Single cell technologies in sequence assembly and genome construction

Diana Lin[1,2]

October 3, 2019

[1] Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[2] Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

What is Single Cell Sequencing?

- Sequencing of the DNA of a single cell, as opposed to bulk tissue cells (multi-cell)

Why do Single Cell Sequencing?

- Resolve cell-to-cell variations
- Identify rare cells in disease progression
- Allows detailed and comprehensive studies of individual cells

Wang et al. 2015. *Molecular Cell.*

Single Cell Sequencing Method

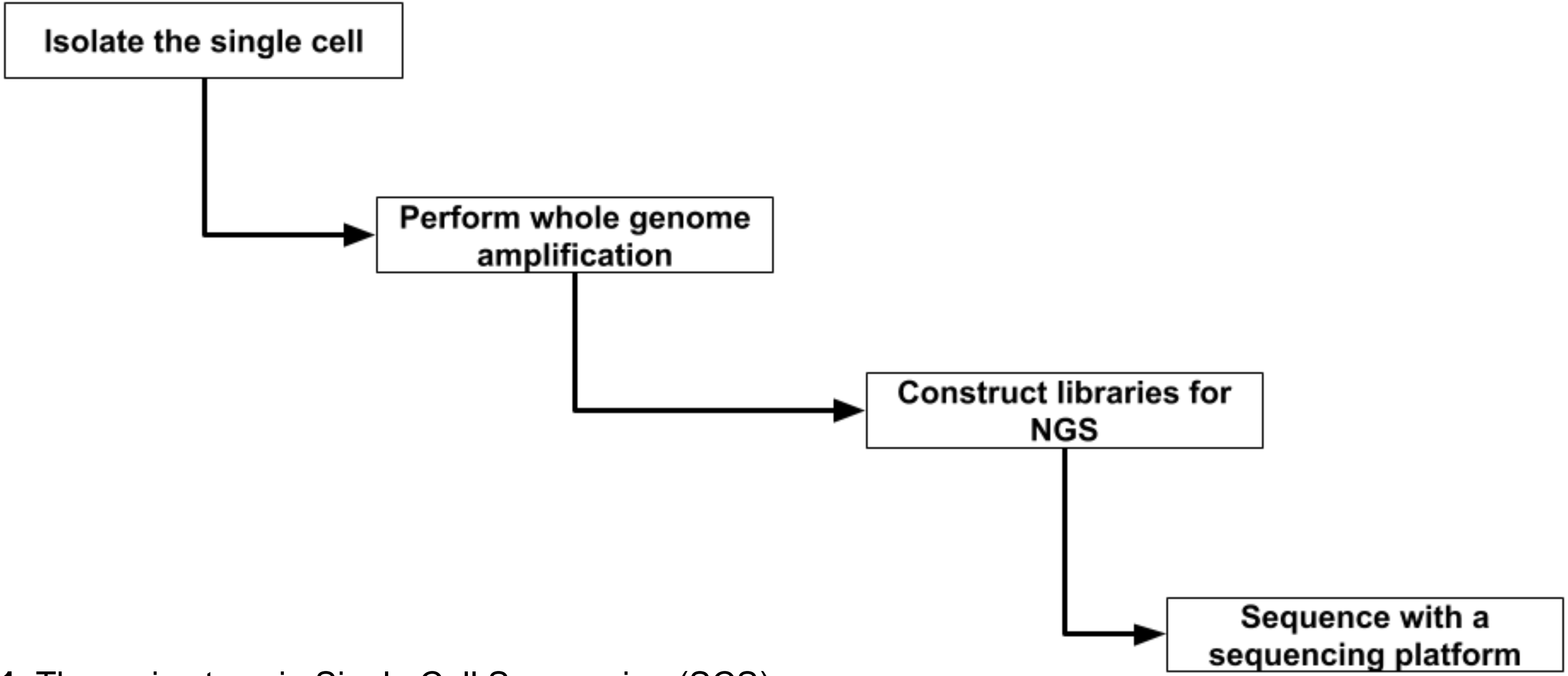

Isolate the single cell
Perform whole genome amplification
Construct libraries for NGS
Sequence with a sequencing platform

Fig 1. The main steps in Single Cell Sequencing (SCS).

Wang et al. 2015. *Molecular Cell.*

Challenges of Single Cell Sequencing

- Limited number of DNA molecules
- This limited amount of input material for whole genome amplification results in technical errors
- Technical errors occur in initial rounds of amplification and then are propagated by all daughter molecules

Wang et al. 2015. *Molecular Cell.*

Whole Genome Amplification (WGA)

Amplification Method	Advantages	Disadvantages
DOP-PCR[1]	● Accurately retains copy number levels	● Generates low physical coverage (~10%) of a single cell genome
MDA[2]	● Achieves high physical coverage (>90%) from a single cell genome	● Non-uniform coverage and causes distortions in read depth ● Poor method to measure DNA copy number

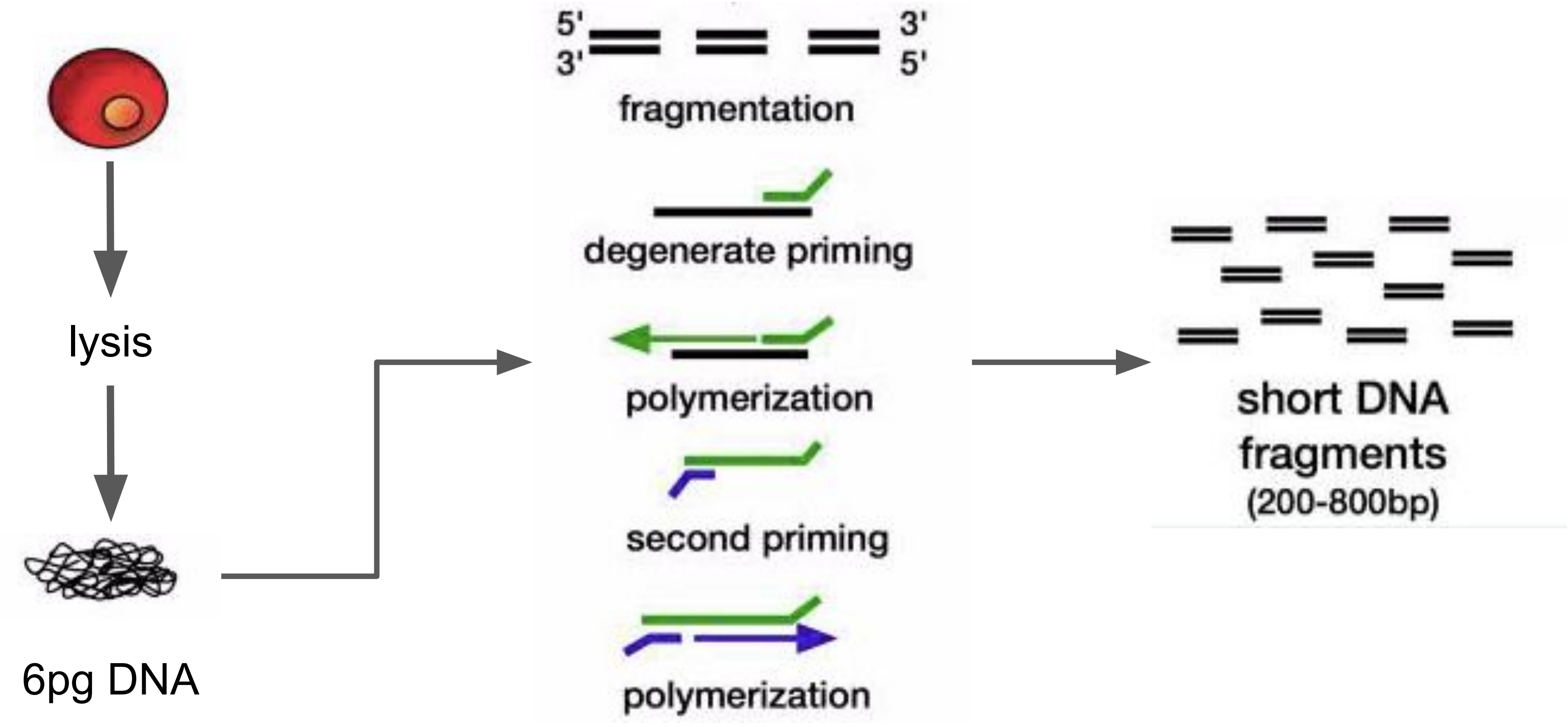
Table 1. Main methods for Whole Genome Amplification.

Wang et al. 2015. *Molecular Cell.*

[1] Degenerate Oligonucleotide Primed PCR

[2] Multiple Displacement Amplification

DOP-PCR



lysis

6pg DNA

fragmentation

degenerate priming

polymerization

second priming

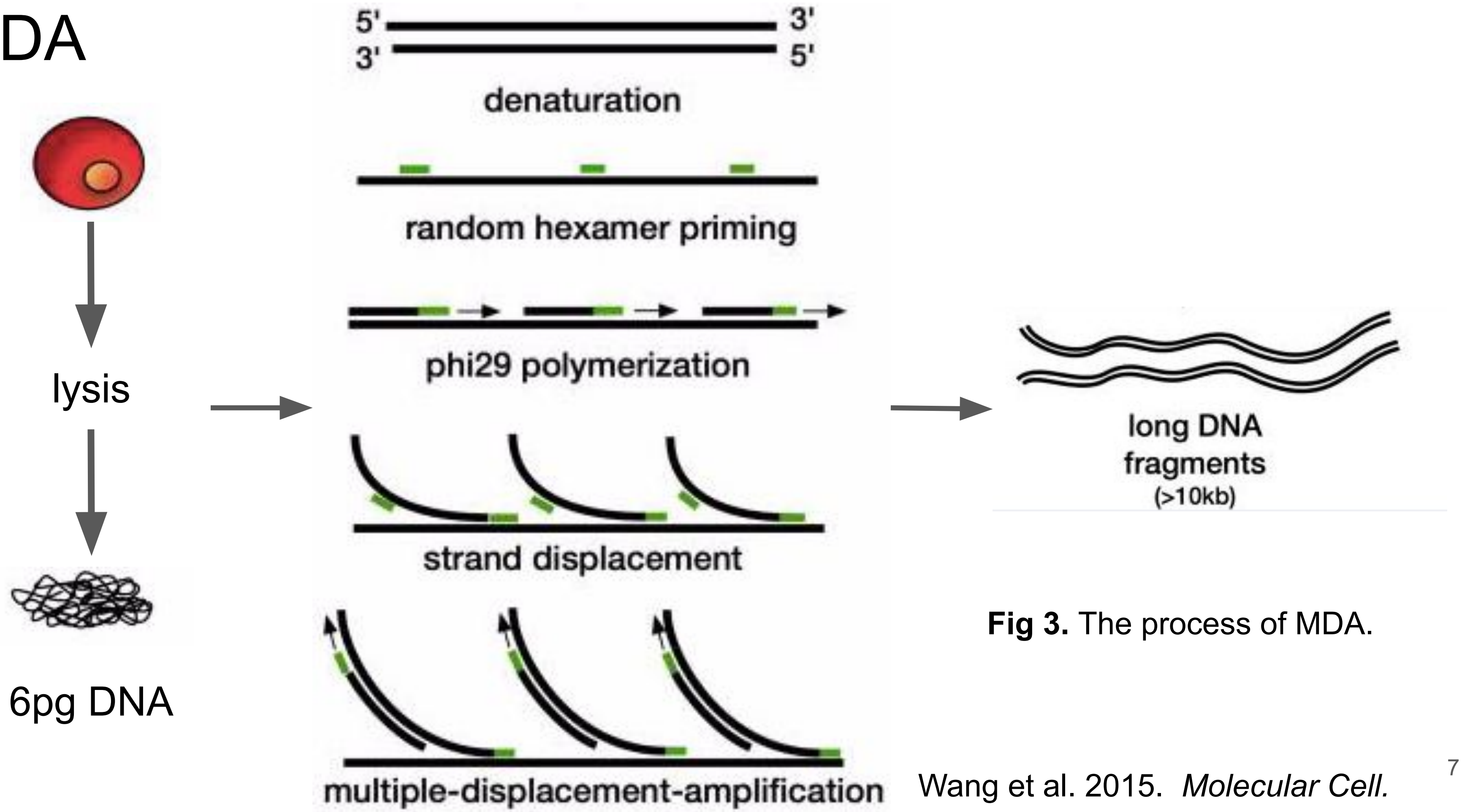
polymerization

short DNA fragments (200-800bp)

Fig 2. The process of DOP-PCR.

Wang et al. 2015. *Molecular Cell.*

MDA



lysis

6pg DNA

Fig 3. The process of MDA.

Wang et al. 2015. *Molecular Cell.*

Technical Errors from WGA

Technical Artifact	Amplification Method	Error Type
coverage non-uniformity	MDA, DOP-PCR	copy number aberrations, false-negative SNVs
false positive amplification error	MDA, DOP-PCR	SNV, indel

Table 2. Technical errors that arise from WGA.

Wang et al. 2015. *Molecular Cell.*

Coverage Non-uniformity

- **Amplification Method:** MDA, DOP-PCR
- **Error Type:** copy number aberrations, false-negative SNVs
- **Description:** Under and over amplifications of different regions of the genome causes copy number aberrations and false-negative SNVs

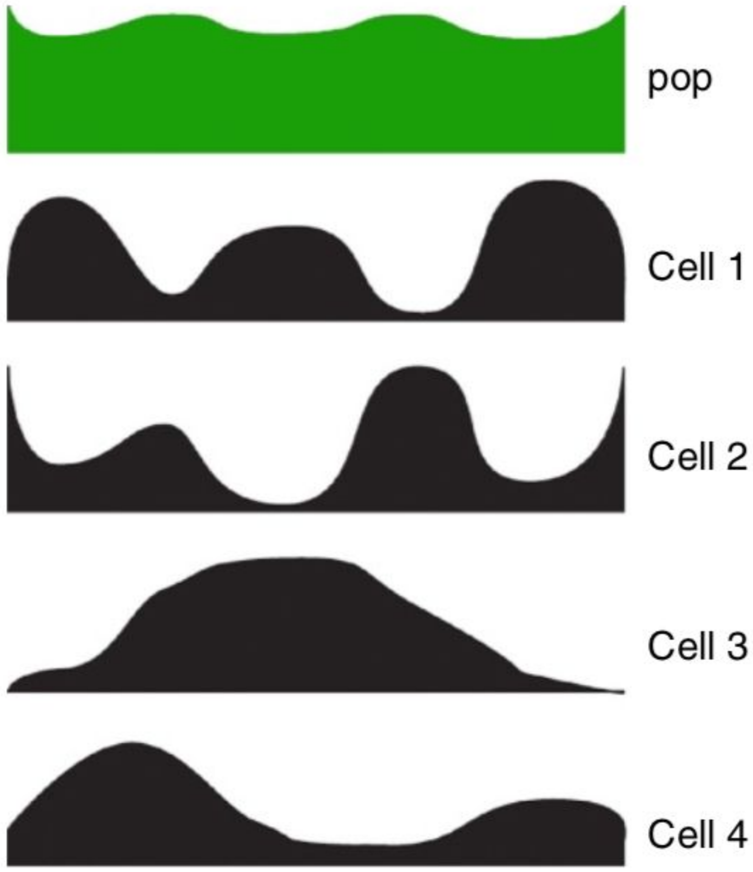


Fig 4. Coverage non-uniformity across single cells of a population.

Navin 2014. *Genome Biology*.

False Positive Amplification Error

- **Amplification Method:** MDA, DOP-PCR
- **Error Type:** SNV, indel
- **Description:** DNA polymerase introduces random false positive errors

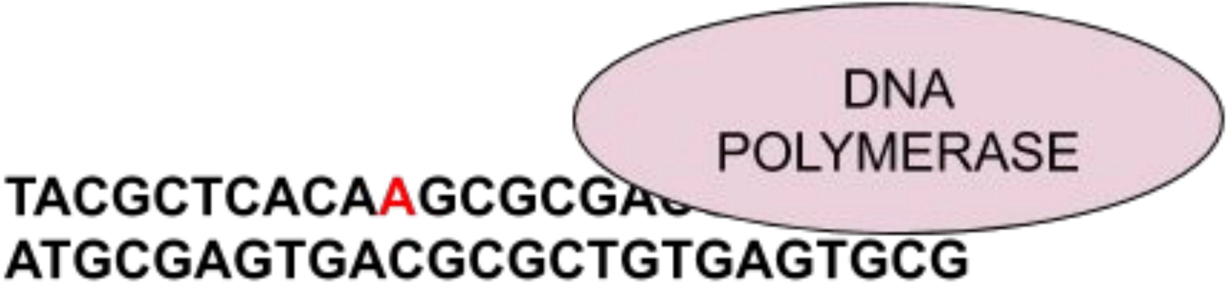


Fig 5. False positive amplification error created by DNA polymerase.

Navin 2014. *Genome Biology*.

Assembly of Microbial Genomes from Single Cells

- SPAdes
    - Constructs paired assembly graphs utilizing read pairs
- EULER+Velvet-SC
    - Uses lower initial coverage cutoff and then progressively increases the cutoff to incorporate more bases

Chitsaz et al. 2011. *Nature Biotechnology.*

SPAdes

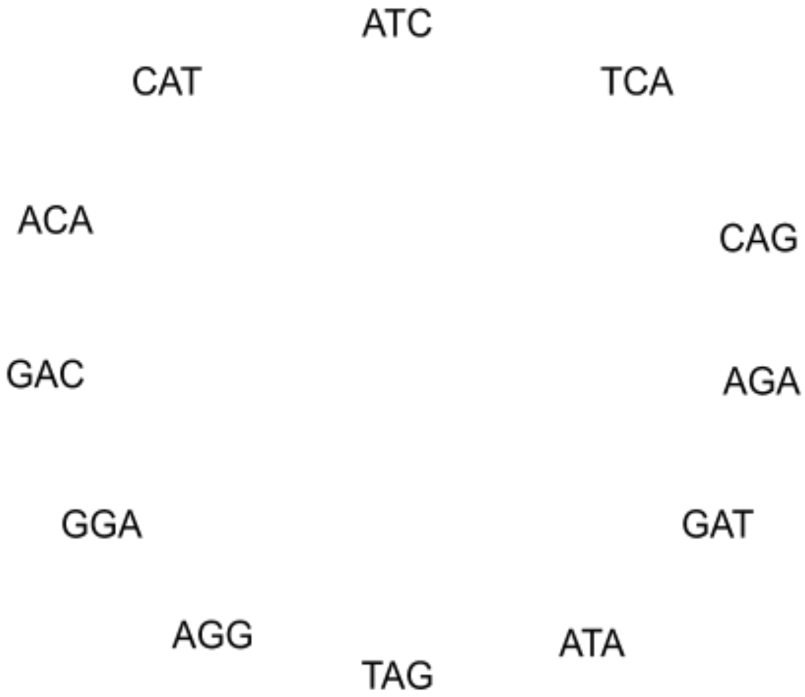


Fig 6a. A simple genome, where the reads are 3 bases long.

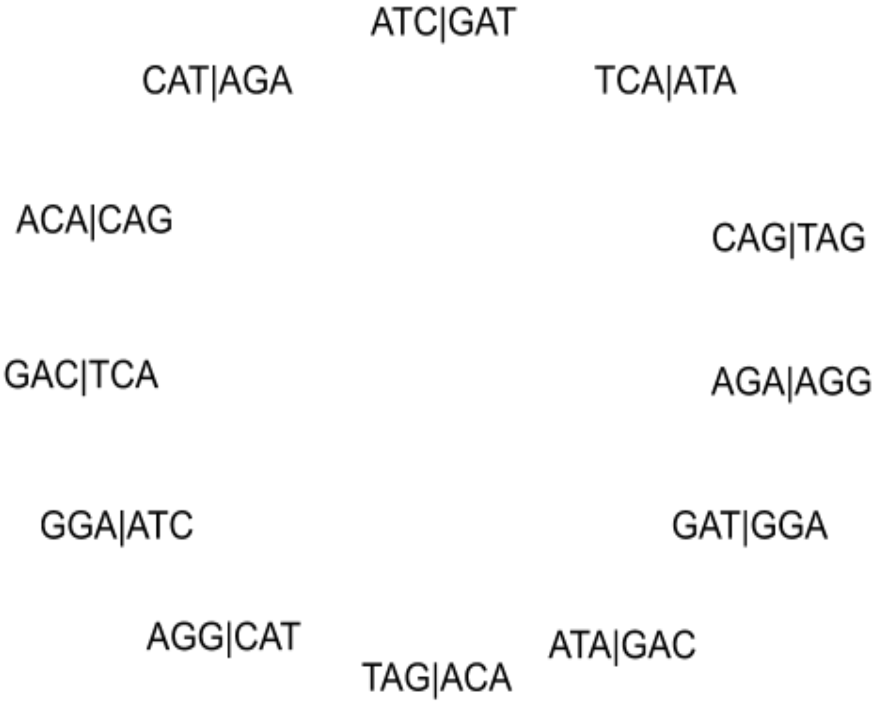


Fig 7a. A simple genome, where the read pairs are 3 bases long.

Bankevich et al. 2012.
Journal of Computational Biology.

SPAdes

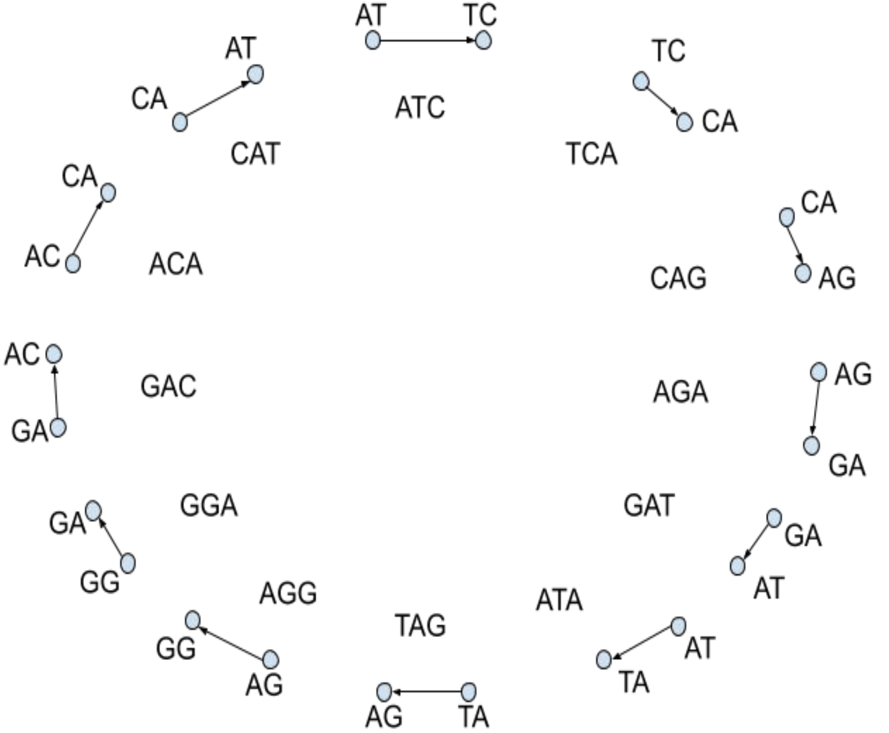


Fig 6b. Split the 3-base reads into k-mers of size = 2 and build DBG.

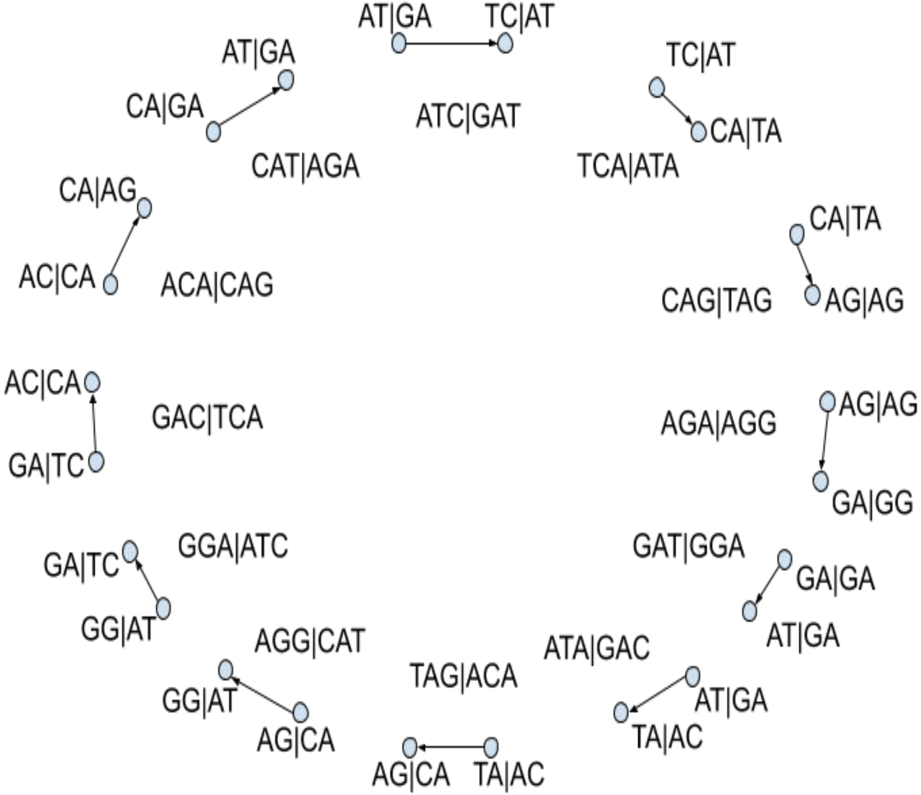


Fig 7b. Split the 3-base read pairs into k-mers of size = 2 and build DBG.

SPAdes

Bankevich et al. 2012.
Journal of Computational Biology.

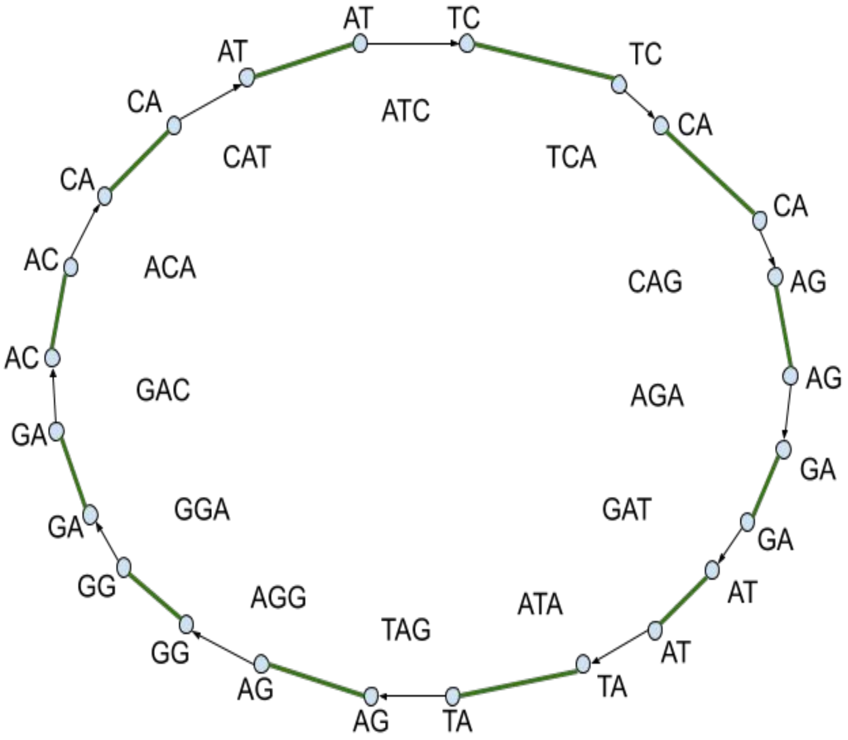


Fig 6c. Connect the adjacent identical k-mers.

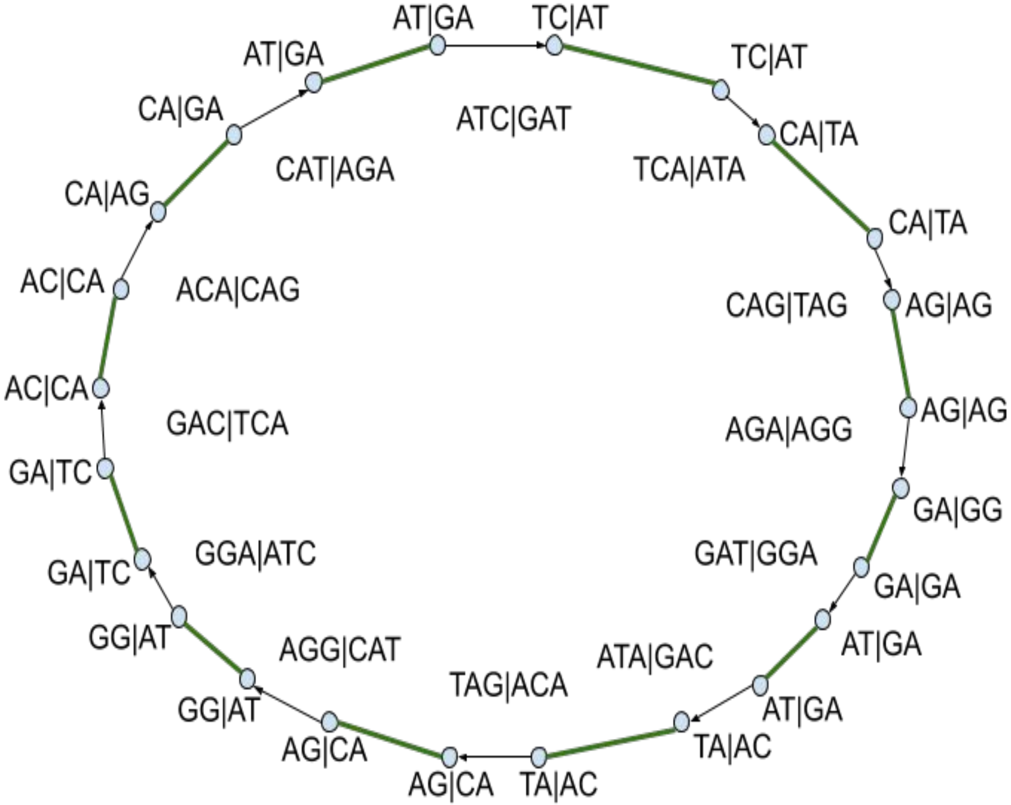


Fig 7c. Connect the adjacent identical k-mer pairs.

Bankevich et al. 2012.
Journal of Computational Biology.

SPAdes

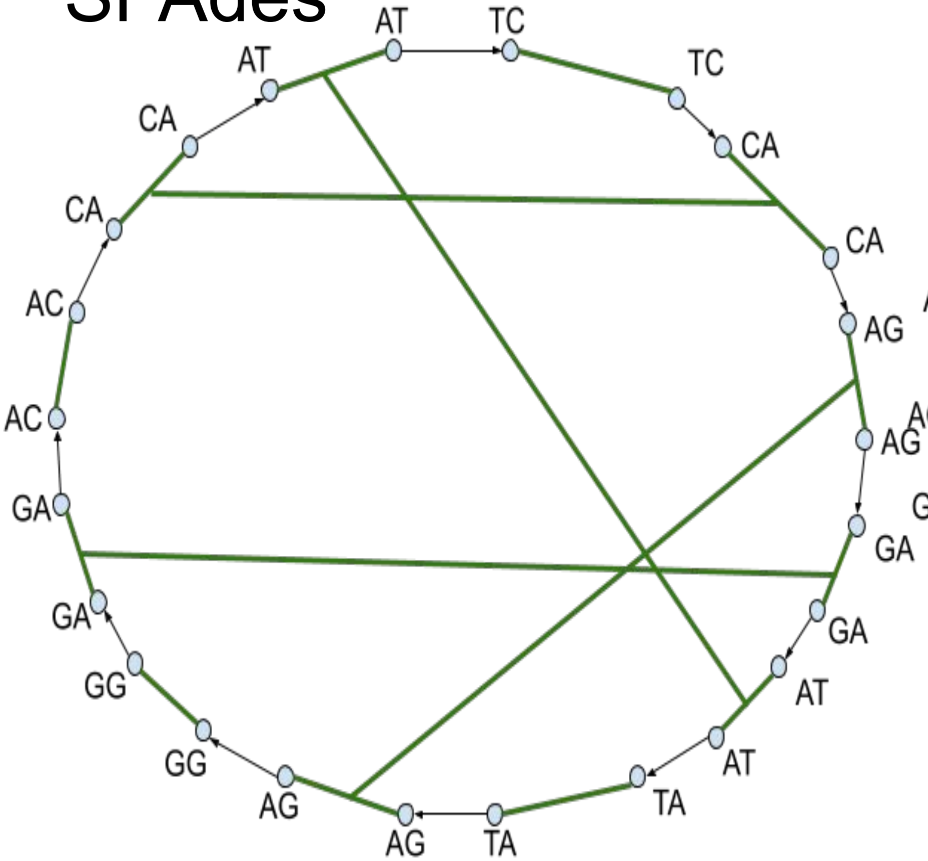


Fig 6d. Connect all the identical k-mers.

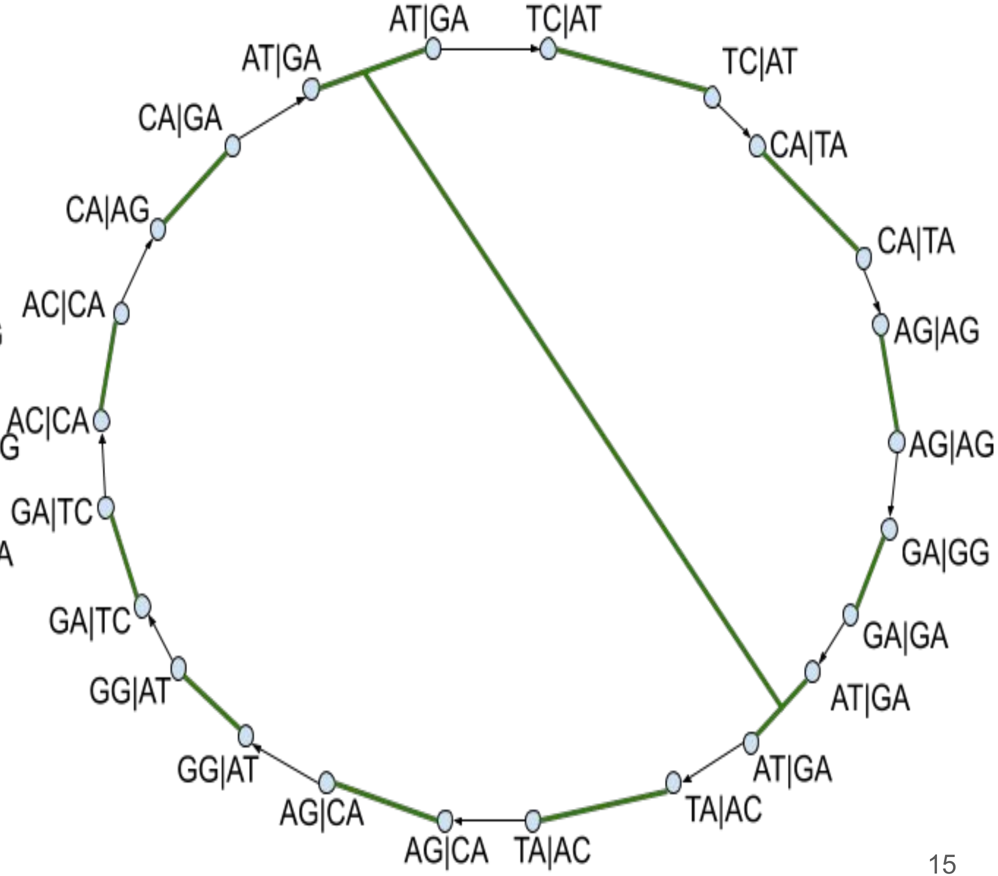


Fig 7d. Connect all the identical k-mers pairs.

SPAdes

Bankevich et al. 2012.
Journal of Computational Biology.

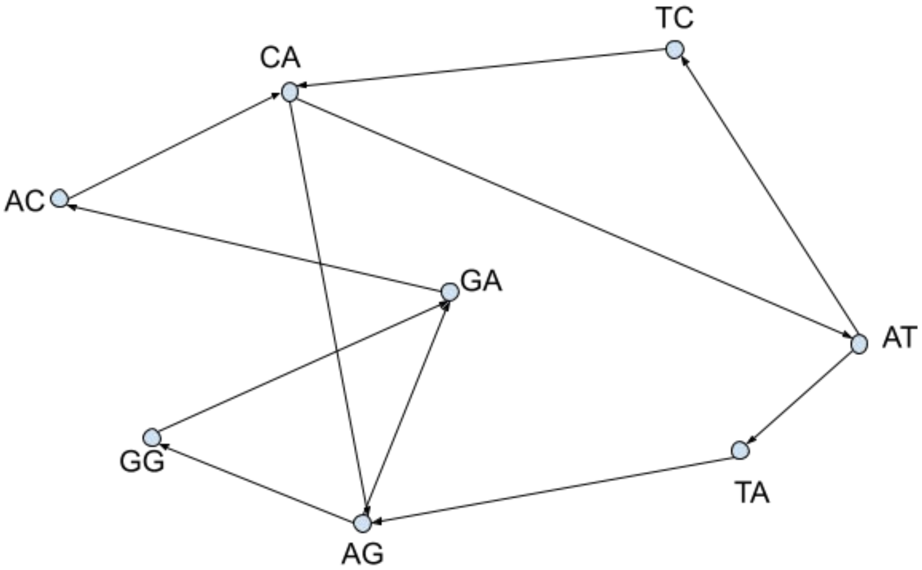


Fig 6e. Simplified DBG.

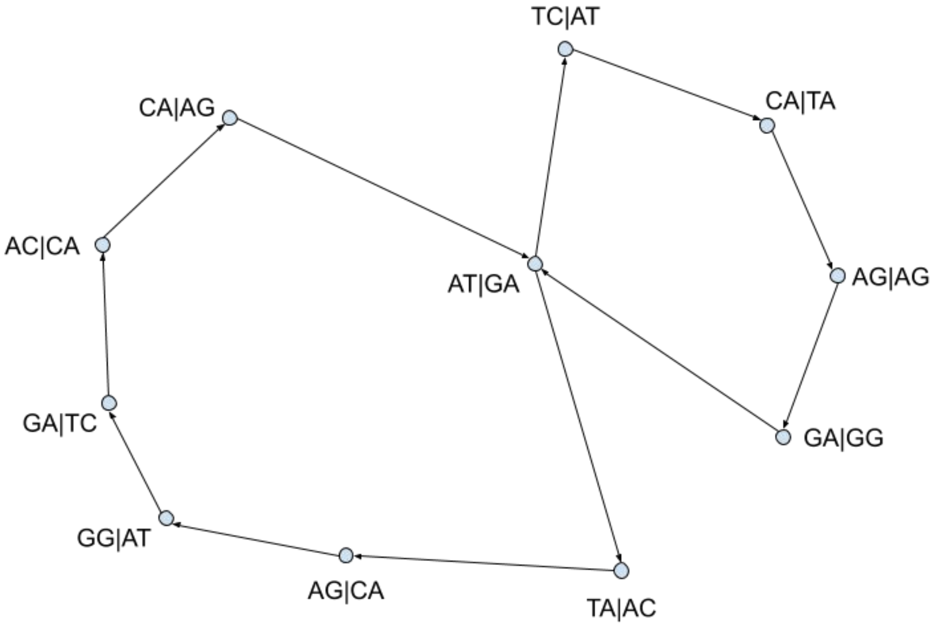


Fig 7e. Simplified paired DBG.

EULER+Velvet-SC

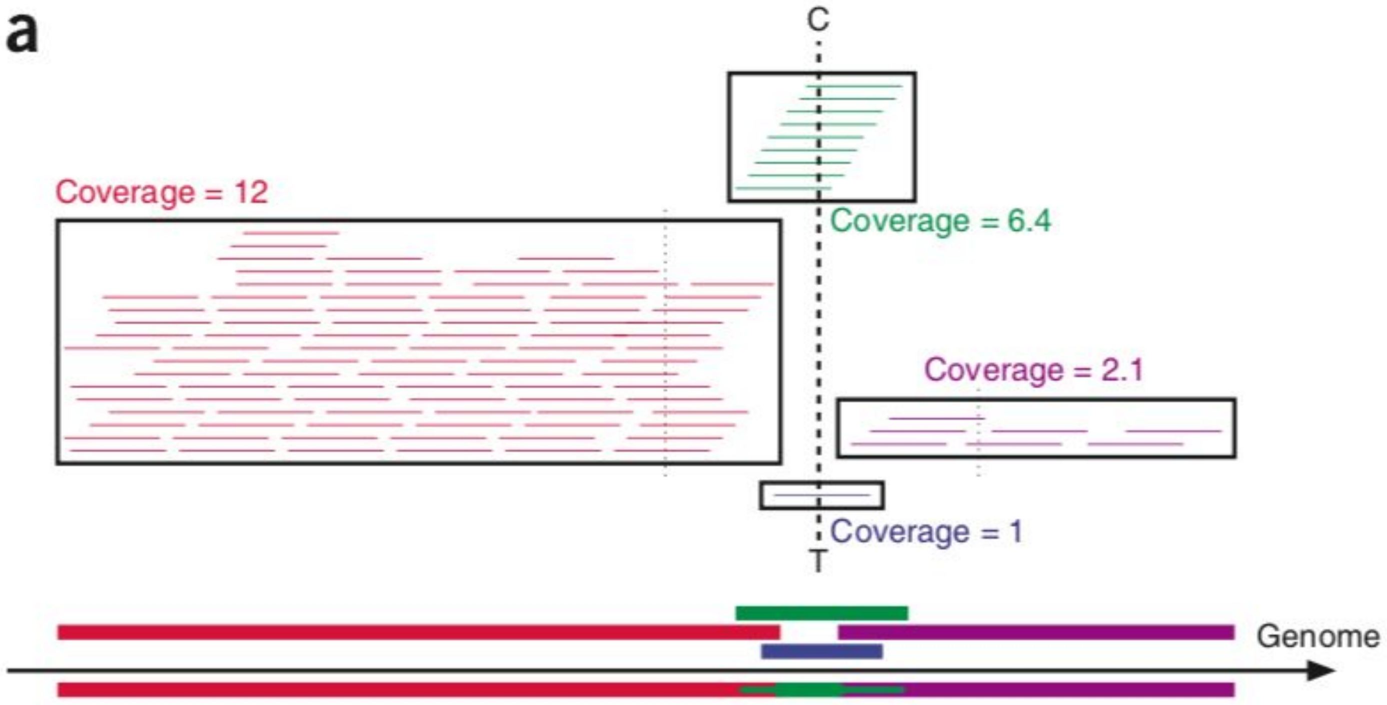


Fig 8a. Uneven coverage of reads to the genome.

Chitsaz et al. 2011. *Nature Biotechnology.*

EULER+Velvet-SC

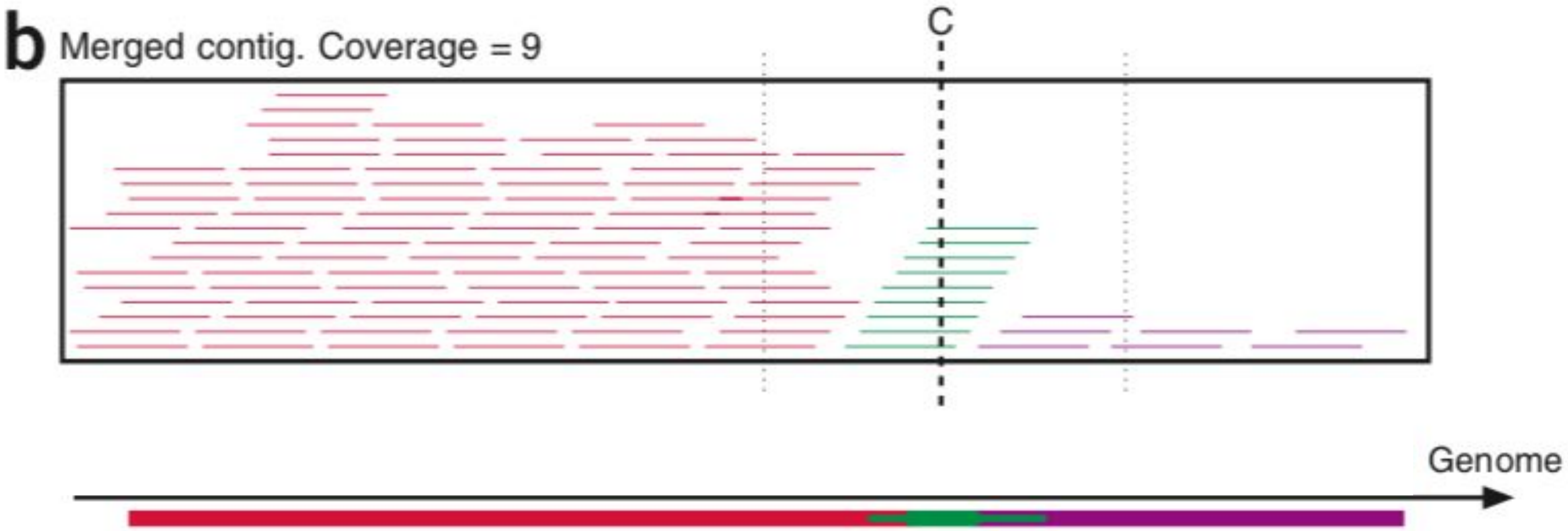


Fig 8b. Merged contig has 9x coverage.

Chitsaz et al. 2011. *Nature Biotechnology.*

Applications

1. Classifying cell types
2. Delineating population diversity
3. Tracing cell lineages
4. Genomic profiling of rare cells

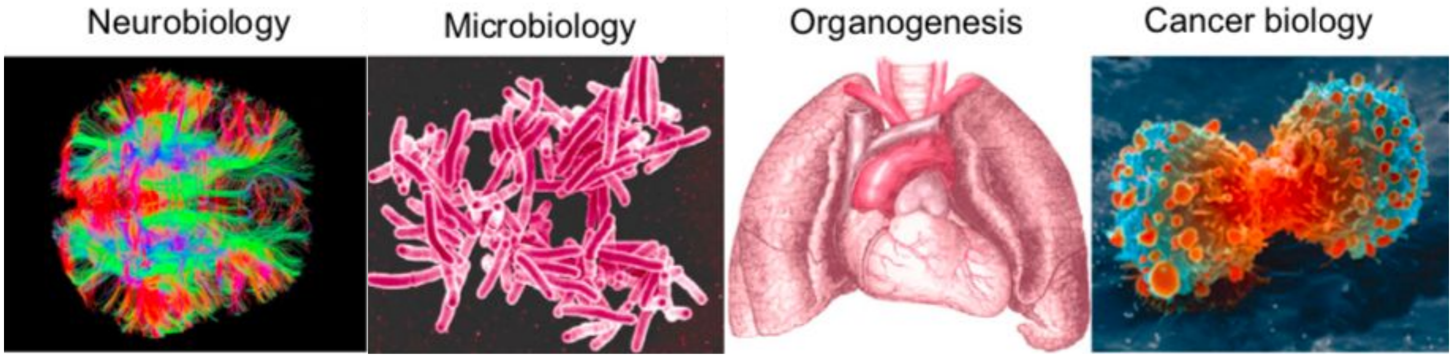
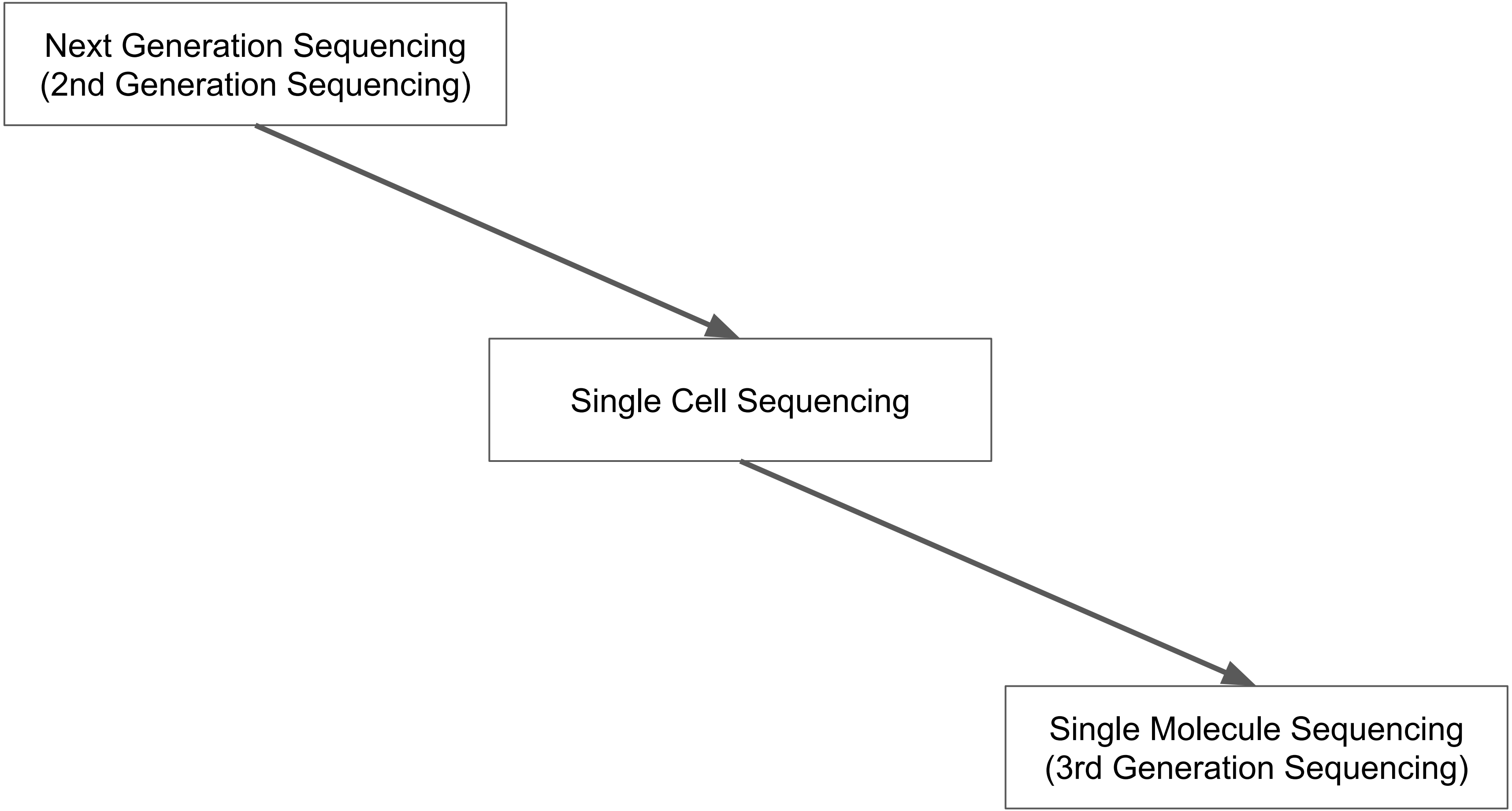


Fig 9. Various applications of SCS across many fields.

Wang et al. 2015. *Molecular Cell.*



Next Generation Sequencing (2nd Generation Sequencing)

Single Cell Sequencing

Single Molecule Sequencing (3rd Generation Sequencing)

Fig 10. The progression of sequencing.

3rd Generation Sequencing

- Characterized by:
  - Single molecule sequencing (SMS)
  - Sequencing in real time

Fig 11. Third generation sequencing technologies.

Dijk et al. 2018. *Trends in Genetics.*

PacBio: SMRT Sequencing

- SMRT: Single Molecule Real Time

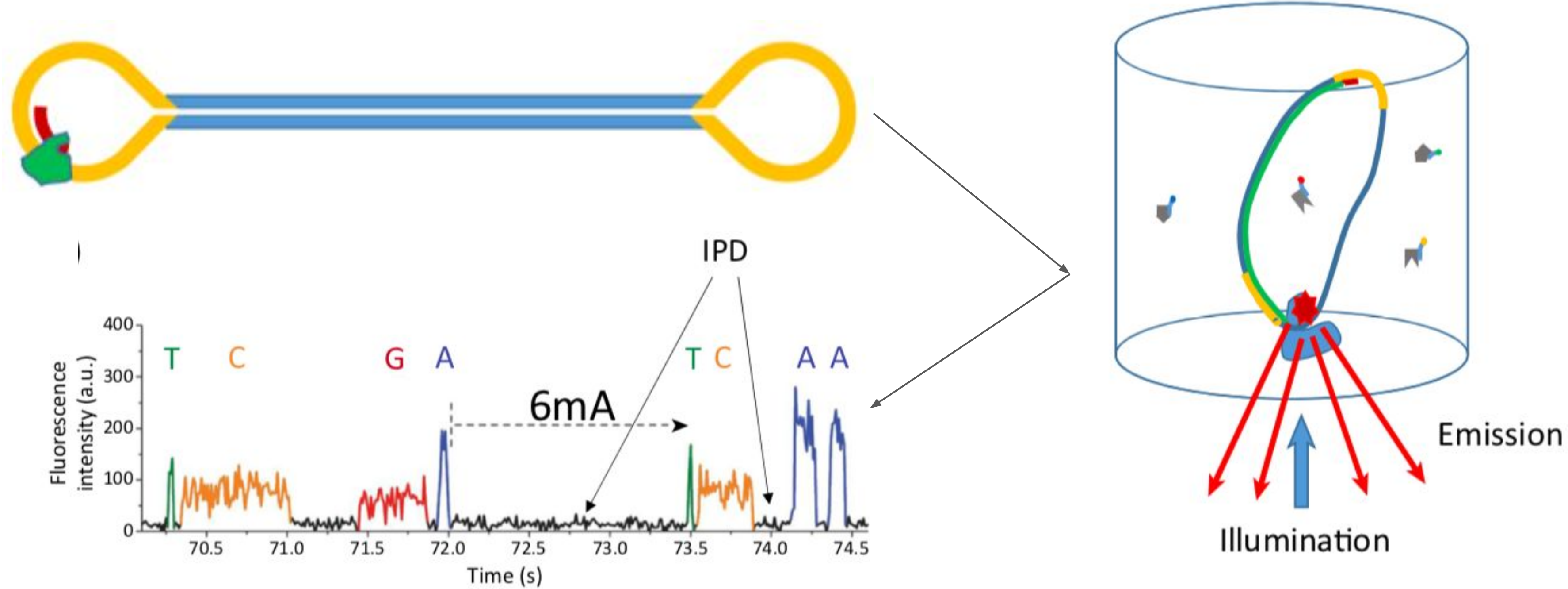


Fig 12. The process of PacBio’s SMRT sequencing.

Dijk et al. 2018. *Trends in Genetics.*

Oxford Nanopore Technologies: ONT Reads

- Nanopore sequencing

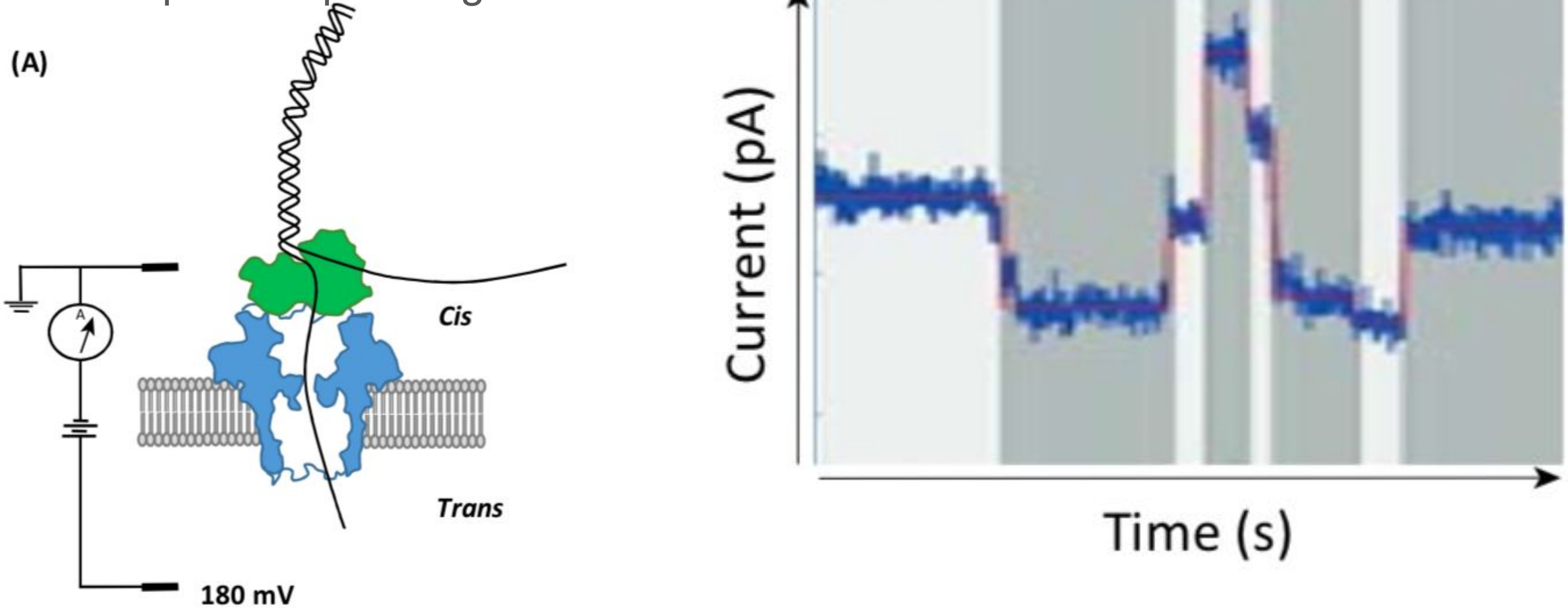
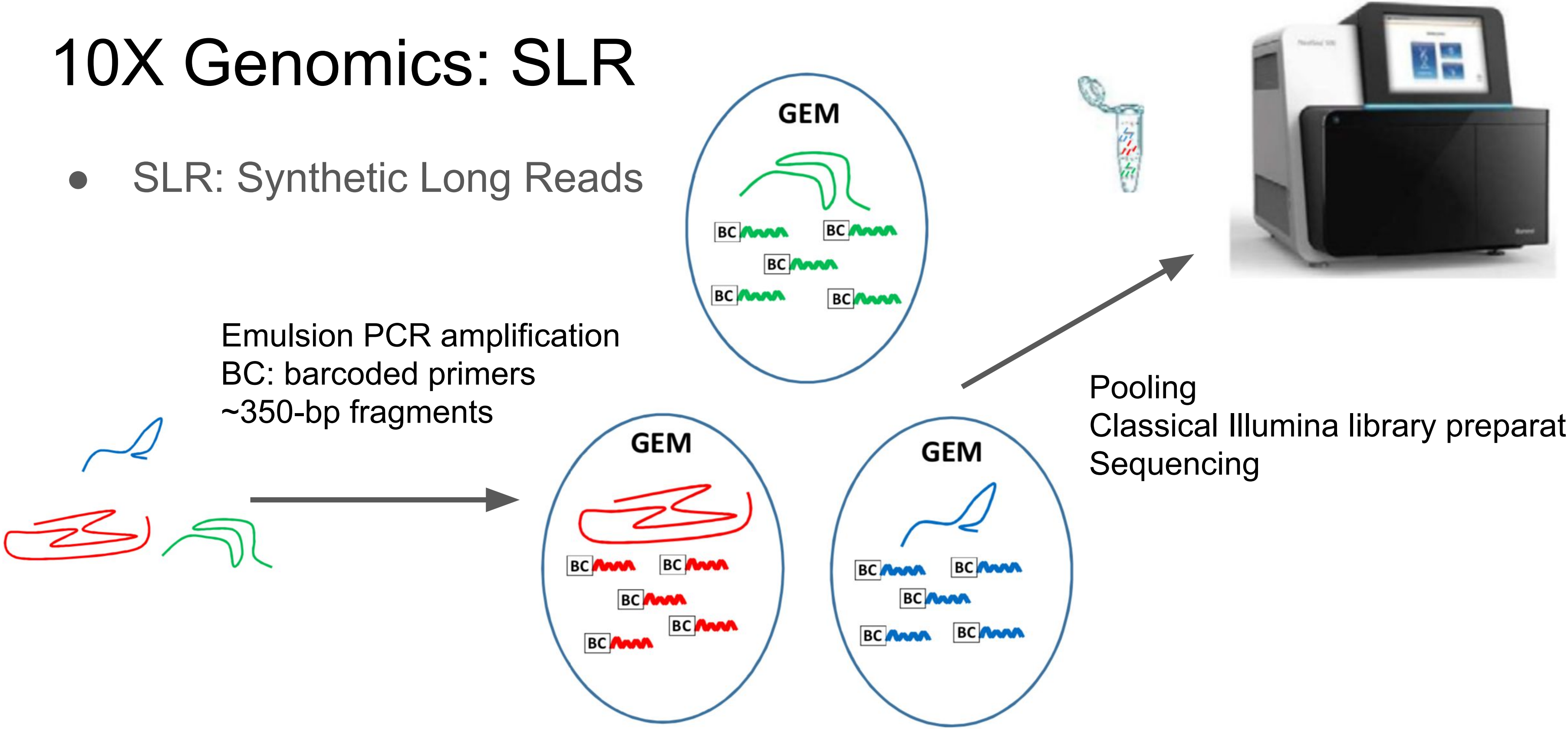


Fig 13. The process of Nanopore sequencing.

Dijk et al. 2018. *Trends in Genetics.*

10X Genomics: SLR

- SLR: Synthetic Long Reads



Emulsion PCR amplification
BC: barcoded primers
~350-bp fragments

Pooling
Classical Illumina library preparation
Sequencing

Fig 14. The process of SLR library prep.

Dijk et al. 2018. *Trends in Genetics.*

Long Read Assemblers

- Canu (PacBio or ONT)
    - Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. (2017).
- Flye (PacBio or ONT)
    - Yu Lin, Jeffrey Yuan, Mikhail Kolmogorov, Max W Shen, Mark Chaisson and Pavel Pevzner, "Assembly of Long Error-Prone Reads Using de Bruijn Graphs", PNAS, 2016 doi:10.1073/pnas.1604560113
- Minimap/Miniasm (PacBio or ONT)
    - Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. (2016).
- Wtdbg2 (PacBio or ONT)
    - Ruan, J. and Li, H. (2019) Fast and accurate long-read assembly with wtdbg2. bioRxiv. doi:10.1101/530972
- Falcon (PacBio)

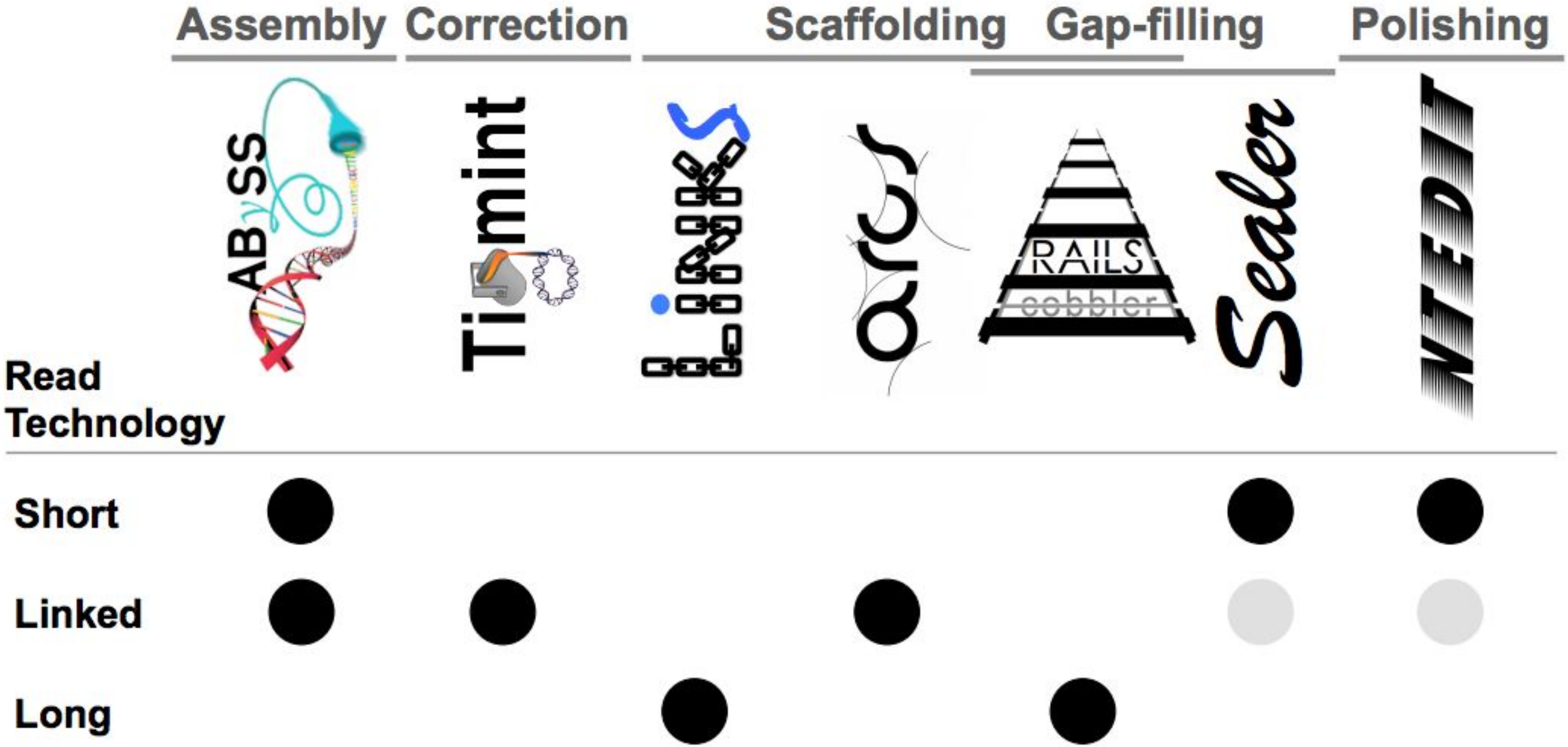
Jung et al. 2019. *Trends in Plant Science.*

Hybrid Assemblers

- DBG2OLC
    - Ye, C. et al. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci. Rep. 6, 31900; doi: 10.1038/srep31900 (2016).
- MaSuRCA
    - Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Yorke JA, Dvorak J, Salzberg S. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the mega-reads algorithm. Genome Research. 2017 Jan 1:066100
- Unicycler
    - Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017.

Jung et al. 2019. *Trends in Plant Science.*

Combining short- and long- read data for assembly



Read Technology	ABySS	Tigmint	LINKS	arcs	RAILS cobbler	Sealer	ntEdit
Short	●					●	●
Linked	●	●		●		○	○
Long			●		●		

Fig 15. Various tools using different read technology in an assembly pipeline. http://github.com/bcgsc

Conclusions

- Single cell sequencing is a technology applicable across various fields with many applications
- There is a need for more bioinformatics tools to filter out technical noise when conducting single cell data analysis
- Third generation sequencing has allowed for the resolution of many genomes with large repetitive elements in both standard (multi-cell) and single-cell studies
- The future of genome assembly is in hybrid assemblies, where the short- and long- read assemblers complement one another’s deficits

Questions?

References

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. 2012. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *J Comput Biol* 19 (5):455-77. doi: 10.1089/cmb.2012.0021.

Chitsaz, H., J. L. Yee-Greenbaum, G. Tesler, M. J. Lombardo, C. L. Dupont, J. H. Badger, M. Novotny, D. B. Rusch, L. J. Fraser, N. A. Gormley, O. Schulz-Trieglaff, G. P. Smith, D. J. Evers, P. A. Pevzner, and R. S. Lasken. 2011. "Efficient de novo assembly of single-cell bacterial genomes from short-read data sets." *Nat Biotechnol* 29 (10):915-21. doi: 10.1038/nbt.1966.

Garvin, Tyler, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S. Atwal, James Hicks, Michael Wigler, and Michael C. Schatz. 2015. "Interactive analysis and assessment of single-cell copy-number variations." *Nature methods* 12 (11):1058-1060. doi: 10.1038/nmeth.3578.

Jayakumar, Vasanthan, and Yasubumi Sakakibara. 2017. "Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data." *Briefings in Bioinformatics* 20 (3):866-876. doi: 10.1093/bib/bbx147.

References

Jung, H., C. Winefield, A. Bombarely, P. Prentis, and P. Waterhouse. 2019. "Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes." *Trends Plant Sci* 24 (8):700-724. doi: 10.1016/j.tplants.2019.05.003.

Lasken, R. S., and T. B. Stockwell. 2007. "Mechanism of chimera formation during the Multiple Displacement Amplification reaction." *BMC Biotechnol* 7:19. doi: 10.1186/1472-6750-7-19.

Navin, Nicholas E. 2014. "Cancer genomics: one cell at a time." *Genome biology* 15 (8):452-452. doi: 10.1186/s13059-014-0452-9.

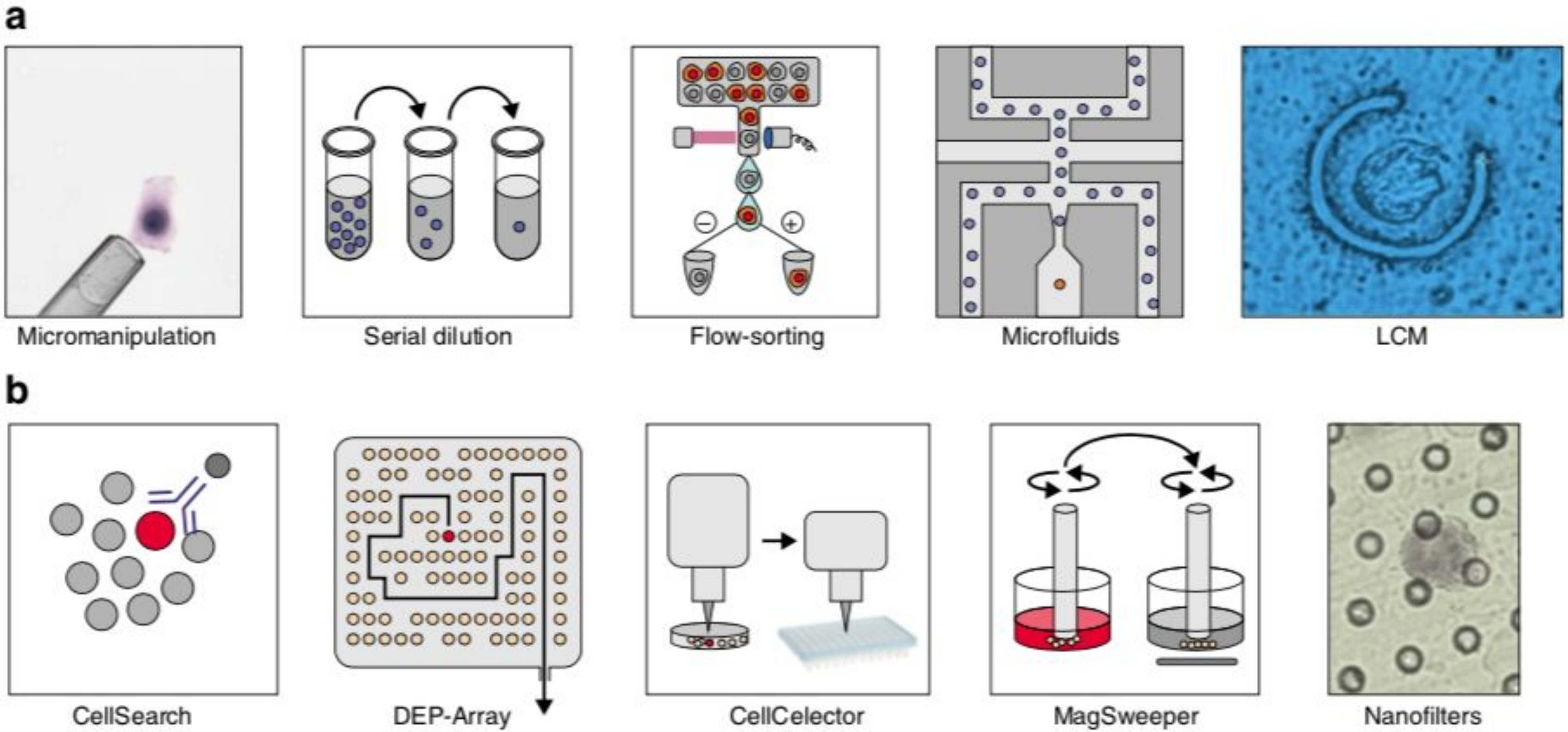
Peng, Y., H. C. Leung, S. M. Yiu, and F. Y. Chin. 2012. "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." *Bioinformatics* 28 (11):1420-8. doi: 10.1093/bioinformatics/bts174.

van Dijk, E. L., Y. Jaszczyszyn, D. Naquin, and C. Thermes. 2018. "The Third Revolution in Sequencing Technology." *Trends Genet* 34 (9):666-681. doi: 10.1016/j.tig.2018.05.008.

Wang, Y., and N. E. Navin. 2015. "Advances and applications of single-cell sequencing technologies." *Mol Cell* 58 (4):598-609. doi: 10.1016/j.molcel.2015.05.005.

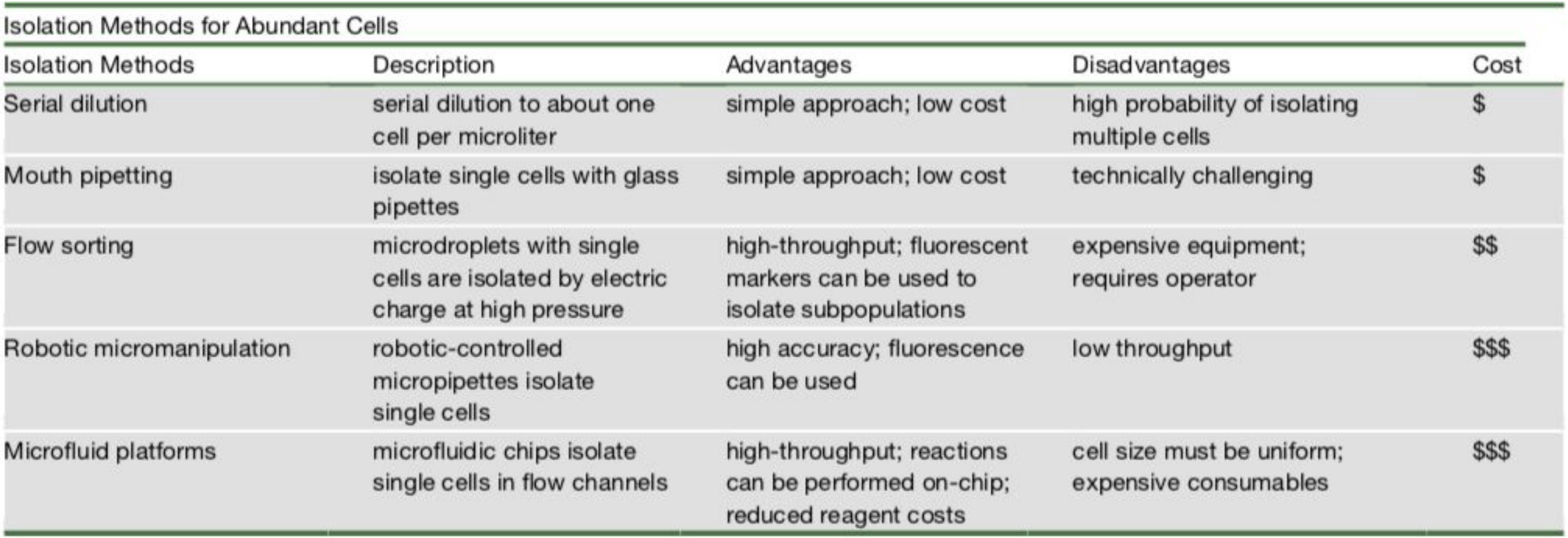
Supplemental Material

Single Cell Isolation



Navin 2014. *Genome Biology*

Isolation Methods for Abundant Cells

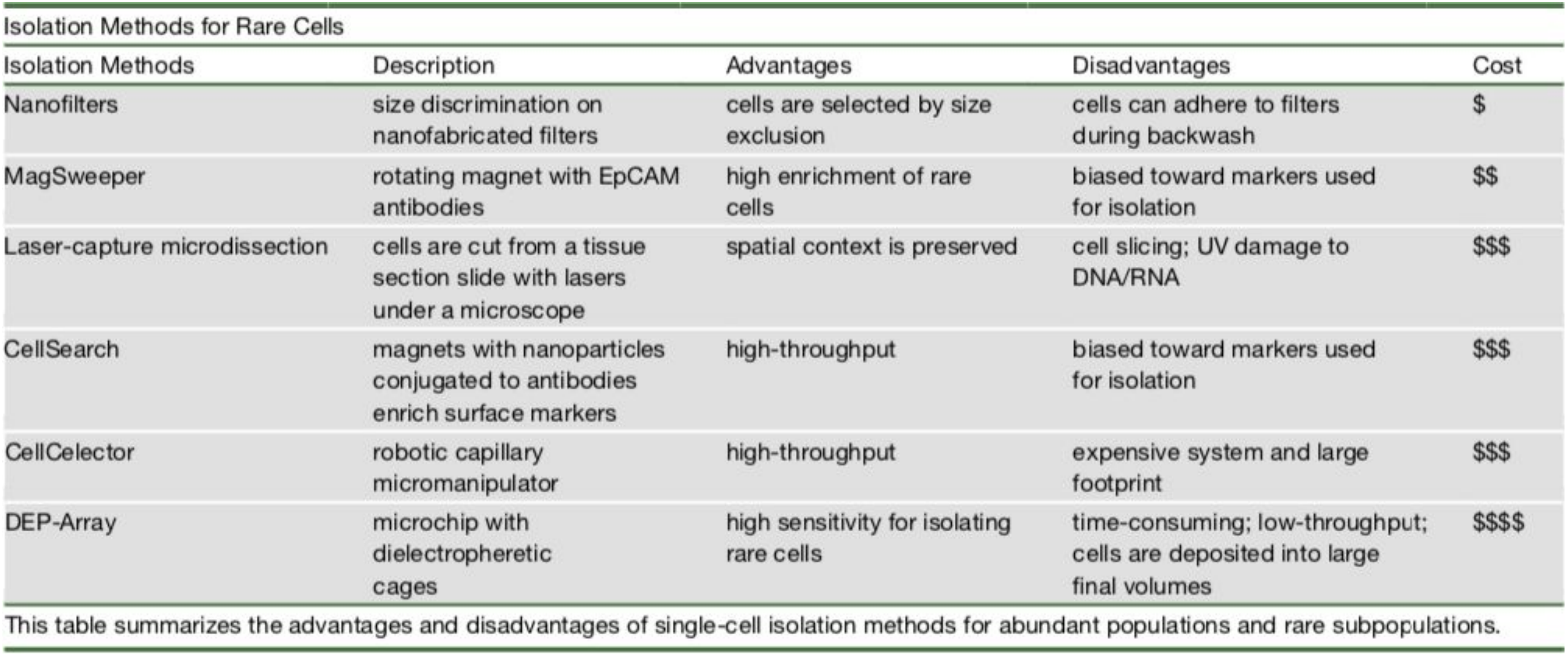


Isolation Methods for Abundant Cells

Isolation Methods	Description	Advantages	Disadvantages	Cost
Serial dilution	serial dilution to about one cell per microliter	simple approach; low cost	high probability of isolating multiple cells	$
Mouth pipetting	isolate single cells with glass pipettes	simple approach; low cost	technically challenging	$
Flow sorting	microdroplets with single cells are isolated by electric charge at high pressure	high-throughput; fluorescent markers can be used to isolate subpopulations	expensive equipment; requires operator	$$
Robotic micromanipulation	robotic-controlled micropipettes isolate single cells	high accuracy; fluorescence can be used	low throughput	$$$
Microfluid platforms	microfluidic chips isolate single cells in flow channels	high-throughput; reactions can be performed on-chip; reduced reagent costs	cell size must be uniform; expensive consumables	$$$

Wang et al. 2015 *Molecular Cell.*

Isolation Methods for Rare Cells



Isolation Methods for Rare Cells

Isolation Methods	Description	Advantages	Disadvantages	Cost
Nanofilters	size discrimination on nanofabricated filters	cells are selected by size exclusion	cells can adhere to filters during backwash	$
MagSweeper	rotating magnet with EpCAM antibodies	high enrichment of rare cells	biased toward markers used for isolation	$$
Laser-capture microdissection	cells are cut from a tissue section slide with lasers under a microscope	spatial context is preserved	cell slicing; UV damage to DNA/RNA	$$$
CellSearch	magnets with nanoparticles conjugated to antibodies enrich surface markers	high-throughput	biased toward markers used for isolation	$$$
CellCelector	robotic capillary micromanipulator	high-throughput	expensive system and large footprint	$$$
DEP-Array	microchip with dielectropheretic cages	high sensitivity for isolating rare cells	time-consuming; low-throughput; cells are deposited into large final volumes	$$$$

This table summarizes the advantages and disadvantages of single-cell isolation methods for abundant populations and rare subpopulations.

Wang et al. 2015 *Molecular Cell.*

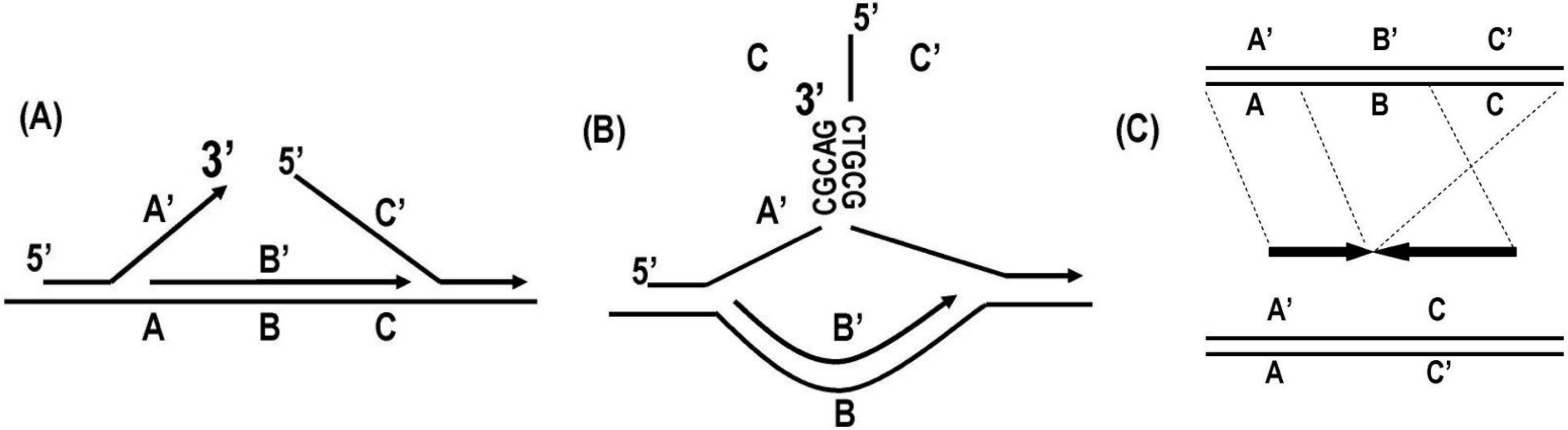
Technical Errors from WGA

Technical Artifact	Amplification Method	Error Type
chimeric molecules	MDA	false-positive inversions
coverage non-uniformity	MDA, DOP-PCR	copy number aberrations, false-negative SNVs
false positive amplification error	MDA, DOP-PCR	SNV, indel
allelic dropout	MDA, DOP-PCR	false-negative errors
pileup regions	DOP-PCR	copy number amplifications

Wang et al. 2015. *Molecular Cell.*

Chimeric Molecules

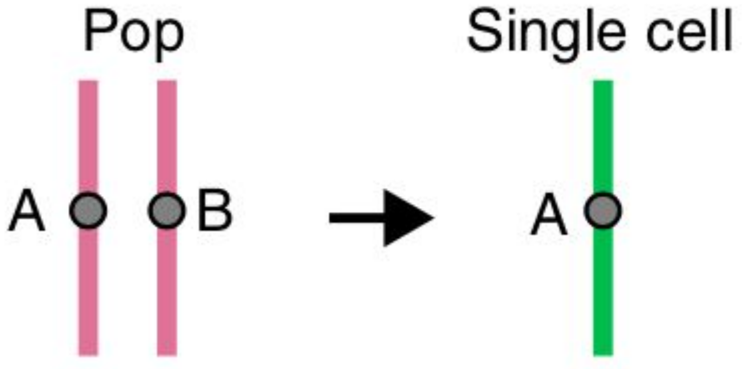
- **Amplification Method:** MDA
- **Error Type:** false-positive inversions
- **Description**: When the 3’ end of a newly synthesized molecule hybridizes with the 5’ end of a newly synthesized molecule causing inversions



Lasken et al. 2007. *BMC Biotechnology*.

Allelic Dropout

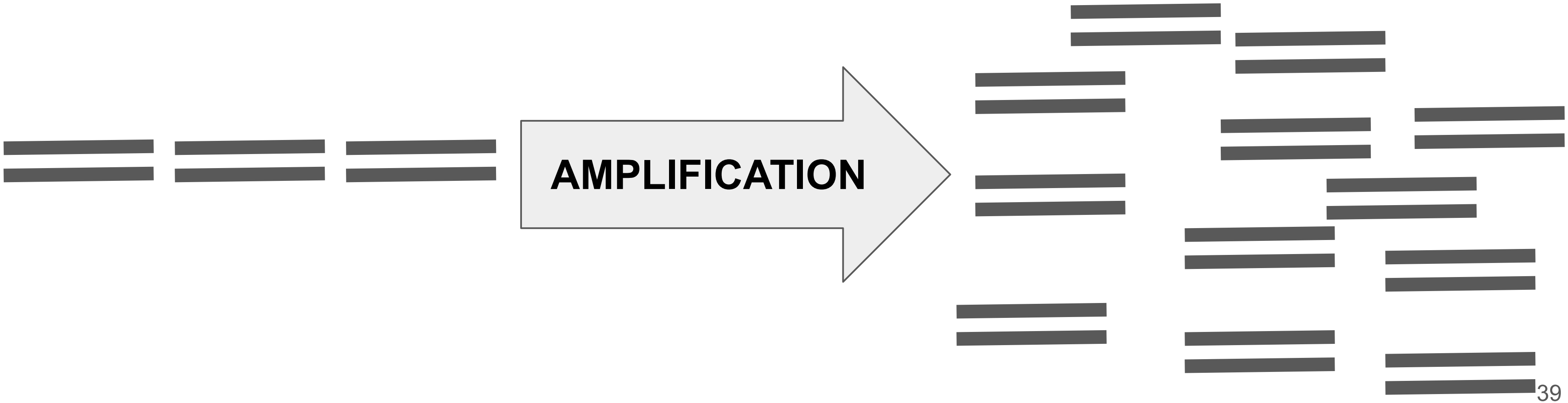
- **Amplification Method:** MDA, DOP-PCR
- **Error Type**: False-negative errors
- **Description**: Heterozygous (AB) variants undergo dropout during WGA leading to homozygous (AA or BB) genotypes



Navin 2014. *Genome Biology*.

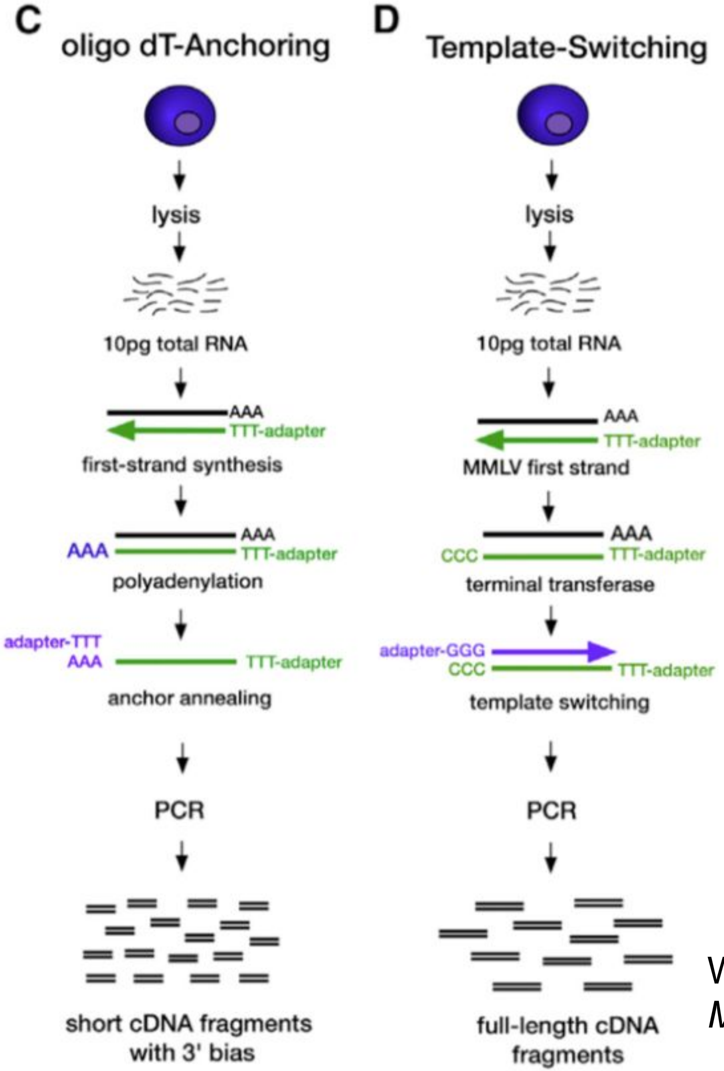
Pileup Regions

- **Amplification Method:** DOP-PCR
- **Error Type**: copy number amplifications
- **Description:** massive over-amplifications of focal genomic regions occur during DOP-PCR



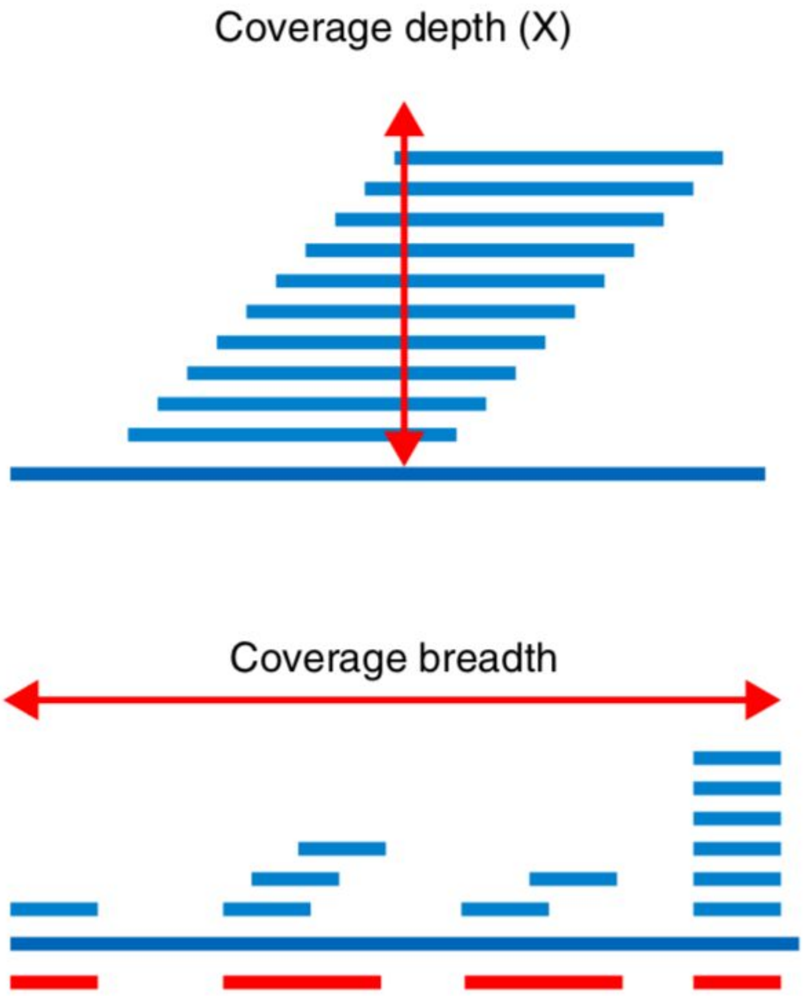
Wang et al. 2015. *Molecular Cell.*

Whole Transcriptome Amplification (WTA)



Wang, et. al. 2015 *Molecular Cell.*

Coverage



Navin 2014. *Genome Biology*.

Technical Errors from WTA

Technical Artifact	Amplification Method	Error Type	Description
amplification distortion	dt-anchor, Template-Switching	erroneous expression values	over/under amplification during WTA leads to erroneous expression values
transcript dropout	dt-anchor, Template-Switching, UMI	false-negative unexpressed genes	failure to amplify a transcript during WTA
3’ bias	dt-anchors	failure of RT polymerase to fully synthesize the first cDNA strand	Strong bias toward amplification of 3’ end of RNA transcripts

Wang et al. 2015. *Molecular Cell.*

Challenges in Filtering Technical Noise

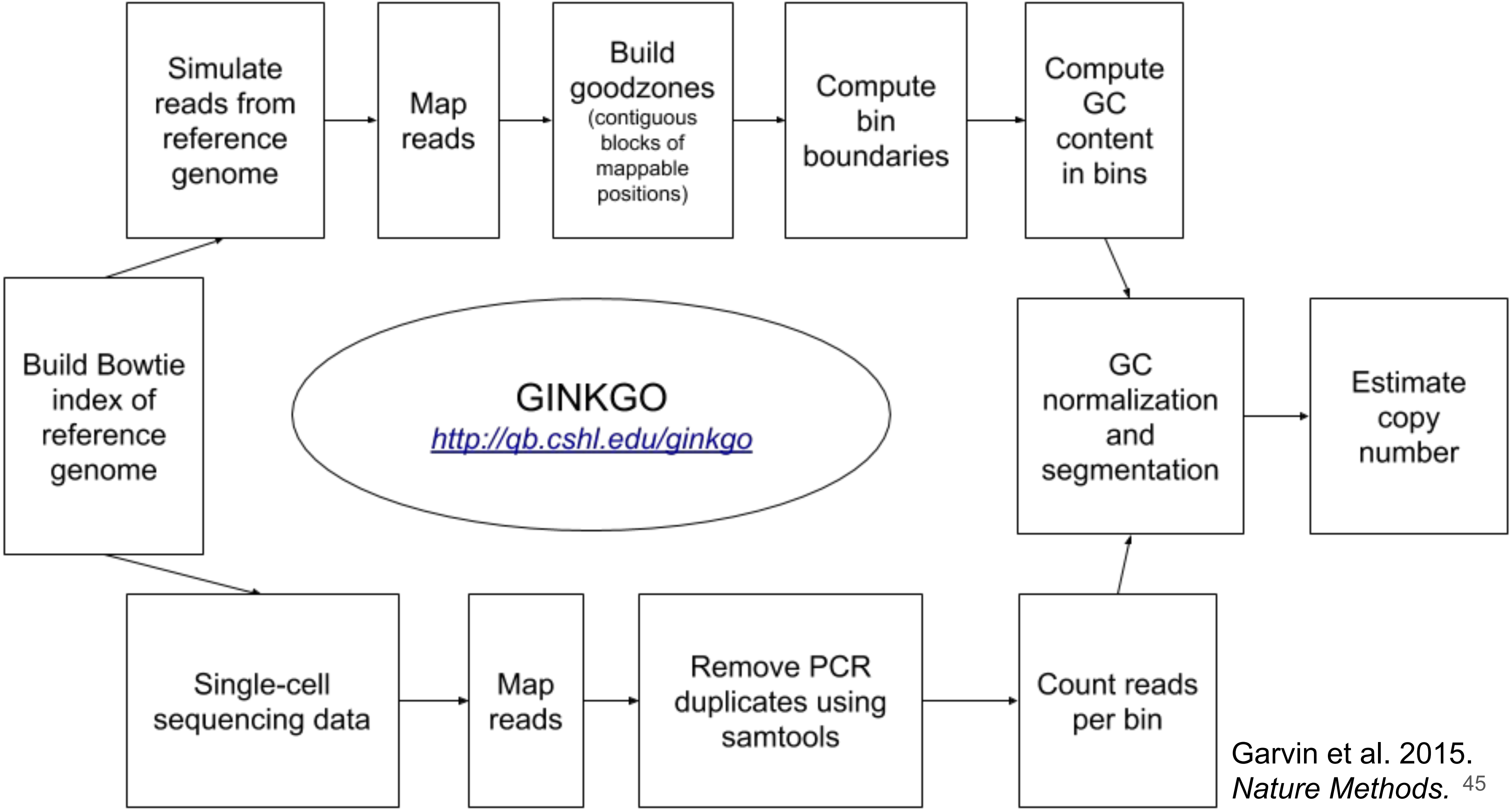
- Copy number aberrations
  - **Reference genome** is required
  - Use reference genome to make simulated reads
- Coverage non-uniformity
  - Adjust coverage cut-off threshold
- SNVs
  - **Reference genome** is required
  - Alignment to the reference genome

Wang et al. 2015. *Molecular Cell.*

Ginkgo

- Quantifies single cell copy number profiles from read count data
- A variable-binning algorithm
  - Normalizes errors in mappability
  - Change bin size based on expected number of reads
  - Requires a **reference genome**

Wang et al. 2015. *Molecular Cell.*



Build Bowtie index of reference genome

Simulate reads from reference genome

Map reads

Build goodzones (contiguous blocks of mappable positions)

Compute bin boundaries

Compute GC content in bins

GC normalization and segmentation

Estimate copy number

Single-cell sequencing data

Map reads

Remove PCR duplicates using samtools

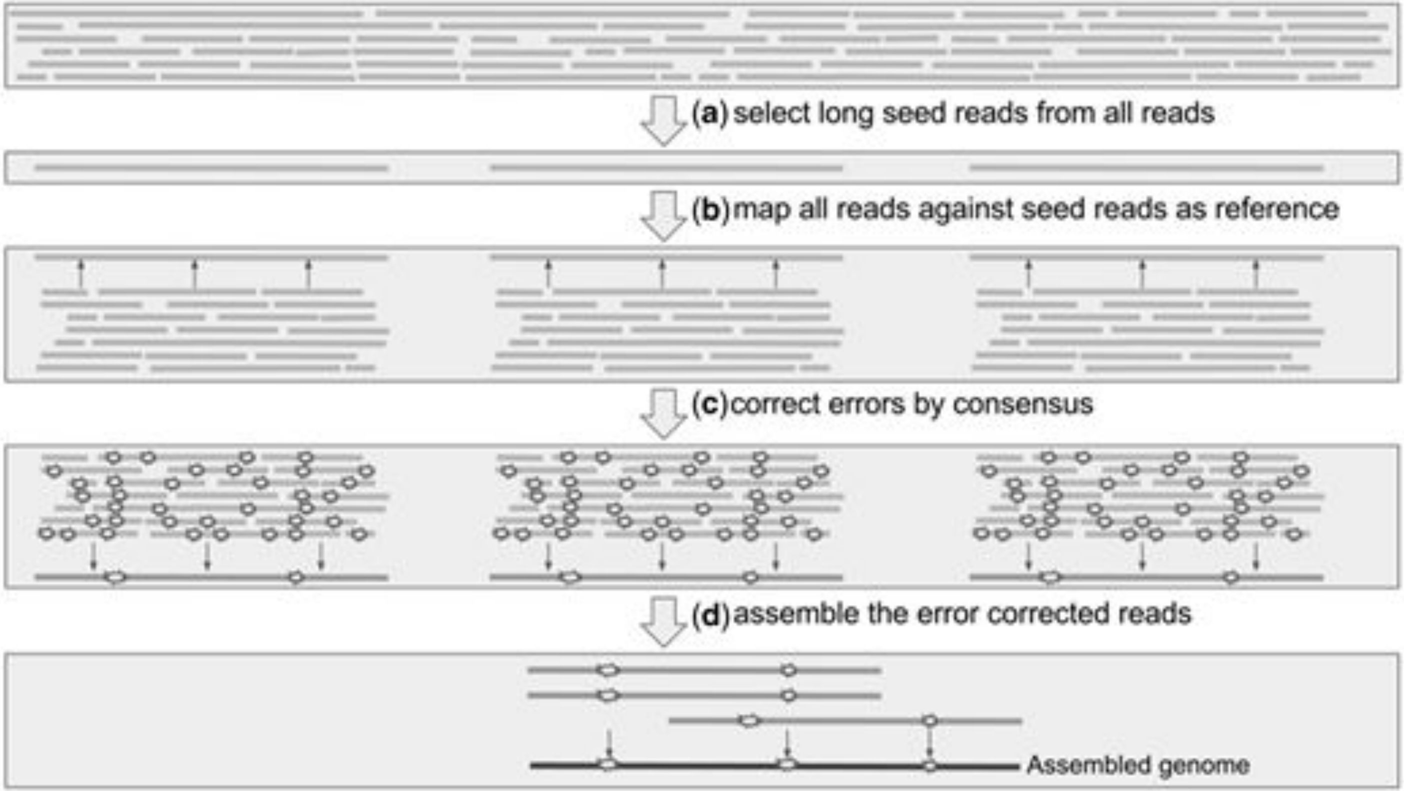
Count reads per bin

GINKGO

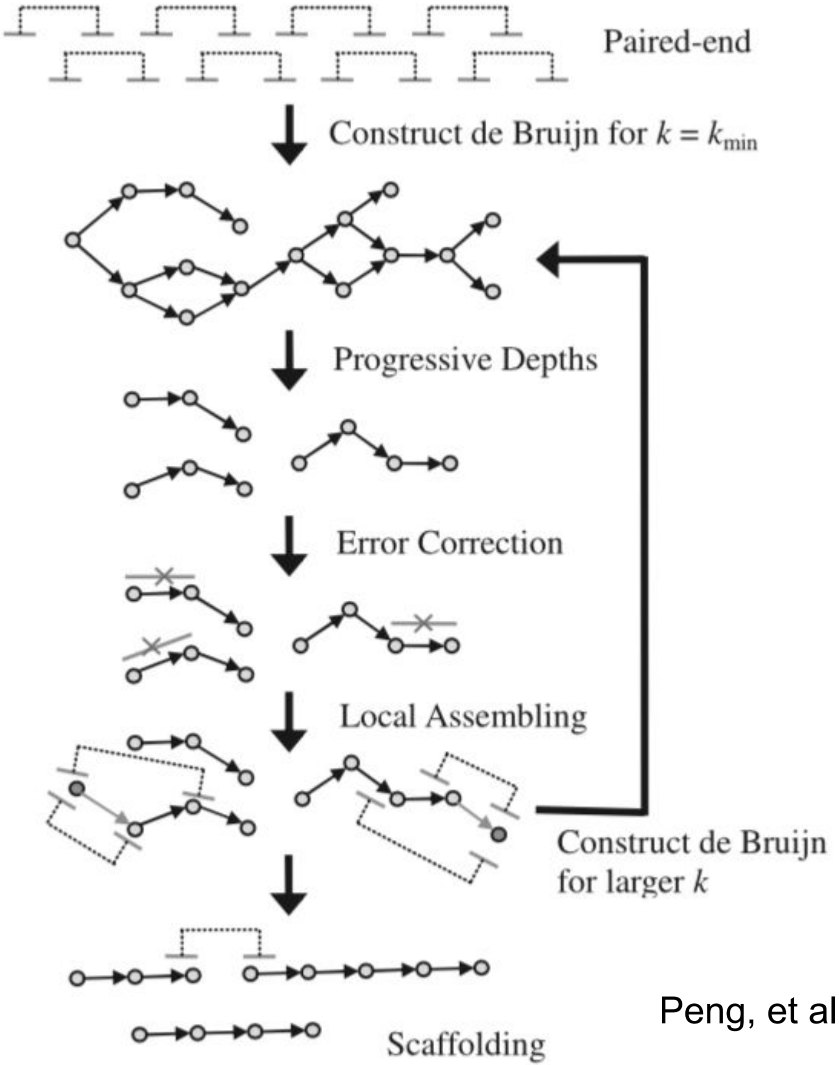
http://qb.cshl.edu/ginkgo

Garvin et al. 2015. *Nature Methods.*

Overlap Consensus Graphs


(a) select long seed reads from all reads
(b) map all reads against seed reads as reference
(c) correct errors by consensus
(d) assemble the error corrected reads
Assembled genome

IDBA-UD


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

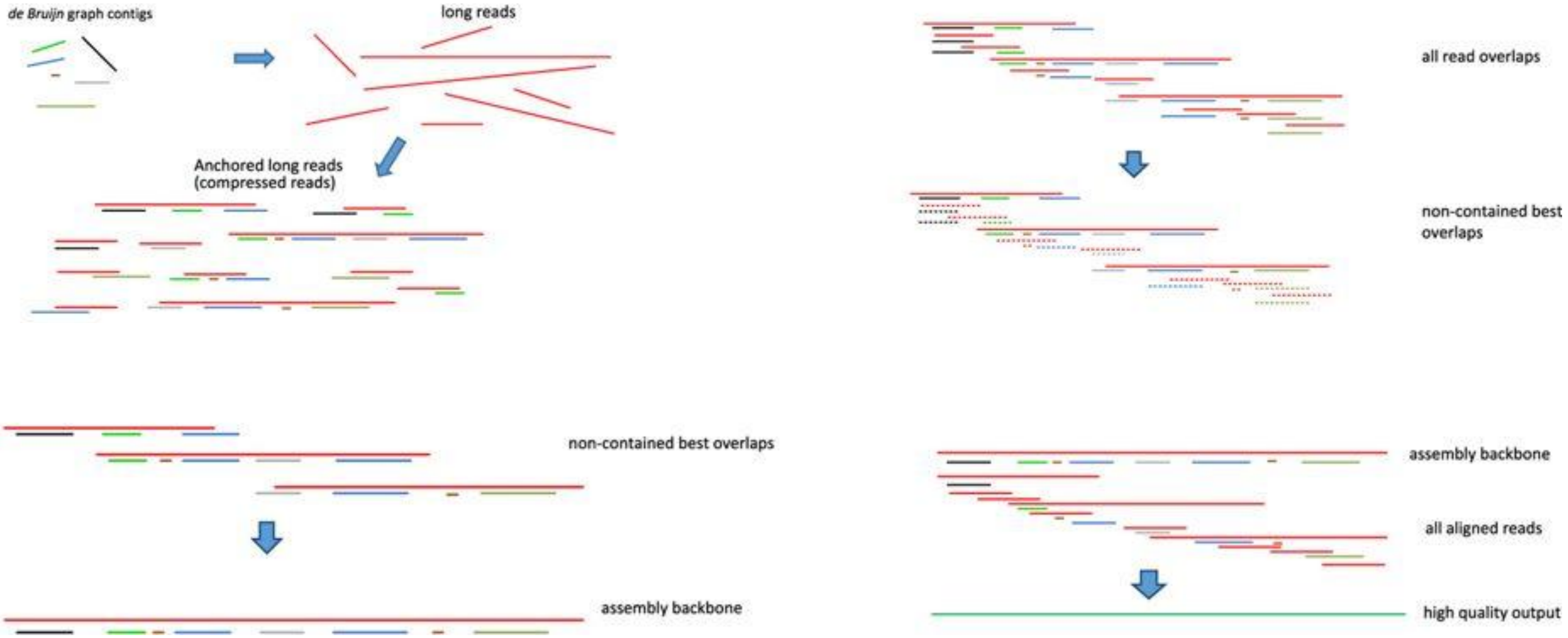
Peng, et al. 2012. *Bioinformatics.*

SPAdes

Stage 1	Assembly graph construction using *the multisized de Bruijn graph*, implementing new bulge/tip removal algorithms, detection/removal of chimeric reads, construction of *distance histograms*, backtracking of performed graph operations
Stage 2	Derivation of accurate distance estimates between *k*-mers in the genome using joint analysis of distance histograms and paths in assembly graph
Stage 3	Construction of *paired assembly graph*
Stage 4	Construction of DNA sequences of contigs and the mapping of reads to contigs by backtracking graph simplifications

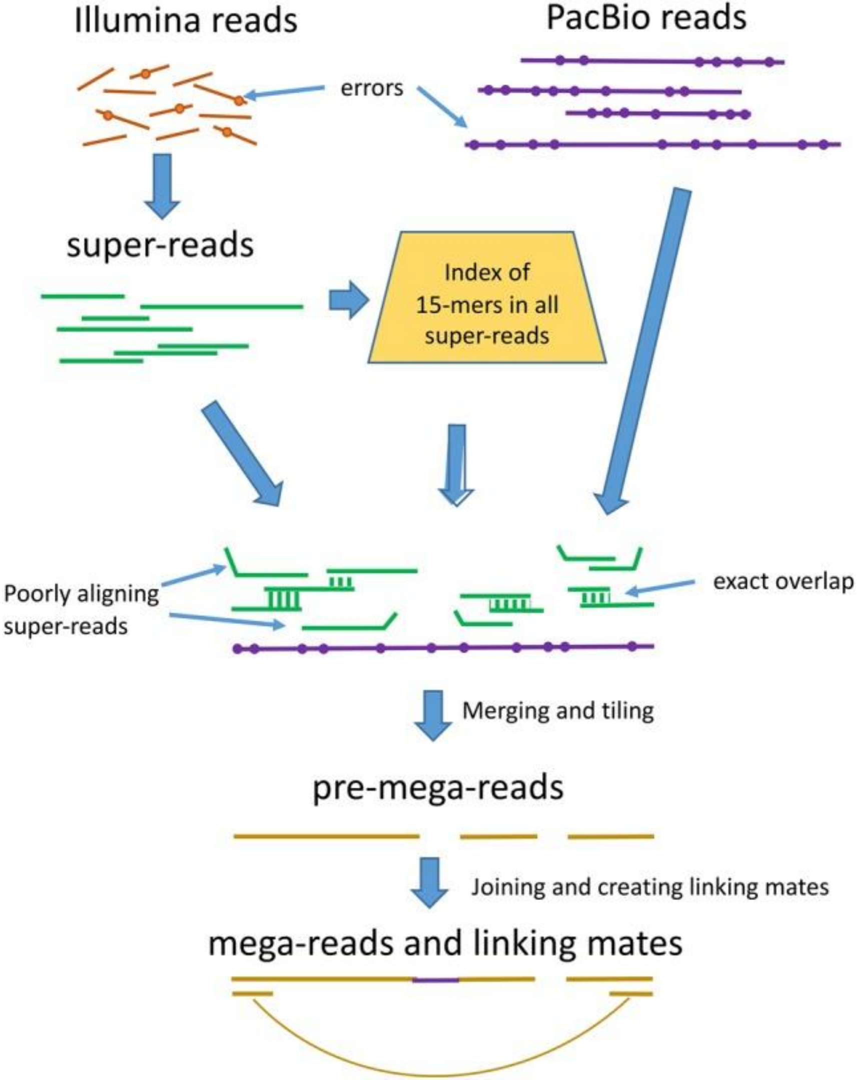
Bankevich et al. 2012.
Journal of Computational Biology.

DBG2OLC



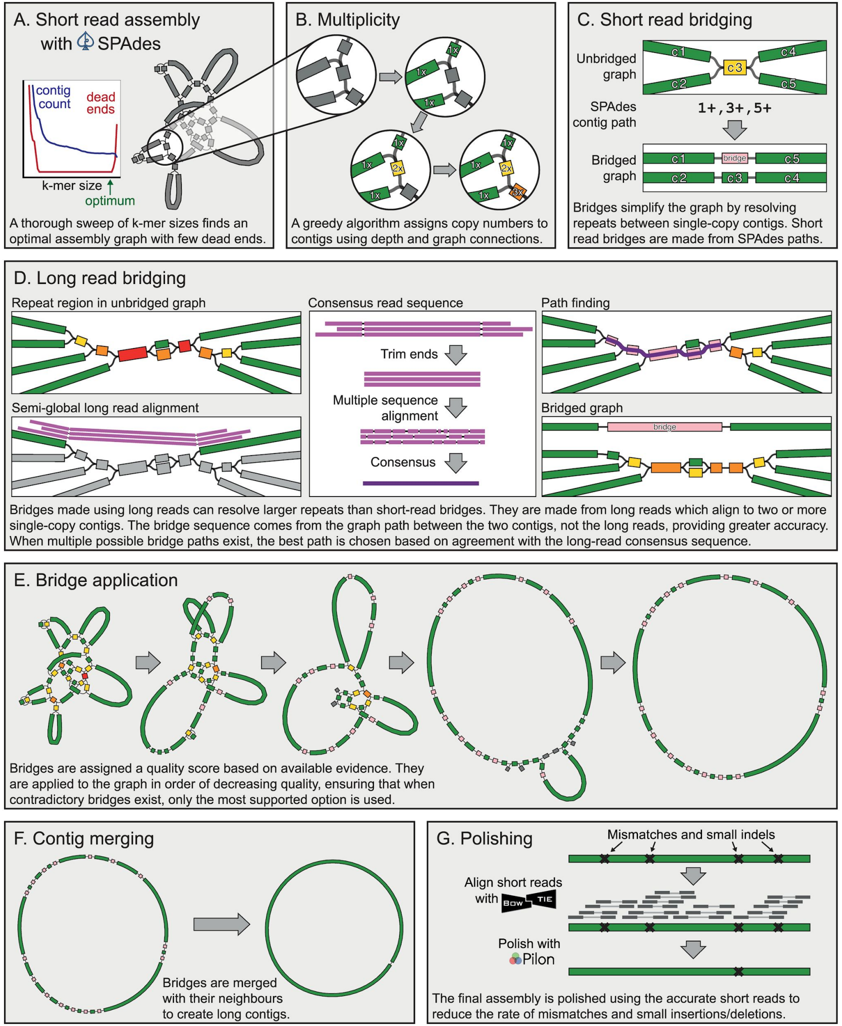
Ye et al. 2016. *Scientific Reports.*

MaSuRCA


Illumina reads
PacBio reads
errors
super-reads
Index of 15-mers in all super-reads
Poorly aligning super-reads
exact overlap
Merging and tiling
pre-mega-reads
Joining and creating linking mates
mega-reads and linking mates

Zimin et al. 2017. *Genome Research.*

Unicycler



A. Short read assembly with SPAdes

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

B. Multiplicity

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

C. Short read bridging

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

D. Long read bridging

Repeat region in unbridged graph

Semi-global long read alignment

Consensus read sequence

Path finding

Bridged graph

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.

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

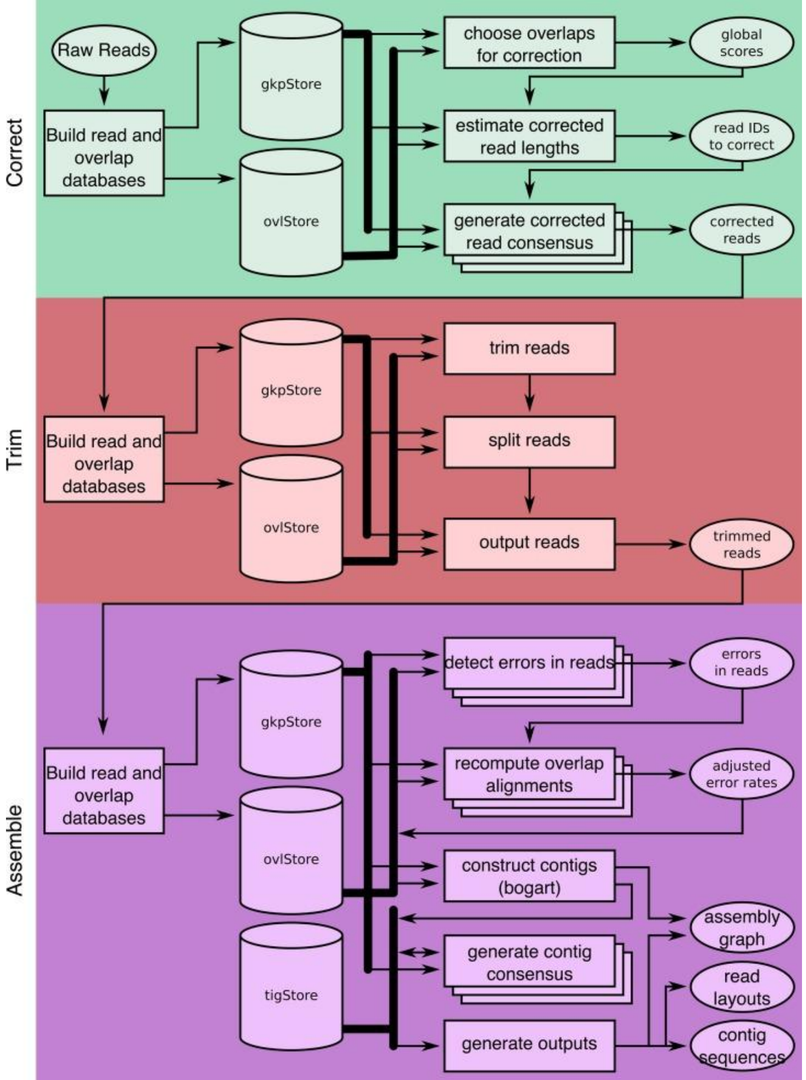
Bridges are merged with their neighbours to create long contigs.

G. Polishing

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

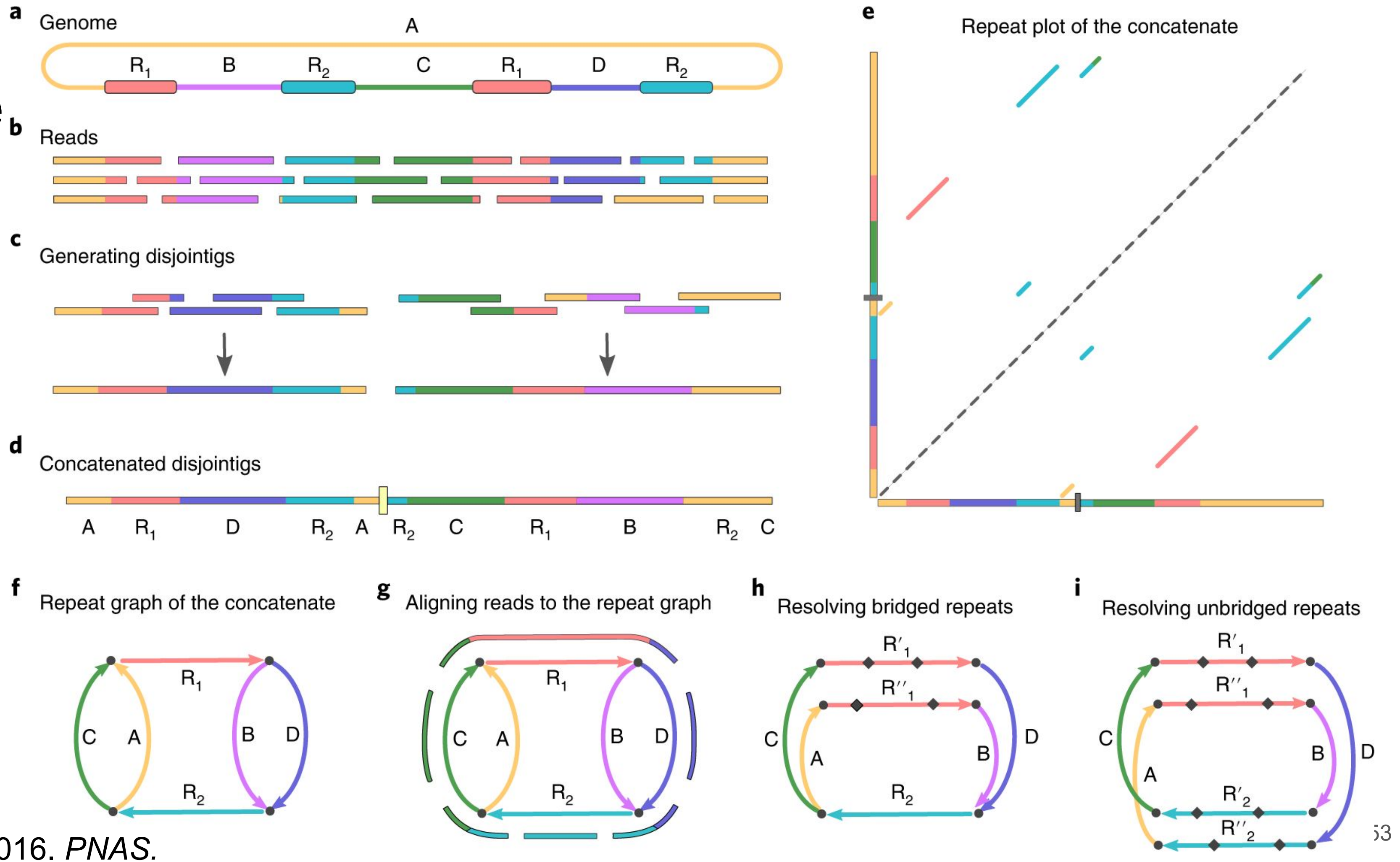
Wick et al. 2017. *PLoS Computational Biology.*

Canu



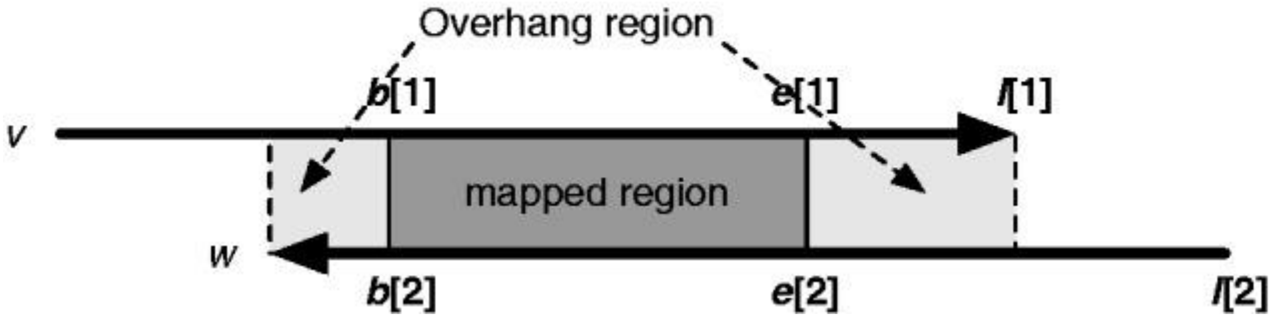
Koren et al. 2017. *Genome Research.*

Flye


a
Genome
A
R1
B
R2
C
R1
D
R2
b
Reads
c
Generating disjointigs
d
Concatenated disjointigs
A
R1
D
R2
A
R2
C
R1
B
R2
C
e
Repeat plot of the concatenate
f
Repeat graph of the concatenate
R1
C
A
B
D
R2
g
Aligning reads to the repeat graph
R1
C
A
B
D
R2
h
Resolving bridged repeats
R′1
R′′1
C
A
B
D
R2
i
Resolving unbridged repeats
R′1
R′′1
C
A
B
D
R′2
R′′2

Lin et al. 2016. *PNAS.*

minimap/miniasm



Li et al. 2016. *Bioinformatics.*