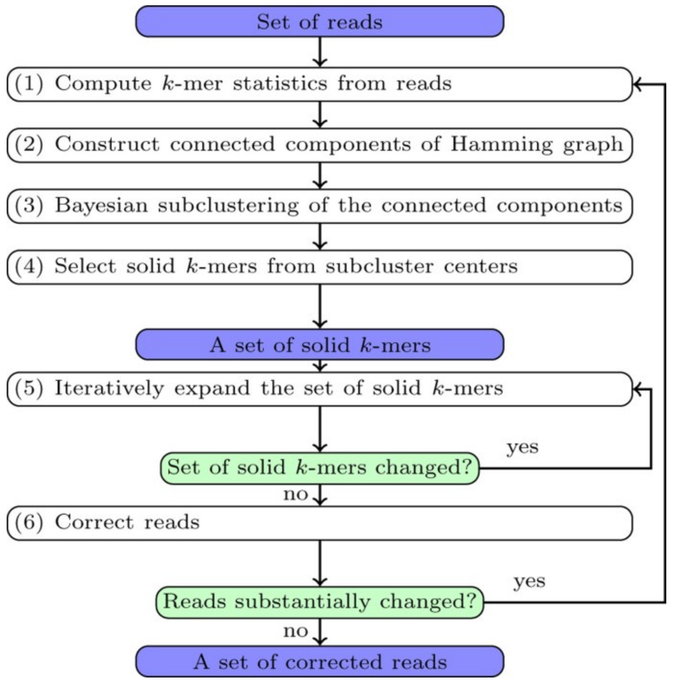
HG : Hamming Graph, black line : HG1, blue line : HG2

Image from Nikolenko *et al., BMC Genomics.* 2013

BayesHammer

Assembly in SPAdes

- Difficult error correction due to coverage evenness issue
- **BayesHammer**: Hamming graph + sub-clustering



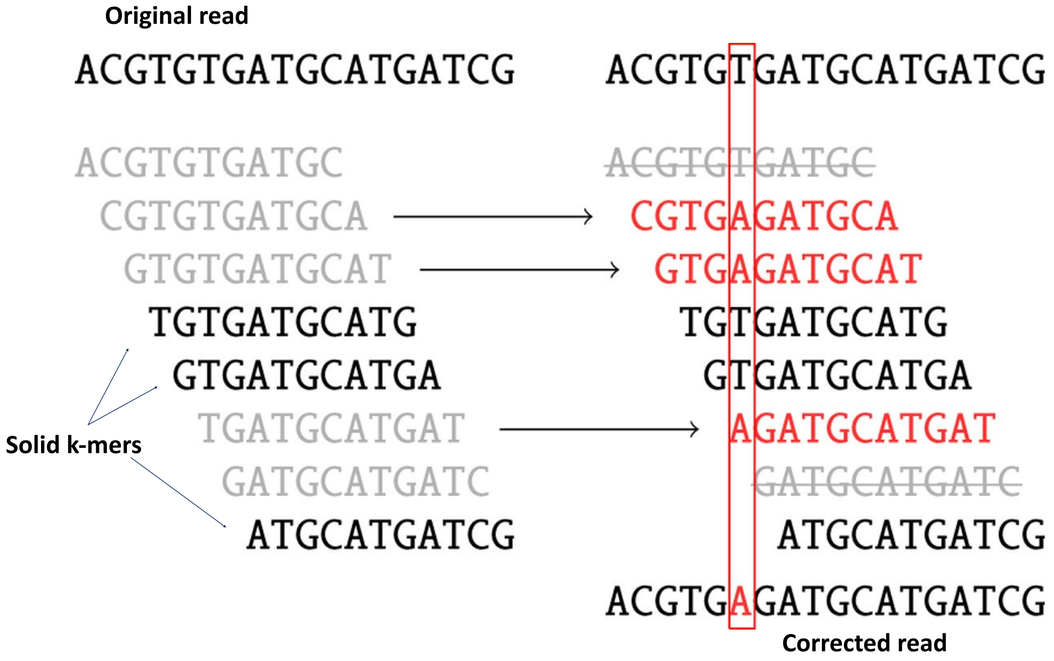


Image from Nikolenko *et al., BMC Genomics.* 2013

Overview

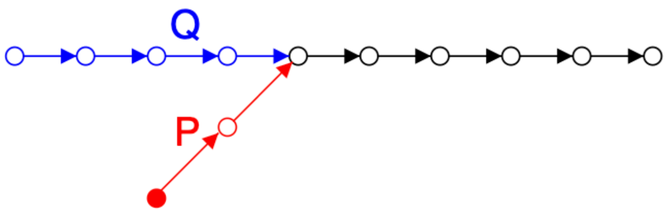
Assembly in SPAdes

Stage I	Assembly graph construction
Stage II	k-bimer adjustment
Stage III	Paired assembly graph construction
Stage IV	Contig construction

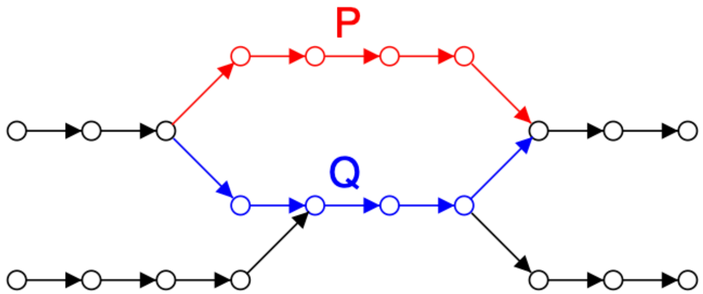
Stage I : Assembly Graph Construction

Assembly in SPAdes

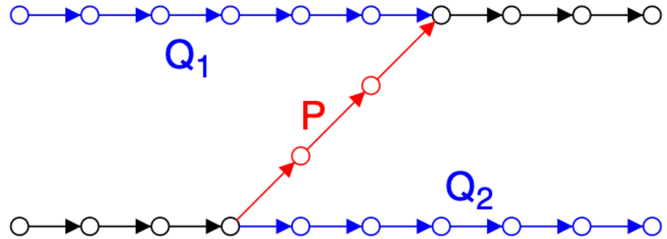
- No error correction? Then, no reliable graph you get.
- Some representative errors on reads



tip



bulge

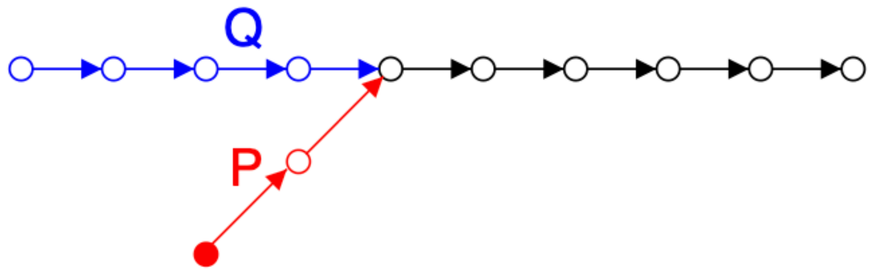


chimeric reads

Stage I : Assembly Graph Construction

Assembly in SPAdes

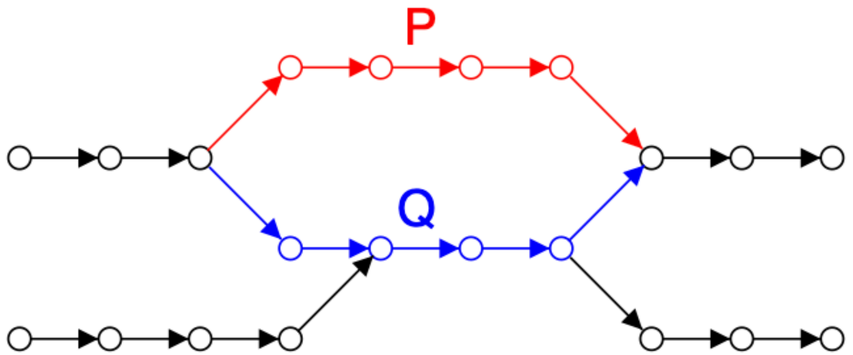
- *tip* correction
- **MCS**: tip clipping (based on coverage)
- **SCS**: gap closing



Stage I : Assembly Graph Construction

Assembly in SPAdes

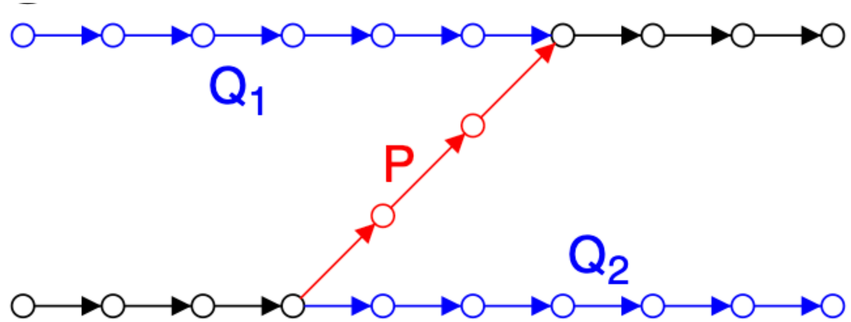
- *bulge* correction
- **SCS**: cannot directly remove (bulge corremoval)



Stage I : Assembly Graph Construction

Assembly in SPAdes

- *chimeric connection* correction
- **MCS**: this not commonly occur
- **SCS**: (gradual chimeric h-path removal)



ACGCTGATCCGATTA

CGCTGATCCGATTAG

CGCTGCCCTAGCATCGGACG

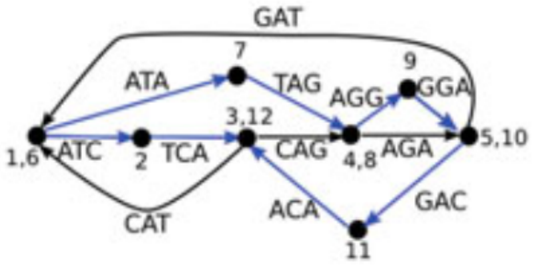
CCCTAGCATCGGAC

CCTAGCATCGGACG

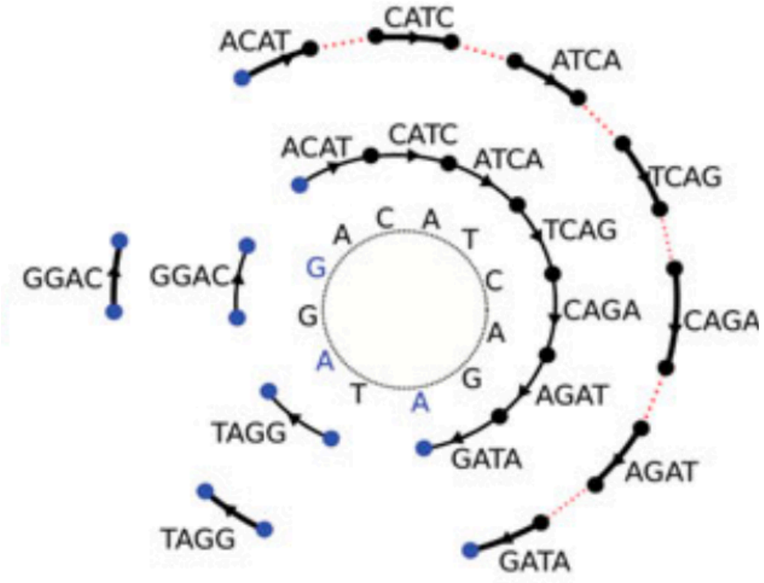
Multisized de Bruijn graph

Introduction

- k greatly affects quality of single-cell assembly



Small



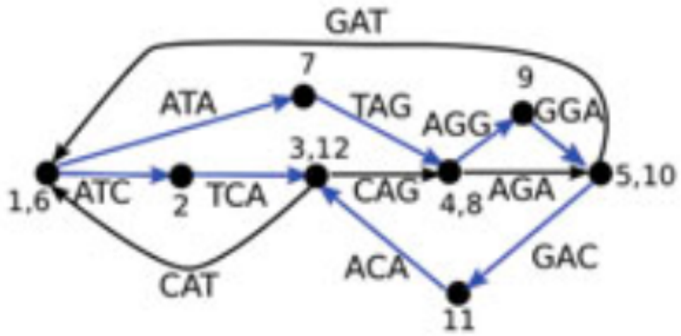
Large

(this paper)

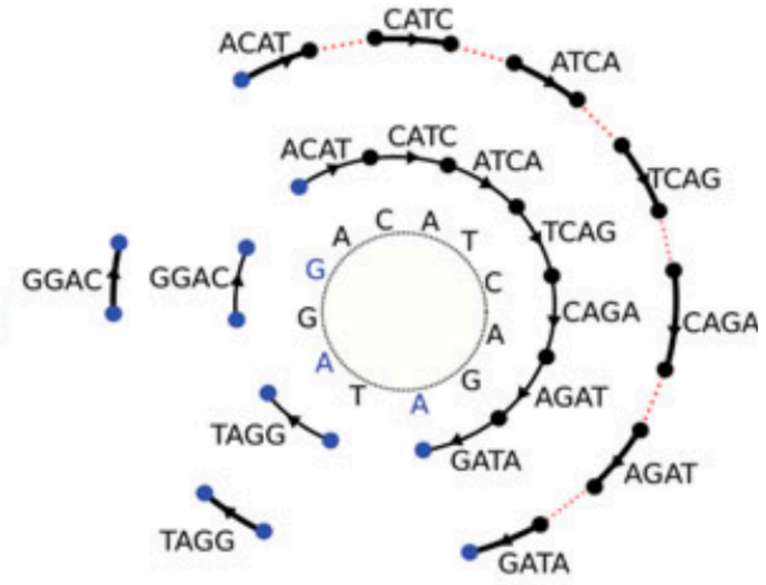
Multisized de Bruijn graph

Introduction

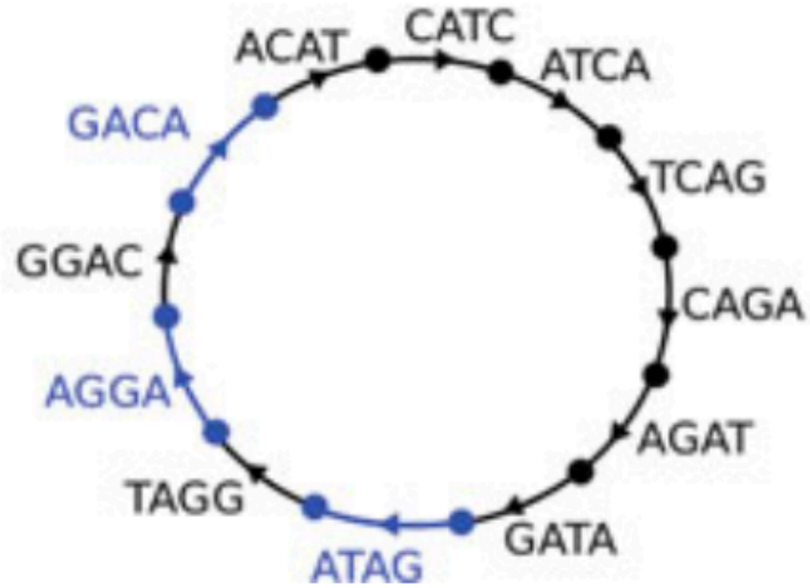
- k greatly affects quality of single-cell assembly



Small



Large

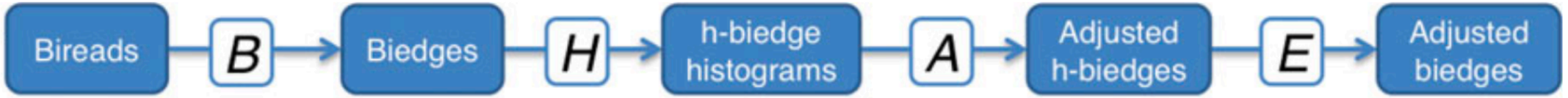


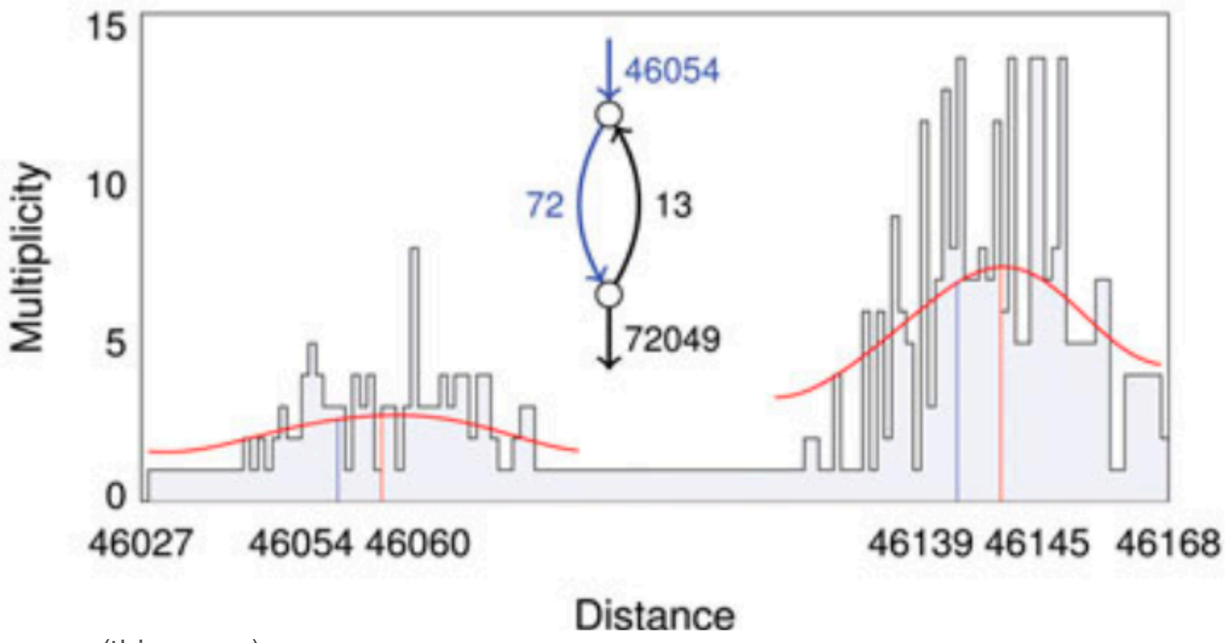
Multisized

(this paper)

Stage II : k-bimer adjustment

Assembly in SPAdes





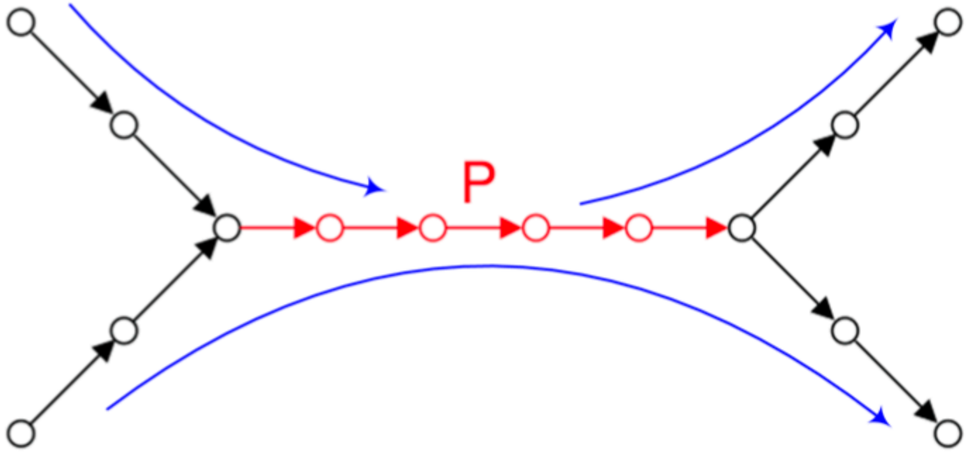
(this paper)

- Break reads into k-mers to study the set of k-mers
- Estimate the distance by clustering

Stage III : Paired Assembly Graph

Assembly in SPAdes

- Repeats result in paths with multiple entrances and multiple exits

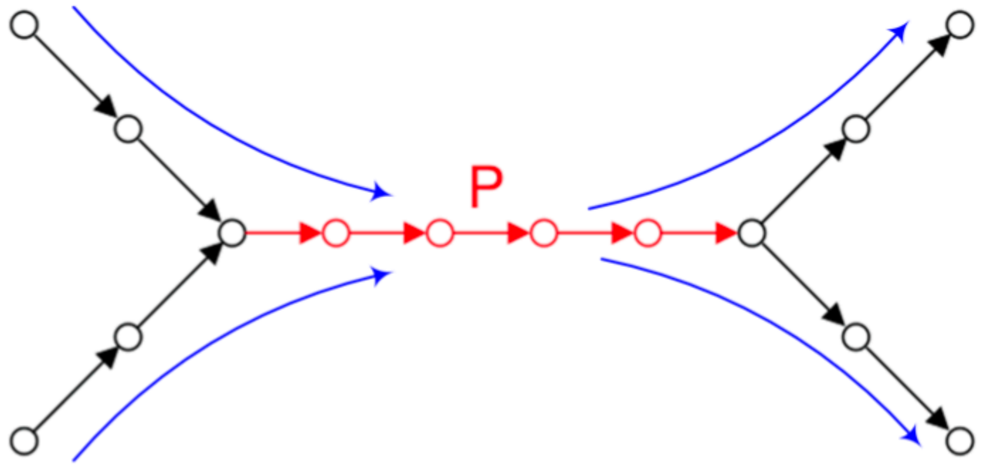


- Paired de Bruijn graph whose vertices correspond to pairs of k-mers

Stage III : Paired Assembly Graph

Assembly in SPAdes

- Repeats result in paths with multiple entrances and multiple exits

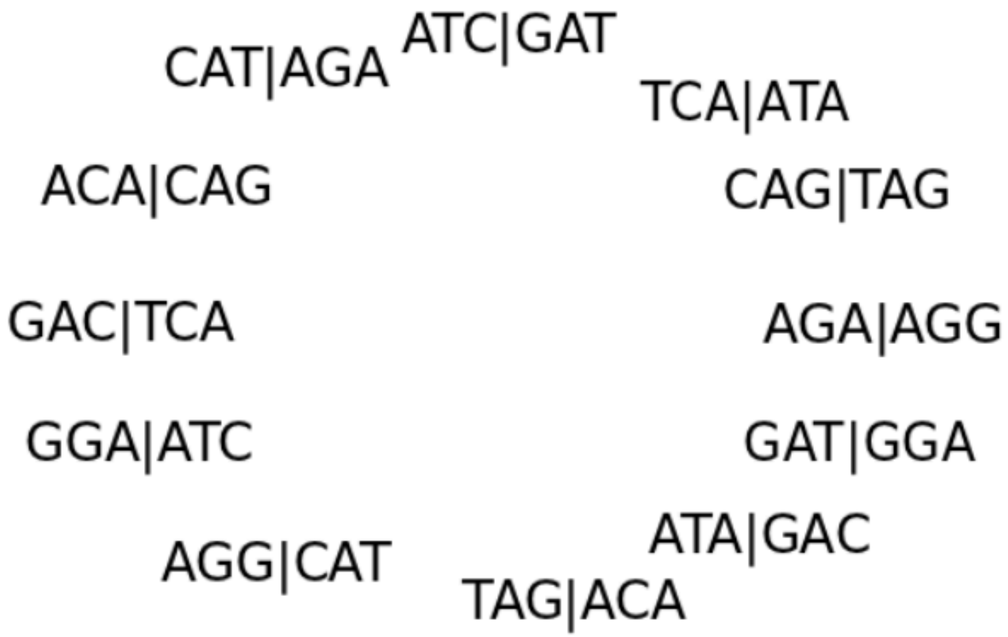


- Paired de Bruijn graph whose vertices correspond to pairs of k-mers

Stage III : Paired Assembly Graph

Assembly in SPAdes

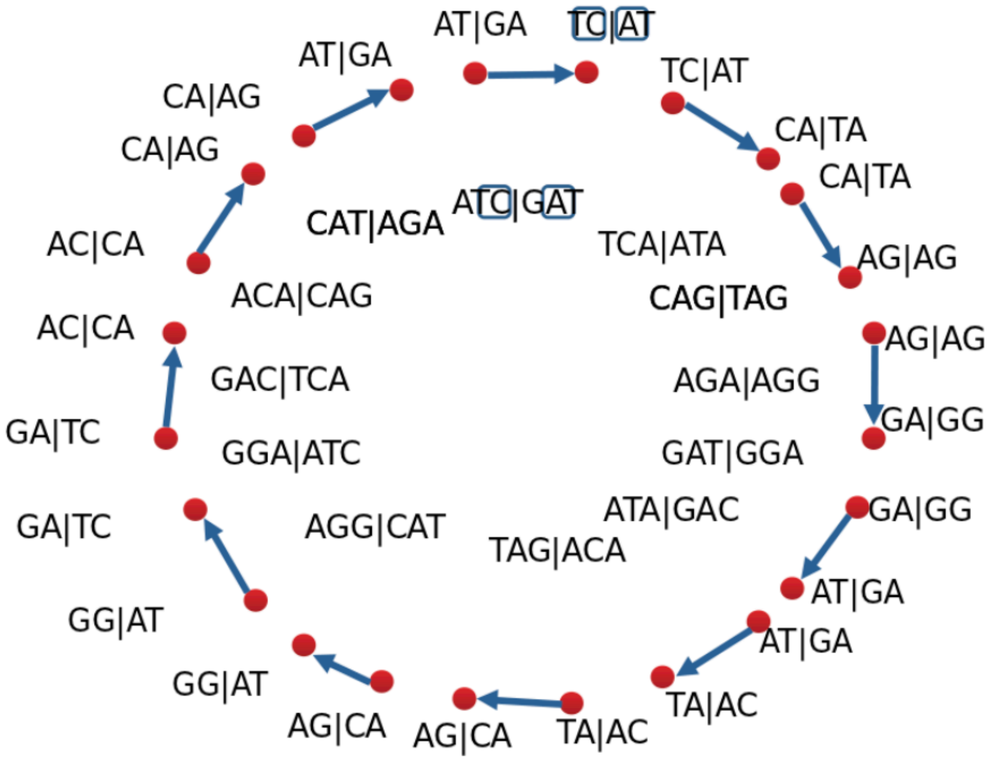
- Paired de Bruijn graph whose vertices correspond to pairs of k-mers



Stage III : Paired Assembly Graph

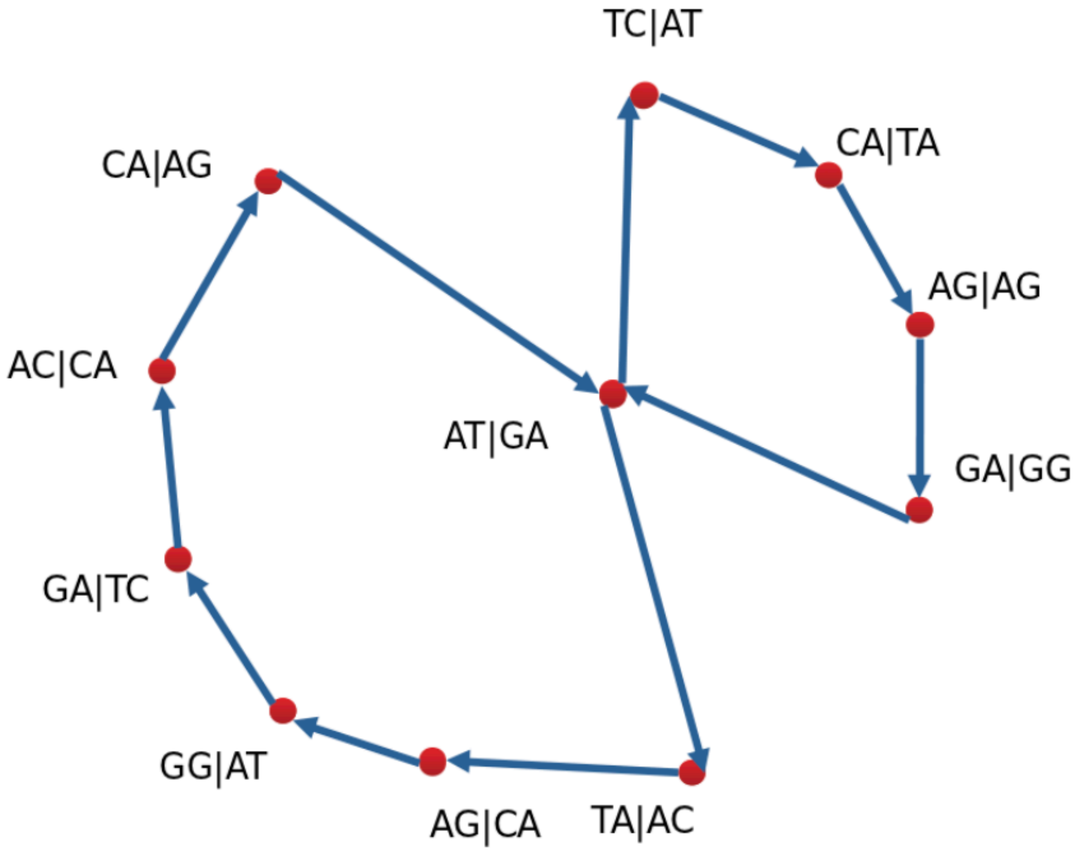
Assembly in SPAdes

- Key idea is the same with standard DBG

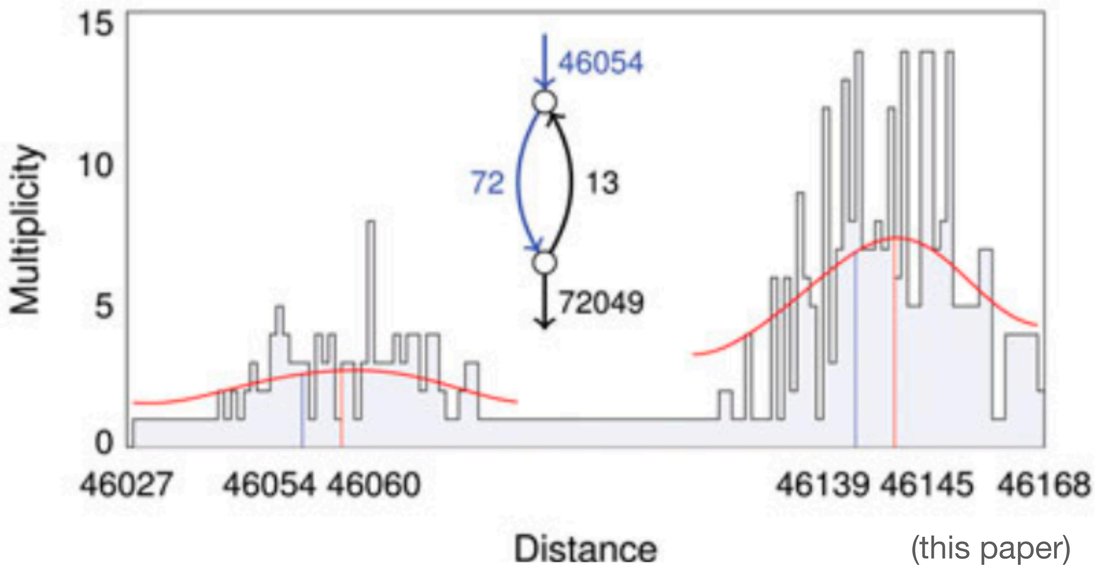


Stage III : Paired Assembly Graph

Assembly in SPAdes



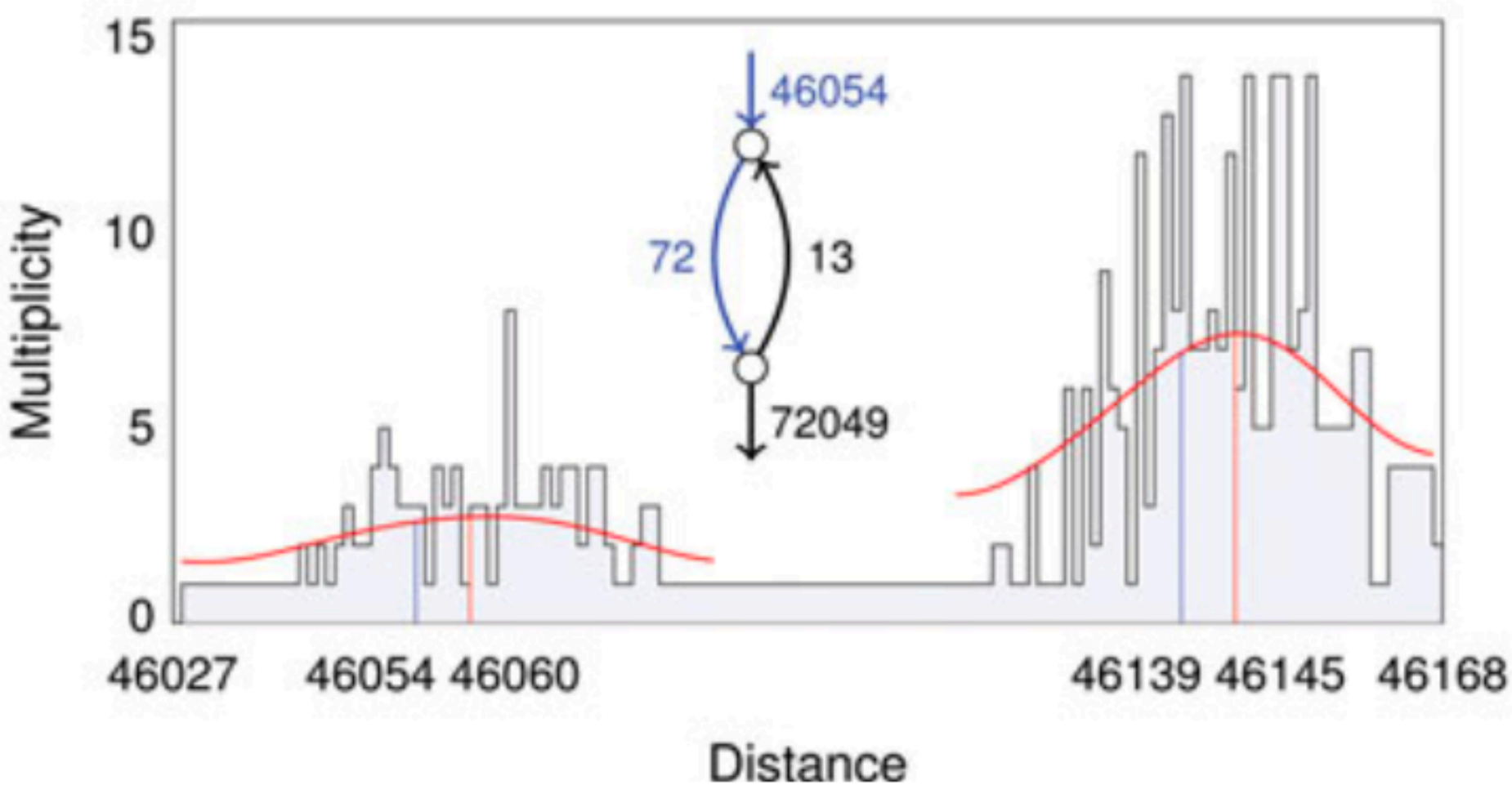
- Repeats handled by combination of paired DBG and distance estimates



(this paper)

Stage III : Paired Assembly Graph

Assembly in SPAdes

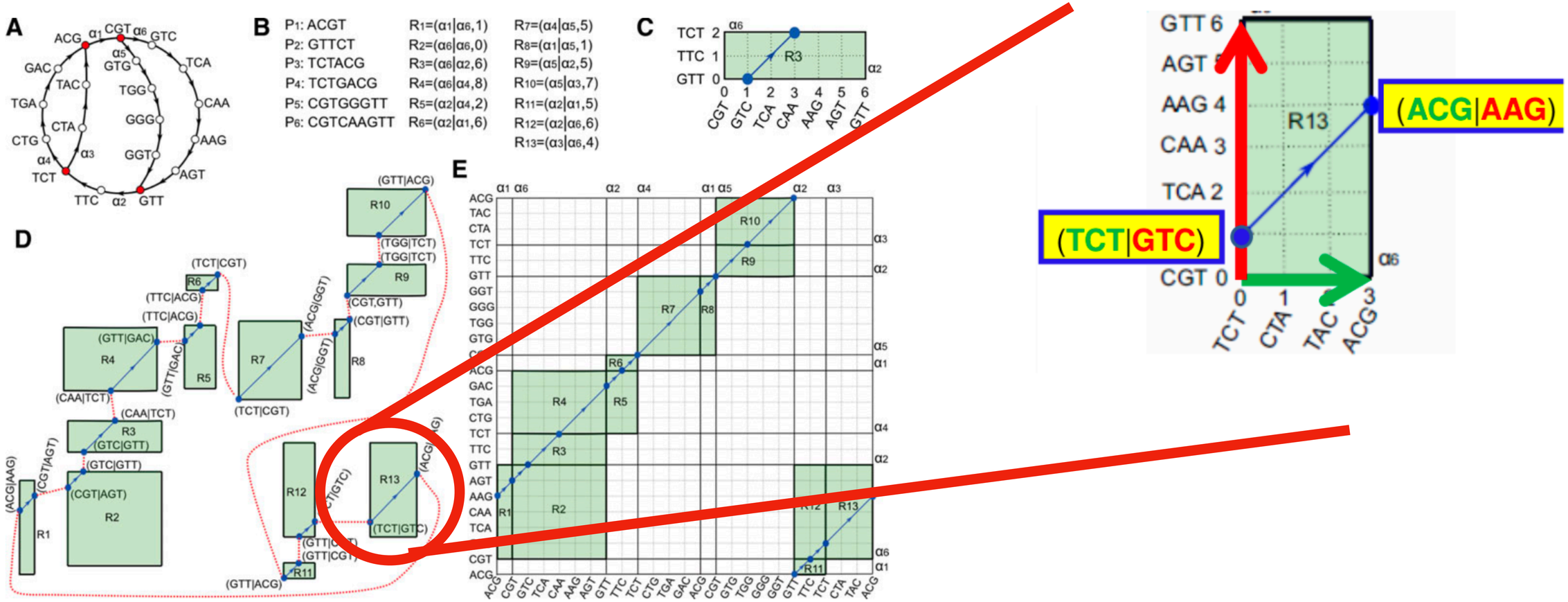


(this paper)

Stage III : Paired Assembly Graph

Assembly in SPAdes

- Rectangle graph

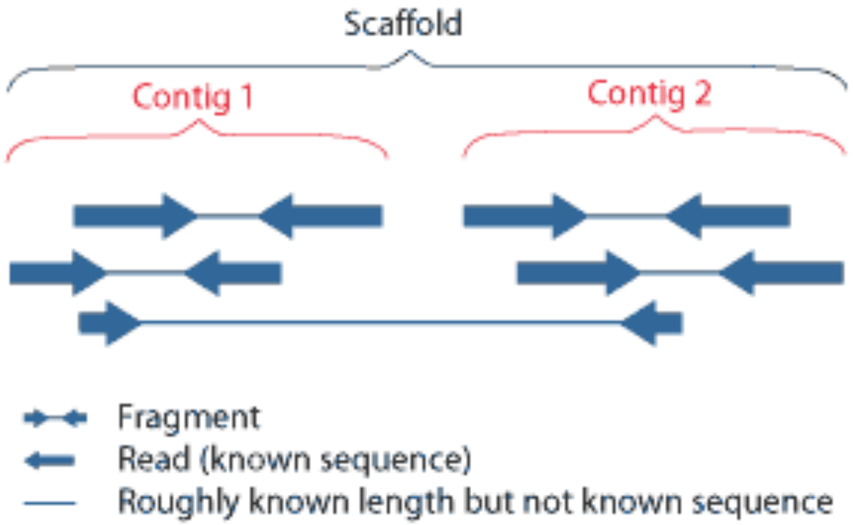


(this paper)

Stage IV : Contig Construction

Assembly in SPAdes

- Well studied in the context of Sanger sequencing



- Backtracking edges relocated during graph simplification

Results

- Properties
- Benchmarks

Properties

Results

```
spades.py [options] -o <output_dir>
```

- Input data

-SPAdes takes as input paired-end reads, mate-pairs and single (unpaired) reads in FASTA and FASTQ.

-For IonTorrent data SPAdes also supports unpaired reads in unmapped BAM format

-In order to run read error correction, reads should be in FASTQ or BAM format.

-The following types of libraries are requested.

  -Illumina or IonTorrent paired-end/high-quality mate-pairs/unpaired reads

  -PacBio CCS reads

-Specifying data for hybrid assembly

  -PacBio CLR reads for CCS reads assembly

  -Oxford Nanopore reads and Sanger reads

Properties

Results

```
spades.py [options] -o <output_dir>
```

- Output

-All output files are saved in the output_dir path specified by the user.

-The reads corrected by BayesHammer -> <output_dir>/corrected/*.fastq.gz

-The final scaffolds (recommended for use as resulting sequences) -> <output_dir>/scaffolds.fasta

-The final contigs -> <output_dir>/contigs.fasta

-Assembly graph and scaffolds paths in GFA 1.0 format -> <output_dir>/assembly_graph_with_scaffolds.gfa

-Assembly graph in FASTG format -> <output_dir>/assembly_graph.fastg

Benchmark

Results

Assembler[a]	*# contigs*	*N50 (bp)*	*Largest (bp)*[b]	*Total (bp)*[c]	*Covered (%)*[d]	*MA*[e]	*MM*[f]	*CG*[g]
Normal multicell sample of *E. coli* (ECOLI-MC)								
EULER-SR	295	**110153**	221409	4598020	99.5	10	5.2	4232
IDBA	**191**	50818	164392	4566786	99.5	4	1.0	4201
SOAPdenovo	192	62512	172567	4529677	97.7	1	26.1	4141
Velvet	198	78602	196677	4570131	**99.9**	4	1.2	4223
Velvet-SC	350	52522	166115	4571760	**99.9**	**0**	1.3	4165
E+V-SC	339	54856	166115	4571406	**99.9**	**0**	2.9	4172
SPAdes-single	445	59666	166117	4578486	**99.9**	**0**	**0.7**	4246
SPAdes	195	86590	**222950**	4608505	**99.9**	2	3.7	**4268**

(this paper)

Benchmark

Results

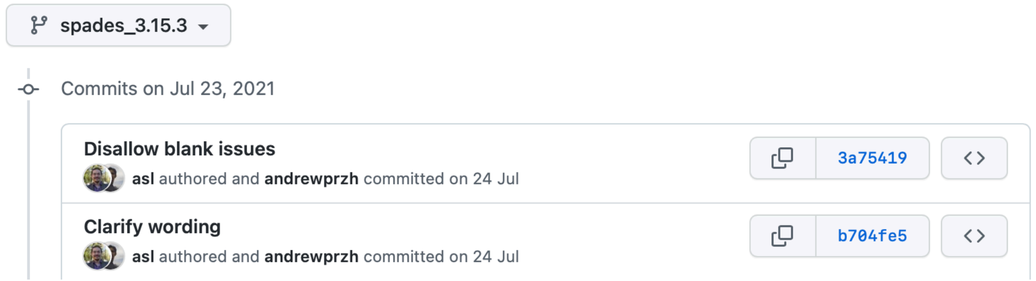
Assembler[a]	*# contigs*	*N50 (bp)*	*Largest (bp)*[b]	*Total (bp)*[c]	*Covered (%)*[d]	*MA*[e]	*MM*[f]	*CG*[g]
Single-cell *E. coli* (ECOLI-SC)								
EULER-SR	1344	26662	126616	4369634	87.8	21	11.0	3457
SOAPdenovo	1240	18468	87533	4237595	82.5	13	99.5	3059
Velvet	**428**	22648	132865	3533351	75.8	2	**1.9**	3117
Velvet-SC	872	19791	121367	4589603	93.8	2	**1.9**	3654
E+ V−SC	501	32051	132865	4570583	93.8	2	6.7	3809
SPAdes-single	1164	42492	166117	4781576	**96.1**	**1**	6.2	3888
SPAdes	1024	**49623**	**177944**	4790509	**96.1**	**1**	5.2	**3911**

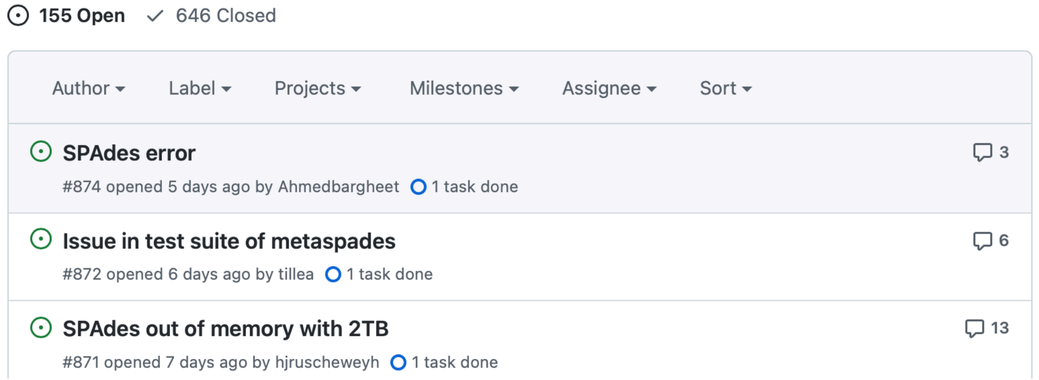
(this paper)

Recent Updates

Results

- Active community with feedback





- Variations for specific purpose

Colab