Session 2.3 – Assembly

Sara Monzón Fernández

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

04-16 Noviembre 2019, 2ª Edición
Programa Formación Continua, ISCIII

Assembly

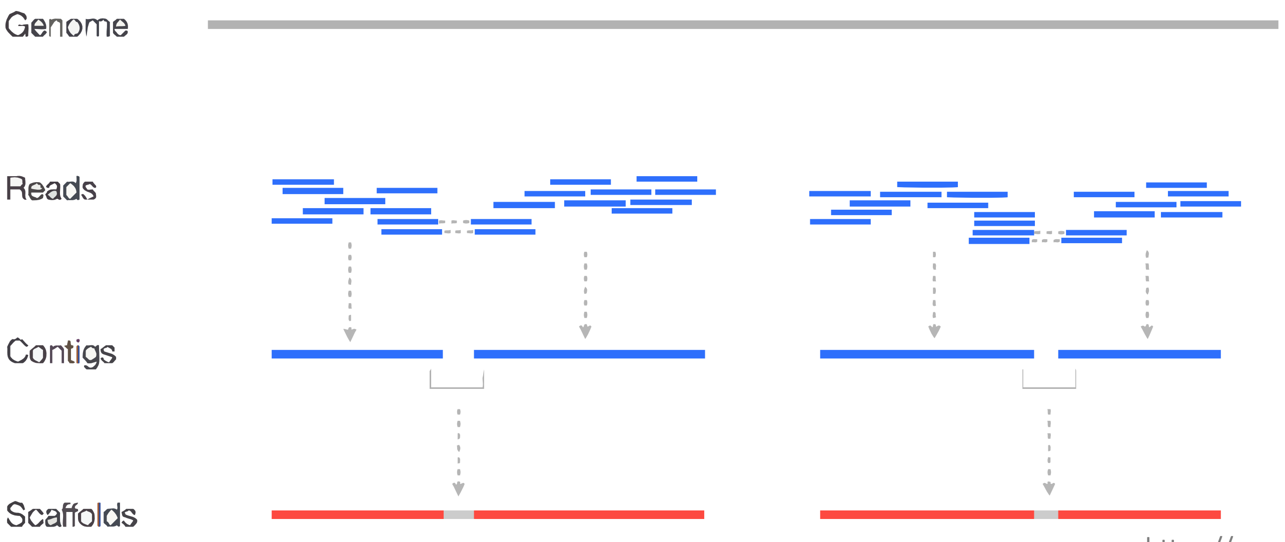
Reconstruct a representation of the original DNA from shorter DNA sequences or small fragments known as reads

- ***De novo:*** with no previous knowledge of the genome to be assembled. It overlap the end of the end of each read in order to créate a longer sequence.

- ***Assembly with reference:*** A similar but not identical genome guides the assembly process. Map reads over supplied genome.

Assembly: contig y scaffold

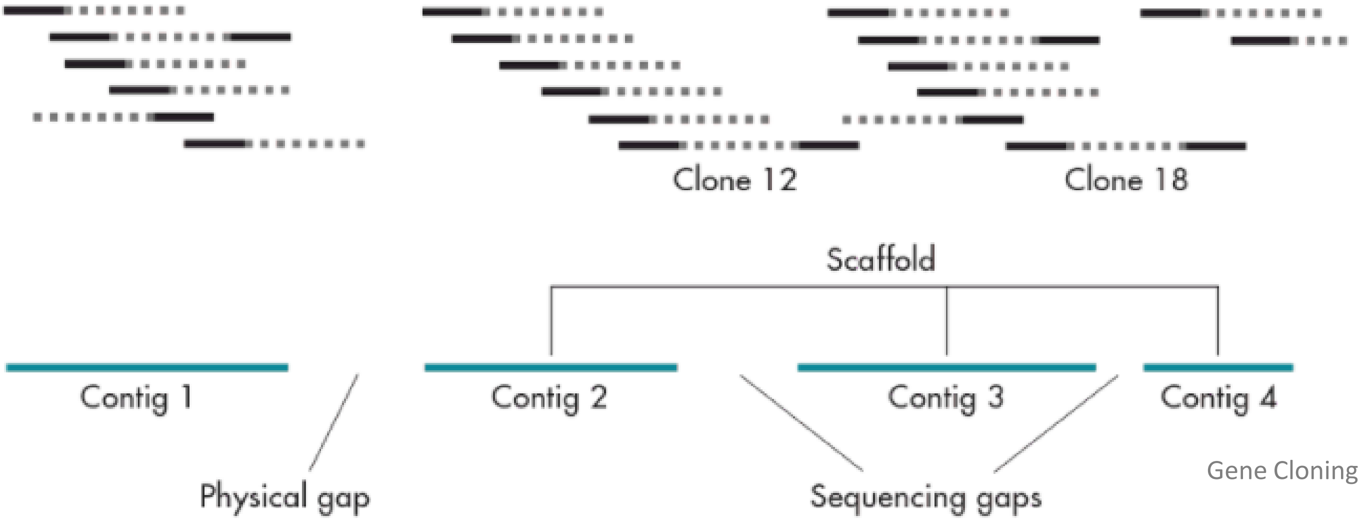
- **Contig:** continuous sequence made up of overlaping shorter sequences
- **Scaffold:** two or more contigs located and rearranged according to spatial information(pair-end, mate pair, reference)



https://www.biostars.org/p/253222/

Assembly: gaps

- **Sequencing gaps:** Position and orientation known by spatial information
- **Physical gaps:** No information about adjacent contigs

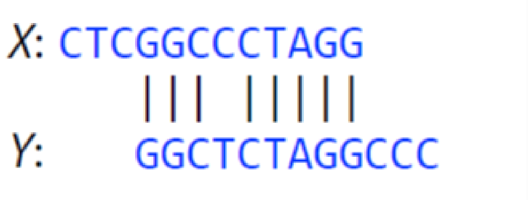


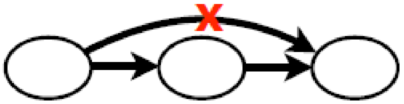
Gene Cloning, Lodge *et al*.

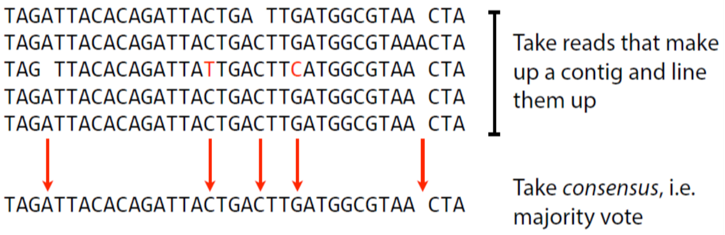
Assembly: Algorithms

- **Overlap, Layout, Consensus (OLC - overlap graph):**
  - O - first overlaps among all the reads are found
  - L - then it carries out a layout of all the reads and overlaps information on a graph
    - Removes redundant and low quality overlaps
  - C - and finally the consensus sequence is inferred

Ex. Newbler, Mira, Celera Assembler, CAP3, PCAP, Phrap, Phusion.







```
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
```

Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote

https://pt.slideshare.net/anton_alexandrov/combining-de-bruijn-graph-overlap-graph-and-microassembly/12?smtNoRedir=1

Assembly: Algorithms

- **De Brujin Graph (DBG: k-mer graph)**

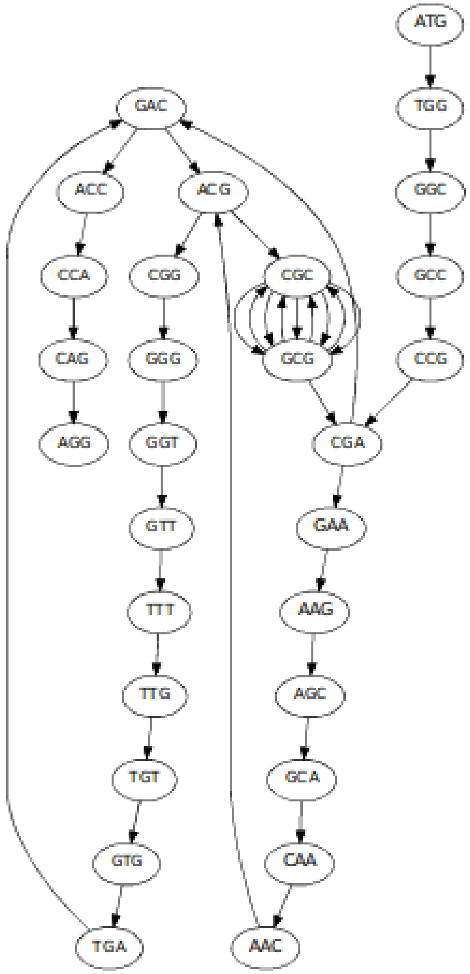
Chopping reads into much shorter k-mers (fixed length fragments) and then using all the k-mers to form a DBG and infer the contigs.

- Nodes in the graph are k-mers

- Edges represent consecutive k-mers (which overlap by k-n symbols)

Ex. SPAdes, ABySS, Velvet, AllPaths, Soap….

https://medium.com/@han_chen

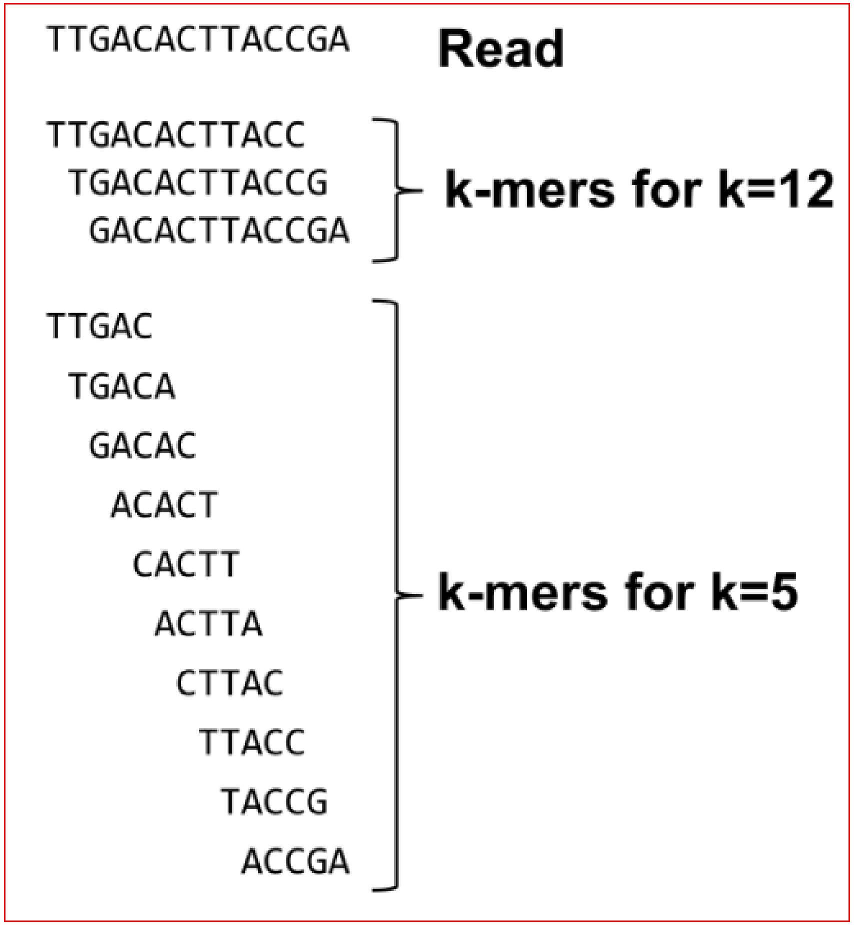


Algorithms: DBG

- **Why choosing DBG:**
  - Sequencing bias
  - Sequence errors
  - Sequence length
- **DBG Flaws:**
  - Millions of pieces
    - Much, much shorter than the genome
    - Lots of them look similar
  - Missing pieces
    - Some parts can't be sequenced easily
    - Dirty Pieces - Multiplex
    - Lots of errors in reads
  - Repeats
    - If they are longer than the read length
    - Causes nodes to be shared, locality confusion

https://galaxyproject.github.io/training-material/topics/assembly/tutorials/debruijn-graph-assembly/slides.html#23

Algorithms: DBG

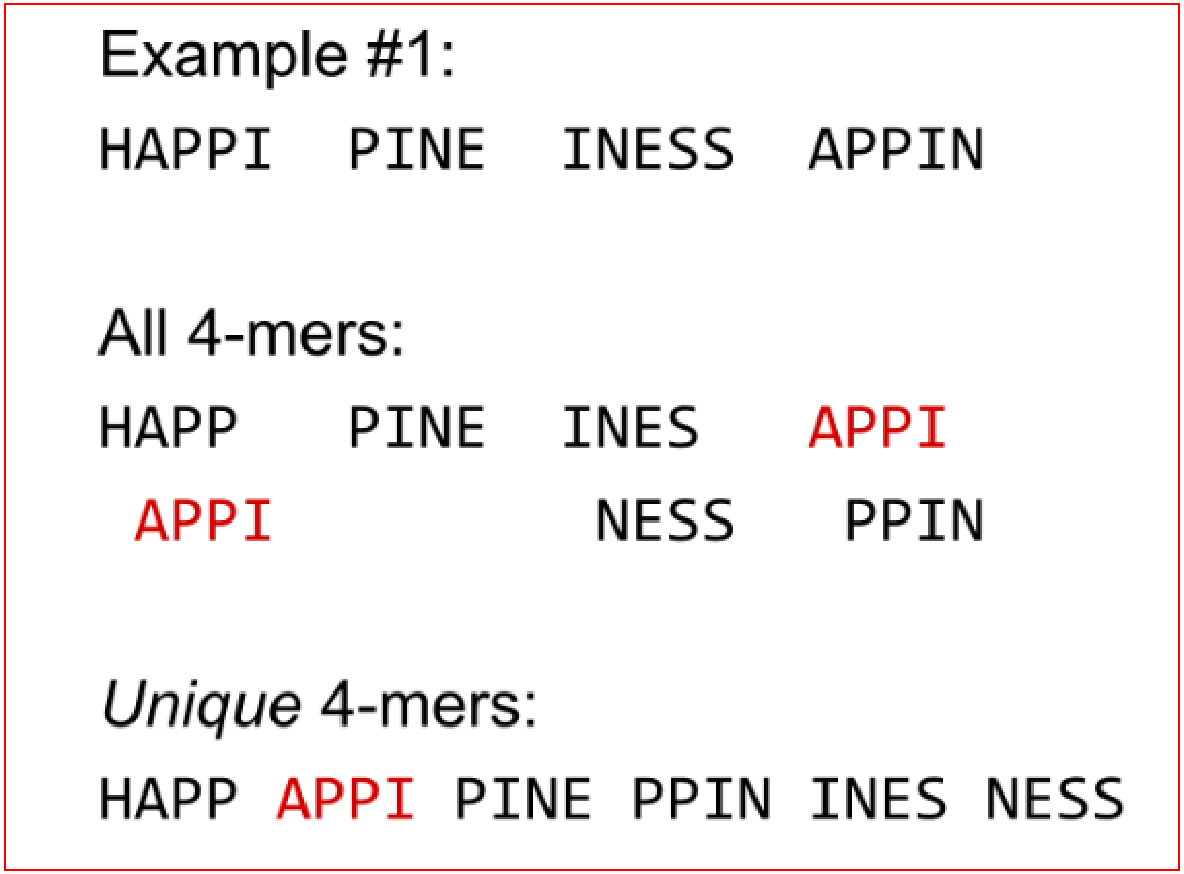


```
TTGACACTTACCGA        Read

TTGACACTTACC
 TGACACTTACCG         k-mers for k=12
  GACACTTACCGA

TTGAC
 TGACA
  GACAC
   ACACT
    CACTT
     ACTTA            k-mers for k=5
      CTTAC
       TTACC
        TACCG
         ACCGA
```

Algorithms: DBG



Example #1:

HAPPI PINE INESS APPIN

All 4-mers:

HAPP PINE INES APPI

APPI NESS PPIN

Unique 4-mers:

HAPP APPI PINE PPIN INES NESS

Algorithms: DBG

Example #1:

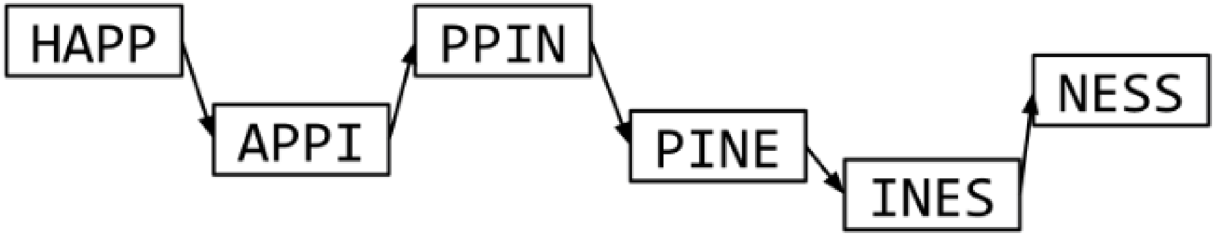
HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

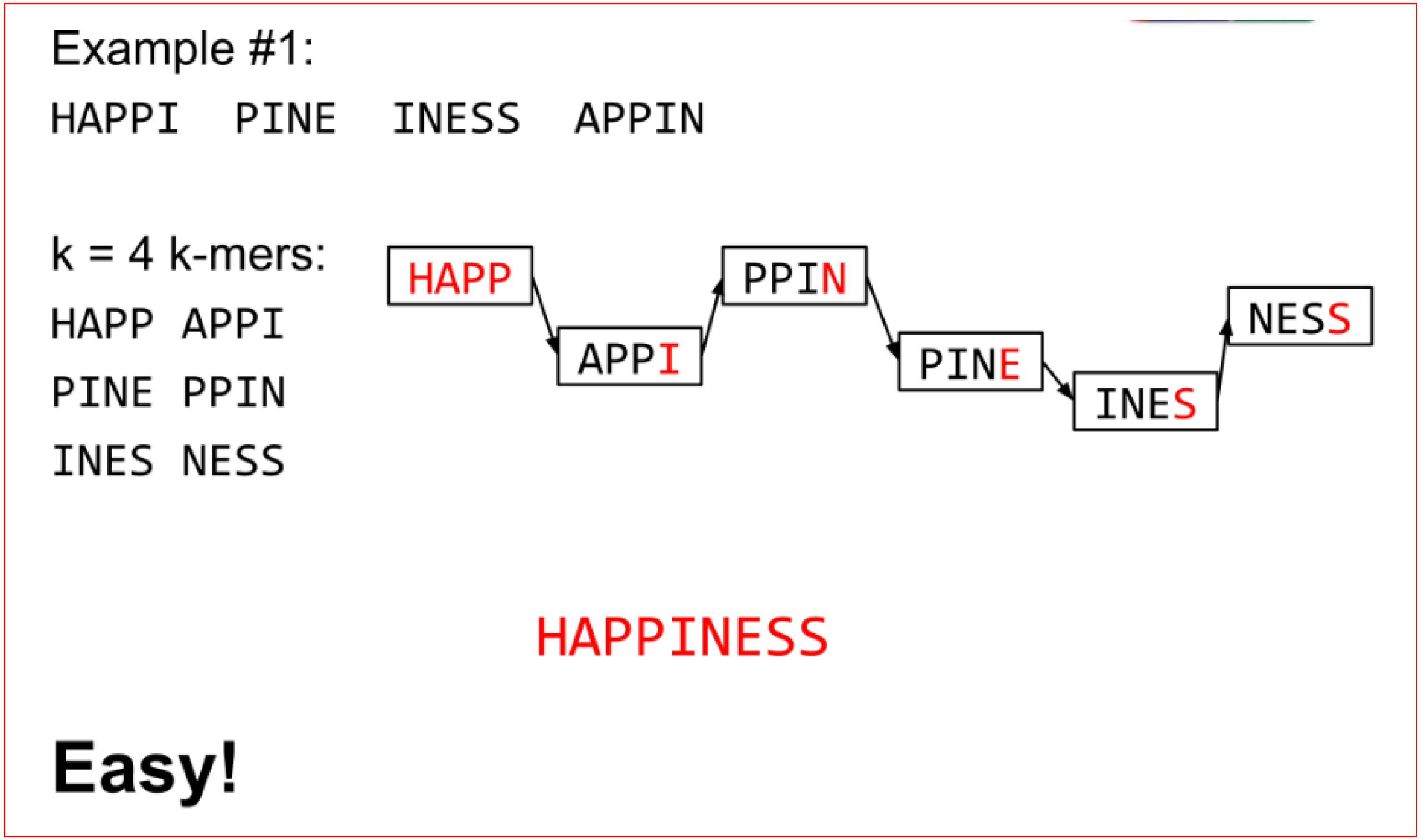
PINE PPIN

INES NESS



HAPP APPI PPIN PINE INES NESS

Algorithms: DBG



Example #1:

HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

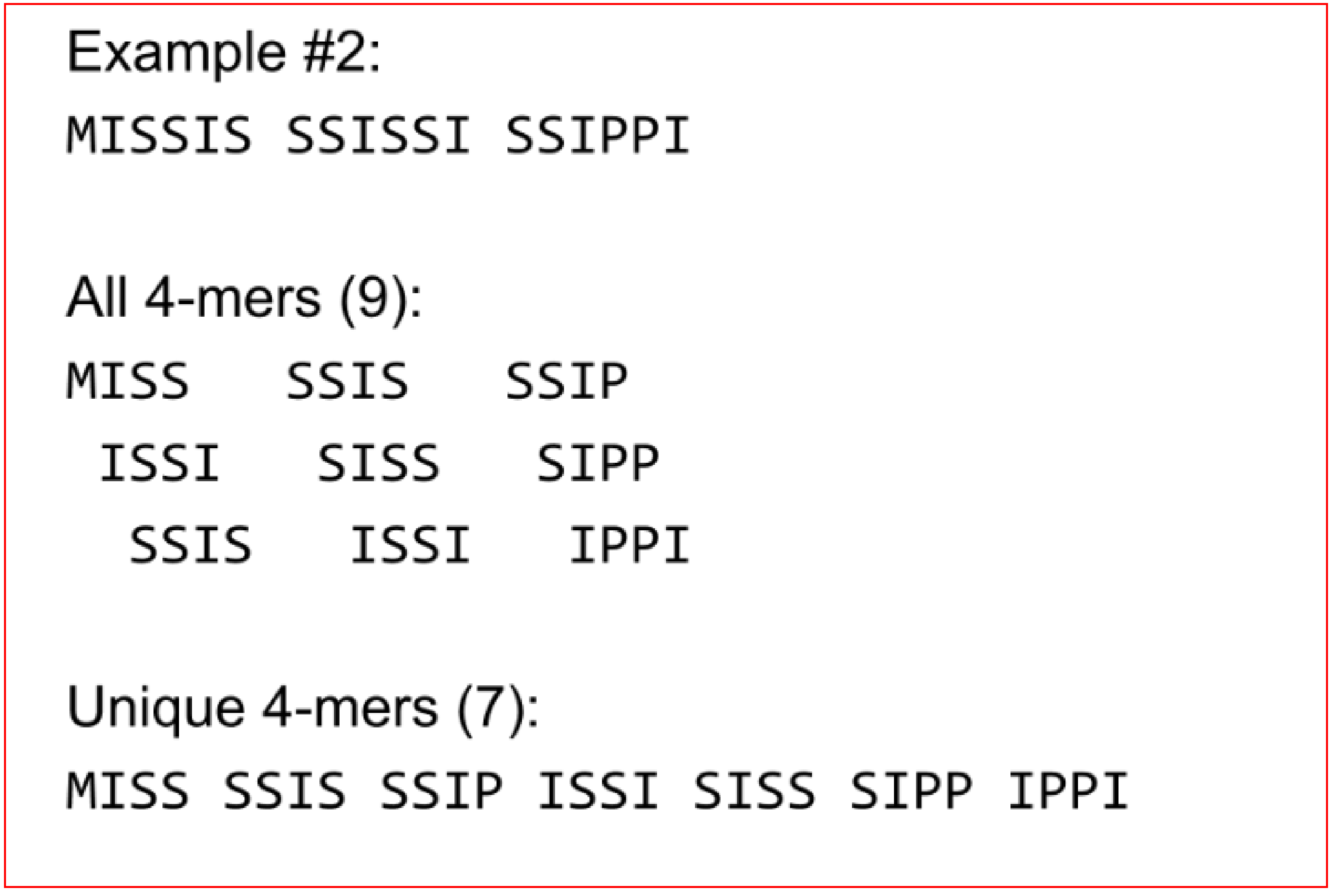
PINE PPIN

INES NESS

HAPPINESS

Easy!

Algorithms: DBG



Example #2:

```
MISSIS SSISSI SSIPPI
```

All 4-mers (9):

```
MISS    SSIS    SSIP
 ISSI    SISS    SIPP
  SSIS    ISSI    IPPI
```

Unique 4-mers (7):

```
MISS SSIS SSIP ISSI SISS SIPP IPPI
```

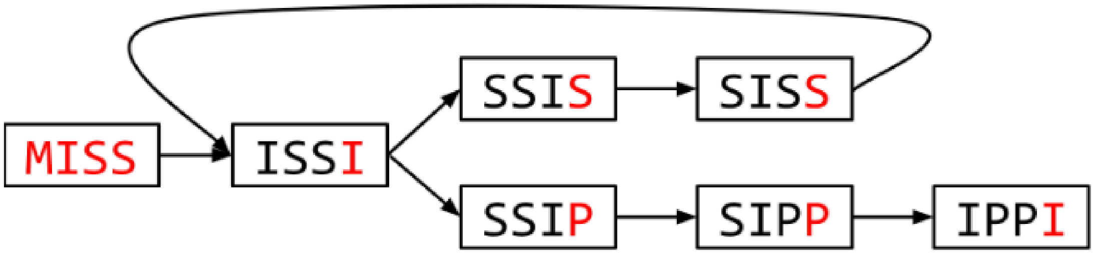
Algorithms: DBG

Example #2:

MISSIS SSISSI SSIPPI

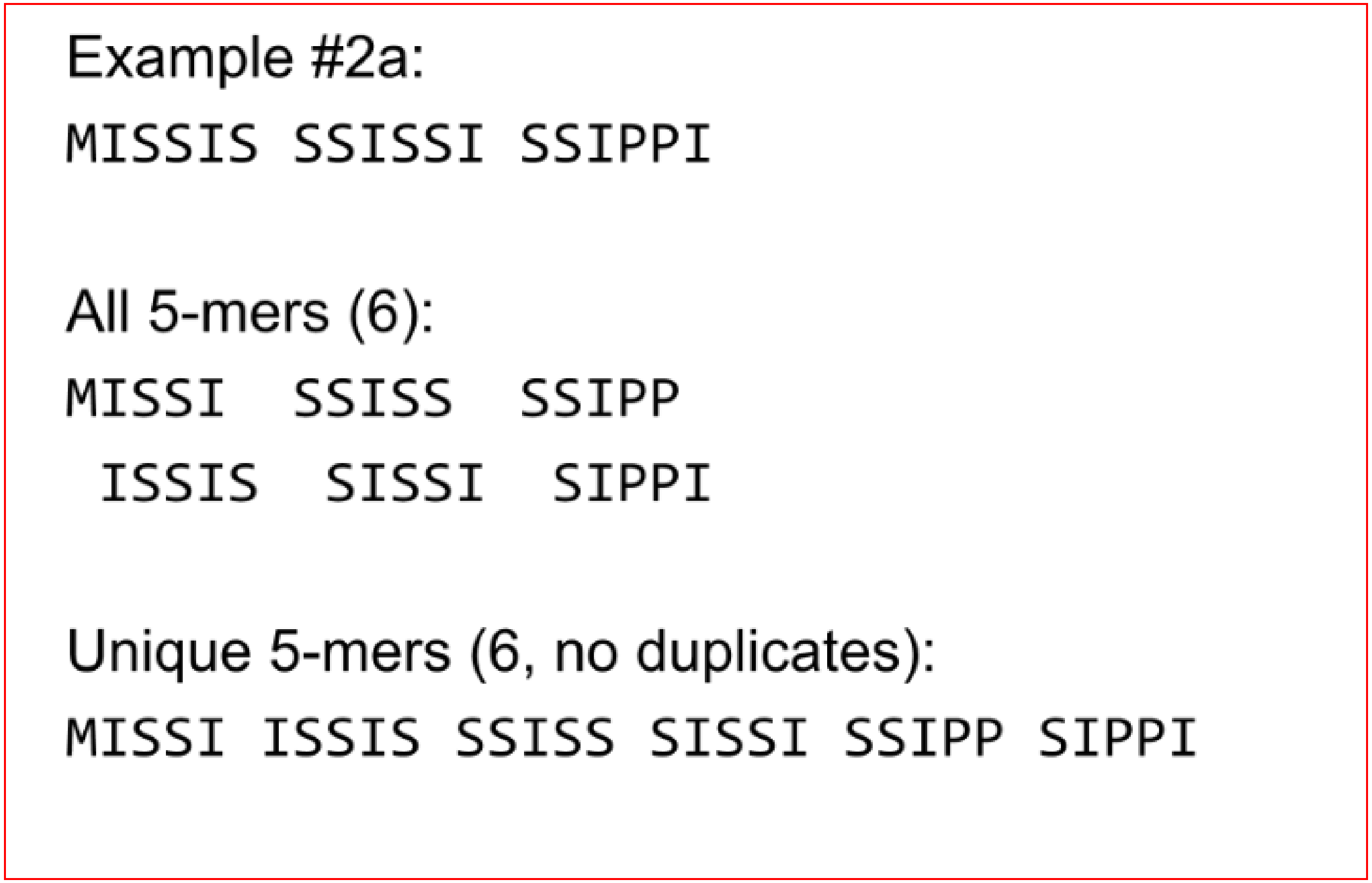
All 4-mers:

MISS ISSI SSIS SISS SSIP SIPP IPPI



MISSISSIPPI or MISSISSISSISSIPPI or …

Algorithms: DBG



Example #2a:

```
MISSIS SSISSI SSIPPI
```

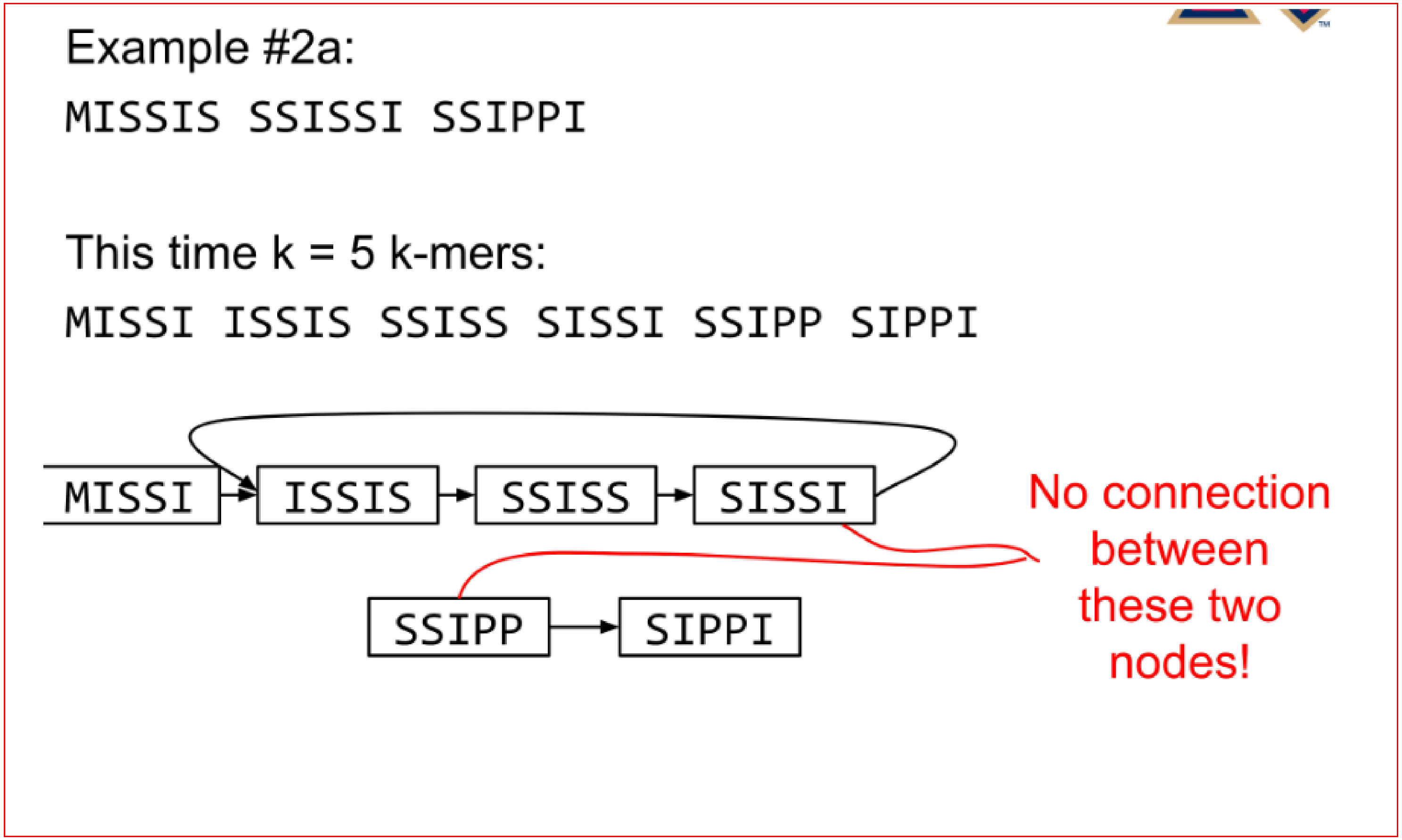
All 5-mers (6):

```
MISSI   SSISS   SSIPP
 ISSIS   SISSI   SIPPI
```

Unique 5-mers (6, no duplicates):

```
MISSI ISSIS SSISS SISSI SSIPP SIPPI
```

Algorithms: DBG



Example #2a:

MISSIS SSISSI SSIPPI

This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI

No connection between these two nodes!

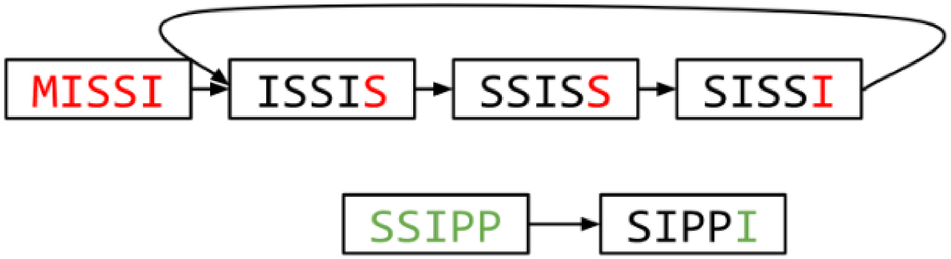
Algorithms: DBG

Example #2a:

MISSIS SSISSI SSIPPI

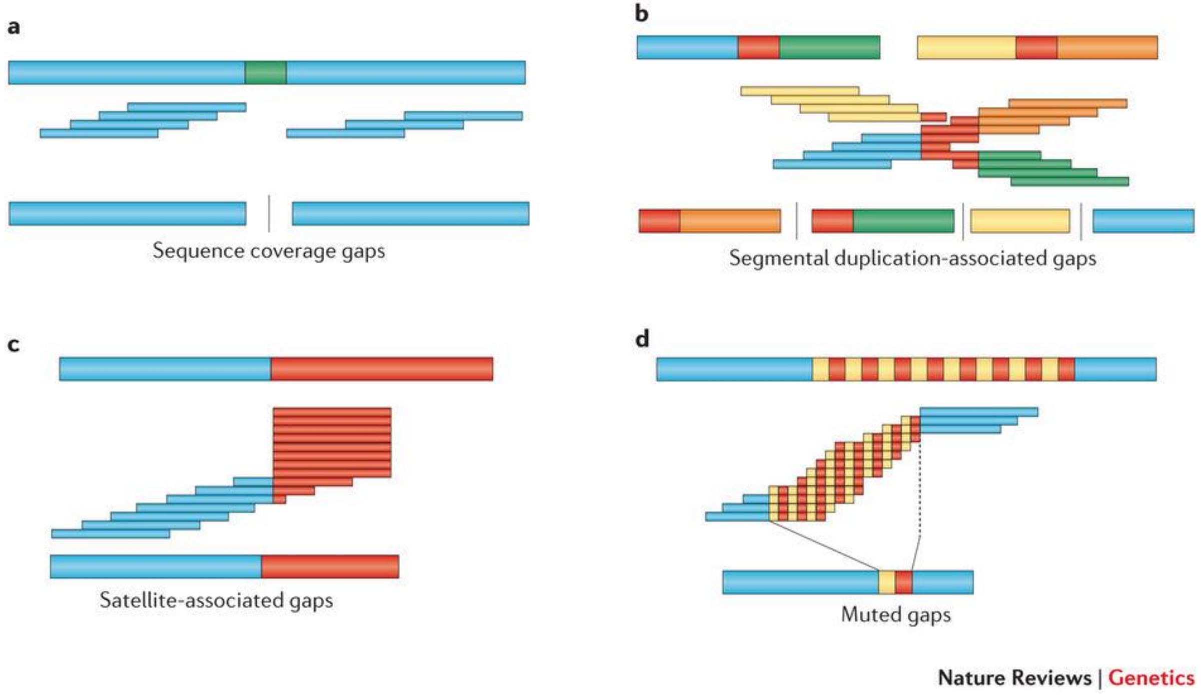
This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI



MISSISSIS SSIPPI

Assembly: Errors


a
Sequence coverage gaps
b
Segmental duplication-associated gaps
c
Satellite-associated gaps
d
Muted gaps
Nature Reviews | Genetics

- **A. Gaps – non sequenced region**
- **B. Long repeats**
  - Cuimera
- **Collapsed repetitive regions**
  - **C. Terminal**
  - **D. Intersticial**

Genetic variation and the de novo assembly of human genomes
Chaisson *et al.*

Assembly: Scaffolding

- **From draft:**

  Order contigs (Nucmer, if there is reference it can be used to align and guide)

  Fill the GAPs (GapFiller, fill sequencing gap (not physical gap)

  Solve repeated sequence ambiguities (Expander)

  Resequence with different library:

  - Longer fragments and/or distance

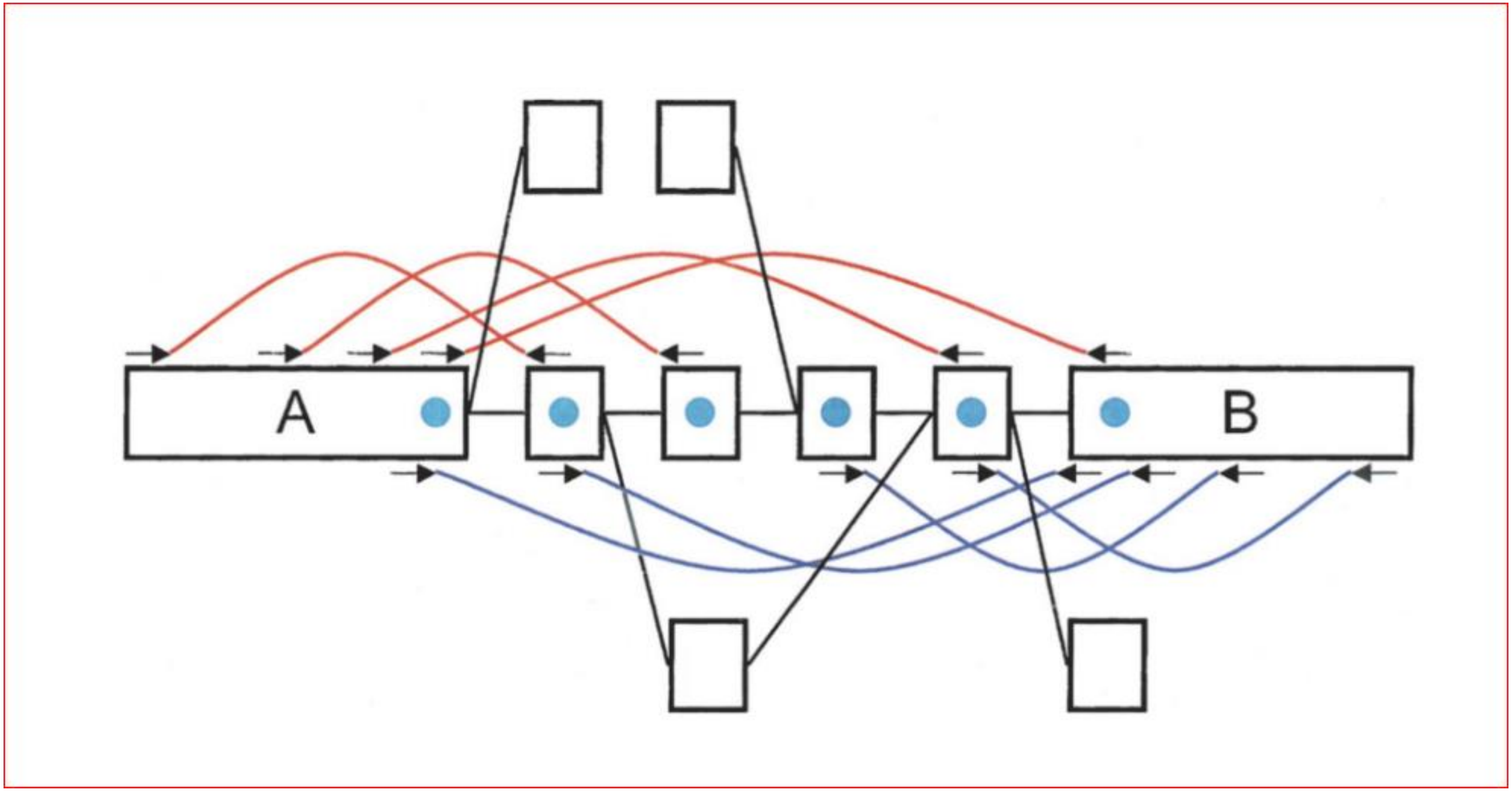
- **Tools for assembly improvement**

  SSPACE (Scaffolding) REAPR (evaluate scaffolding, breaking incorrect scaffolds)

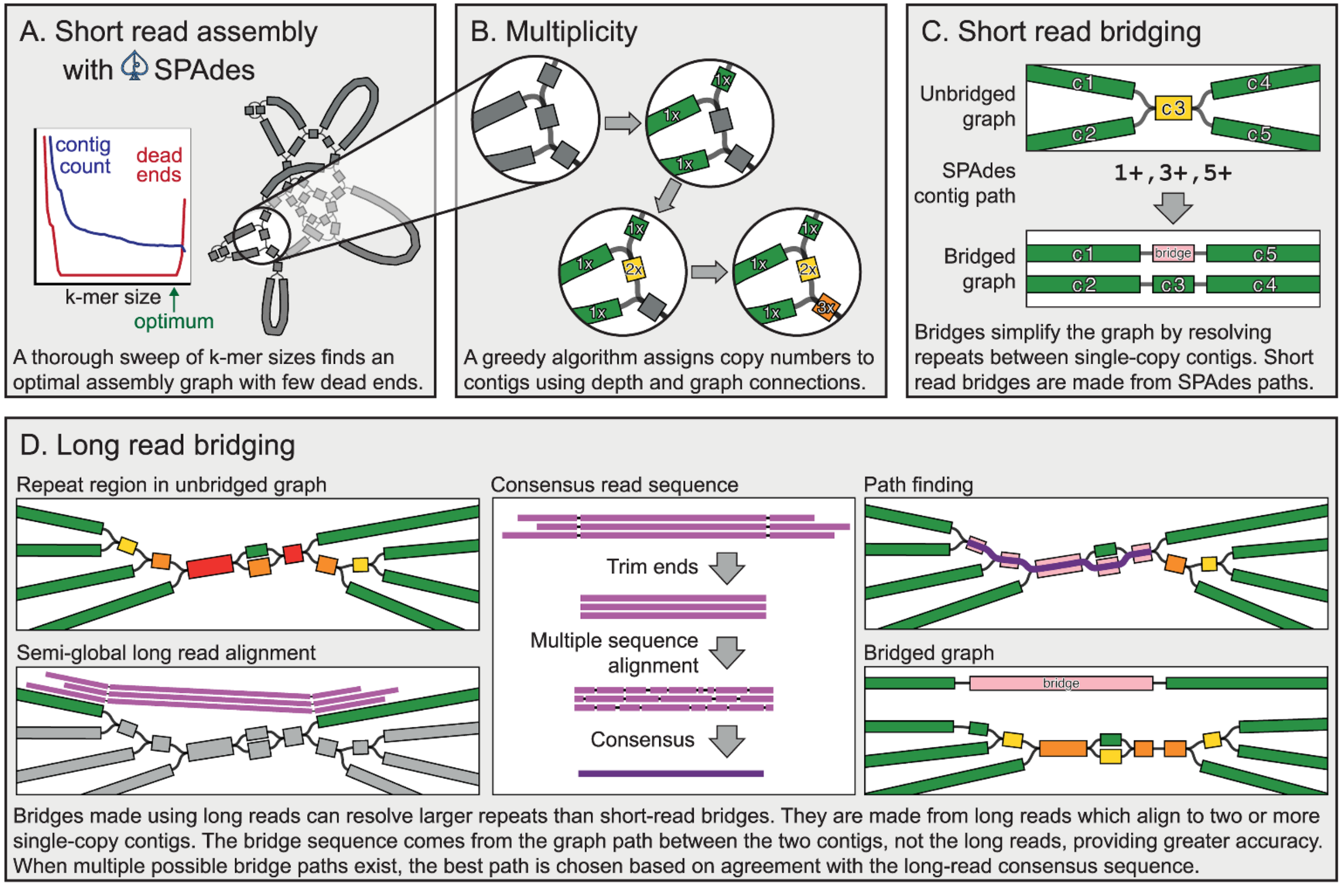
- **Assembly visualyzing**

  Artemis, ACT (compare two or more sequences), Icarus (Quast)

Assembly: Scaffolding



Unicycler



A thorough sweep of k-mer sizes finds an optimal assembly graph with few dead ends.

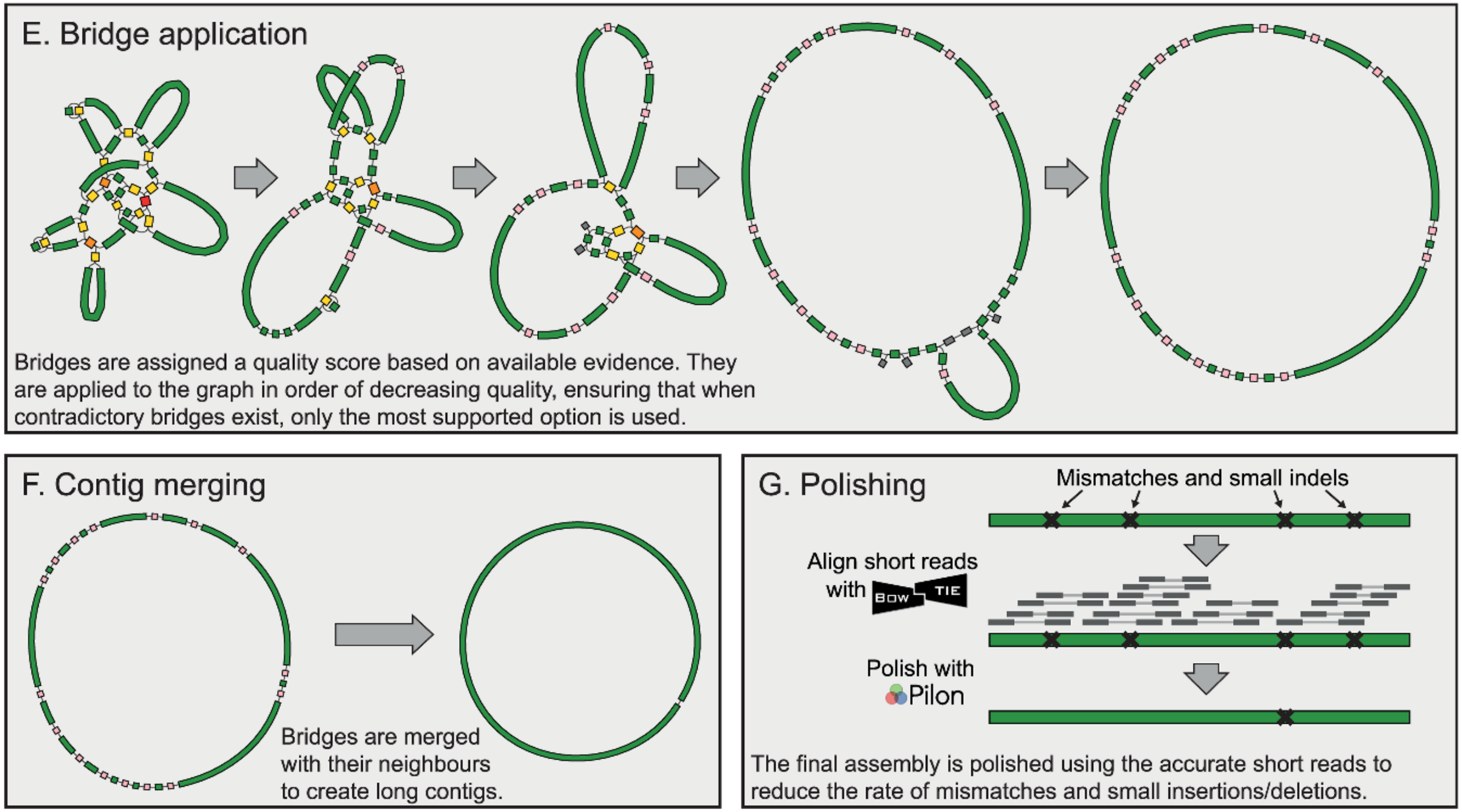
A greedy algorithm assigns copy numbers to contigs using depth and graph connections.

Bridges simplify the graph by resolving repeats between single-copy contigs. Short read bridges are made from SPAdes paths.

Bridges made using long reads can resolve larger repeats than short-read bridges. They are made from long reads which align to two or more single-copy contigs. The bridge sequence comes from the graph path between the two contigs, not the long reads, providing greater accuracy. When multiple possible bridge paths exist, the best path is chosen based on agreement with the long-read consensus sequence.

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595

Unicycler



E. Bridge application

Bridges are assigned a quality score based on available evidence. They are applied to the graph in order of decreasing quality, ensuring that when contradictory bridges exist, only the most supported option is used.

F. Contig merging

Bridges are merged with their neighbours to create long contigs.

G. Polishing

Mismatches and small indels

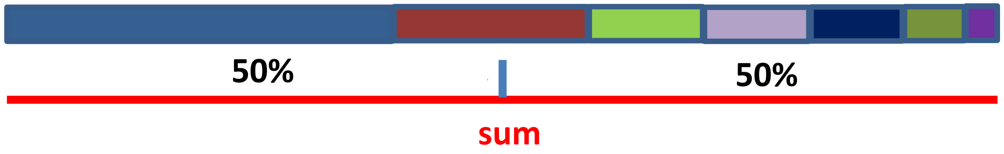
Align short reads with BOWTIE

Polish with Pilon

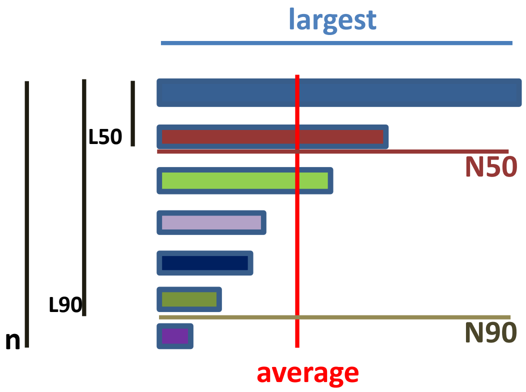
The final assembly is polished using the accurate short reads to reduce the rate of mismatches and small insertions/deletions.

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595

Assembly: Metrics



- sum = total bases number
- n = contigs number
- average = average contig length
- largest = largest contig
- N50 = length of the shortest contig where 50% of sum is held
- L50 = number of contigs which have 50% of the genome
- N90 = length of the shortest contig where 90% of sum is held.
- L90 = number of contigs which have 90% of the genome



Assembly: Evaluation

- Software that evaluate differets algorithms & parameters
  iMetAMOS, *Koren et al., BMCBioinformatics 2014, 15:126*
  GAGE-B, *Magoc et al.,* Bioinformatics 2013,29(14):1718-25

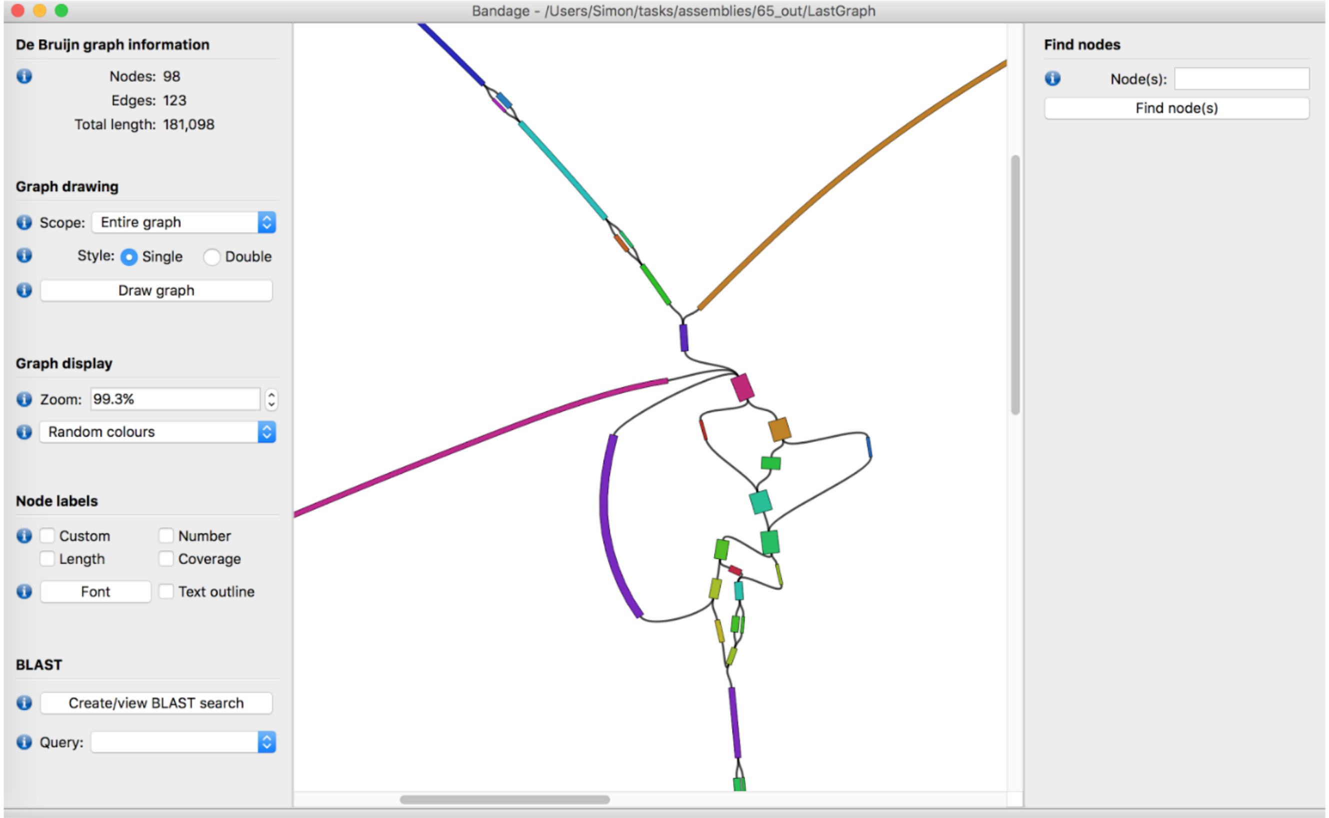
- **Graph evaluation**: Bandage, Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015)

- **Assembly evaluation**: Quast, *Gurevich et al., Bioinformatics 2013, 29:8*

- **Metrics for a good assembly:**
  Large N50
  Sum closest to expected
  Low n
  Low L50

Assembly: Evaluation - Bandage

- Graph evaluation: Bandage, Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015)



Bandage - /Users/Simon/tasks/assemblies/65_out/LastGraph

De Bruijn graph information

Nodes: 98
Edges: 123
Total length: 181,098

Graph drawing

Scope: Entire graph
Style: ● Single ○ Double

Graph display

Zoom: 99.3%
Random colours

Node labels

□ Custom □ Number
□ Length □ Coverage

□ Text outline

BLAST

Query:

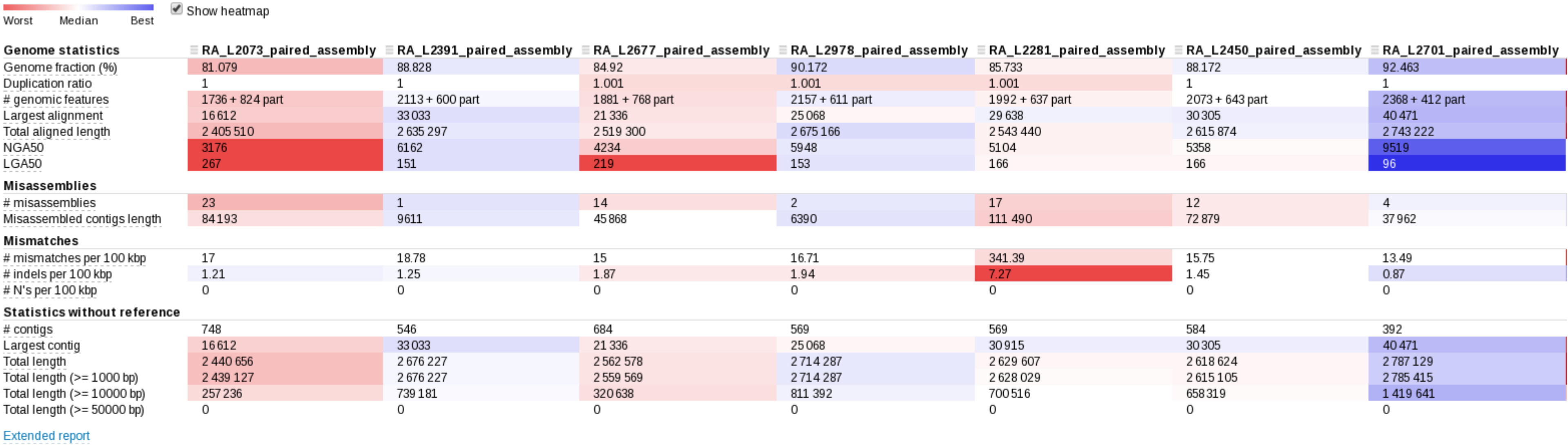
Find nodes

Node(s):

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES Instituto de Salud Carlos III >_BU-ISCIII

Assembly: Evaluation - Quast

- Assembly evaluation: Quast, *Gurevich et al., Bioinformatics 2013, 29:8*

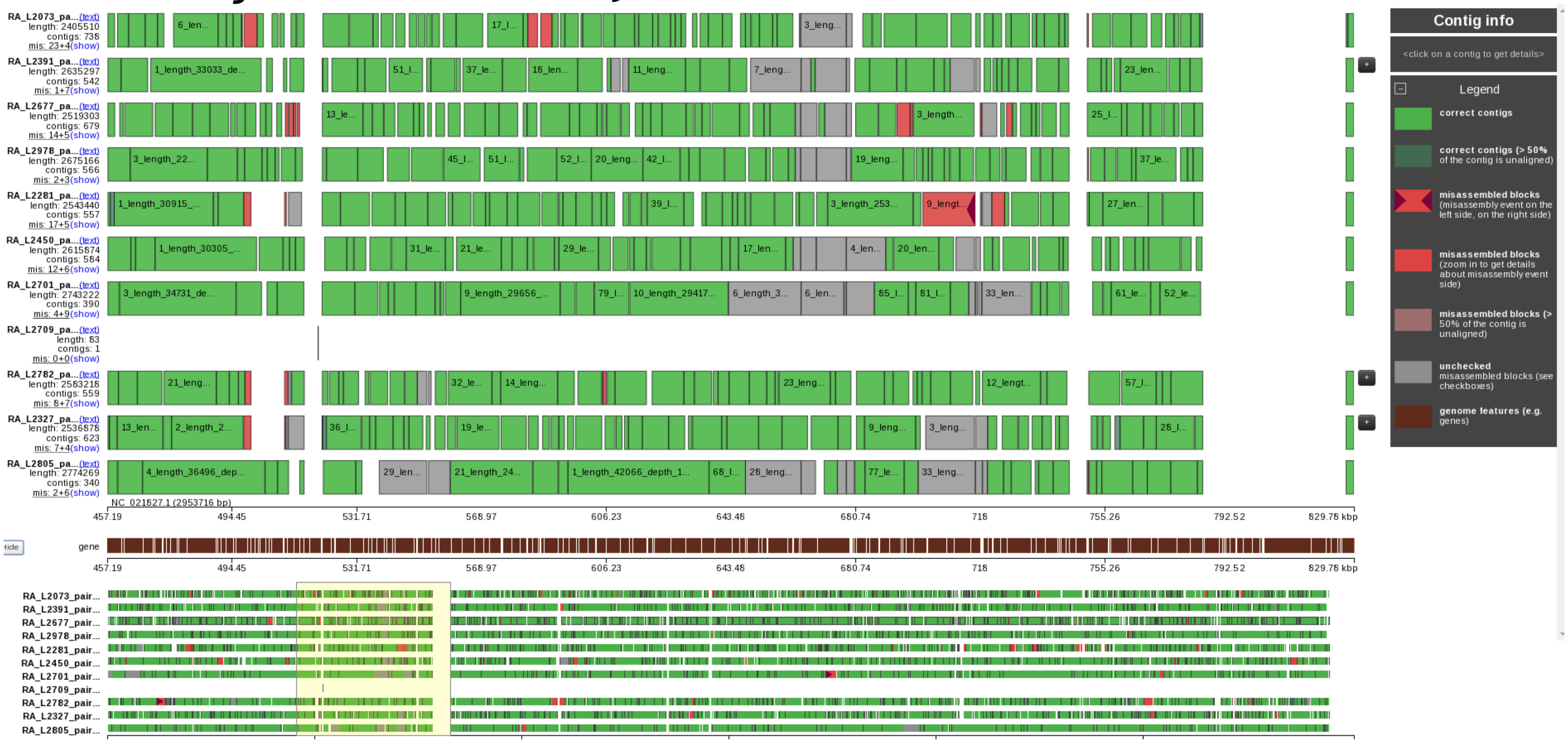


Genome statistics	RA_L2073_paired_assembly	RA_L2391_paired_assembly	RA_L2677_paired_assembly	RA_L2978_paired_assembly	RA_L2281_paired_assembly	RA_L2450_paired_assembly	RA_L2701_paired_assembly
Genome fraction (%)	81.079	88.828	84.92	90.172	85.733	88.172	92.463
Duplication ratio	1	1	1.001	1.001	1.001	1	1
# genomic features	1736 + 824 part	2113 + 600 part	1881 + 768 part	2157 + 611 part	1992 + 637 part	2073 + 643 part	2368 + 412 part
Largest alignment	16 612	33 033	21 336	25 068	29 638	30 305	40 471
Total aligned length	2 405 510	2 635 297	2 519 300	2 675 166	2 543 440	2 615 874	2 743 222
NGA50	3176	6162	4234	5948	5104	5358	9519
LGA50	267	151	219	153	166	166	96
Misassemblies							
# misassemblies	23	1	14	2	17	12	4
Misassembled contigs length	84 193	9611	45 868	6390	111 490	72 879	37 962
Mismatches							
# mismatches per 100 kbp	17	18.78	15	16.71	341.39	15.75	13.49
# indels per 100 kbp	1.21	1.25	1.87	1.94	7.27	1.45	0.87
# N's per 100 kbp	0	0	0	0	0	0	0
Statistics without reference							
# contigs	748	546	684	569	569	584	392
Largest contig	16 612	33 033	21 336	25 068	30 915	30 305	40 471
Total length	2 440 656	2 676 227	2 562 578	2 714 287	2 629 607	2 618 624	2 787 129
Total length (>= 1000 bp)	2 439 127	2 676 227	2 559 569	2 714 287	2 628 029	2 615 105	2 785 415
Total length (>= 10000 bp)	257 236	739 181	320 638	811 392	700 516	658 319	1 419 641
Total length (>= 50000 bp)	0	0	0	0	0	0	0

Extended report

Assembly: Evaluation - Quast

- Assembly evaluation: Quast, *Gurevich et al., Bioinformatics 2013, 29:8*



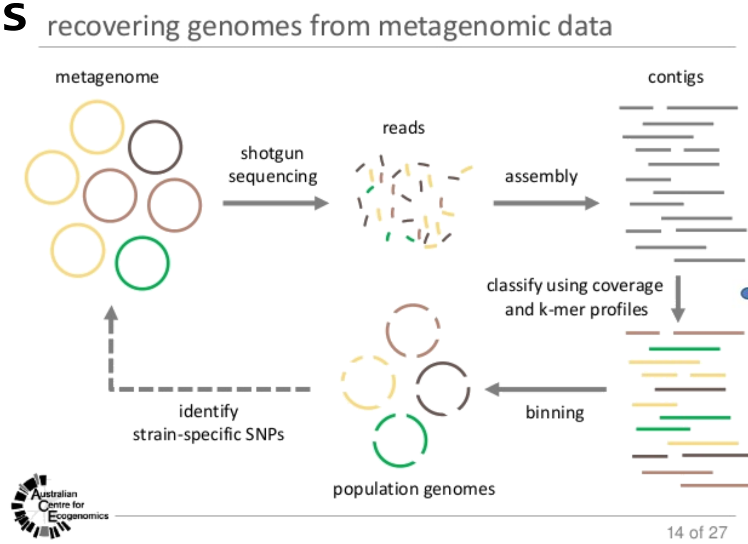
Assembly: Assemblers

Name	Type	Technologies	Author	Presented /Last updated	Licence*	Homepage
DNASTAR Lasergene Genomics Suite	(large) genomes, exomes, transcriptomes, metagenomes, ESTs	Illumina, ABI SOLiD, Roche 454, Ion Torrent, Solexa, Sanger	DNASTAR	2007 / 2016	C	link
Newbler	genomes, ESTs	454, Sanger	454/Roche	2004/2012	C	link
Canu	Small and large, haploid/diploid genomes	PacBio/Oxford Nanopore reads	Koren et al.[8]	2001 / 2018	OS	link
SPAdes	(small) genomes, single-cell	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	Bankevich, A et al.	2012 / 2017	OS	link
Velvet	(small) genomes	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007 / 2011	OS	link

***Licences:** OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics

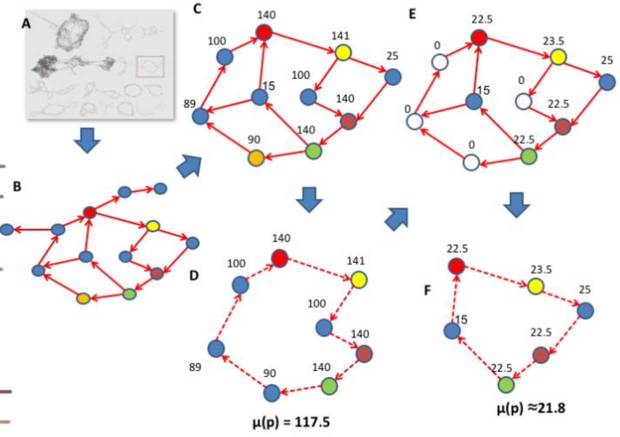
Assembly: Specials assemblers

- **Diploid genomes**
- **Metagenomics**
- **Plasmids**
- **Transcriptome**

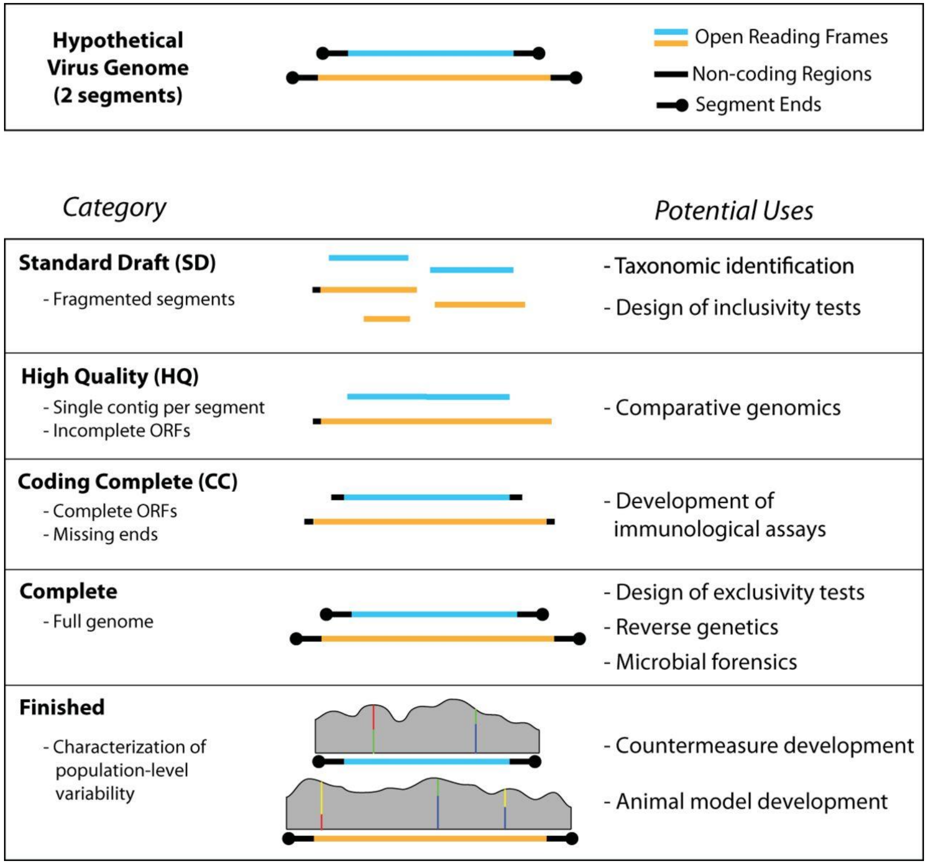


recovering genomes from metagenomic data

metagenome
shotgun sequencing
reads
assembly
contigs
classify using coverage and k-mer profiles
binning
population genomes
identify strain-specific SNPs
Australian Centre for Ecogenomics
14 of 27



Assembly: Categories



Hypothetical Virus Genome (2 segments)

Category	*Potential Uses*
Standard Draft (SD) - Fragmented segments	- Taxonomic identification - Design of inclusivity tests
High Quality (HQ) - Single contig per segment - Incomplete ORFs	- Comparative genomics
Coding Complete (CC) - Complete ORFs - Missing ends	- Development of immunological assays
Complete - Full genome	- Design of exclusivity tests - Reverse genetics - Microbial forensics
Finished - Characterization of population-level variability	- Countermeasure development - Animal model development

Standards for Sequencing Viral Genomes in the Era of HighThroughput Sequencing. Ladner *et al*.