Genome assembly strategies

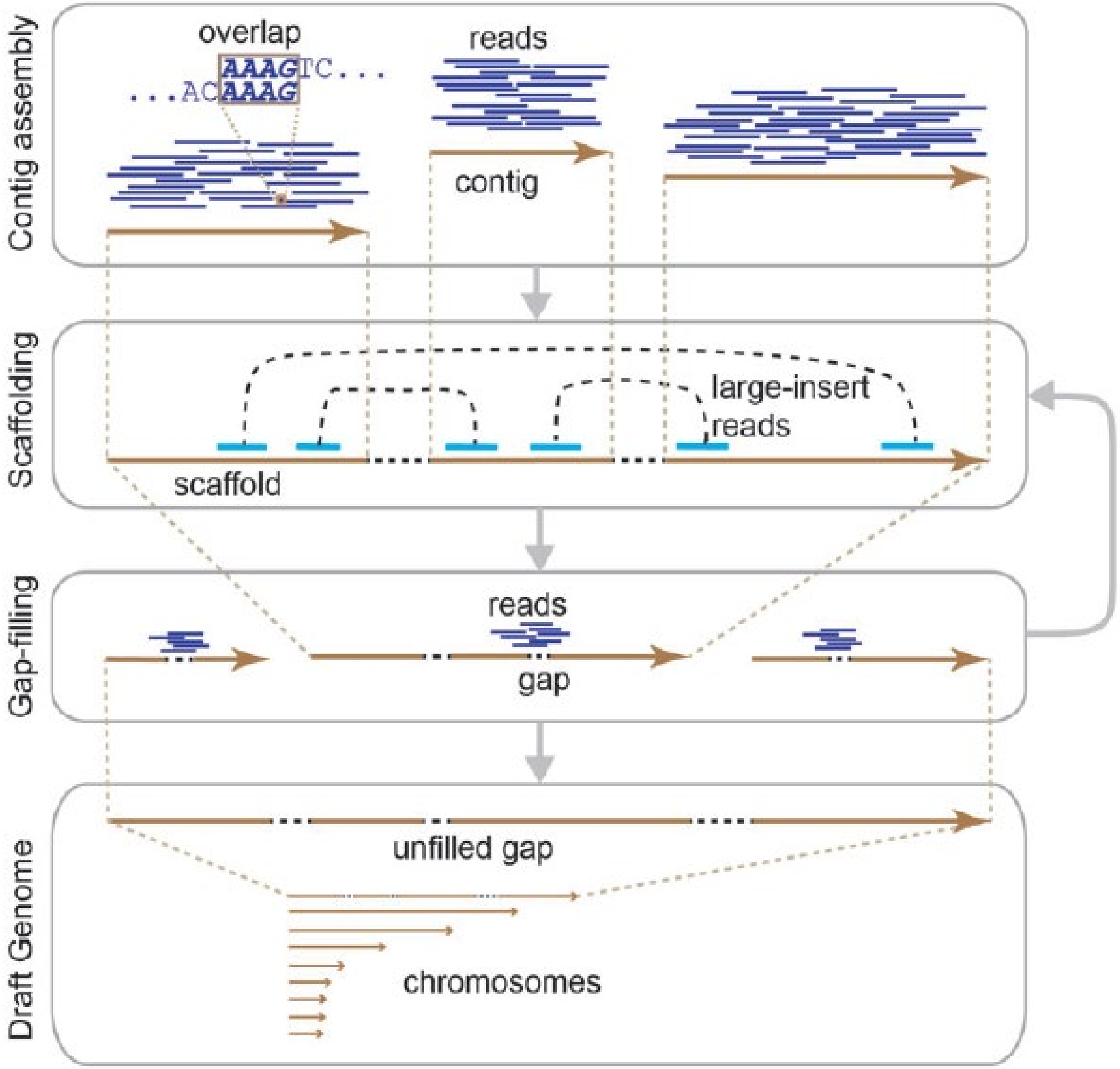
Arturo Vera Ponce de Leon

May 2019

veraponcedeleon.1@osu.edu

Genome assembly

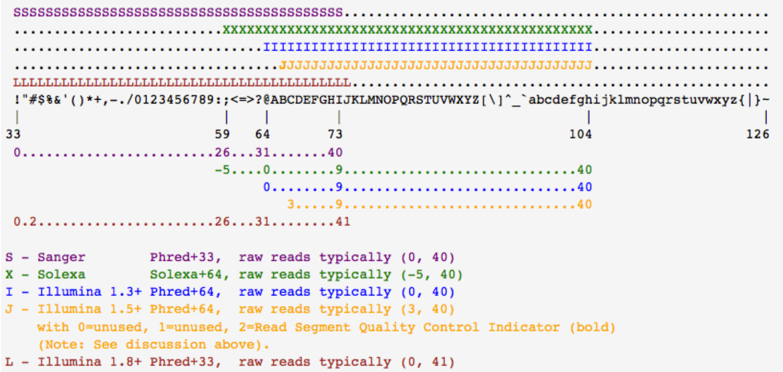
Genome assembly



Briefings in Bioinformatics, 2016, 1–18

Reads (fastq)

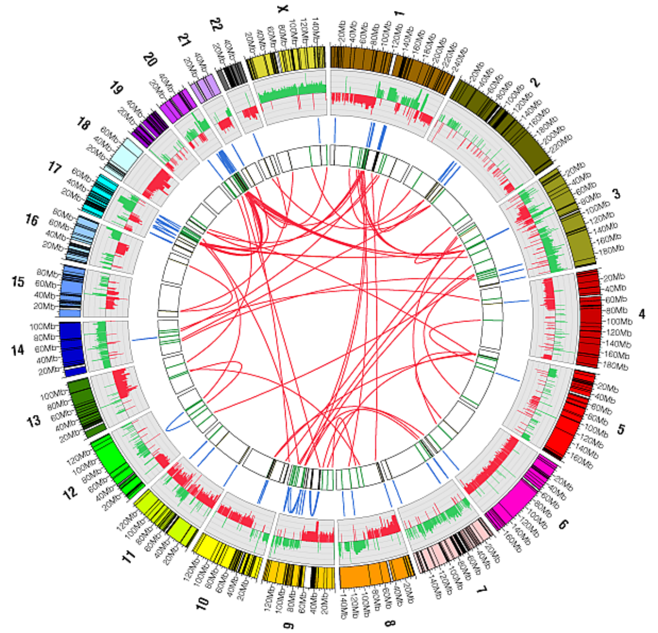
Summarizing



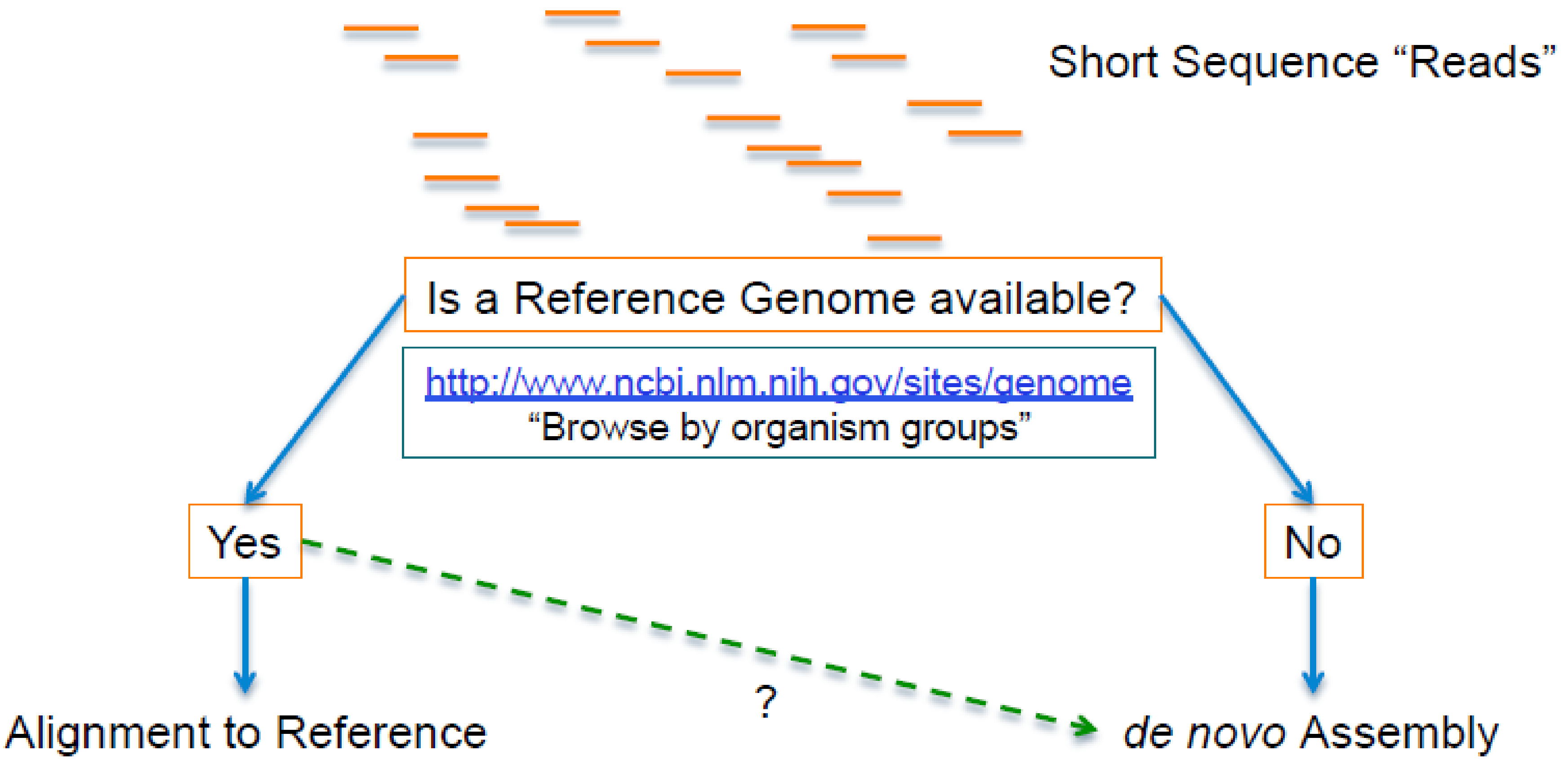
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Chromosome



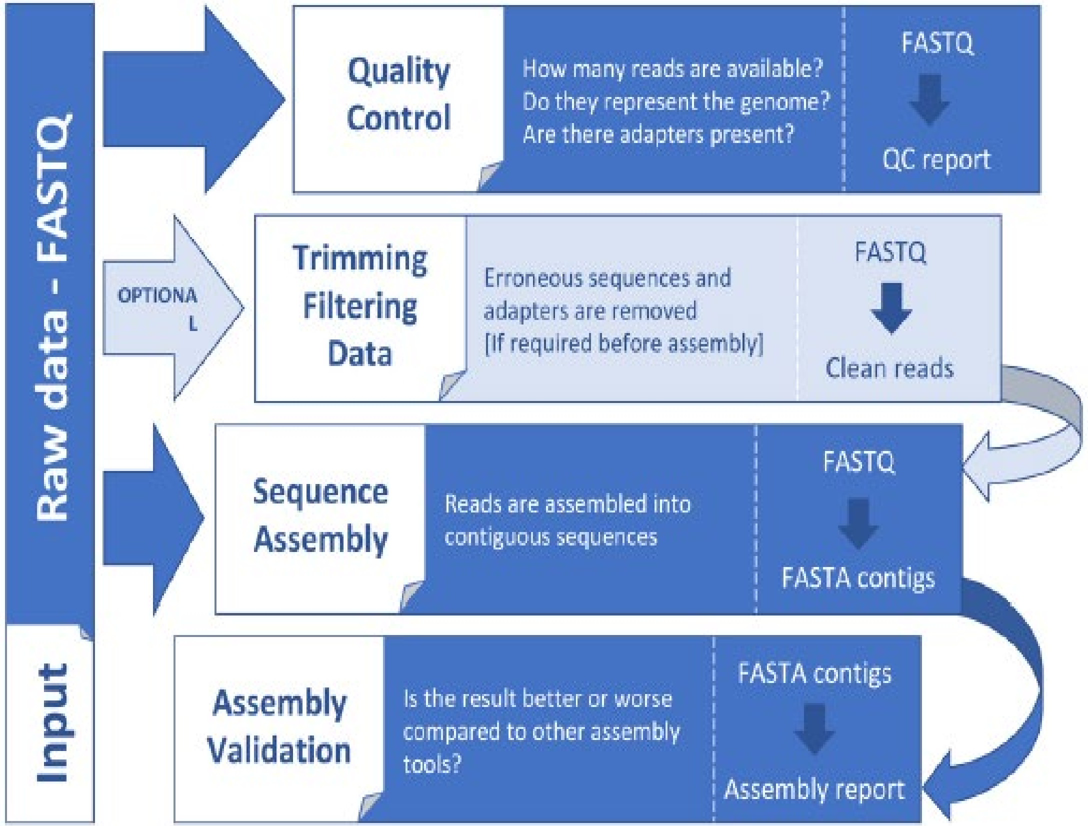
Genome *de-novo* assembly or reference mapping



Short Sequence "Reads"
Is a Reference Genome available?
http://www.ncbi.nlm.nih.gov/sites/genome
"Browse by organism groups"
Yes
No
?
Alignment to Reference
de novo Assembly

Genome Assembly

Work flow to classic genome assembly strategies



How many reads are available?
Do they represent the genome?
Are there adapters present?

Erroneous sequences and adapters are removed [If required before assembly]

Reads are assembled into contiguous sequences

Is the result better or worse compared to other assembly tools?

Figure 2. General steps in a genome assembly workflow. Input and output data are indicated for each step.

F1000Research 2018, 7(ELIXIR):148

Software used in this lecture

fastQC

TrimGalore

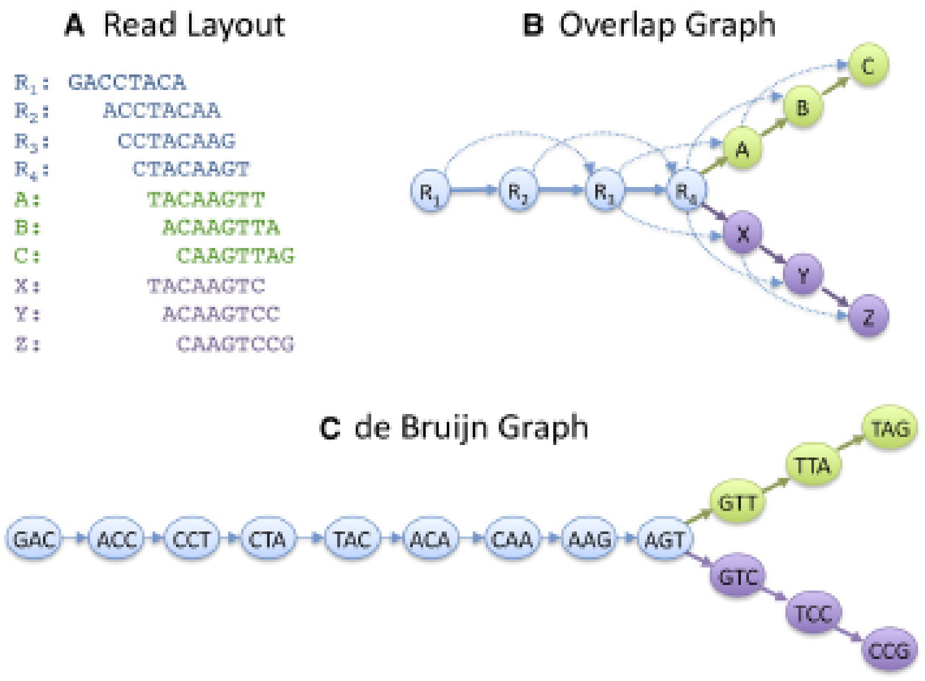
IDBA
SPADES

QUAST
BUSCO
CheckM

Strategies to genome assembly

- Algorithms

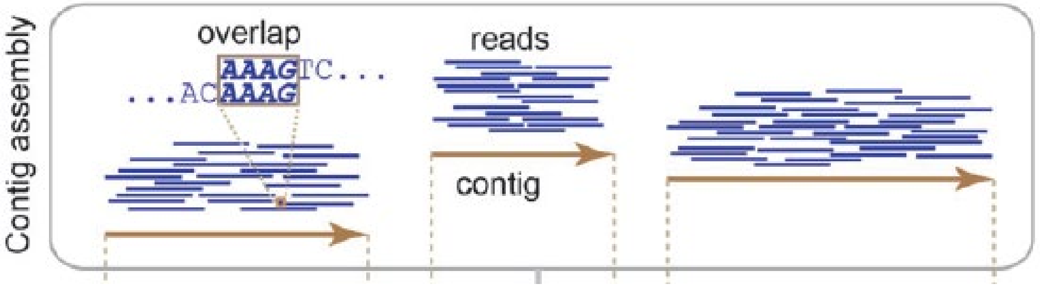
1. Greedy
2. Overlap-layout-consensus (OLC)
3. De Bruijn Graph

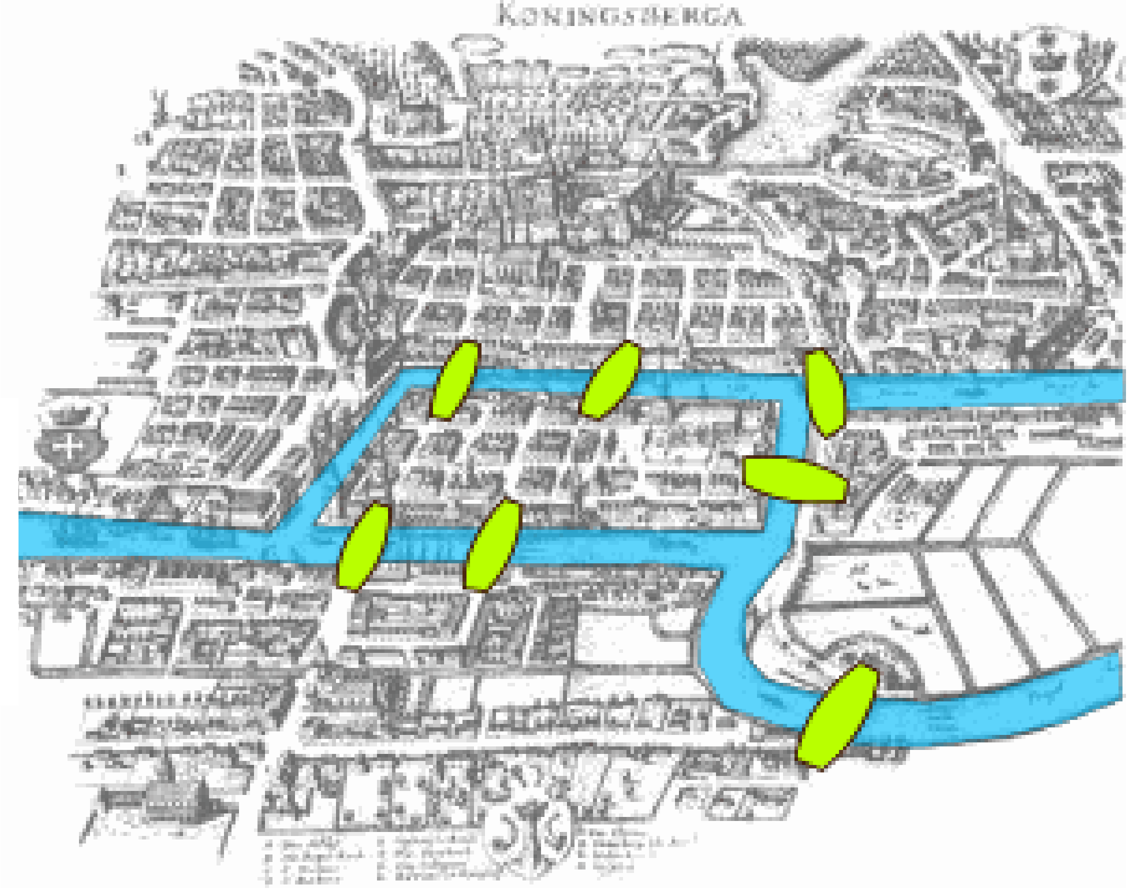


Schatz M C et al. Genome Res. 2010;20:1165-1173

Steeps to genome assembly using De-Bruijn graph

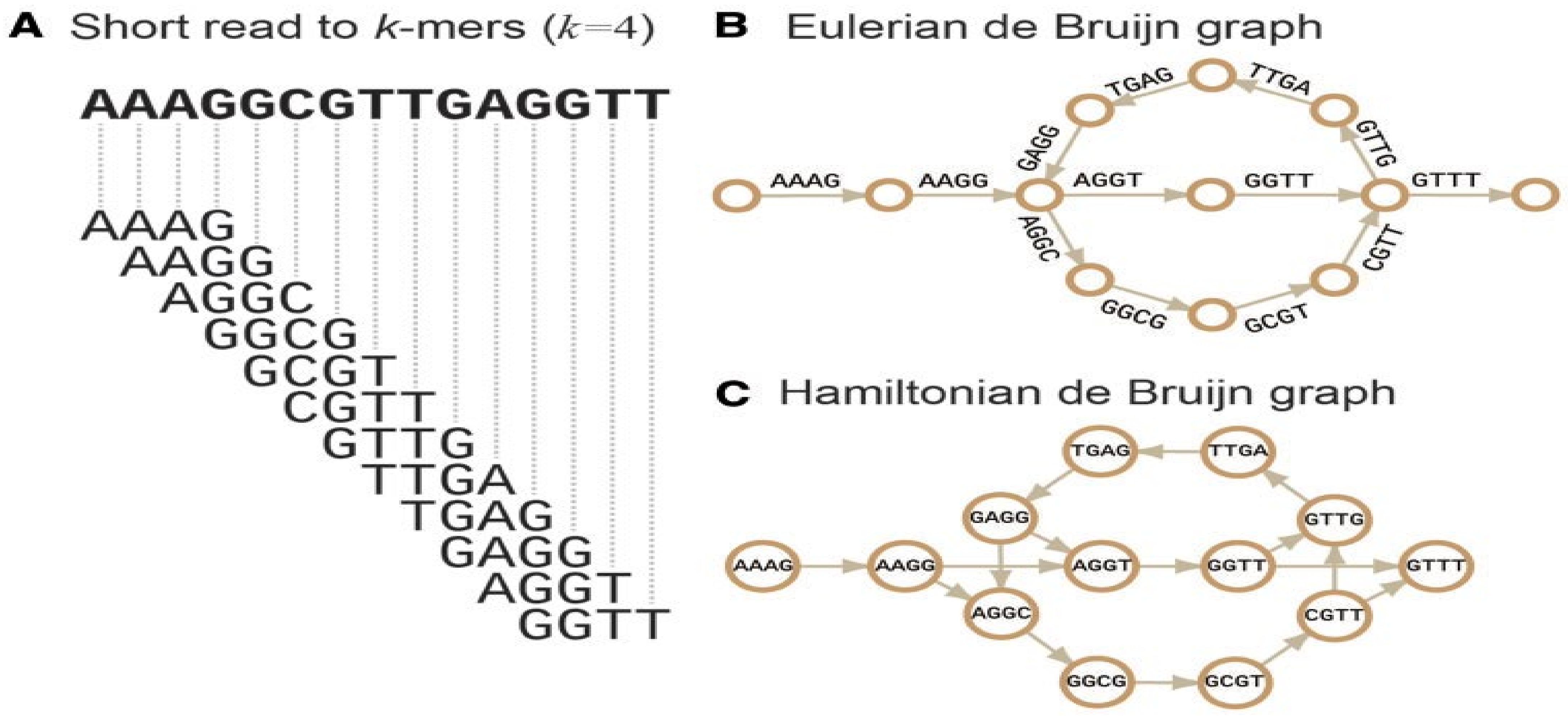
The basic strategy for de novo assembly for short NGS reads comprises three steps: (i) contig assembly, (ii) scaffolding and (iii) gap filling.





Seven Bridges of Königsberg

The K-mers: divide and conquer



The K-mers: divide and conquer

- It breaks reads into successive k-mers and the graph maps the k-mers

- Each k-mer is a node and edges are drawn between each k-mer in a read.

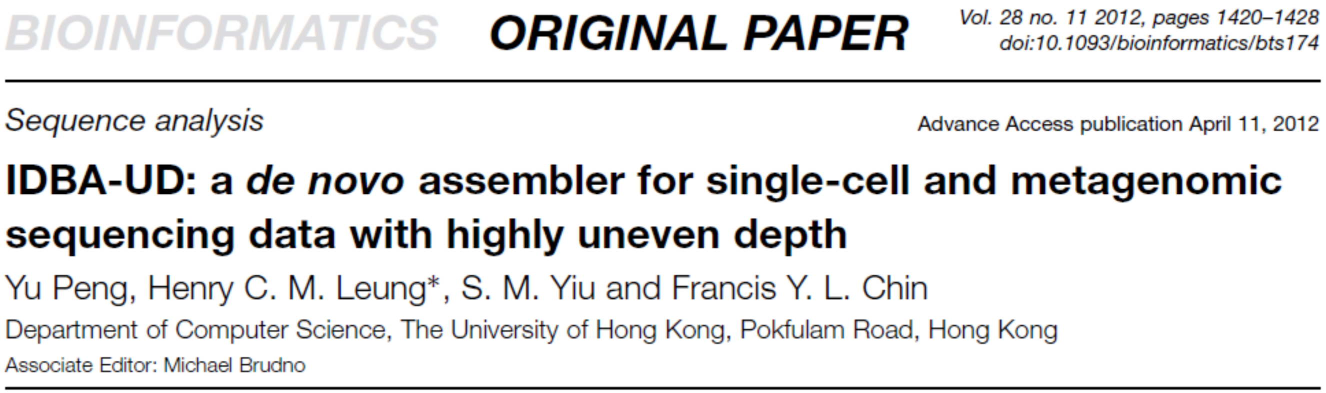
- Repeat sequences create a fork in the graph; alternative sequences create a bubble.

- The k-mer size can only be determined by “trial and error”.

- A small value of K will create a complex graph but a large value of K may miss small overlaps. A good starting point would be a k-mer size that is 2/3 the size of the read

- Good for short reads or small genomes. With long reads and/or large genomes, may require lots of RAM (e.g., ~0.5 TB for human)

Let´s go to assembly some bacterial genomes



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 28 no. 11 2012, pages 1420–1428
doi:10.1093/bioinformatics/bts174

Sequence analysis

Advance Access publication April 11, 2012

IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth

Yu Peng, Henry C. M. Leung*, S. M. Yiu and Francis Y. L. Chin

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong

Associate Editor: Michael Brudno

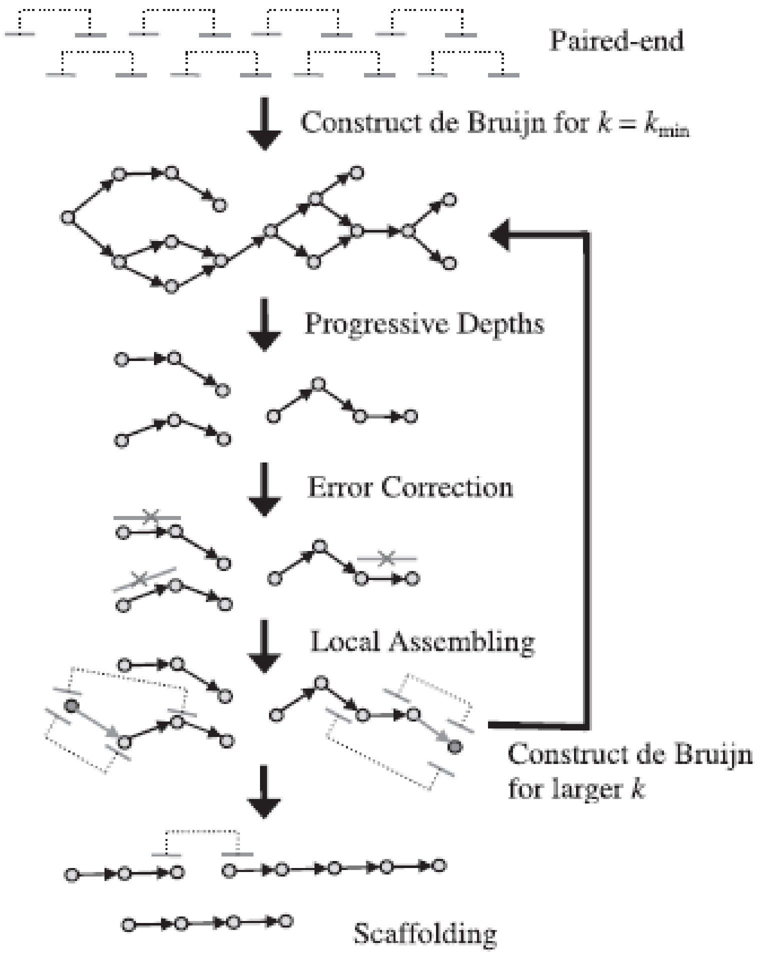

Paired-end
Construct de Bruijn for $k = k_{\min}$
Progressive Depths
Error Correction
Local Assembling
Construct de Bruijn for larger k
Scaffolding

Fig. 1. Flowchart of IDBA-UD

Evaluating the assembly

§ **Genome assembly results:**

- **contig size and number of contigs produced**
- **scaffold size and number**
- **N50 and N90**

§ **Coverage**

§ **GC Content**

§ **Genome annotation**

- repeats analysis and annotation
- protein-coding gene annotation (including gene structure prediction and gene function annotation)
- non-coding RNA gene annotation (including annotation of microRNA, tRNA, rRNA, and other ncRNA)
- transposon and tandem repeats annotation

§ Comparative genomics and evolution (chromosome structure, conserved gene families)

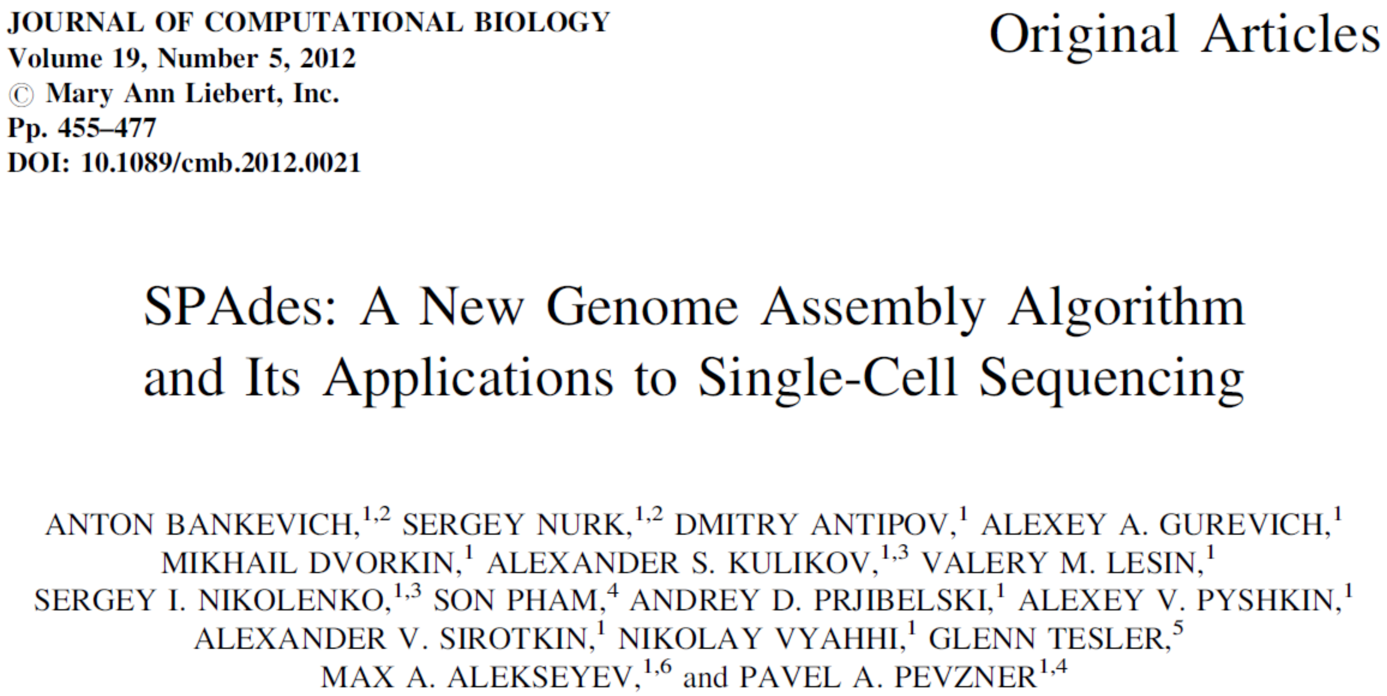
Basic stats

Basic statistics
N50 the length of the shortest contig such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs.

Contig size (bp)
3000
2000 N50
1200
800
600 N90
400
Total: **8000**

N90 = the length of the shortest contig such that the sum of contigs of equal length or longer is at least 90% of the total length of all contigs.

SPADEs



JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 5, 2012
© Mary Ann Liebert, Inc.
Pp. 455–477
DOI: 10.1089/cmb.2012.0021

Original Articles

SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing

ANTON BANKEVICH,[1,2] SERGEY NURK,[1,2] DMITRY ANTIPOV,[1] ALEXEY A. GUREVICH,[1]
MIKHAIL DVORKIN,[1] ALEXANDER S. KULIKOV,[1,3] VALERY M. LESIN,[1]
SERGEY I. NIKOLENKO,[1,3] SON PHAM,[4] ANDREY D. PRJIBELSKI,[1] ALEXEY V. PYSHKIN,[1]
ALEXANDER V. SIROTKIN,[1] NIKOLAY VYAHHI,[1] GLENN TESLER,[5]
MAX A. ALEKSEYEV,[1,6] and PAVEL A. PEVZNER[1,4]