Reviews in
Computational Biology

Phylogeny-guided Genome Assembly

Christophe Dessimoz

May 9th, 2011

ETH Zürich inf | Informatik Computer Science

Outline

- **Background on Genome Assembly**
    - next generation sequencing
    - comparative assembly
    - de-novo assembly
    - read mapping

Outline

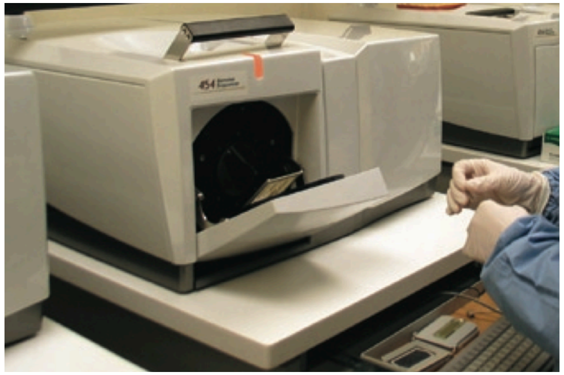
- **Background on Genome Assembly**
    - next generation sequencing
    - comparative assembly
    - de-novo assembly
    - read mapping
- **Phylogeny-based Genome Assembly**
    - Multiple reference genomes
    - Gene Library
    - Meta assembly
    - Comparative genomics

Outline

- **Background on Genome Assembly**
  - next generation sequencing
  - comparative assembly
  - de-novo assembly
  - read mapping
- **Phylogeny-based Genome Assembly**
  - Multiple reference genomes
  - Gene Library
  - Meta assembly
  - Comparative genomics
- **Perspectives**

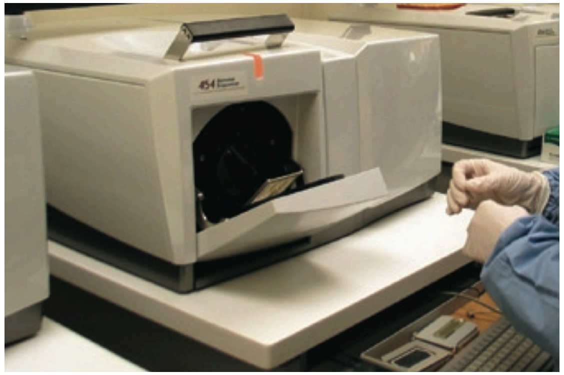
Key Point

We observe the emergence of new type of methods for genome assembly based on multiple reference genomes in their phylogenetic context.



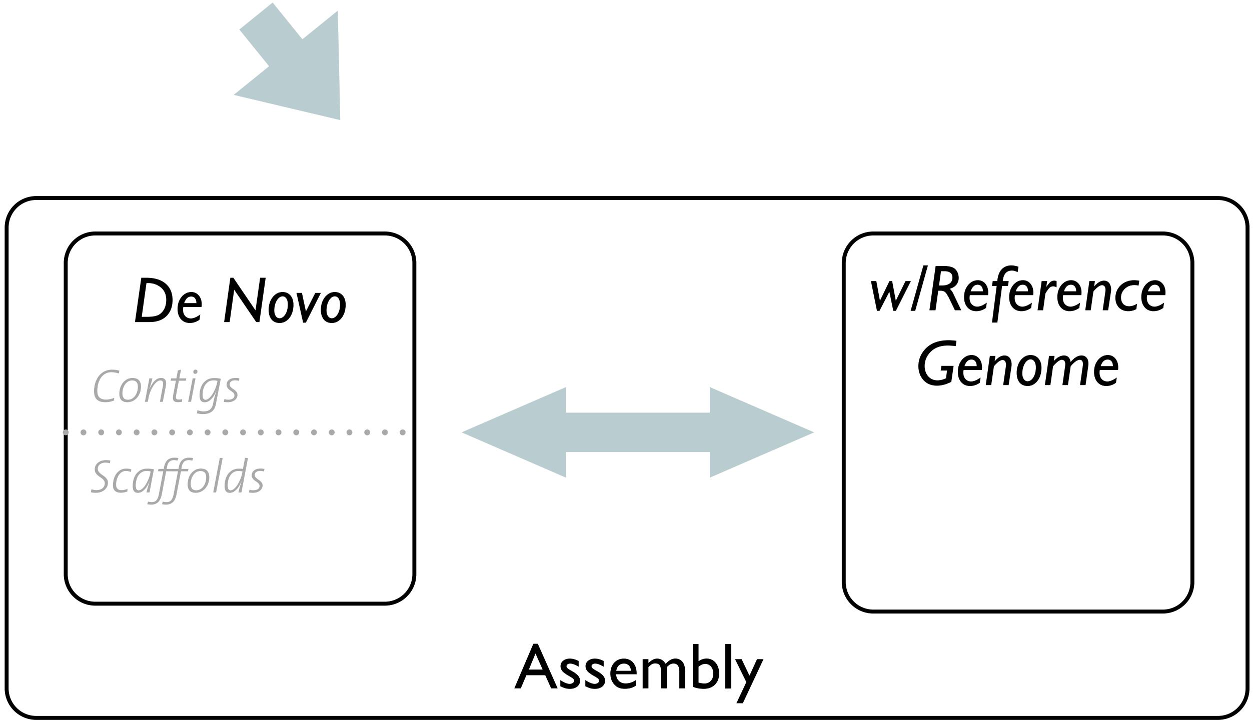
Schuster, Nature Methods 2008

Sequencing



Schuster, Nature Methods 2008

Sequencing



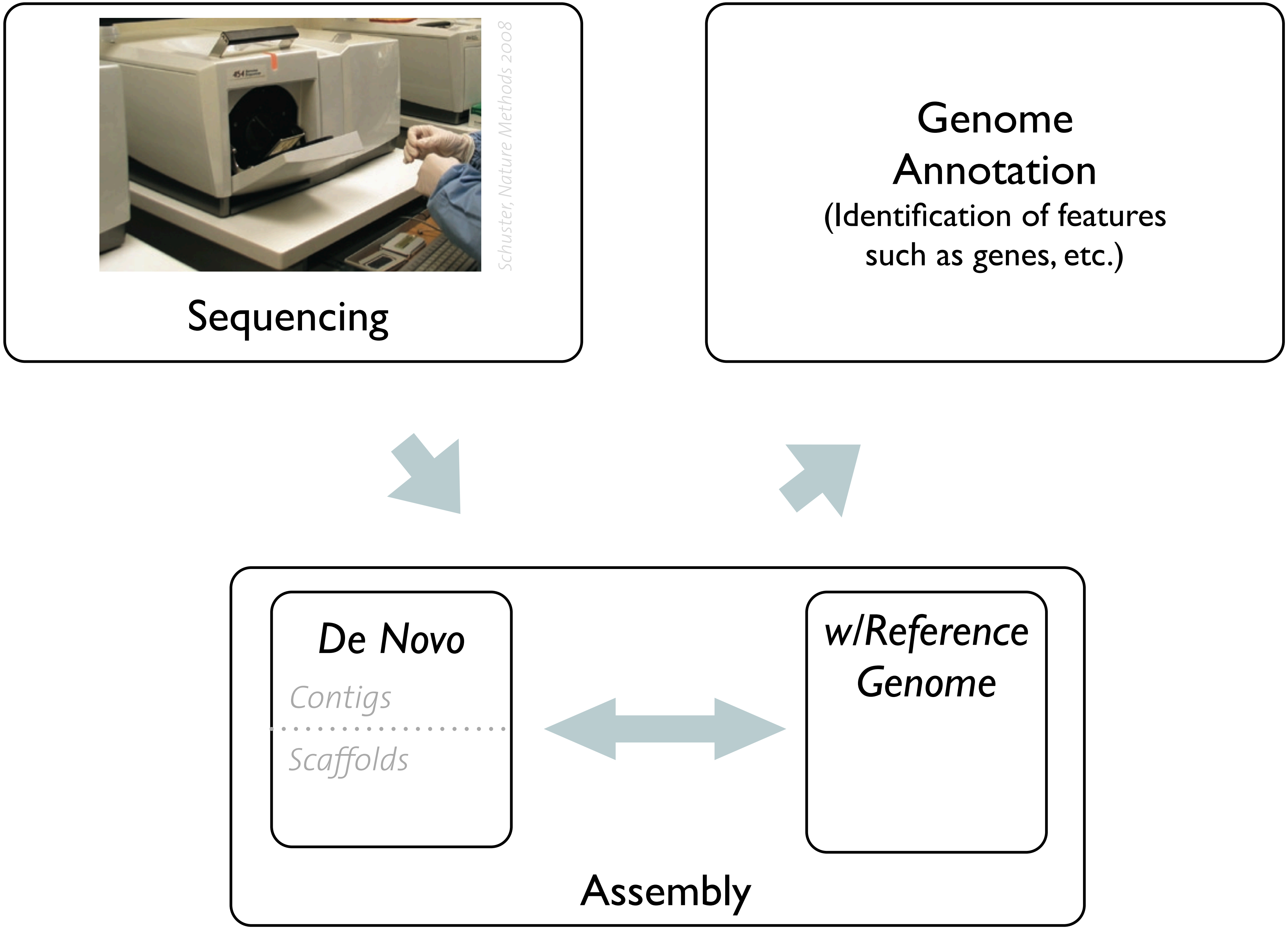
De Novo

Contigs

Scaffolds

w/Reference Genome

Assembly



Schuster, Nature Methods 2008

Sequencing

Genome Annotation
(Identification of features such as genes, etc.)

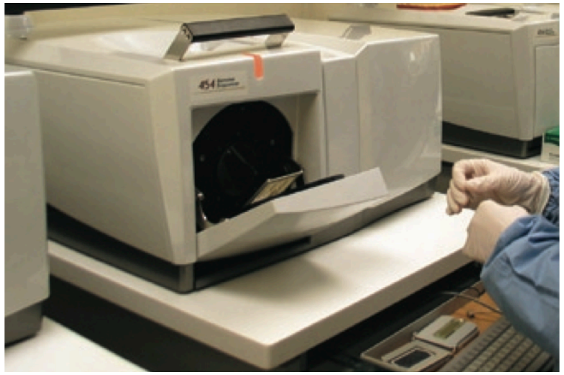
De Novo

Contigs

Scaffolds

w/Reference Genome

Assembly



Schuster, Nature Methods 2008

Sequencing

Genome Annotation
(Identification of features such as genes, etc.)

De Novo

Contigs

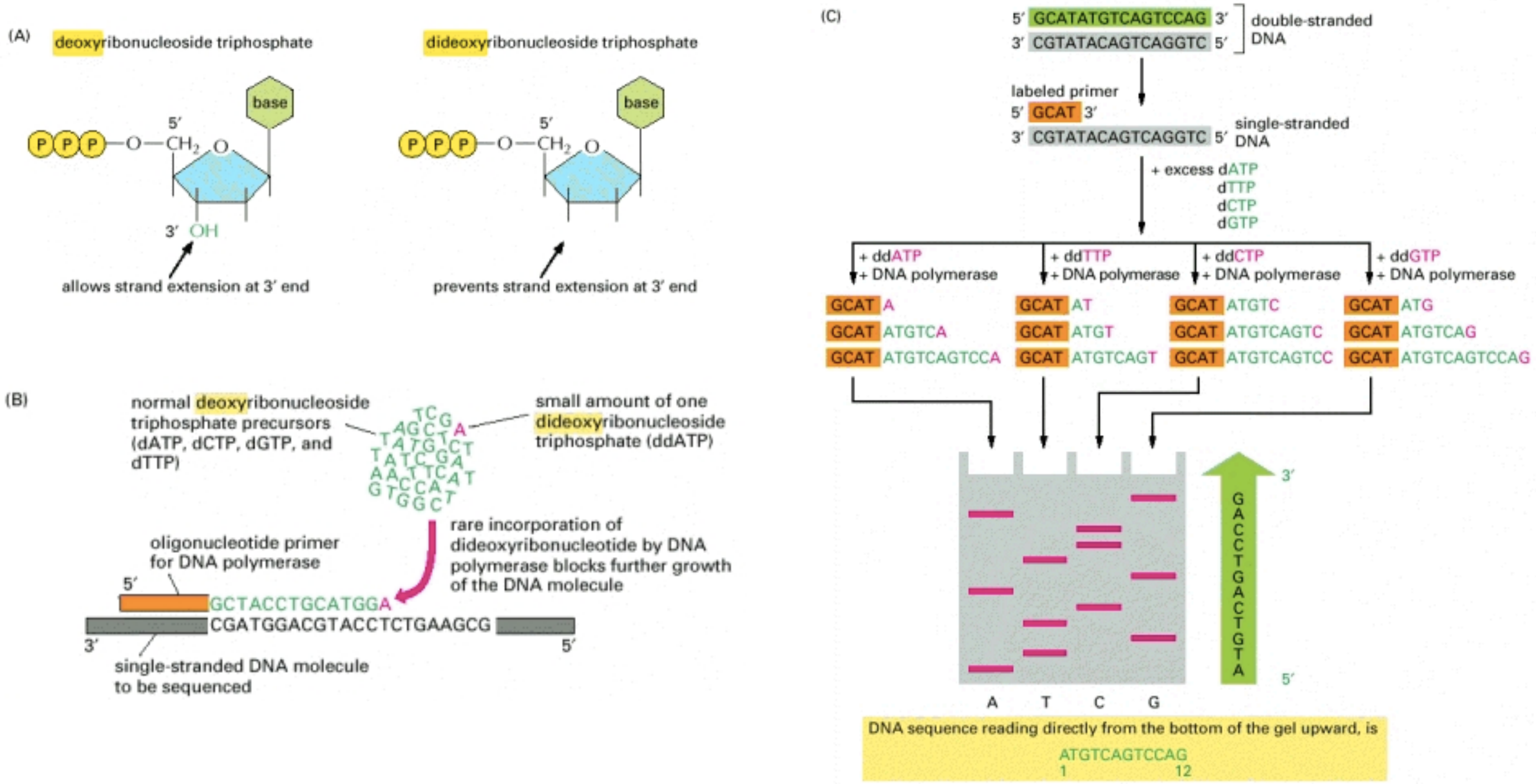
Scaffolds

Phylogeny-based

w/Reference Genome

Assembly

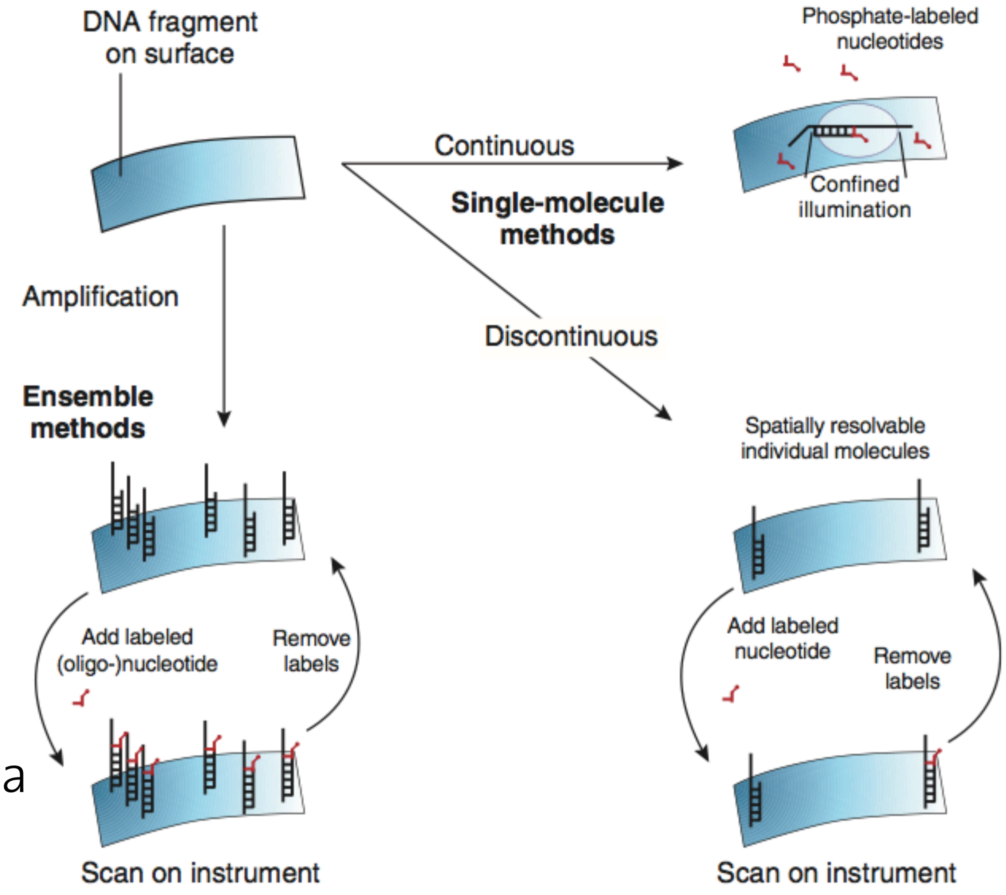
Sanger Sequencing



(A)
deoxyribonucleoside triphosphate
dideoxyribonucleoside triphosphate
P P P
O
5′
CH_2
base
3′ OH
allows strand extension at 3′ end
prevents strand extension at 3′ end
(B)
normal deoxyribonucleoside triphosphate precursors (dATP, dCTP, dGTP, and dTTP)
small amount of one dideoxyribonucleoside triphosphate (ddATP)
rare incorporation of dideoxyribonucleotide by DNA polymerase blocks further growth of the DNA molecule
oligonucleotide primer for DNA polymerase
5′
GCTACCTGCATGGA
CGATGGACGTACCTCTGAAGCG
3′
5′
single-stranded DNA molecule to be sequenced
(C)
5′ GCATATGTCAGTCCAG 3′
3′ CGTATACAGTCAGGTC 5′
double-stranded DNA
labeled primer
5′ GCAT 3′
3′ CGTATACAGTCAGGTC 5′
single-stranded DNA
+ excess dATP
dTTP
dCTP
dGTP
+ ddATP
+ DNA polymerase
+ ddTTP
+ DNA polymerase
+ ddCTP
+ DNA polymerase
+ ddGTP
+ DNA polymerase
GCAT A
GCAT ATGTCA
GCAT ATGTCAGTCCA
GCAT AT
GCAT ATGT
GCAT ATGTCAGT
GCAT ATGTC
GCAT ATGTCAGTC
GCAT ATGTCAGTCC
GCAT ATG
GCAT ATGTCAG
GCAT ATGTCAGTCCAG
A T C G
3′
GACCTGACTGTA
5′
DNA sequence reading directly from the bottom of the gel upward, is
ATGTCAGTCCAG
1
12

Alberts et al, *Molecular Biology of the Cell*, 2002, Garland Science, 4th Edition

Overview Next Gen.



Pacific Biosciences

“3rd Gen”

Helicos

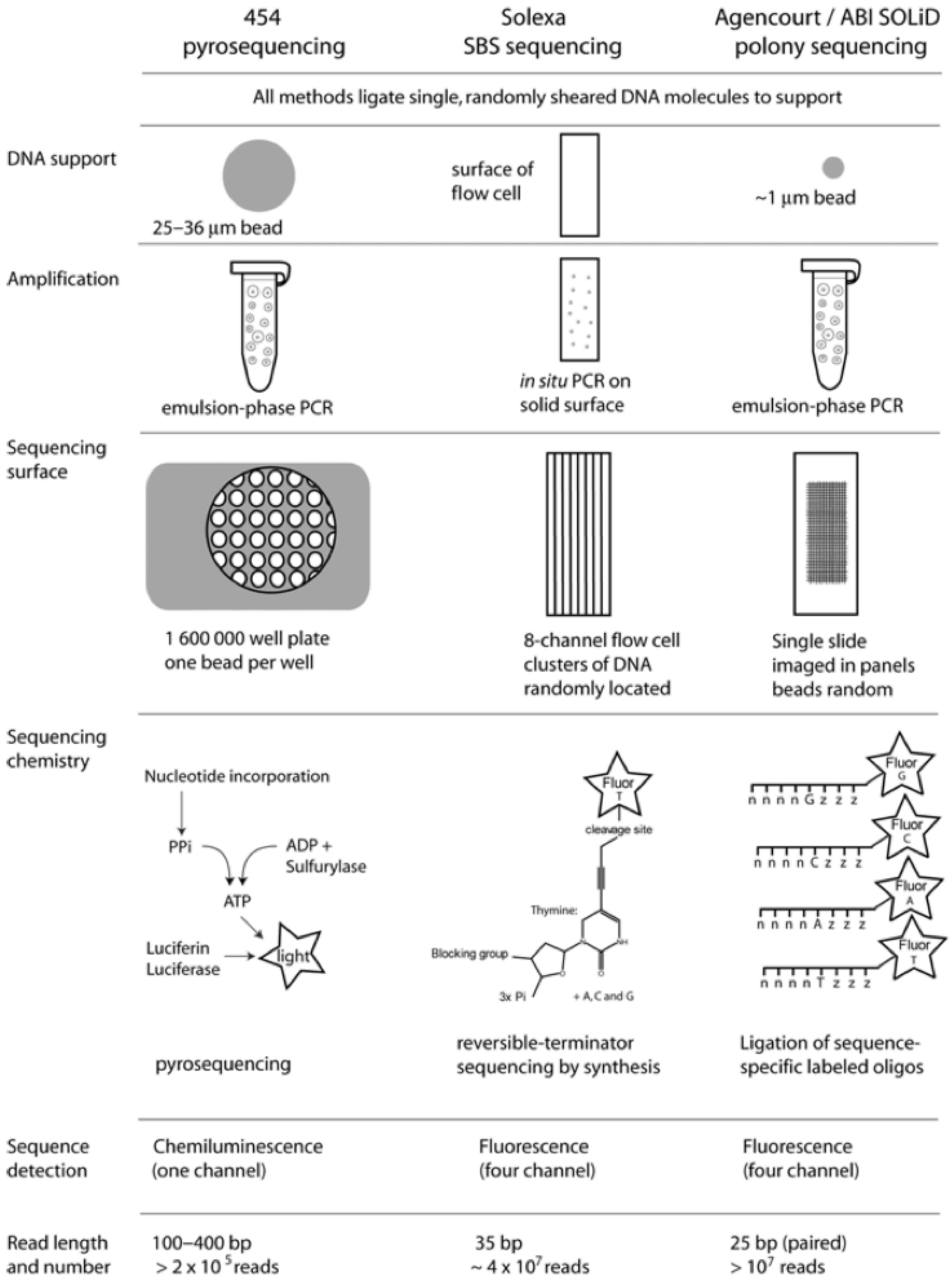
“2nd Gen”

Roche/454

Illumina/Solexa

AB Solid

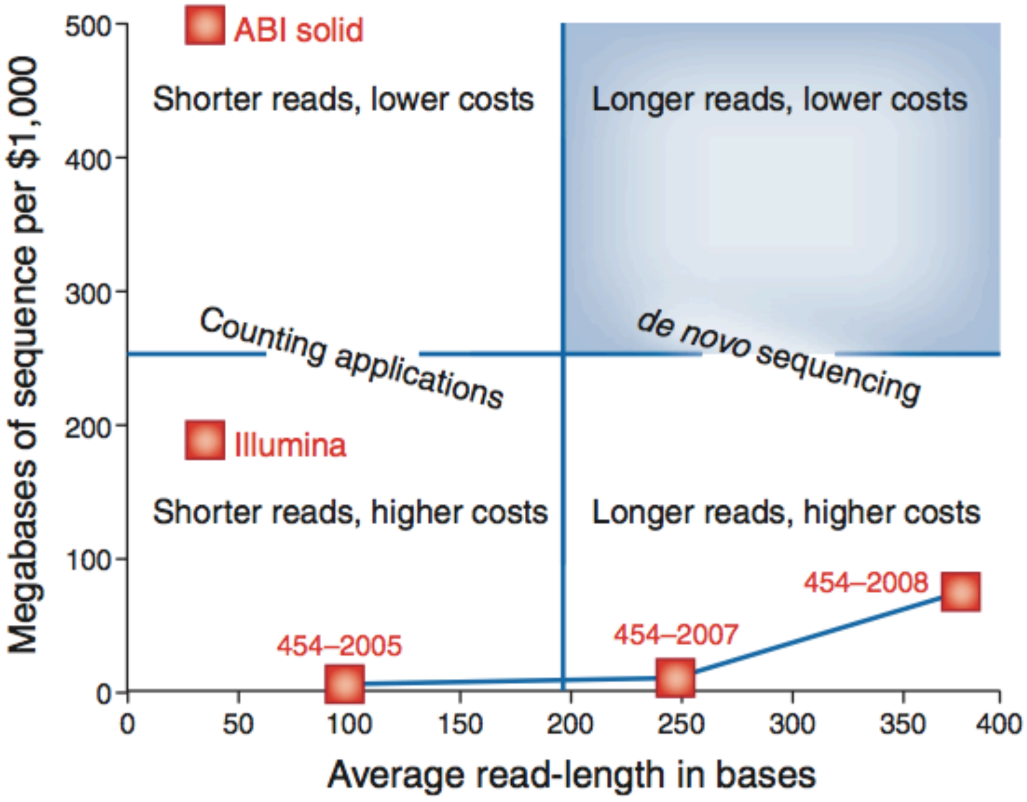
Fuller et al. The challenges of sequencing by synthesis. Nat Biotechnol (2009) vol. 27 (11) pp. 1013-23



	454 pyrosequencing	Solexa SBS sequencing	Agencourt / ABI SOLiD polony sequencing
	All methods ligate single, randomly sheared DNA molecules to support		
DNA support	25–36 µm bead	surface of flow cell	~1 µm bead
Amplification	emulsion-phase PCR	*in situ* PCR on solid surface	emulsion-phase PCR
Sequencing surface	1 600 000 well plate one bead per well	8-channel flow cell clusters of DNA randomly located	Single slide imaged in panels beads random
Sequencing chemistry	pyrosequencing	reversible-terminator sequencing by synthesis	Ligation of sequence-specific labeled oligos
Sequence detection	Chemiluminescence (one channel)	Fluorescence (four channel)	Fluorescence (four channel)
Read length and number	100–400 bp > 2 x 10^5 reads	35 bp ~ 4 x 10^7 reads	25 bp (paired) > 10^7 reads

Hudson. *Sequencing breakthroughs for genomic ecology and evolutionary biology*. Molecular Ecology Notes (2007) vol. 8 (1) pp. 3-17

Read length vs. Cost



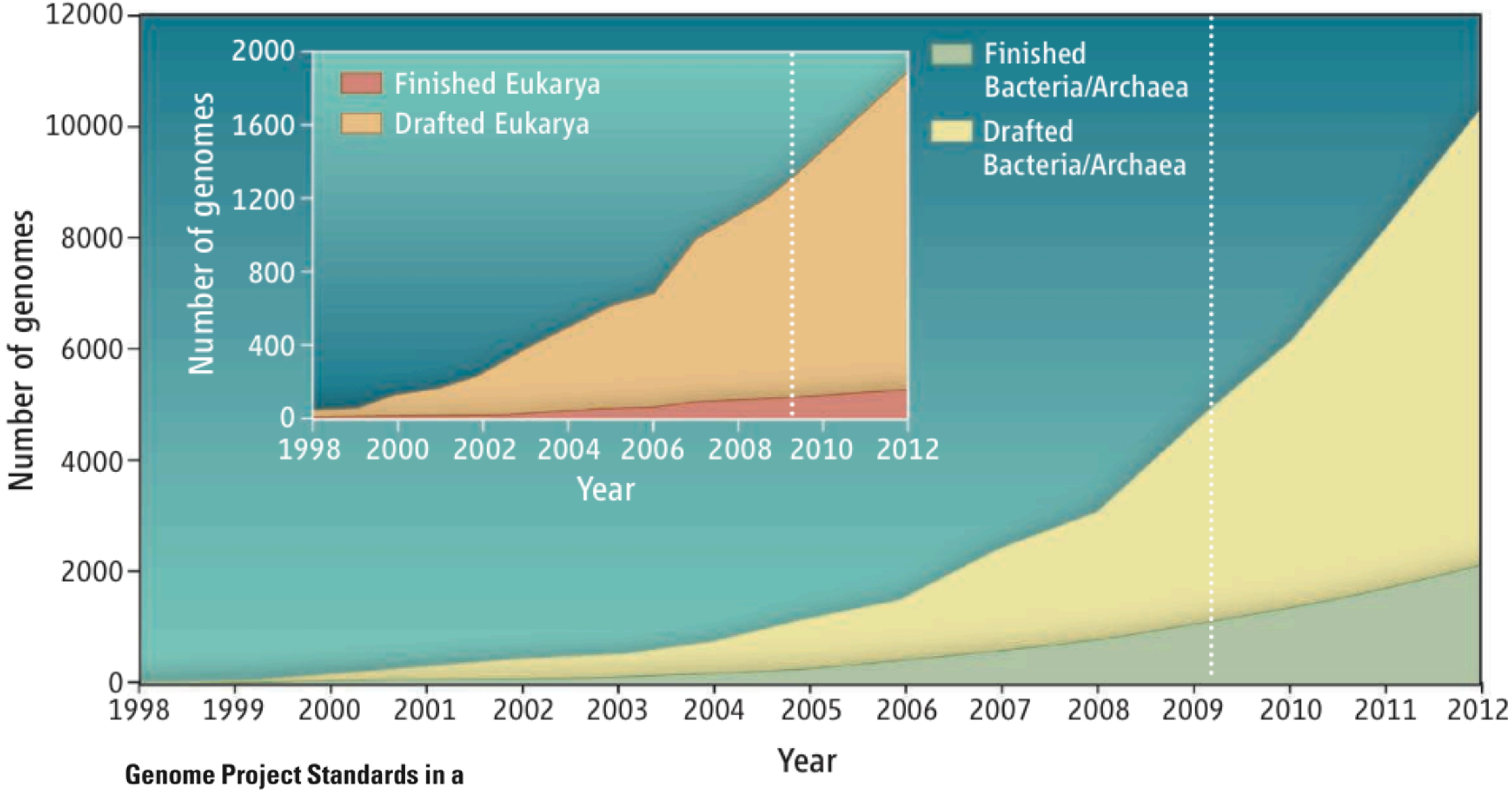
Rothberg and Leamon. The development and impact of 454 sequencing. Nat Biotechnol (2008) vol. 26 (10) pp. 1117-24

What is different?

- Much higher throughput / lower cost
- Sequence individual DNA fragments -> can deal with mixtures (environmental samples, etc.)
- Shorter reads

Mcpherson. Next-generation gap. Nature Methods (2009) vol. 6 (11 Suppl) pp. S2-5

Genome assembly is hard

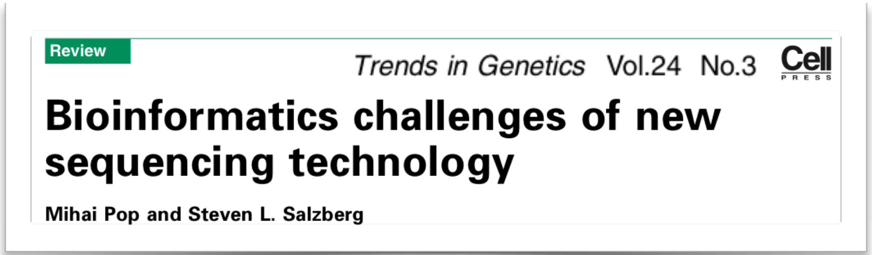


Genome Project Standards in a New Era of Sequencing

P. S. G. Chain,[1,2,3]*†§ D. V. Grafham,[4]†§ R. S. Fulton, [5]† M. G. FitzGerald,[6]† J. Hostetler,[7]† D. Muzny,[8]† J. Ali,[9] B. Birren,[6] D. C. Bruce,[1,10] C. Buhay,[8] J. R. Cole,[3] Y. Ding,[8] S. Dugan,[8] D. Field,[11] G. M. Garrity,[3] R. Gibbs,[8] T. Graves,[5] C. S. Han,[1,10] S. H. Harrison,[3]* S. Highlander,[8] P. Hugenholtz, [1] H. M. Khouri,[12] C. D. Kodira,[6]* E. Kolker,[13,14] N. C. Kyrpides,[1] D. Lang,[12] A. Lapidus,[1] S. A. Malfatti,[12] V. Markowitz,[15] T. Metha,[6] K. E. Nelson,[7] J. Parkhill,[4] S. Pitluck,[1] X. Qin,[8] T. D. Read,[16] J. Schmutz,[17] S. Sozhamannan,[18] P. Sterk,[11] R. L. Strausberg,[7] G. Sutton,[7] N. R. Thomson,[4] J. M. Tiedje,[3] G. Weinstock,[5] A. Wollam,[5] Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium,‡ J. C. Detter[10]†‡

9 OCTOBER 2009 VOL 326 **SCIENCE**

Reviews on Assembly

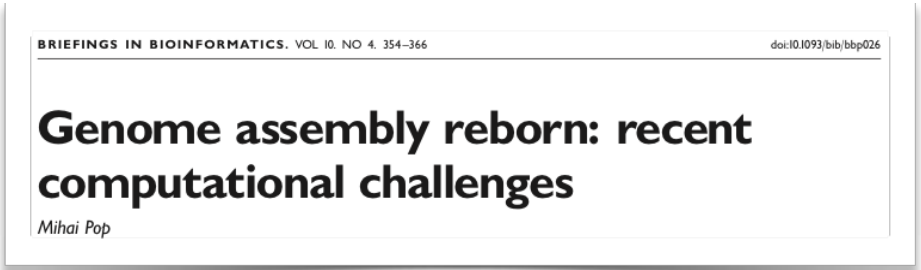


Review

Trends in Genetics Vol.24 No.3 Cell PRESS

Bioinformatics challenges of new sequencing technology

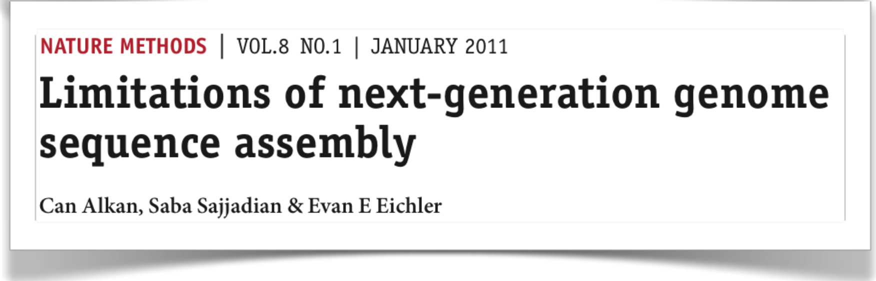
Mihai Pop and Steven L. Salzberg



BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 4. 354–366 doi:10.1093/bib/bbp026

Genome assembly reborn: recent computational challenges

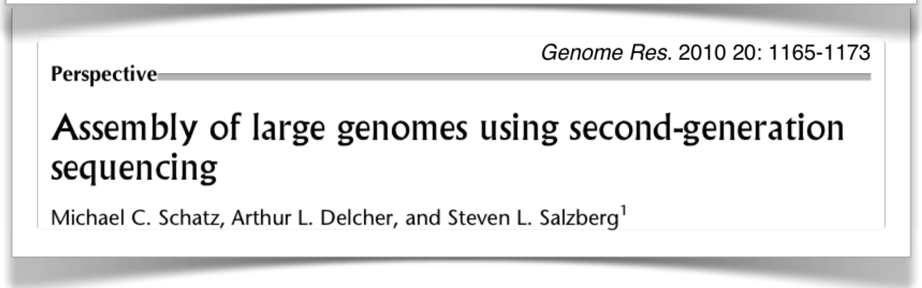
Mihai Pop



NATURE METHODS | VOL.8 NO.1 | JANUARY 2011

Limitations of next-generation genome sequence assembly

Can Alkan, Saba Sajjadian & Evan E Eichler



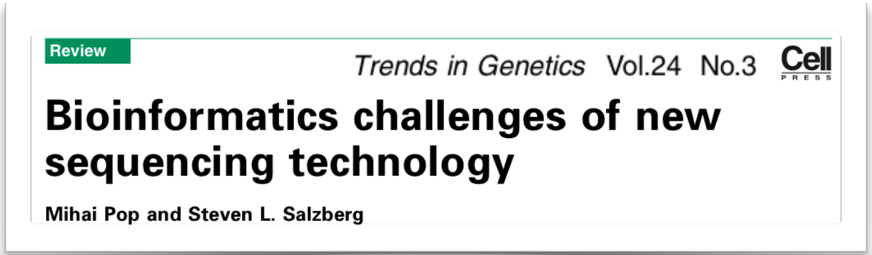
Genome Res. 2010 20: 1165-1173

Perspective

Assembly of large genomes using second-generation sequencing

Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg[1]

Reviews on Assembly

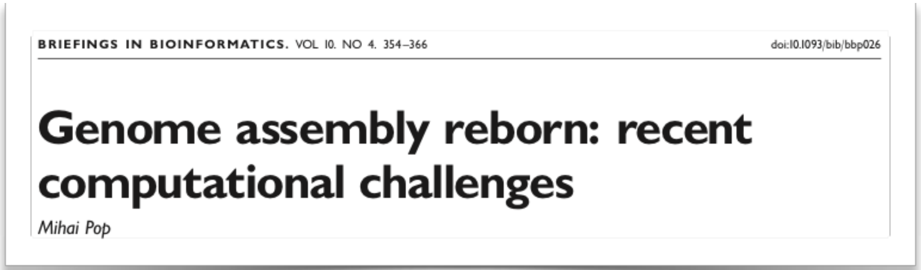


Review

Trends in Genetics Vol.24 No.3 Cell PRESS

Bioinformatics challenges of new sequencing technology

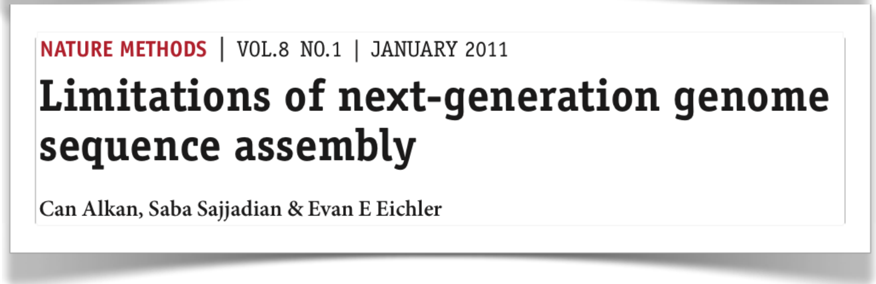
Mihai Pop and Steven L. Salzberg



BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 4. 354–366 doi:10.1093/bib/bbp026

Genome assembly reborn: recent computational challenges

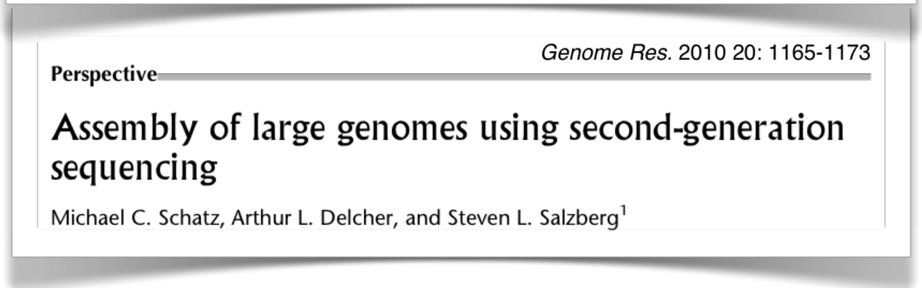
Mihai Pop



NATURE METHODS | VOL.8 NO.1 | JANUARY 2011

Limitations of next-generation genome sequence assembly

Can Alkan, Saba Sajjadian & Evan E Eichler



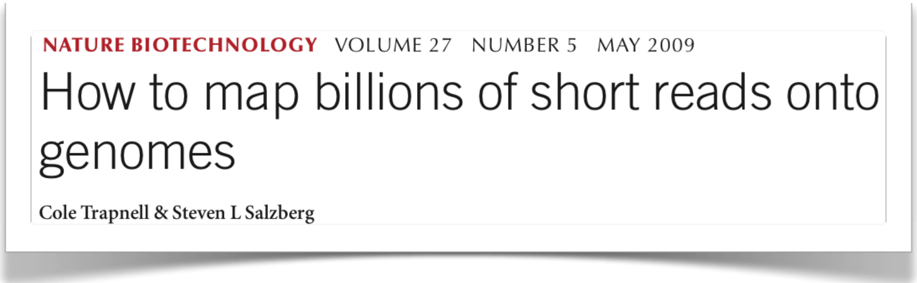
Genome Res. 2010 20: 1165-1173

Perspective

Assembly of large genomes using second-generation sequencing

Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg[1]

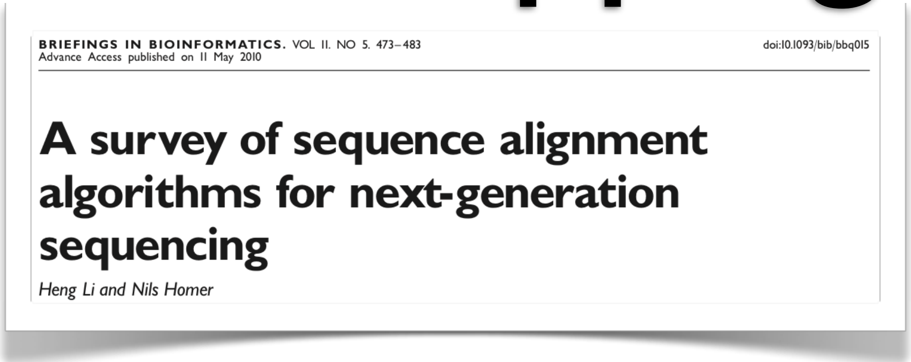
Reviews on Read Mapping



NATURE BIOTECHNOLOGY VOLUME 27 NUMBER 5 MAY 2009

How to map billions of short reads onto genomes

Cole Trapnell & Steven L Salzberg



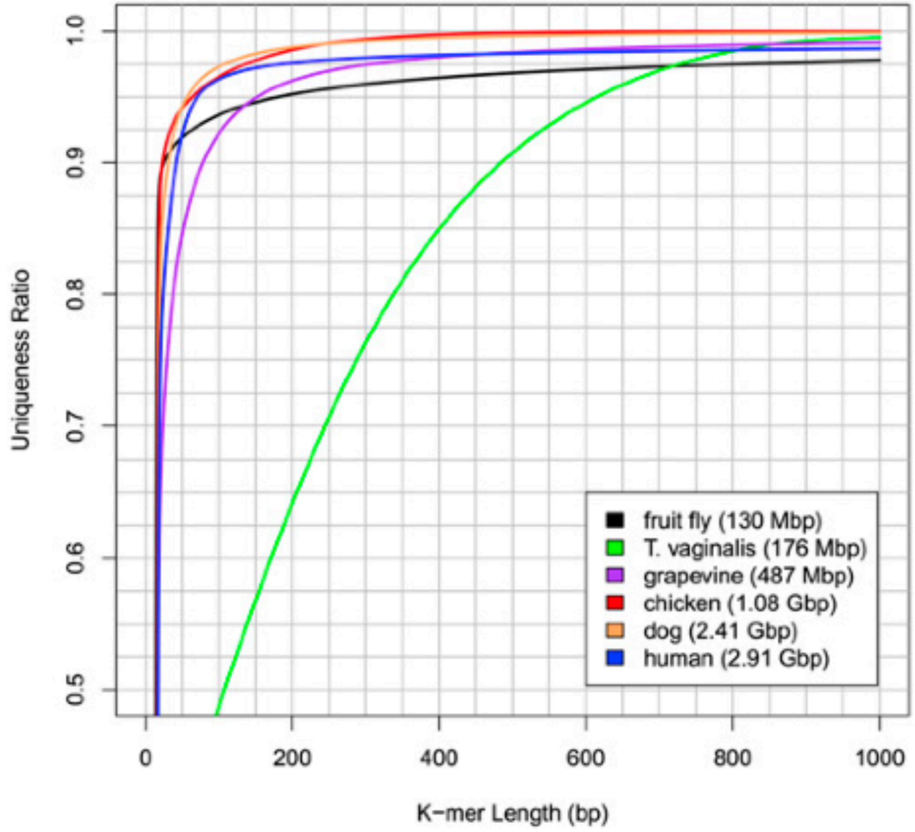
BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 5. 473–483
Advance Access published on 11 May 2010 doi:10.1093/bib/bbq015

A survey of sequence alignment algorithms for next-generation sequencing

Heng Li and Nils Homer

General Problems

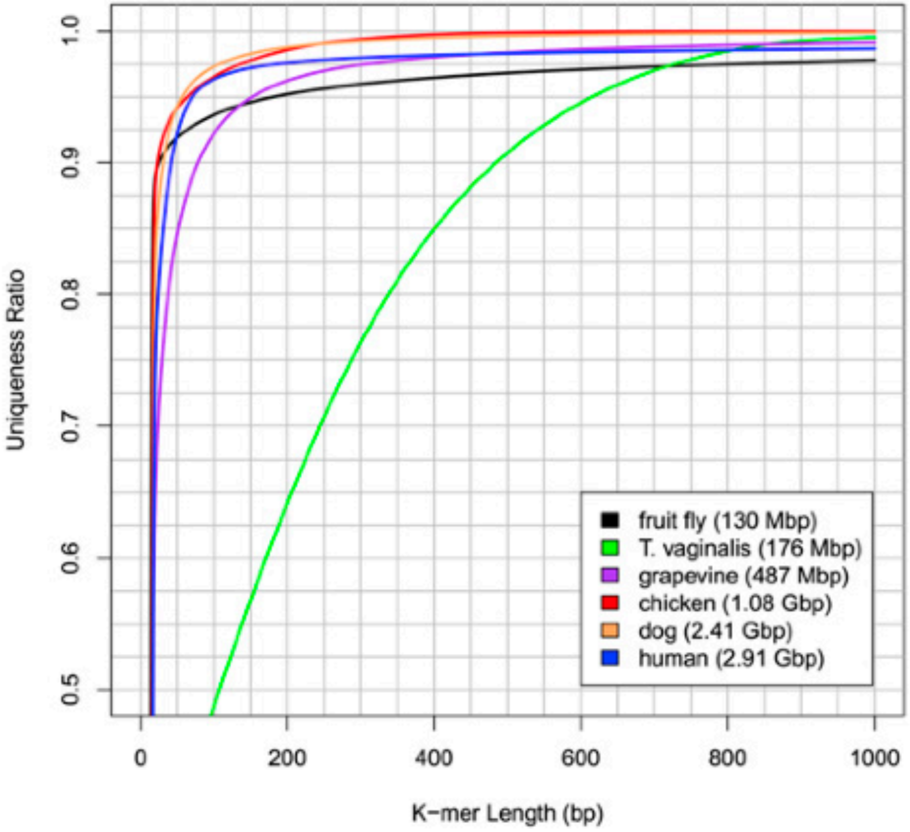
- Repeats



Schatz et al. 2010

General Problems

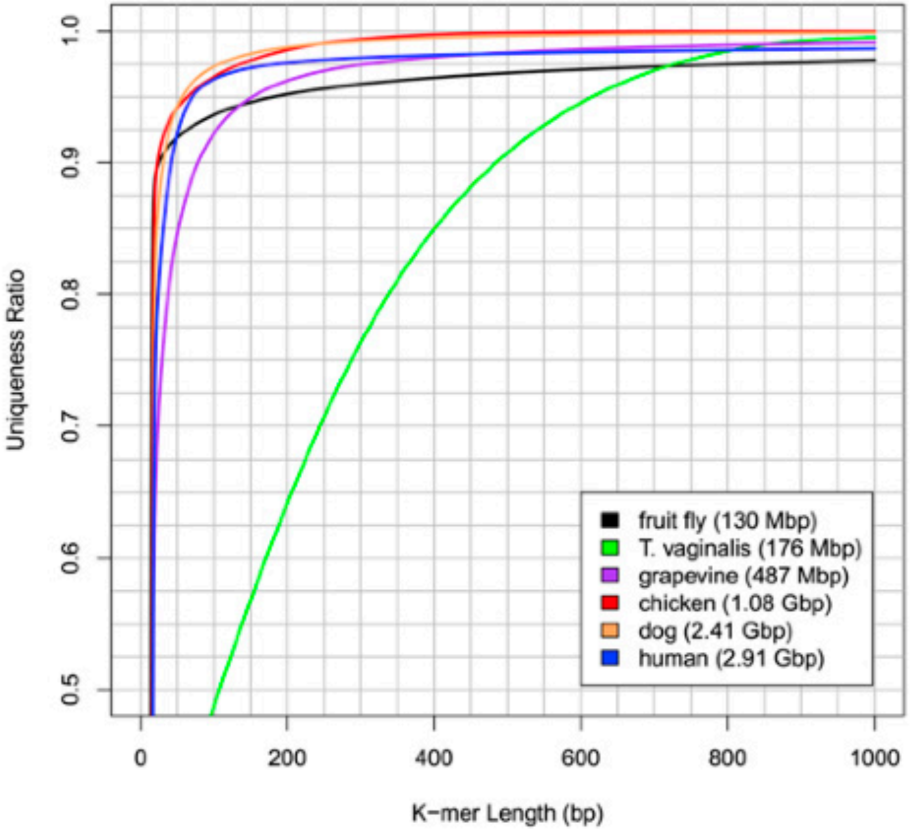
- Repeats
- Sequencing errors



Schatz et al. 2010

General Problems

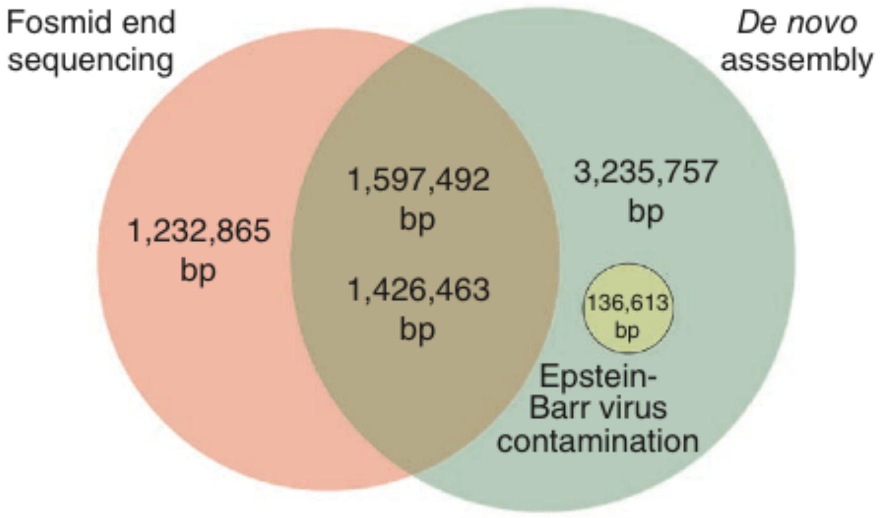
- Repeats
- Sequencing errors
- Polymorphisms



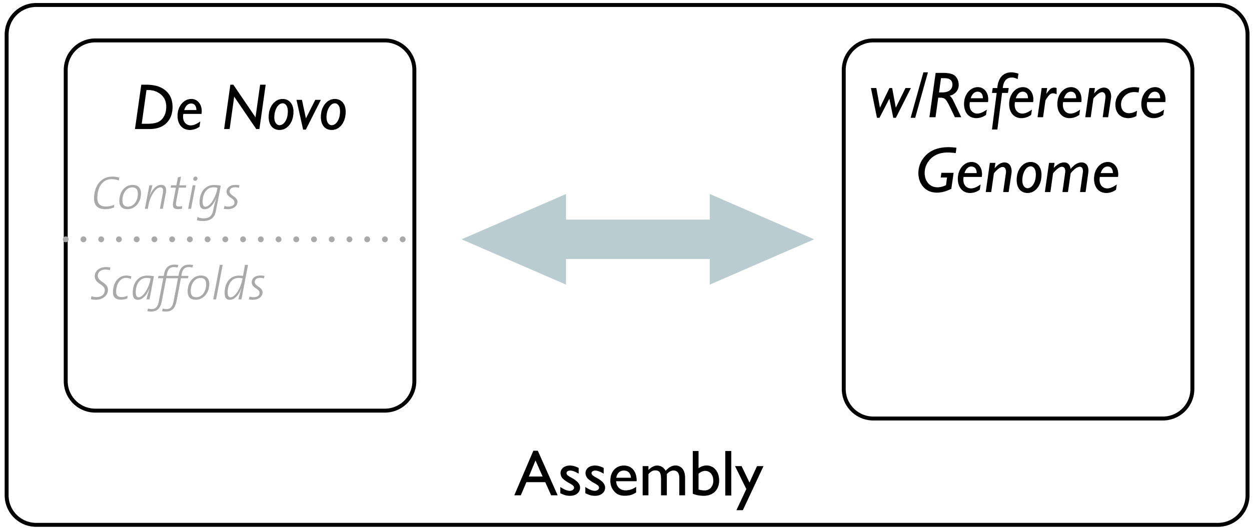
Schatz et al. 2010

General Problems

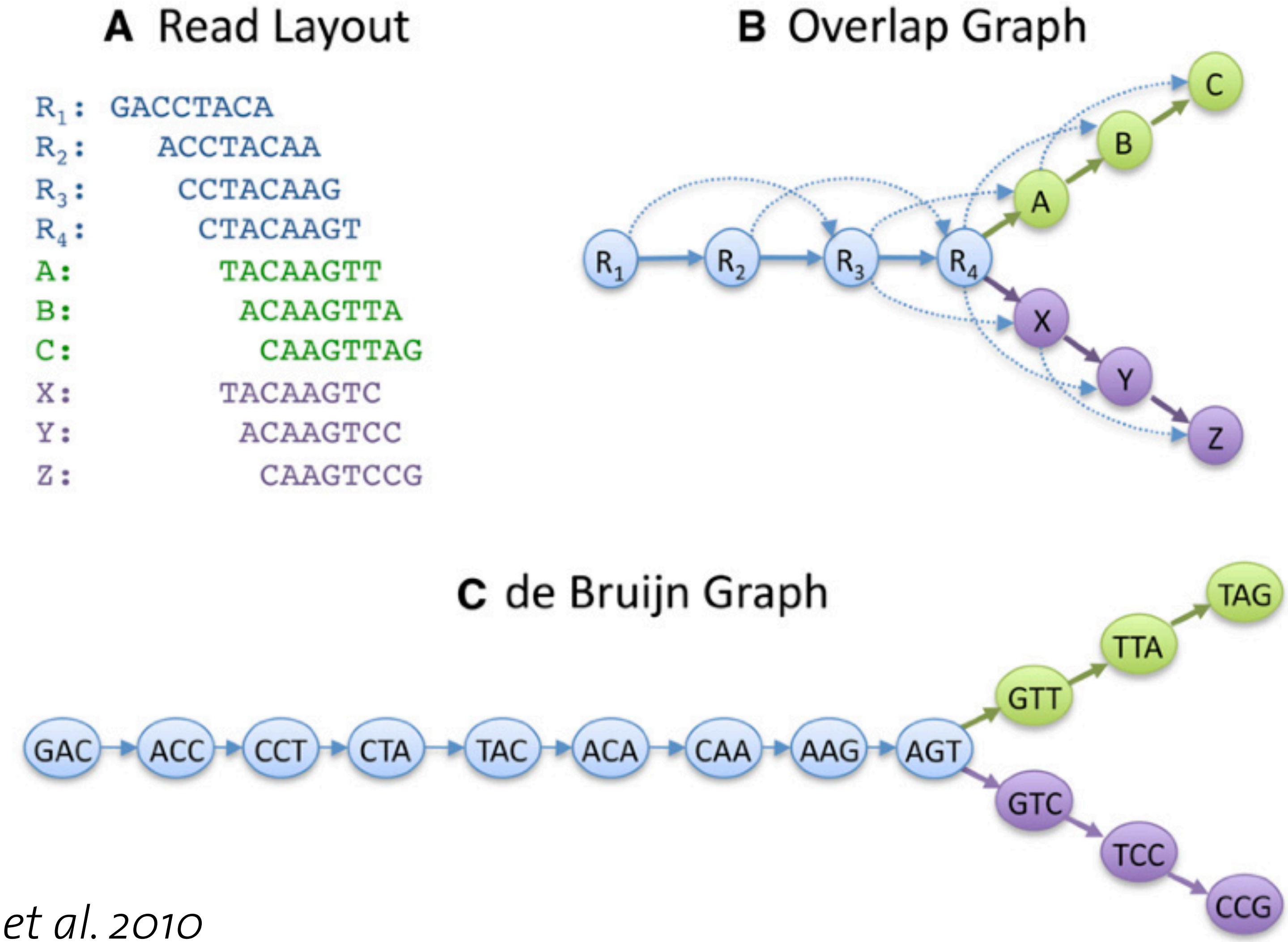
- Repeats
- Sequencing errors
- Polymorphisms
- Contamination



Alkan et al. 2011


De Novo
Contigs
Scaffolds
w/Reference Genome
Assembly

de Novo Assembly



A Read Layout

```
R1: GACCTACA
R2:  ACCTACAA
R3:   CCTACAAG
R4:    CTACAAGT
A:      TACAAGTT
B:       ACAAGTTA
C:        CAAGTTAG
X:      TACAAGTC
Y:       ACAAGTCC
Z:        CAAGTCCG
```

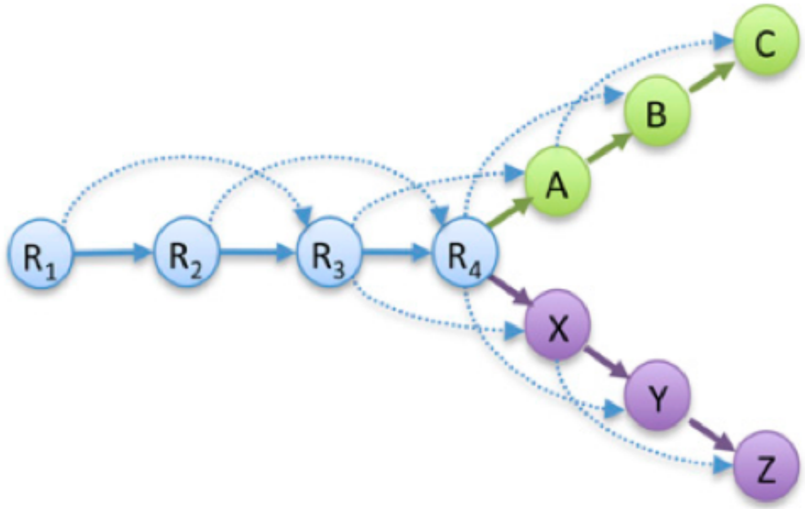
B Overlap Graph

C de Bruijn Graph

Schatz et al. 2010

Overlap Graph

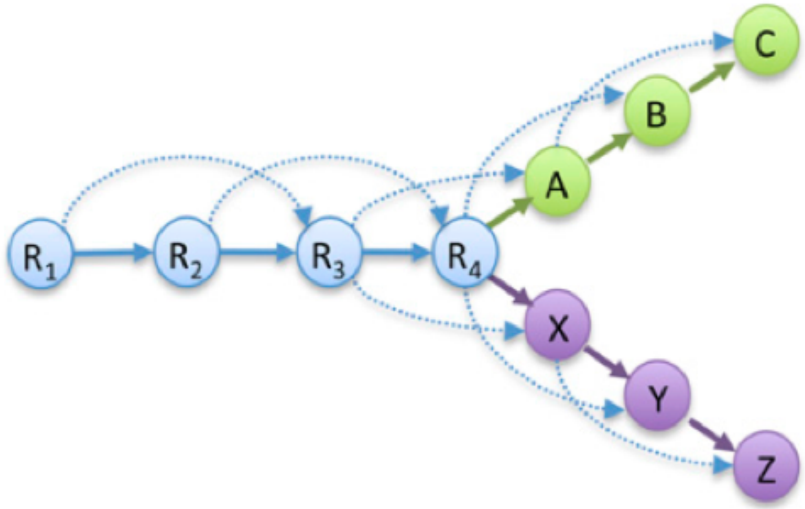
- Identify all pairwise overlaps among contigs (expensive for deep coverage, short reads)



Schatz et al. 2010

Overlap Graph

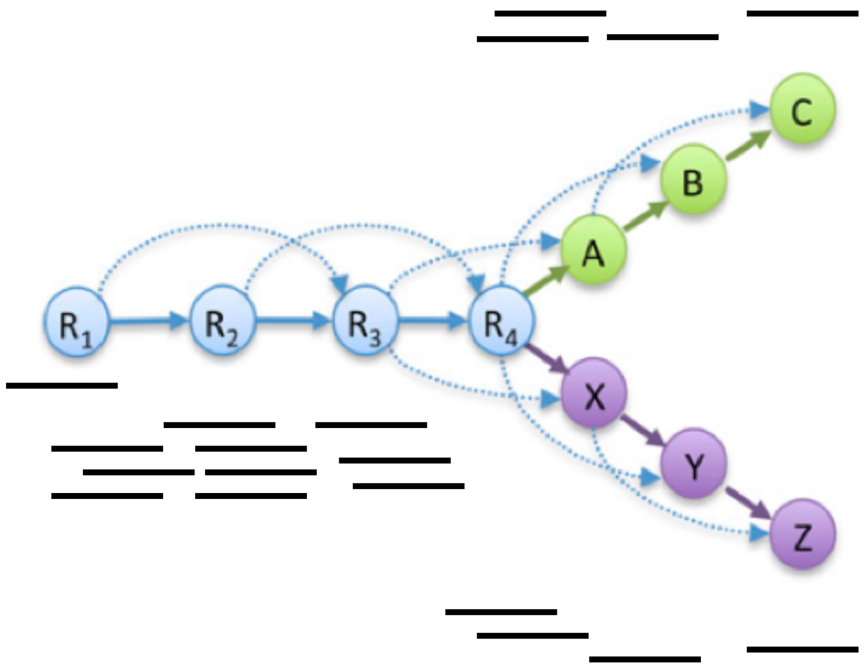
- Identify all pairwise overlaps among contigs (expensive for deep coverage, short reads)
- Error correction



Schatz et al. 2010

Overlap Graph

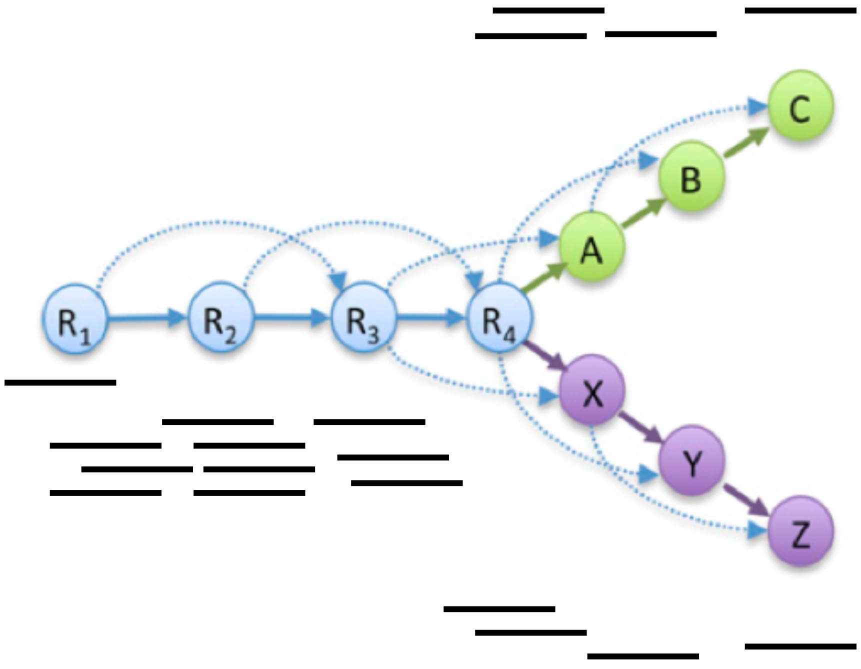
- Identify all pairwise overlaps among contigs (expensive for deep coverage, short reads)
- Error correction
- Contigs with disproportionally many reads are flagged as repeats



Schatz et al. 2010

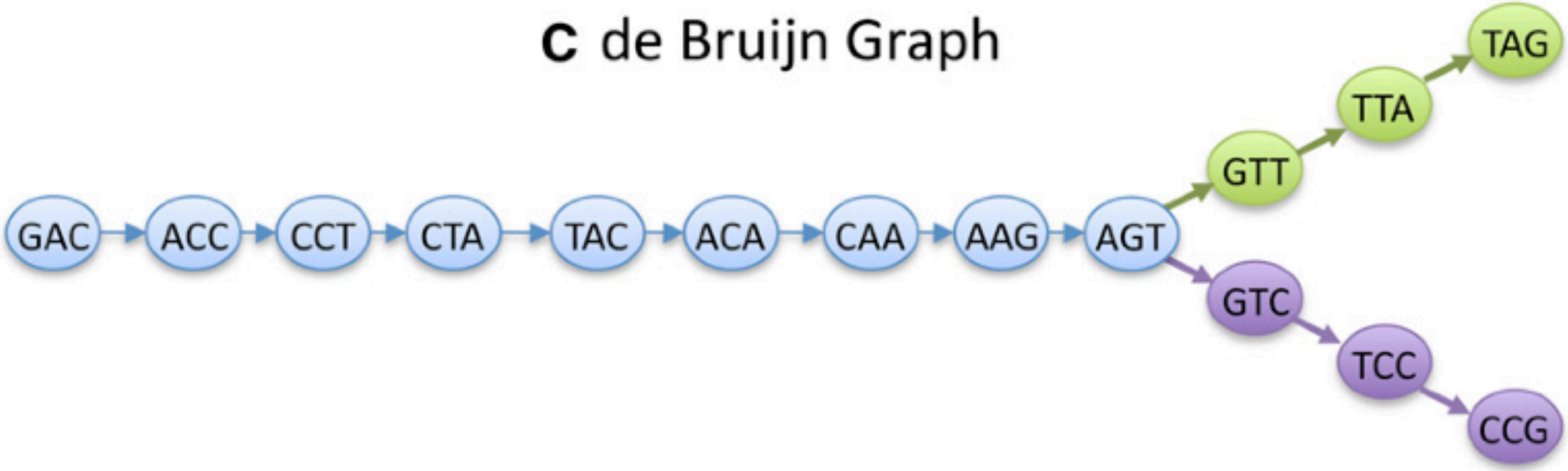
Overlap Graph

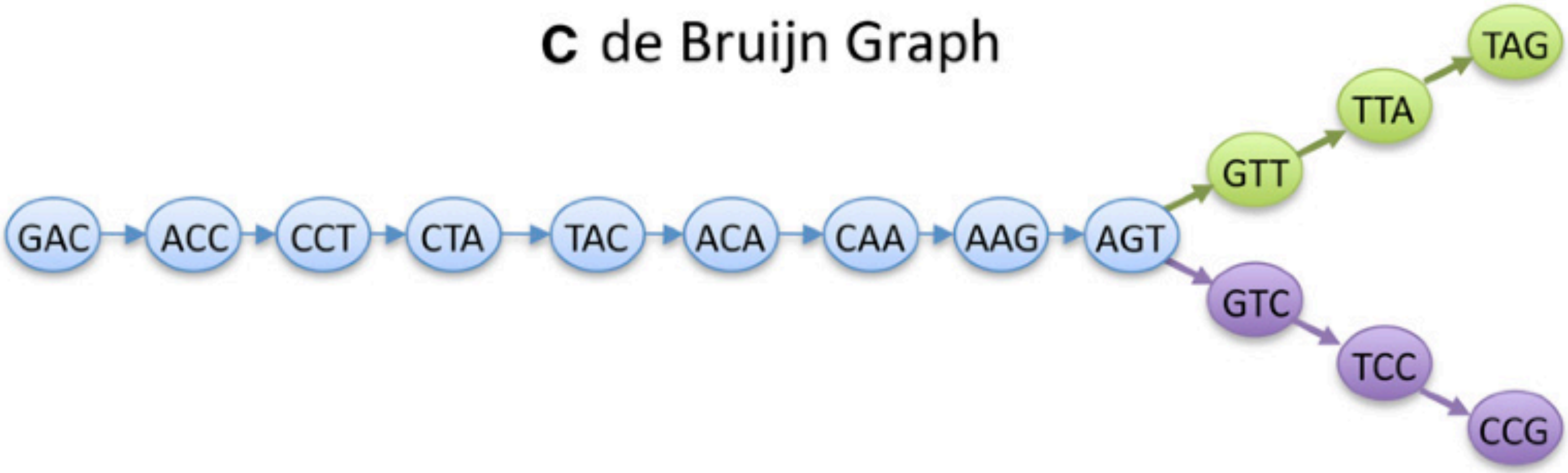
- Identify all pairwise overlaps among contigs (expensive for deep coverage, short reads)
- Error correction
- Contigs with disproportionally many reads are flagged as repeats
- Ideally, should identify Hamiltonian path through all contigs (Traveling salesman problem)



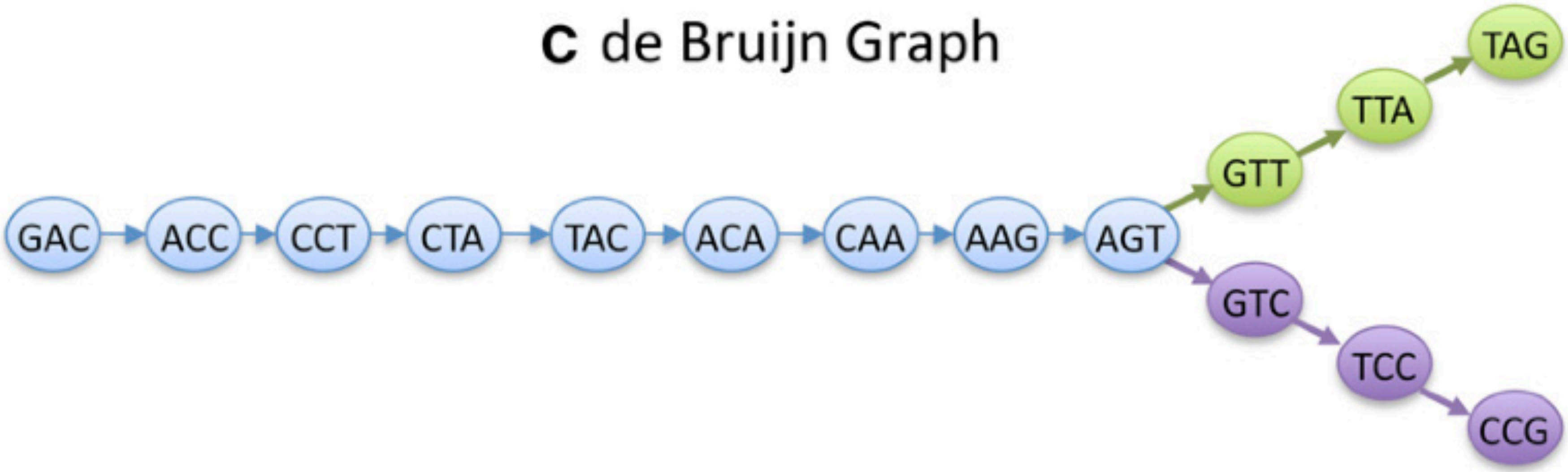
Schatz et al. 2010

C de Bruijn Graph

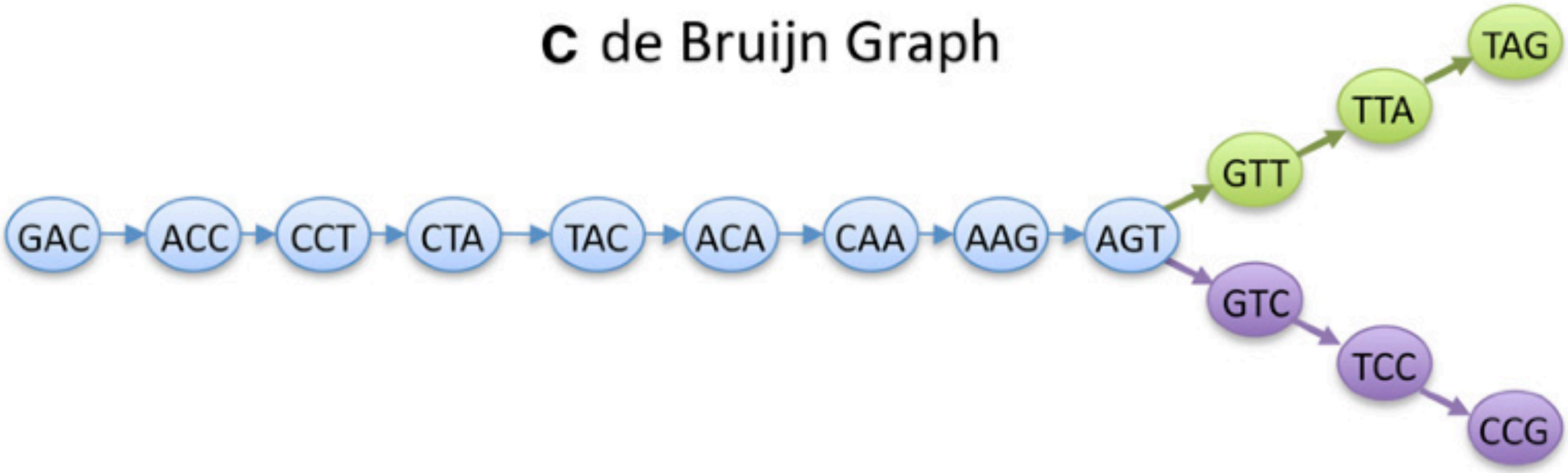




- Decompose reads into k-mers (here k=4)



- Decompose reads into k-mers (here k=4)
- Each k-mer induces an edge in de Bruijn graph (no pairwise overlap computation)

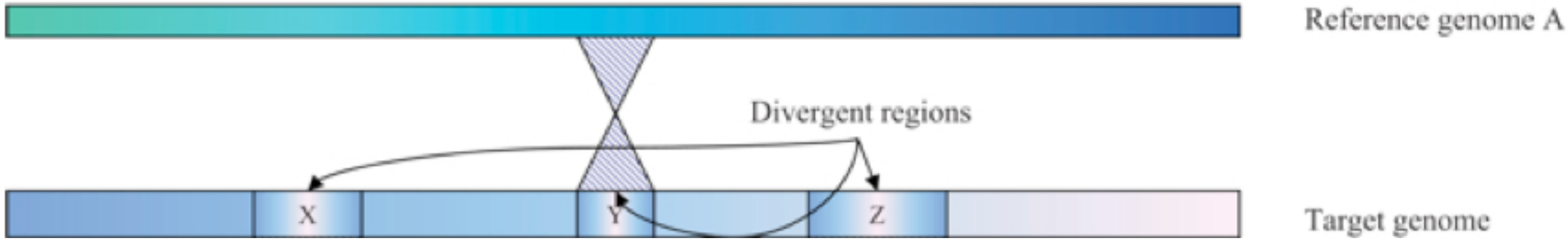


- Decompose reads into k-mers (here k=4)
- Each k-mer induces an edge in de Bruijn graph (no pairwise overlap computation)
- Identify *Eulerian path* (path which uses all edges)

How to bridge gaps? (“Scaffolding”)

- Increase coverage
- Use mate-pairs
- Gap closing through PCR
- Use mRNA library

Comparative Assembly: Map to Reference Genome



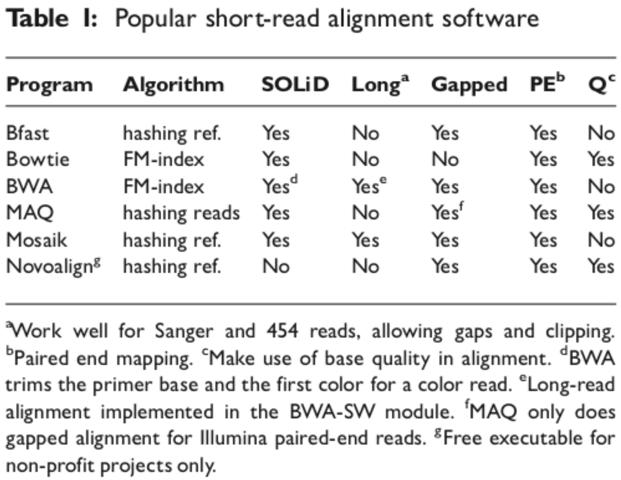
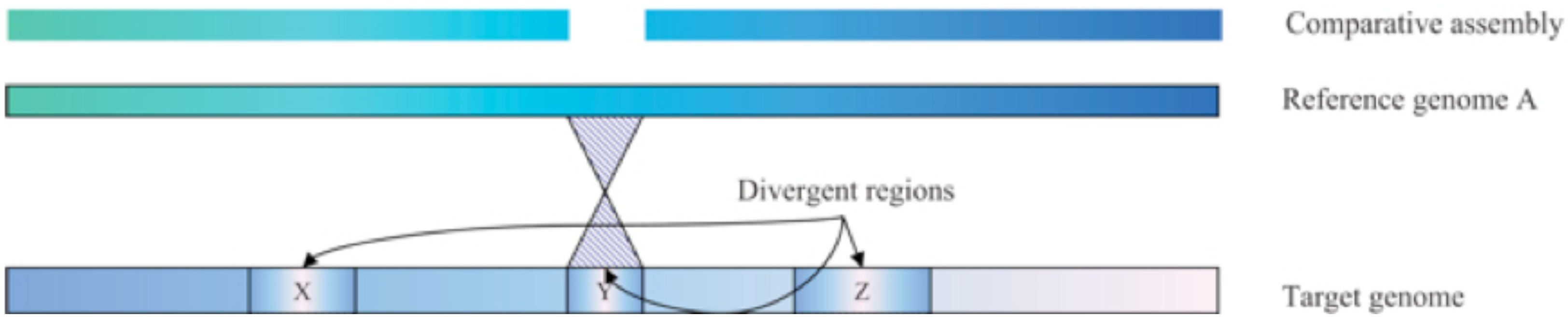


Table 1: Popular short-read alignment software

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a] Work well for Sanger and 454 reads, allowing gaps and clipping. [b] Paired end mapping. [c] Make use of base quality in alignment. [d] BWA trims the primer base and the first color for a color read. [e] Long-read alignment implemented in the BWA-SW module. [f] MAQ only does gapped alignment for Illumina paired-end reads. [g] Free executable for non-profit projects only.

Comparative Assembly: Map to Reference Genome



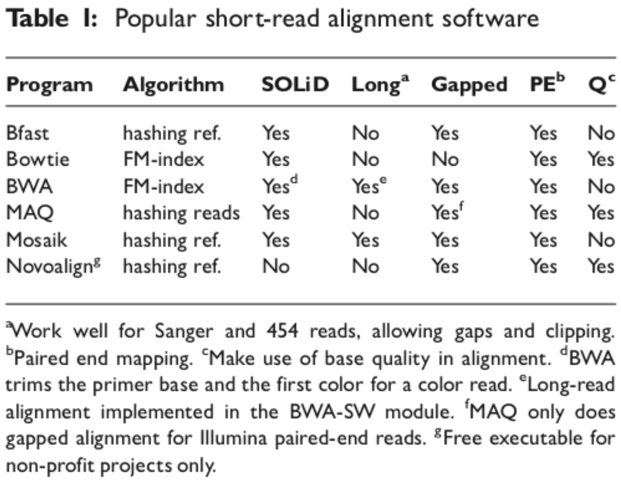
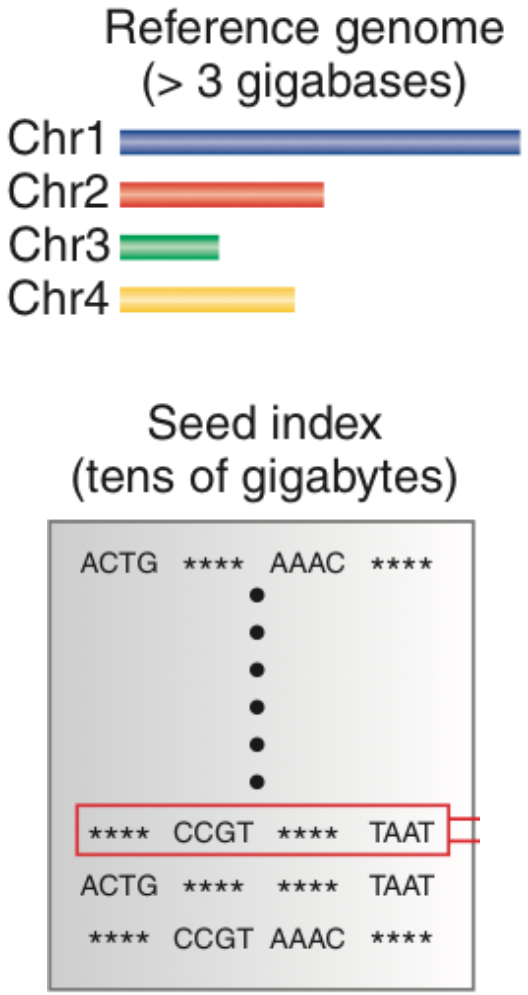


Table 1: Popular short-read alignment software

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a] Work well for Sanger and 454 reads, allowing gaps and clipping. [b] Paired end mapping. [c] Make use of base quality in alignment. [d] BWA trims the primer base and the first color for a color read. [e] Long-read alignment implemented in the BWA-SW module. [f] MAQ only does gapped alignment for Illumina paired-end reads. [g] Free executable for non-profit projects only.

Hash tables



ACTG **** AAAC ****

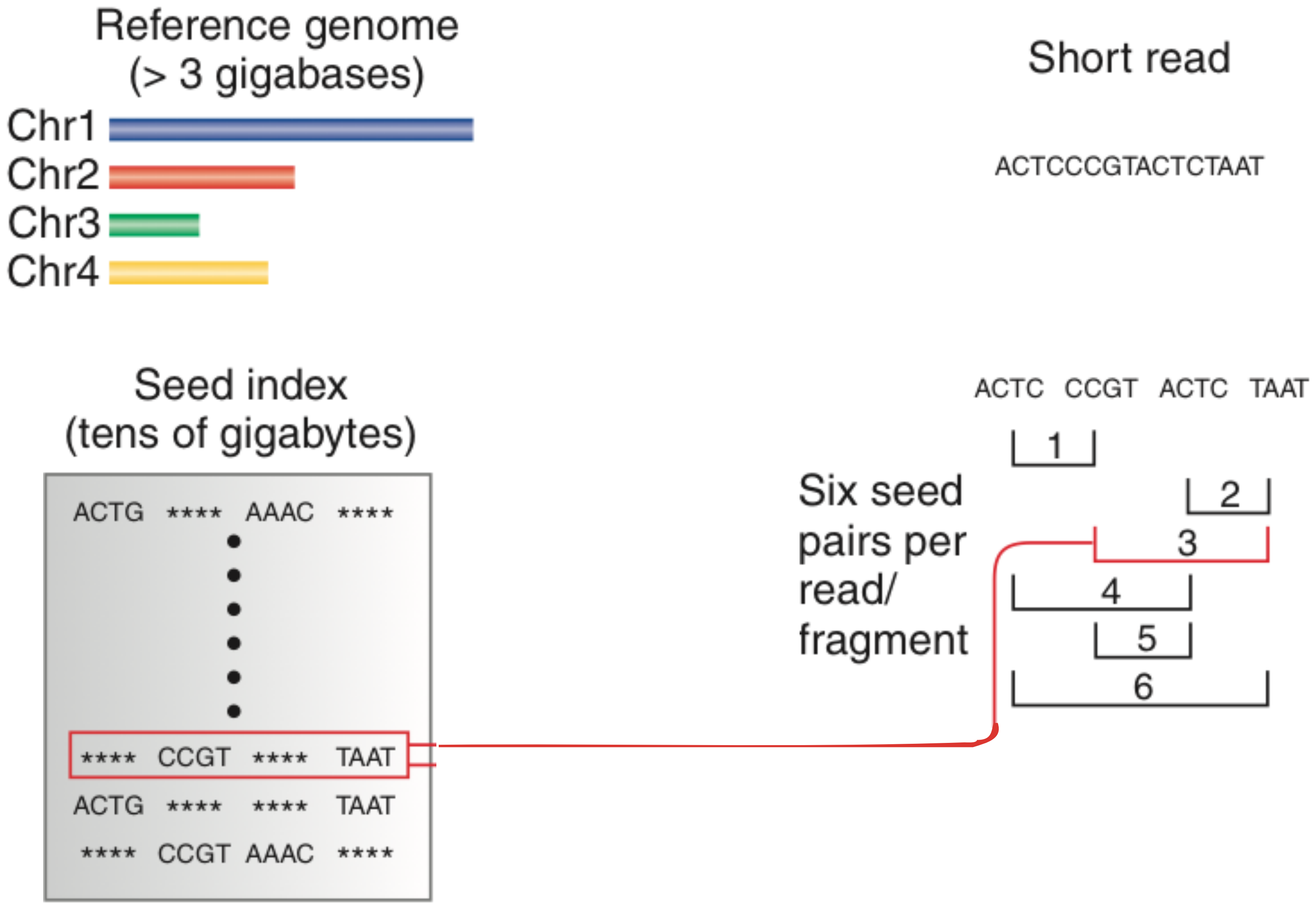
**** CCGT **** TAAT

ACTG **** **** TAAT

**** CCGT AAAC ****

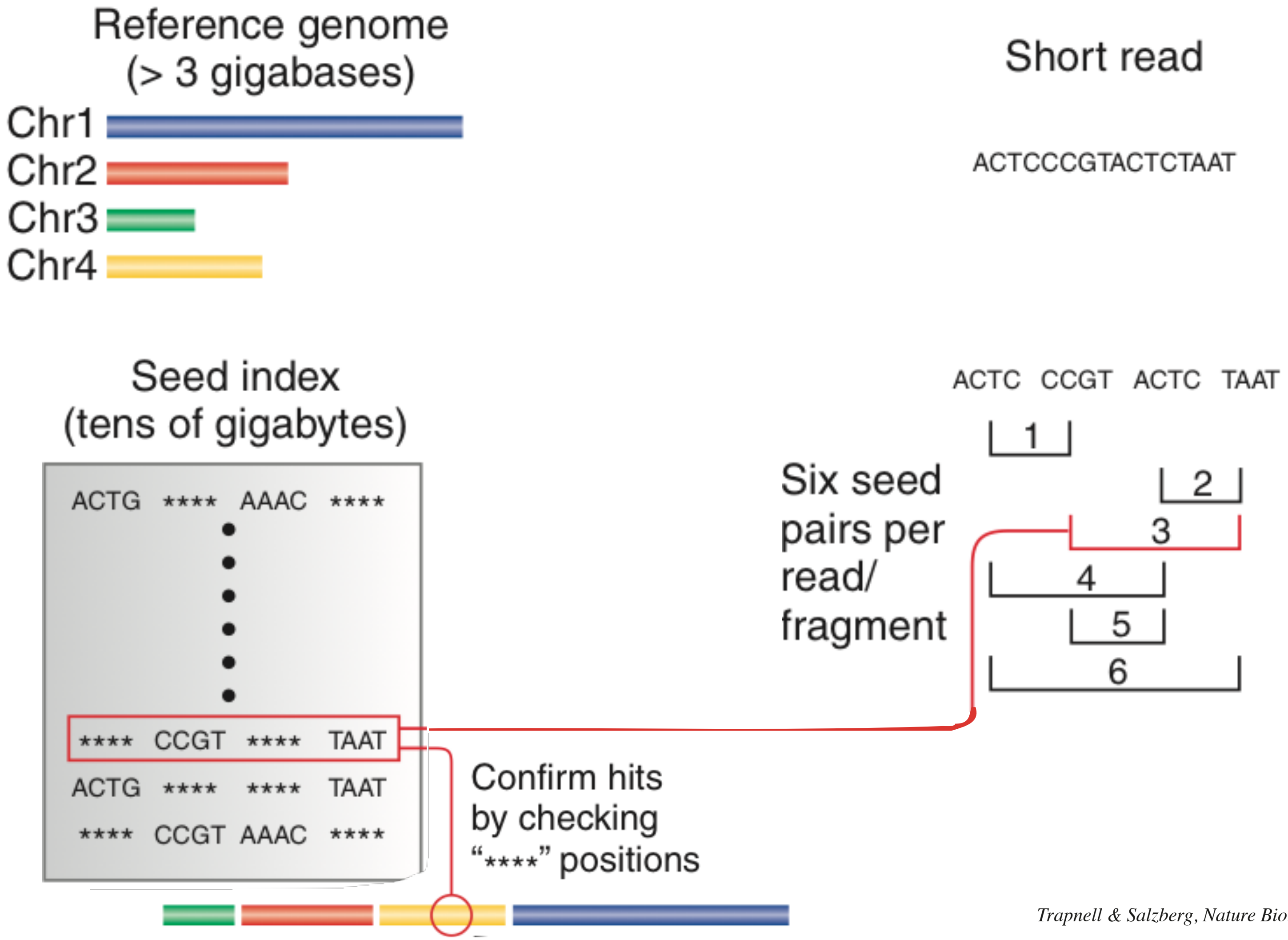
Trapnell & Salzberg, Nature Biotechnology 2009

Hash tables



Trapnell & Salzberg, Nature Biotechnology 2009

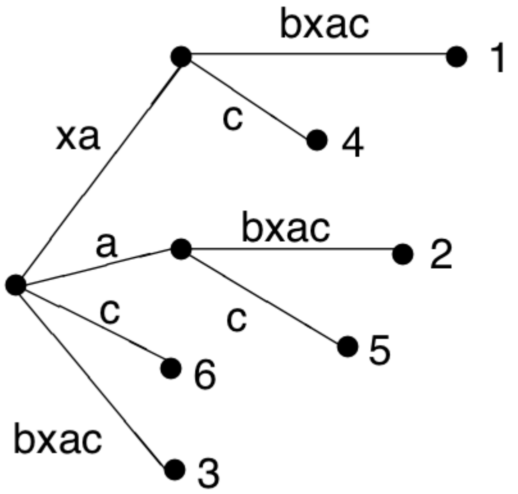
Hash tables



Trapnell & Salzberg, Nature Biotechnology 2009

Suffix Tree/Array

The suffix tree for string $\begin{array}{cccccc}1 & 2 & 3 & 4 & 5 & 6\\ x & a & b & x & a & c\end{array}$:

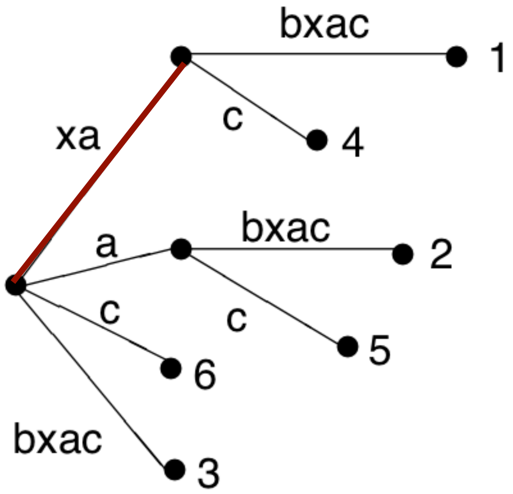


Query: xa

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Suffix Tree/Array

The suffix tree for string $\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ x & a & b & x & a & c \end{array}$:

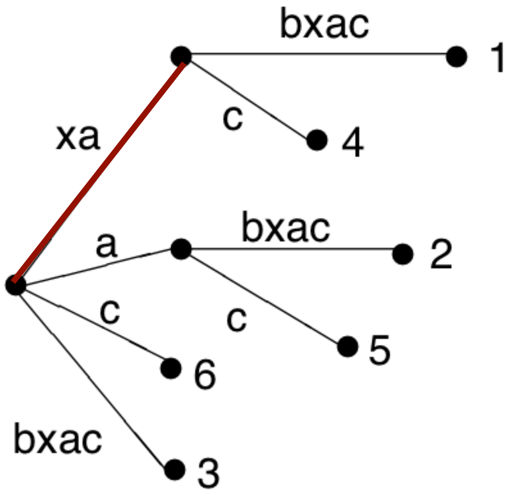


Query: xa

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Suffix Tree/Array

The suffix tree for string $\begin{matrix}1&2&3&4&5&6\\x&a&b&x&a&c\end{matrix}$:



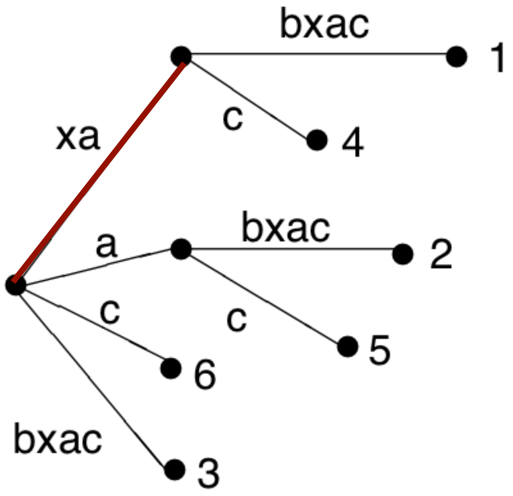
Query: xa

Suffix array: [2 5 3 6 1 4]

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Suffix Tree/Array

The suffix tree for string $\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ x & a & b & x & a & c \end{array}$:



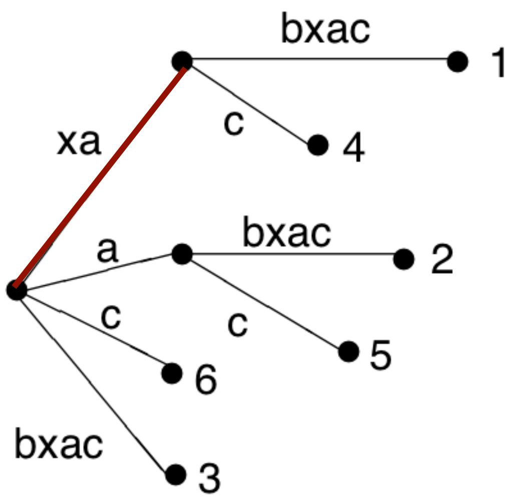
Query: xa

Suffix array: [2 5 3 6 1 4]

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Suffix Tree/Array

The suffix tree for string $\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ x & a & b & x & a & c \end{array}$:



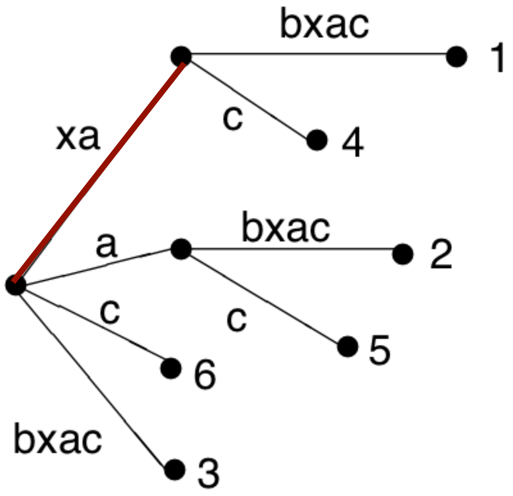
Query: xa

Suffix array: [2 5 3 6 1 4]

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Suffix Tree/Array

The suffix tree for string $\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ x & a & b & x & a & c \end{array}$:



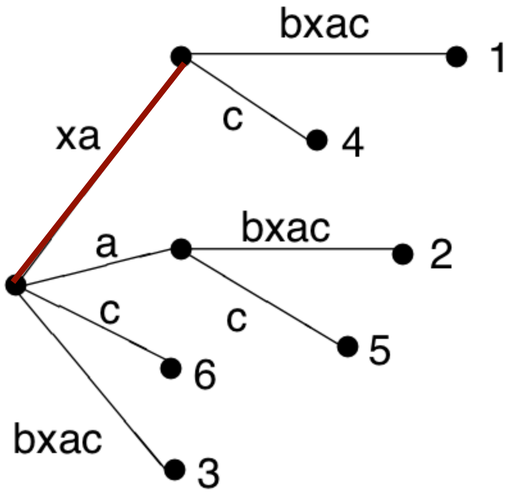
Query: xa

Suffix array: [2 5 3 6 1 4]

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Suffix Tree/Array

The suffix tree for string $\begin{array}{cccccc}1 & 2 & 3 & 4 & 5 & 6 \\ x & a & b & x & a & c\end{array}$:



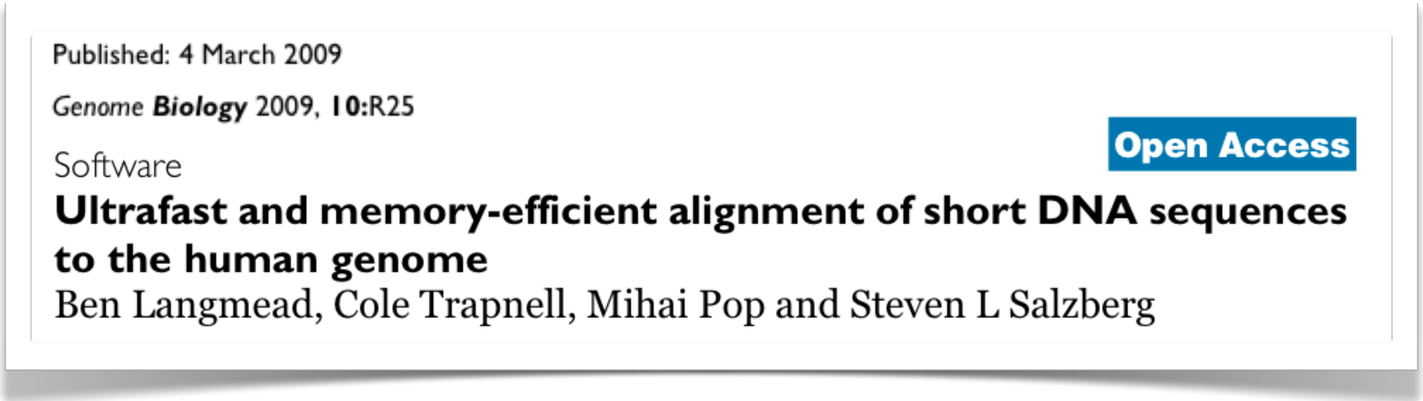
Query: xa

Suffix array: [2 5 3 6 1 4]

(3 Gbase * 64 bit = 24 Gbytes)

Adapted from lecture slides of Dr. Pekka Kilpela (http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides06.pdf)

Burrows-Wheeler Transform



Published: 4 March 2009

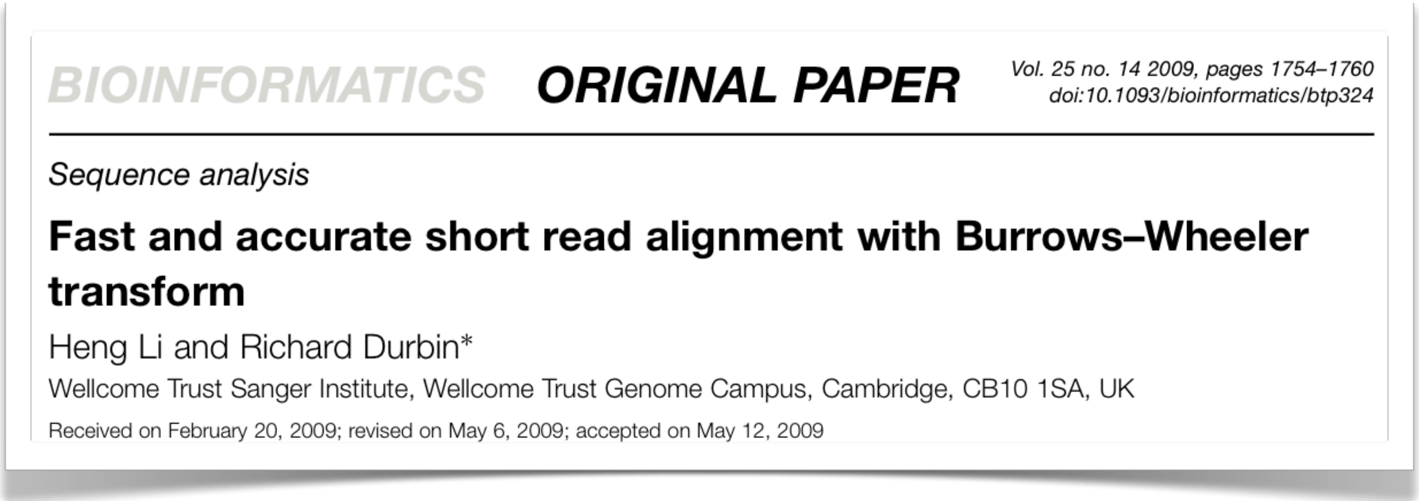
Genome ***Biology*** 2009, **10**:R25

Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg



BIOINFORMATICS ***ORIGINAL PAPER***

Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

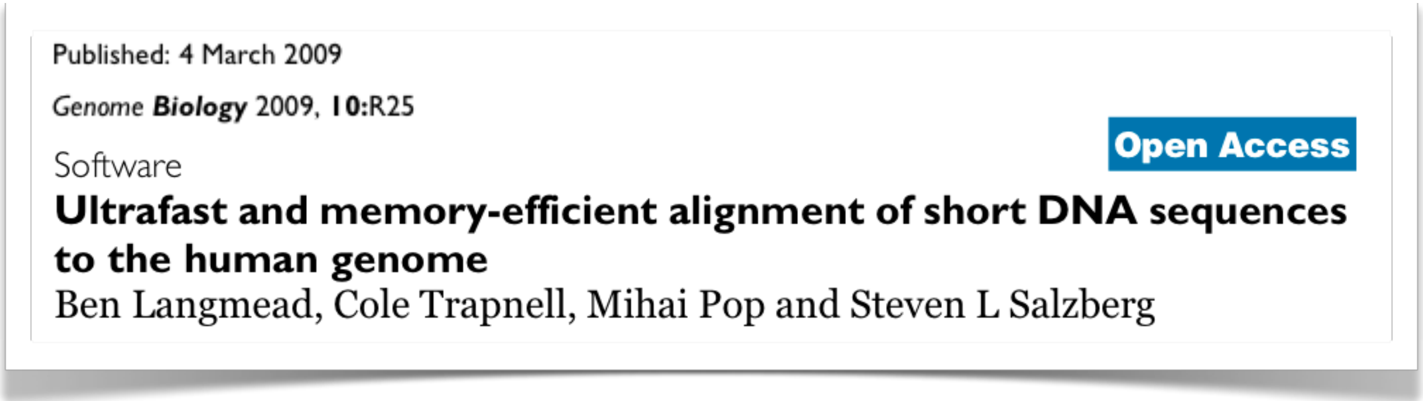
Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Burrows-Wheeler Transform

Cited by 447 (May 2011)



Published: 4 March 2009

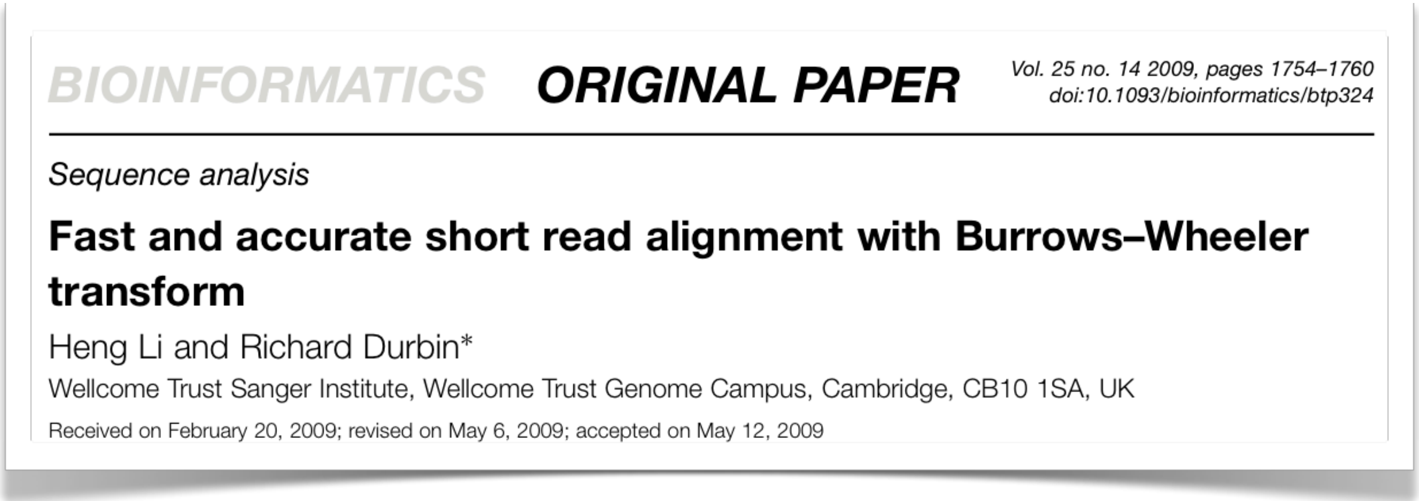
Genome ***Biology*** 2009, **10:**R25

Open Access

Software

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg



BIOINFORMATICS ***ORIGINAL PAPER*** *Vol. 25 no. 14 2009, pages 1754–1760*
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

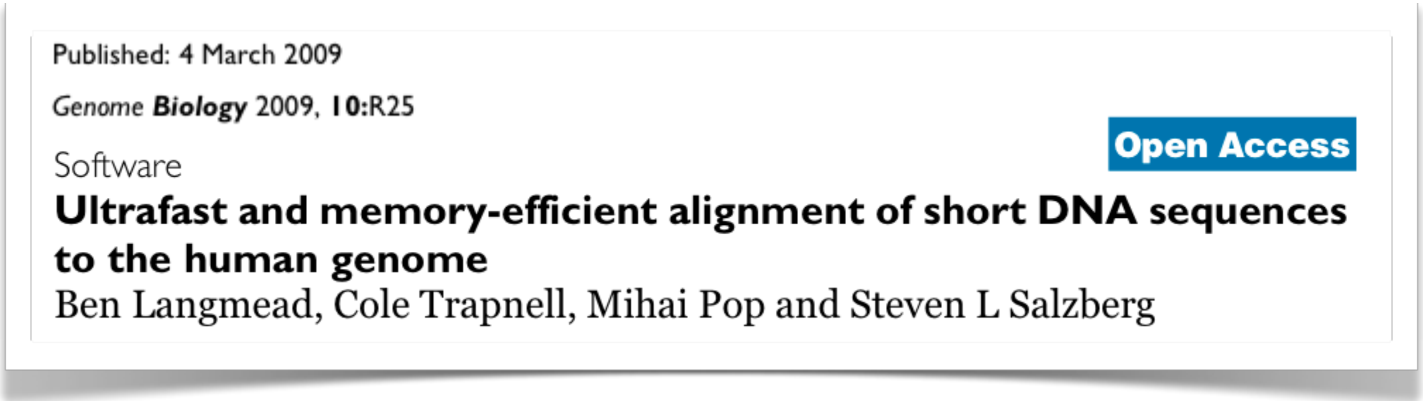
Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Burrows-Wheeler Transform

Cited by 447 (May 2011)



Published: 4 March 2009

Genome **Biology** 2009, **10:**R25

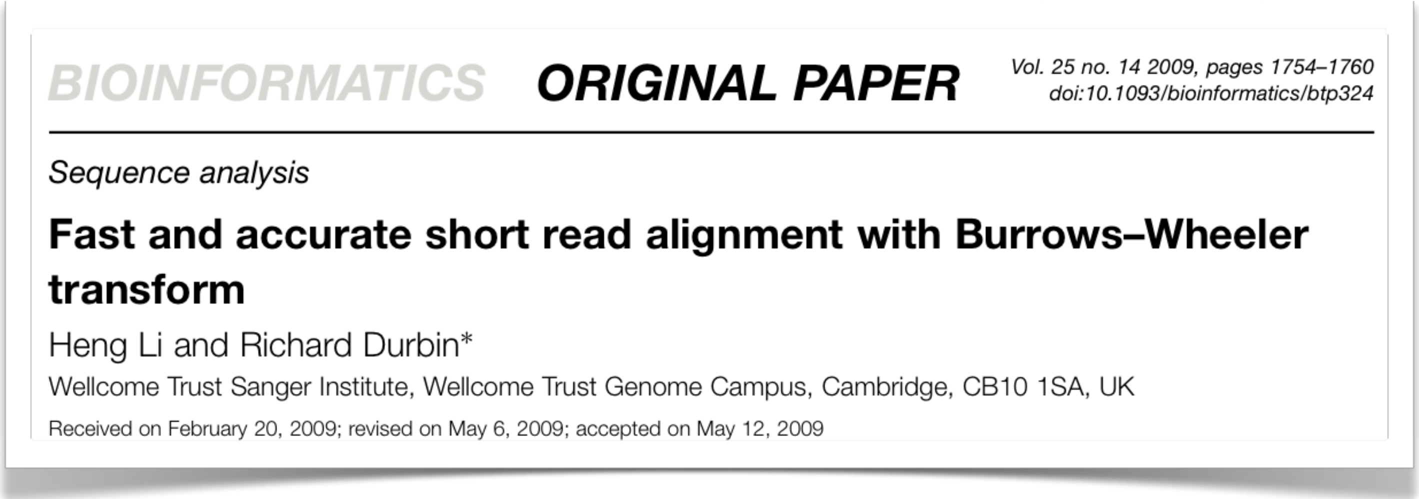
Open Access

Software

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Cited by 246 (May 2011)



BIOINFORMATICS ***ORIGINAL PAPER***

Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

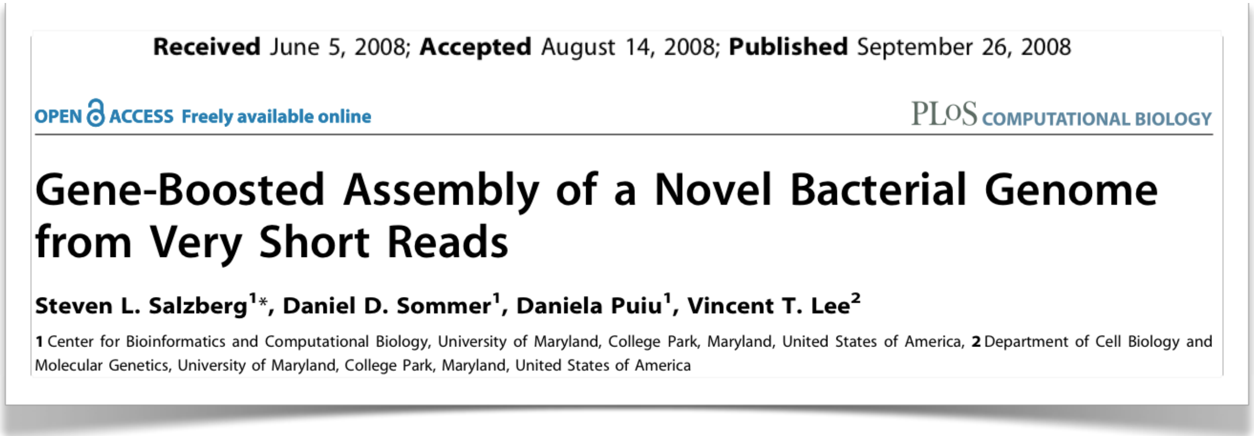
Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Phylogeny-guided Genome Assembly



Received June 5, 2008; **Accepted** August 14, 2008; **Published** September 26, 2008

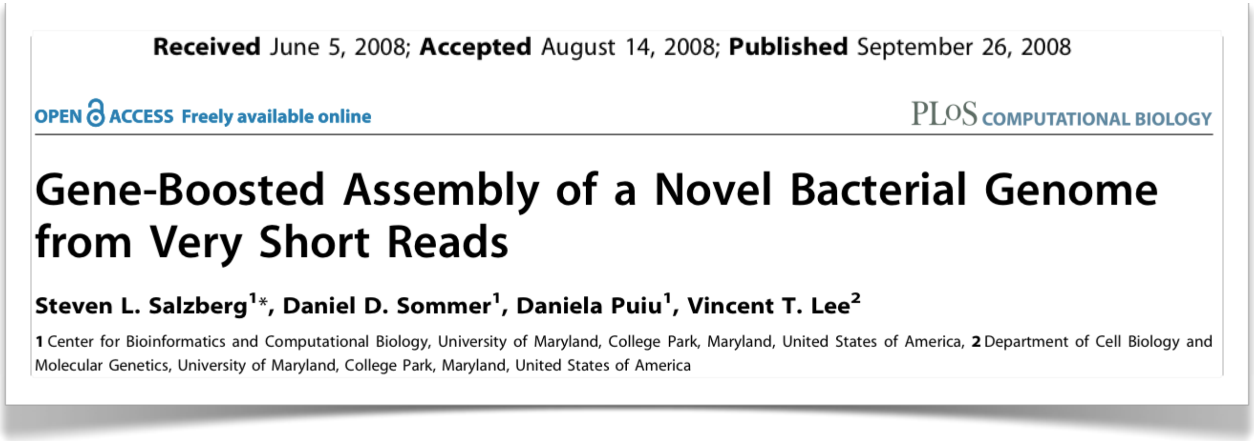
OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Gene-Boosted Assembly of a Novel Bacterial Genome from Very Short Reads

Steven L. Salzberg[1]*, Daniel D. Sommer[1], Daniela Puiu[1], Vincent T. Lee[2]

1 Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America, **2** Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland, United States of America

- Use multiple genomes to try to bridge as many gaps as possible.
- Use library of protein-coding genes to bridge further gaps (protein evolve slower)
- Do *de novo* assembly of unmapped contigs.



Received June 5, 2008; **Accepted** August 14, 2008; **Published** September 26, 2008

OPEN ACCESS Freely available online

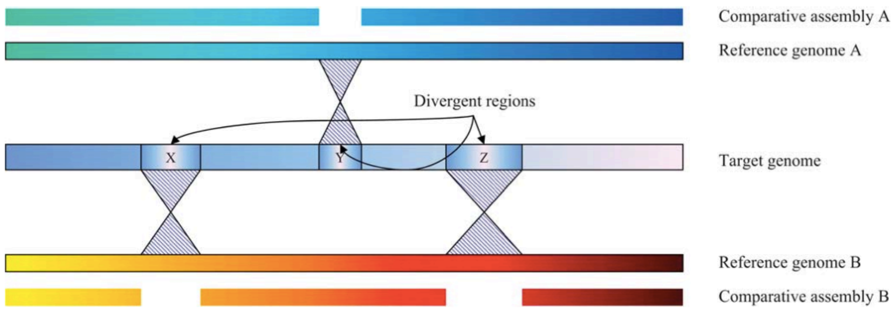
PLoS COMPUTATIONAL BIOLOGY

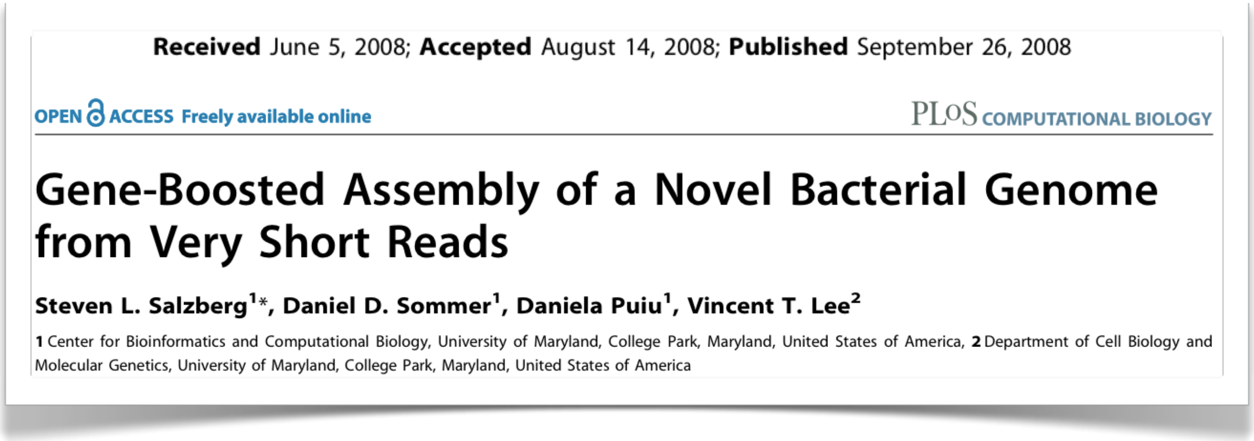
Gene-Boosted Assembly of a Novel Bacterial Genome from Very Short Reads

Steven L. Salzberg[1]*, Daniel D. Sommer[1], Daniela Puiu[1], Vincent T. Lee[2]

[1] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America, [2] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland, United States of America

- Use multiple genomes to try to bridge as many gaps as possible.
- Use library of protein-coding genes to bridge further gaps (protein evolve slower)
- Do *de novo* assembly of unmapped contigs.





Received June 5, 2008; **Accepted** August 14, 2008; **Published** September 26, 2008

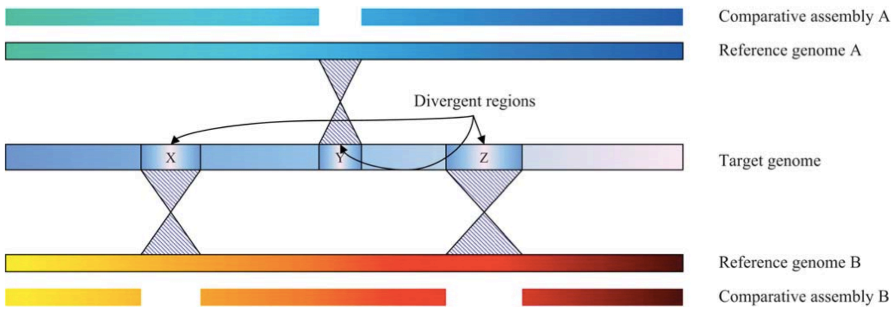
OPEN ACCESS Freely available online PLOS COMPUTATIONAL BIOLOGY

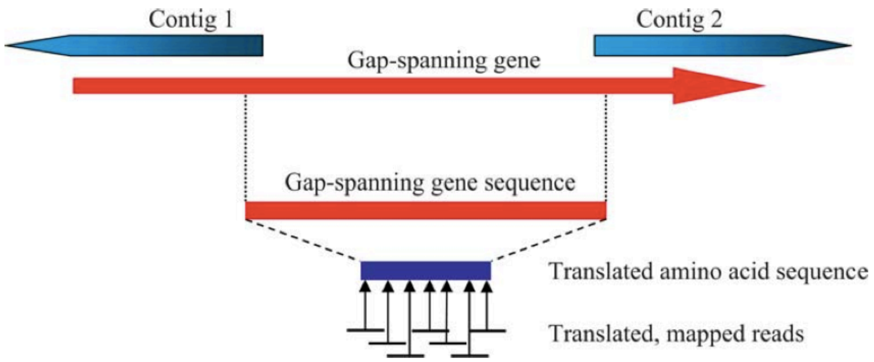
Gene-Boosted Assembly of a Novel Bacterial Genome from Very Short Reads

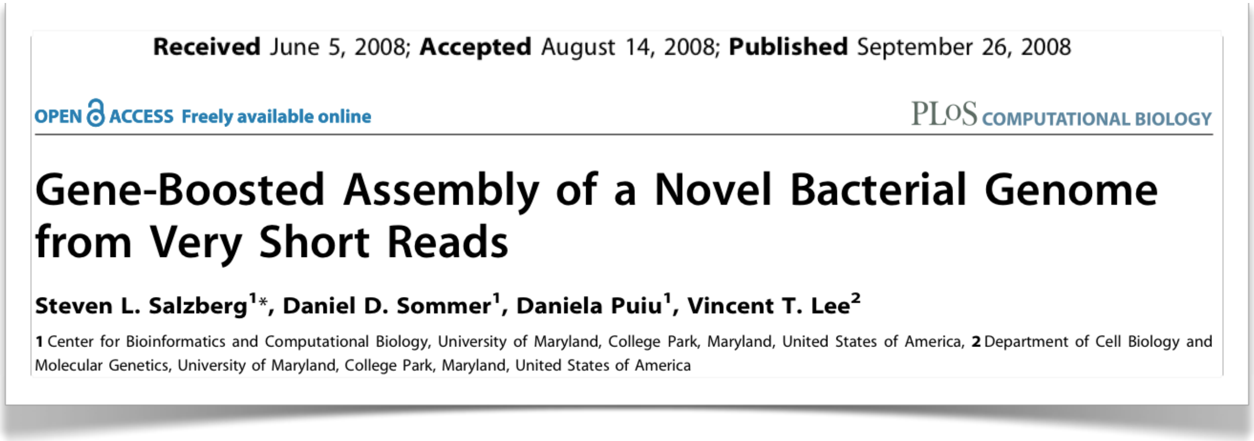
Steven L. Salzberg[1]*, Daniel D. Sommer[1], Daniela Puiu[1], Vincent T. Lee[2]

1 Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America, **2** Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland, United States of America

- Use multiple genomes to try to bridge as many gaps as possible.
- Use library of protein-coding genes to bridge further gaps (protein evolve slower)







Received June 5, 2008; **Accepted** August 14, 2008; **Published** September 26, 2008

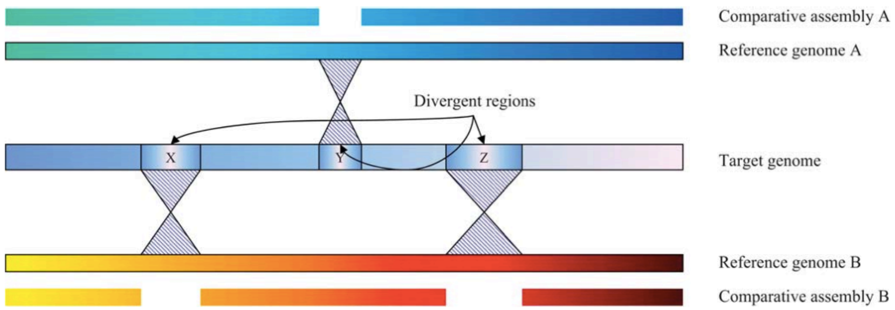
OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

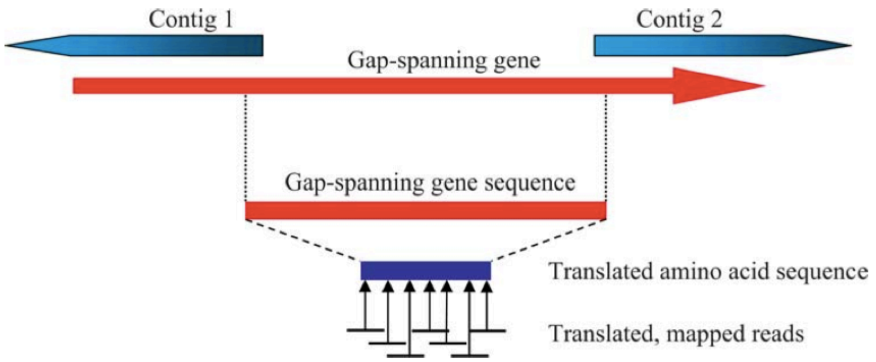
Gene-Boosted Assembly of a Novel Bacterial Genome from Very Short Reads

Steven L. Salzberg[1]*, Daniel D. Sommer[1], Daniela Puiu[1], Vincent T. Lee[2]

[1] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America, [2] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland, United States of America

- Use multiple genomes to try to bridge as many gaps as possible.
- Use library of protein-coding genes to bridge further gaps (protein evolve slower)
- Do *de novo* assembly of unmapped contigs.



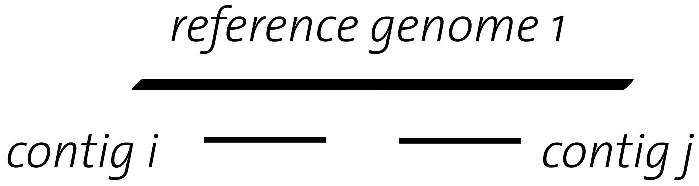


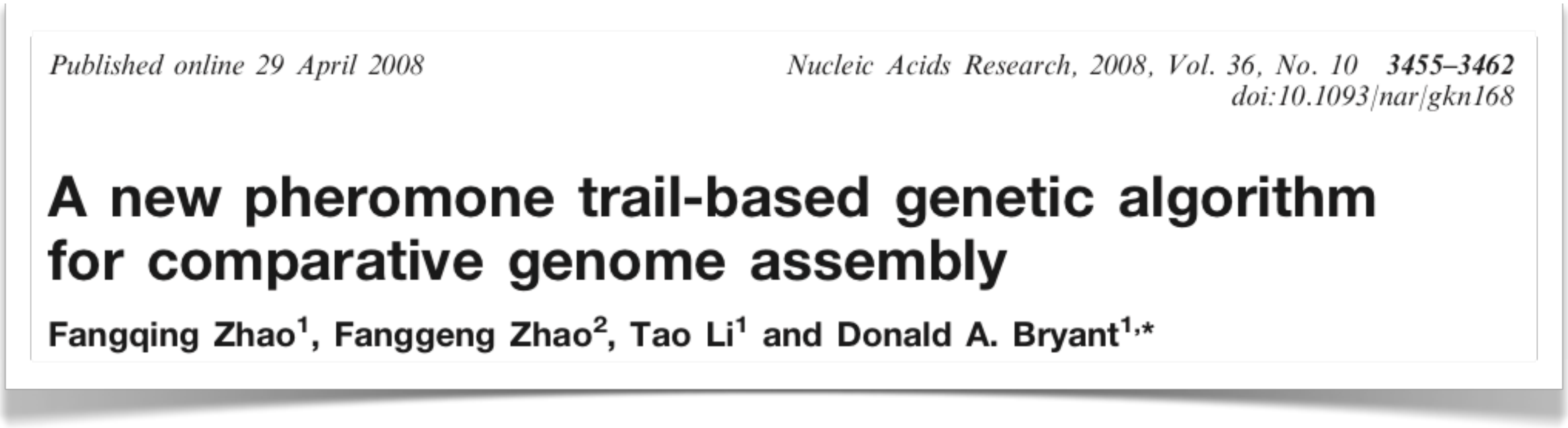
Published online 29 April 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 10* **3455–3462**
doi:10.1093/nar/gkn168

A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]

- **Define a distance matrix ("fitness matrix") between every pair of contig.**





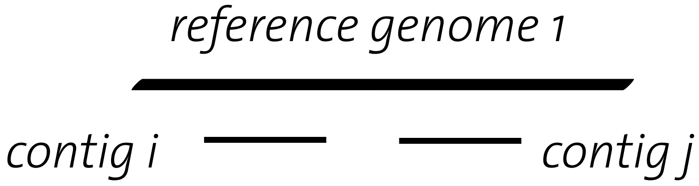
Published online 29 April 2008 Nucleic Acids Research, 2008, Vol. 36, No. 10 **3455–3462**
doi:10.1093/nar/gkn168

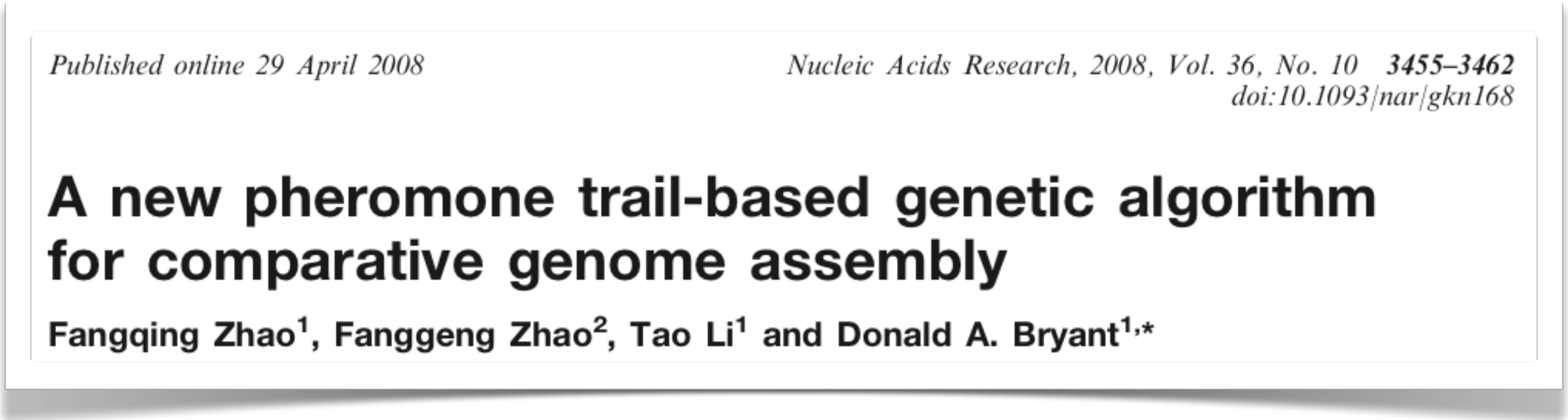
A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]

- **Define a distance matrix ("fitness matrix") between every pair of contig.**

- **Model several reference genome by averaging the fitness matrices obtained with each genome.**





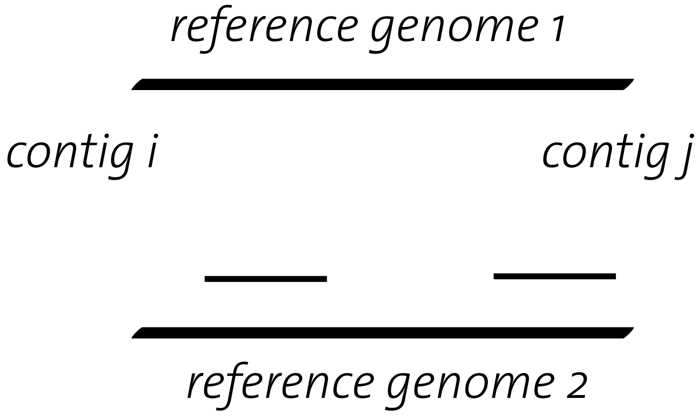
Published online 29 April 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 10* **3455–3462**
doi:10.1093/nar/gkn168

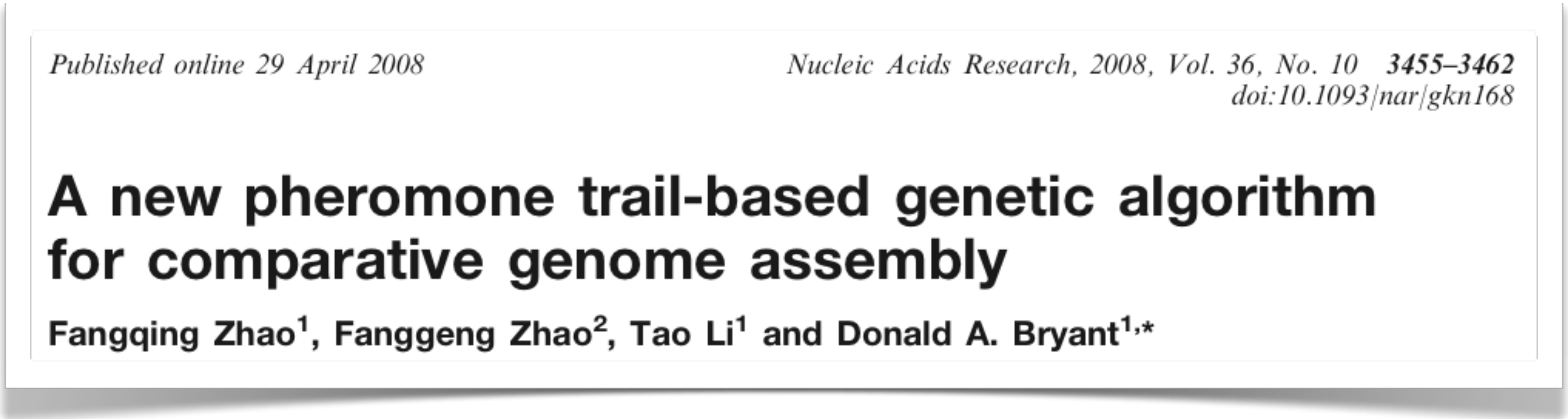
A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]

- **Define a distance matrix ("fitness matrix") between every pair of contig.**

- **Model several reference genome by averaging the fitness matrices obtained with each genome.**





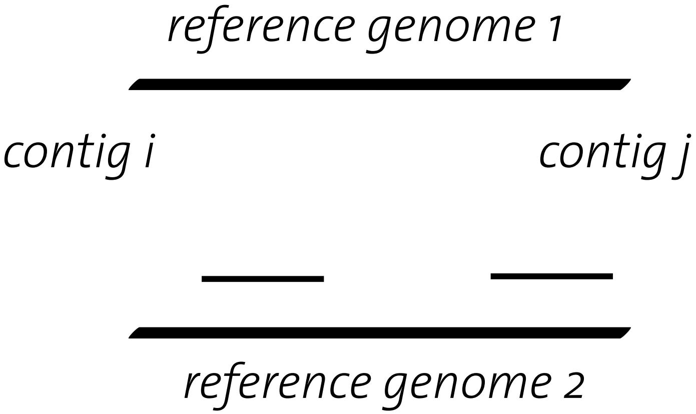
Published online 29 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 10 3455–3462
doi:10.1093/nar/gkn168

A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]

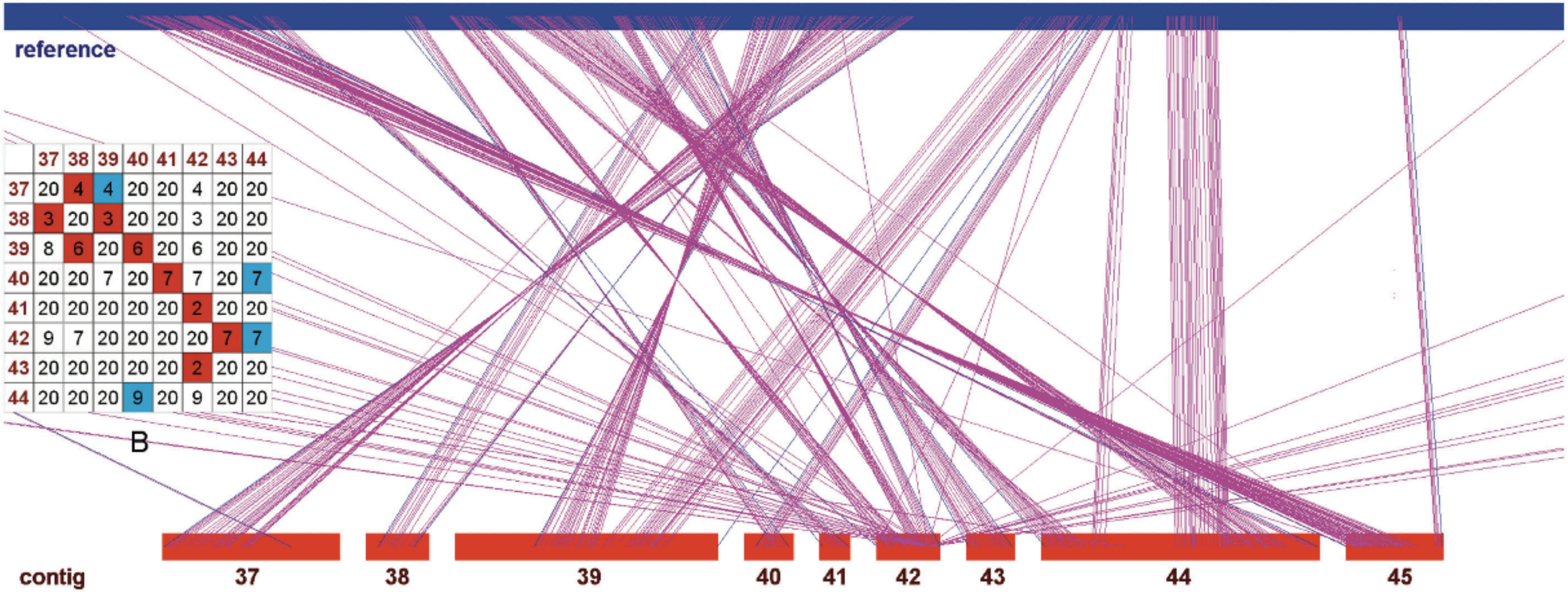
- Define a distance matrix (“fitness matrix”) between every pair of contig.
- Model several reference genome by averaging the fitness matrices obtained with each genome.
- Use a genetic algorithm to identify the best ordering of contig (one with highest “fitness”)



Published online 29 April 2008 Nucleic Acids Research, 2008, Vol. 36, No. 10 3455–3462
doi:10.1093/nar/gkn168

A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]



	37	38	39	40	41	42	43	44
37	20	4	4	20	20	4	20	20
38	3	20	3	20	20	3	20	20
39	8	6	20	6	20	6	20	20
40	20	20	7	20	7	7	20	7
41	20	20	20	20	20	2	20	20
42	9	7	20	20	20	20	7	7
43	20	20	20	20	20	2	20	20
44	20	20	20	9	20	9	20	20

Published online 29 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 10 3455–3462
doi:10.1093/nar/gkn168

A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]

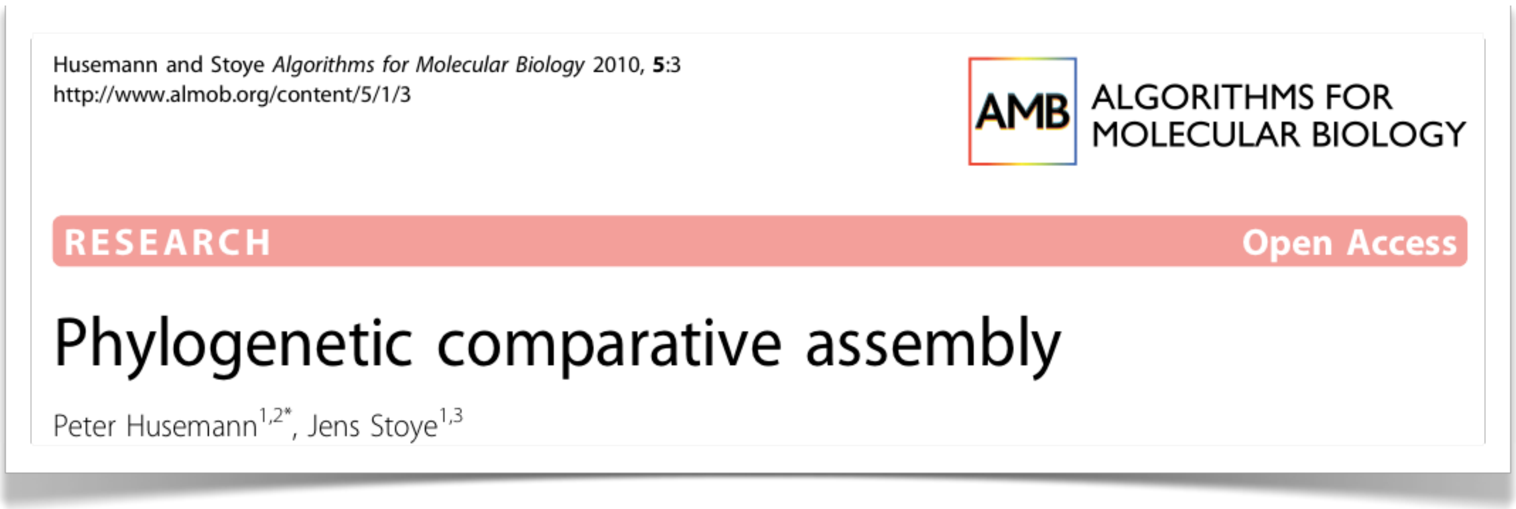
		Reference Plut	
		Best	Average
Clim	PGA	0.378	0.346 ± 0.026
	BLAST-end	0.135	NA
	Projector2	0.162	NA
	OSLay	0.108	NA
Cvib	PGA	0.769	0.738 ± 0.015
	BLAST-end	0.538	NA
	Projector2	0.577	NA
	OSLay	0.423	NA
Cpar	PGA	0.586	0.559 ± 0.018
	BLAST-end	0.172	NA
	Projector2	0.155	NA
	OSLay	0.103	NA

Published online 29 April 2008 Nucleic Acids Research, 2008, Vol. 36, No. 10 3455–3462
doi:10.1093/nar/gkn168

A new pheromone trail-based genetic algorithm for comparative genome assembly

Fangqing Zhao[1], Fanggeng Zhao[2], Tao Li[1] and Donald A. Bryant[1,*]

		Reference Plut		2 or 3 Refs	
		Best	Average	Best	Average
Clim	PGA	0.378	0.346 ± 0.026	0.514[b]	0.443 ± 0.040[b]
				NA	Na
	BLAST-end	0.135	NA	NA	NA
	Projector2	0.162	NA	NA	NA
	OSLay	0.108	NA	NA	NA
Cvib	PGA	0.769	0.738 ± 0.015	0.731[c]	0.731 ± 0.000[c]
	BLAST-end	0.538	NA	NA	NA
	Projector2	0.577	NA	NA	NA
	OSLay	0.423	NA	NA	NA
Cpar	PGA	0.586	0.559 ± 0.018	0.741[d]	0.738 ± 0.007[d]
				NA	NA
	BLAST-end	0.172	NA	NA	NA
	Projector2	0.155	NA	NA	NA
	OSLay	0.103	NA	NA	NA



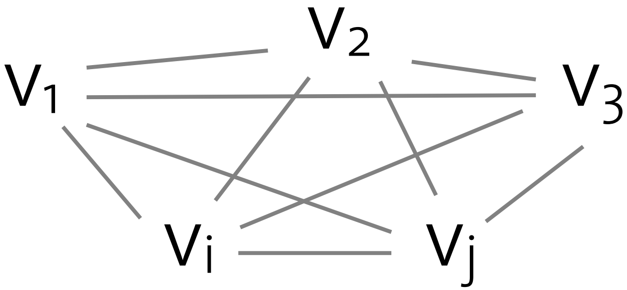
Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

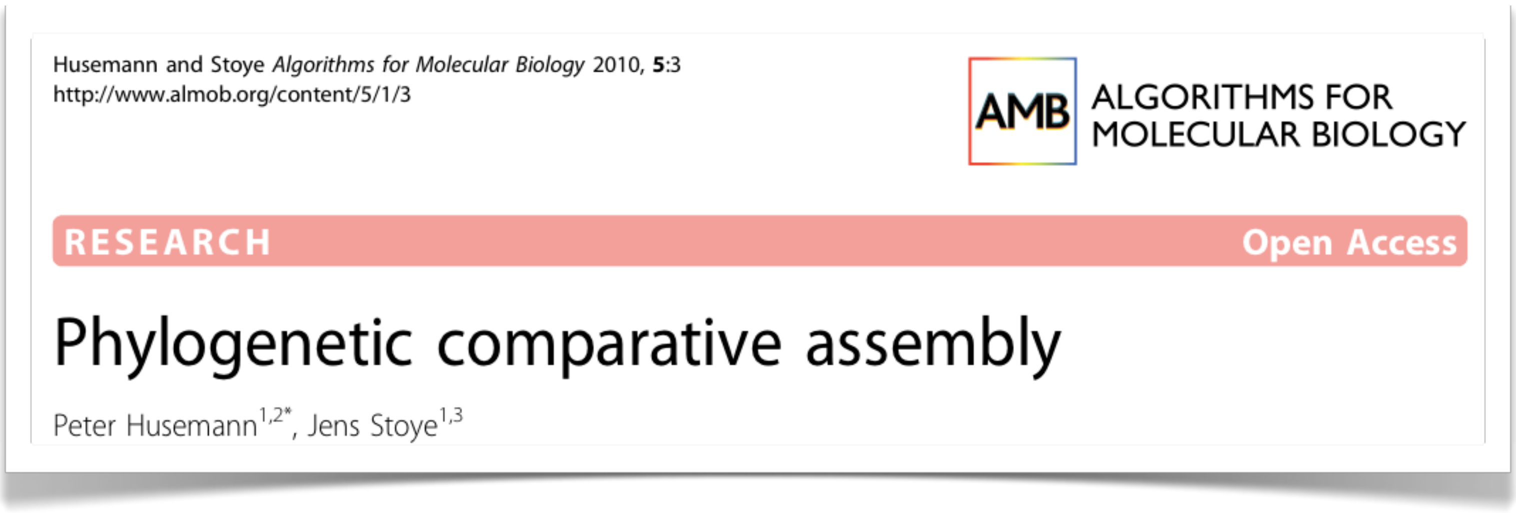
AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Phylogenetic comparative assembly

Peter Husemann[1,2*], Jens Stoye[1,3]





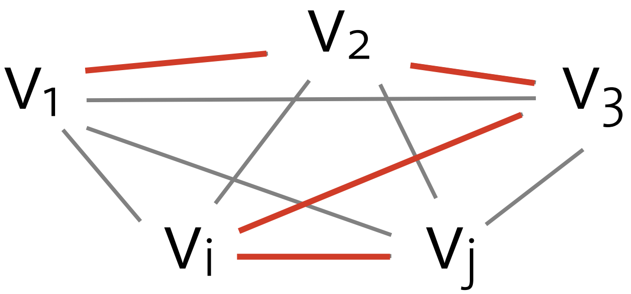
Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

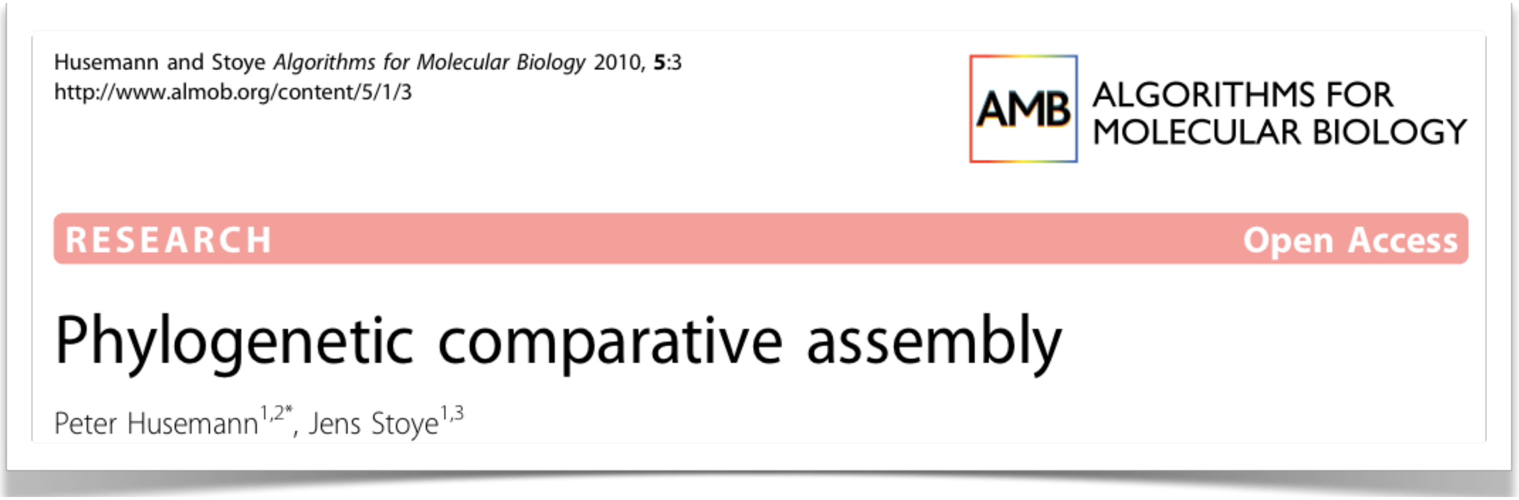
AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Phylogenetic comparative assembly

Peter Husemann[1,2*], Jens Stoye[1,3]





Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

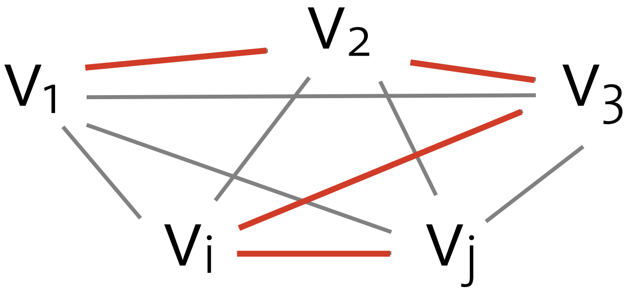
AMB ALGORITHMS FOR MOLECULAR BIOLOGY

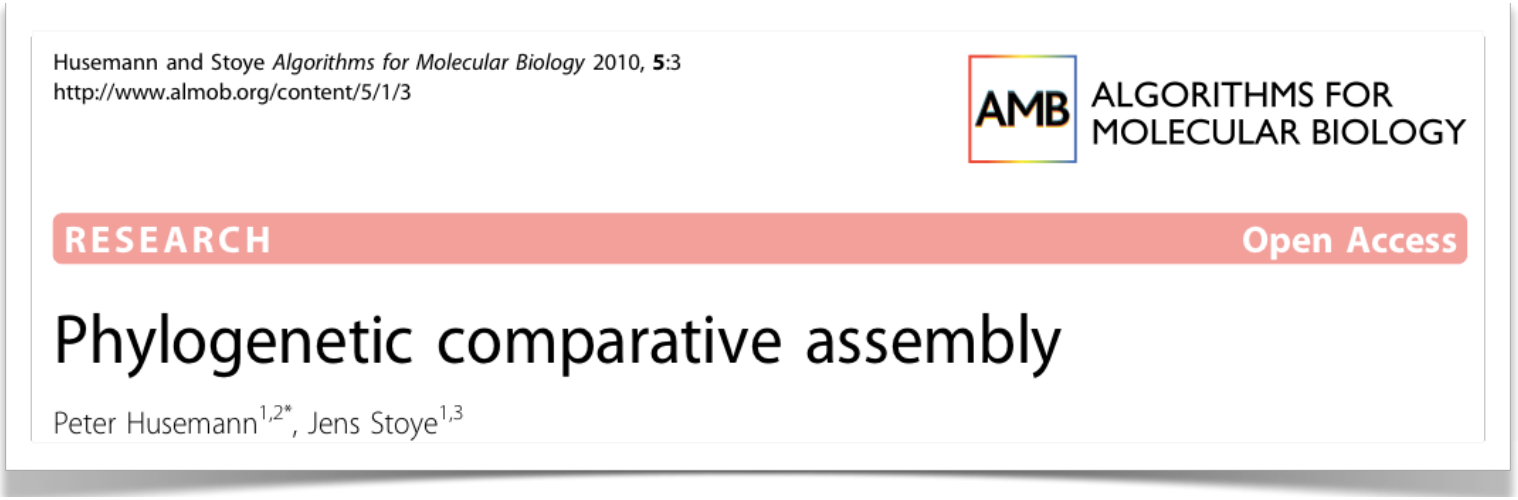
RESEARCH Open Access

Phylogenetic comparative assembly

Peter Husemann[1,2*], Jens Stoye[1,3]

$$w_r(v_i, v_j) = \sum_{m_i^r \in \mathcal{M}_i^r, m_j^r \in \mathcal{M}_j^r} s\Big(d(\pi(m_i^r), \pi(m_j^r)), d_{\mathcal{T}} \Big) \cdot \mathrm{qhits}(m_i^r) \cdot \mathrm{qhits}(m_j^r)$$





Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

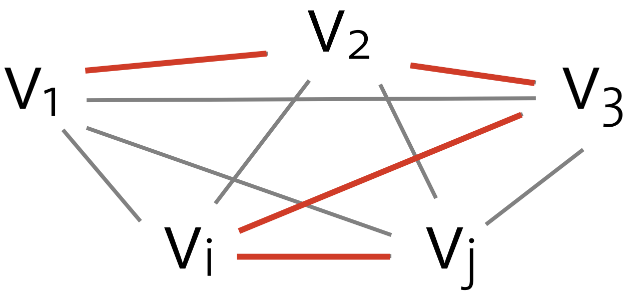
RESEARCH Open Access

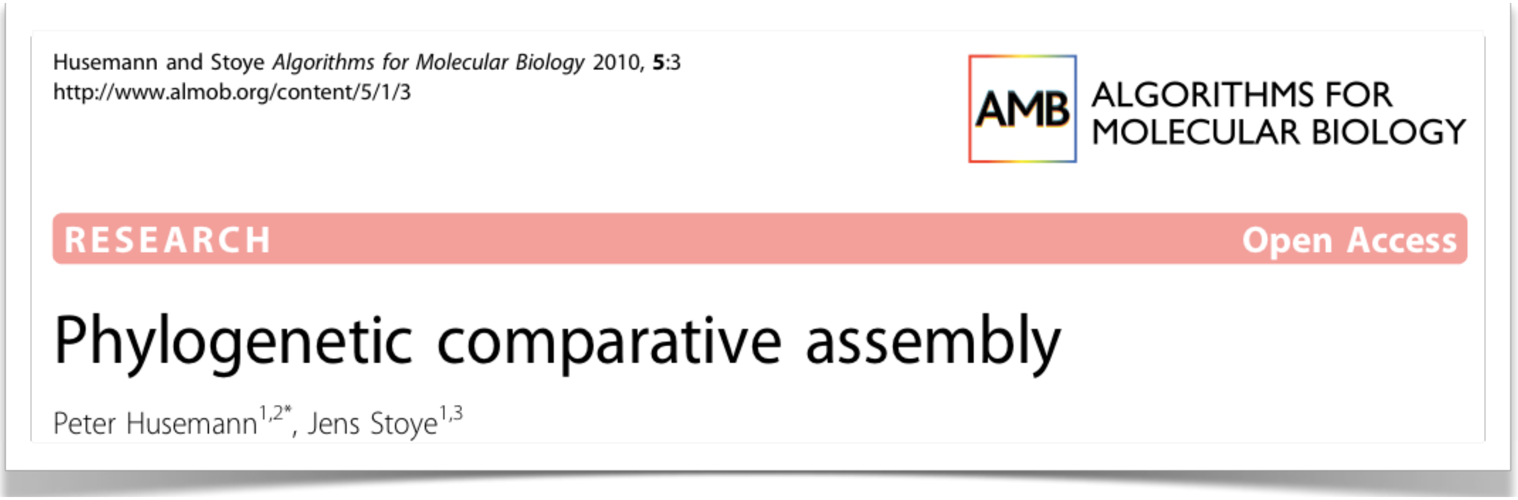
Phylogenetic comparative assembly

Peter Husemann[1,2*], Jens Stoye[1,3]

$$w_r(v_i, v_j) = \sum_{m_i^r \in \mathcal{M}_i^r, m_j^r \in \mathcal{M}_j^r} s\left(d(\pi(m_i^r), \pi(m_j^r)), d_T \right) \cdot \text{qhits}(m_i^r) \cdot \text{qhits}(m_j^r)$$

Weight of edge
between
two contigs





Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH **Open Access**

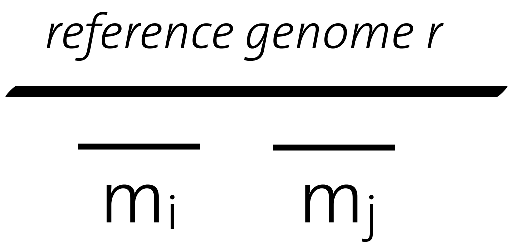
Phylogenetic comparative assembly

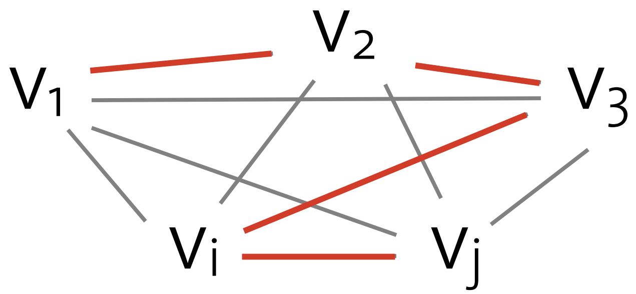
Peter Husemann[1,2*], Jens Stoye[1,3]

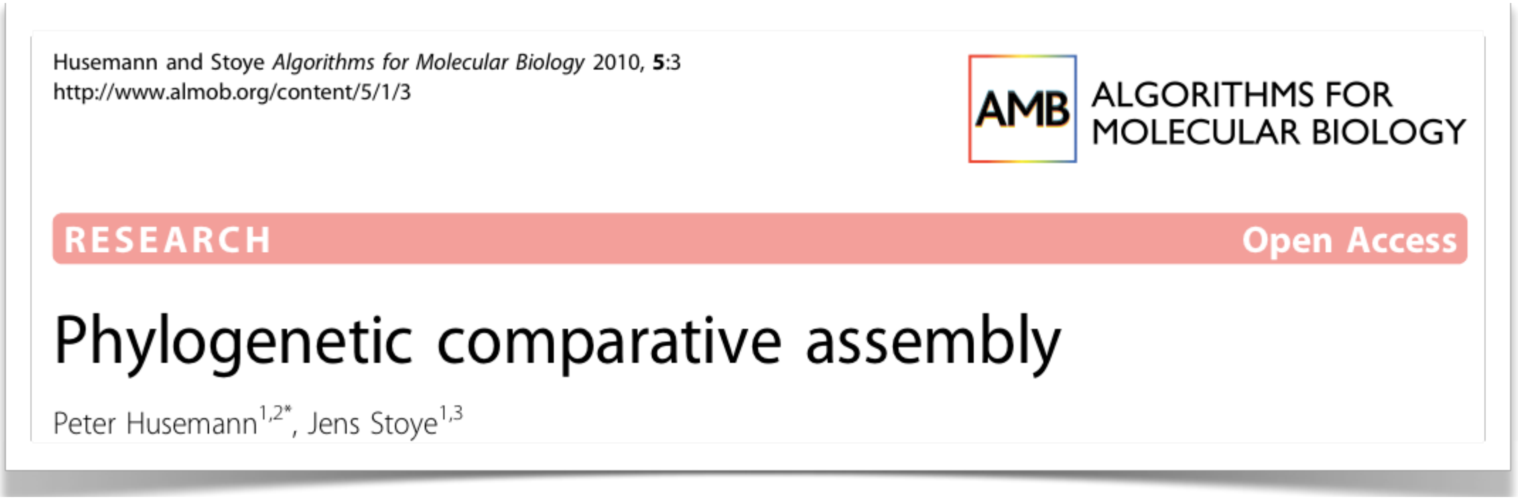
$$w_r(v_i, v_j) = \sum_{m_i^r \in \mathcal{M}_i^r, m_j^r \in \mathcal{M}_j^r} s\left(d(\pi(m_i^r), \pi(m_j^r)), d_{\mathcal{T}} \right) \cdot \mathrm{qhits}(m_i^r) \cdot \mathrm{qhits}(m_j^r)$$

Weight of edge between two contigs

All pairs of matches between $\{v_i, v_j\} \times r$







Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Phylogenetic comparative assembly

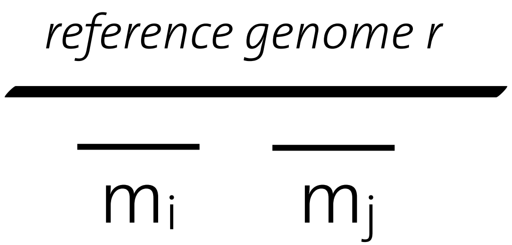
Peter Husemann[1,2*], Jens Stoye[1,3]

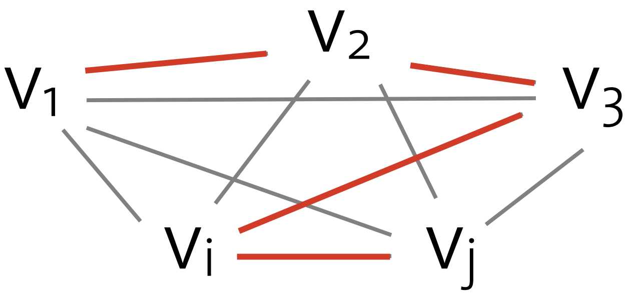
$$w_r(v_i, v_j) = \sum_{m_i^r \in \mathcal{M}_i^r, m_j^r \in \mathcal{M}_j^r} s\left(d(\pi(m_i^r), \pi(m_j^r)), d_T \right) \cdot \mathrm{qhits}(m_i^r) \cdot \mathrm{qhits}(m_j^r)$$

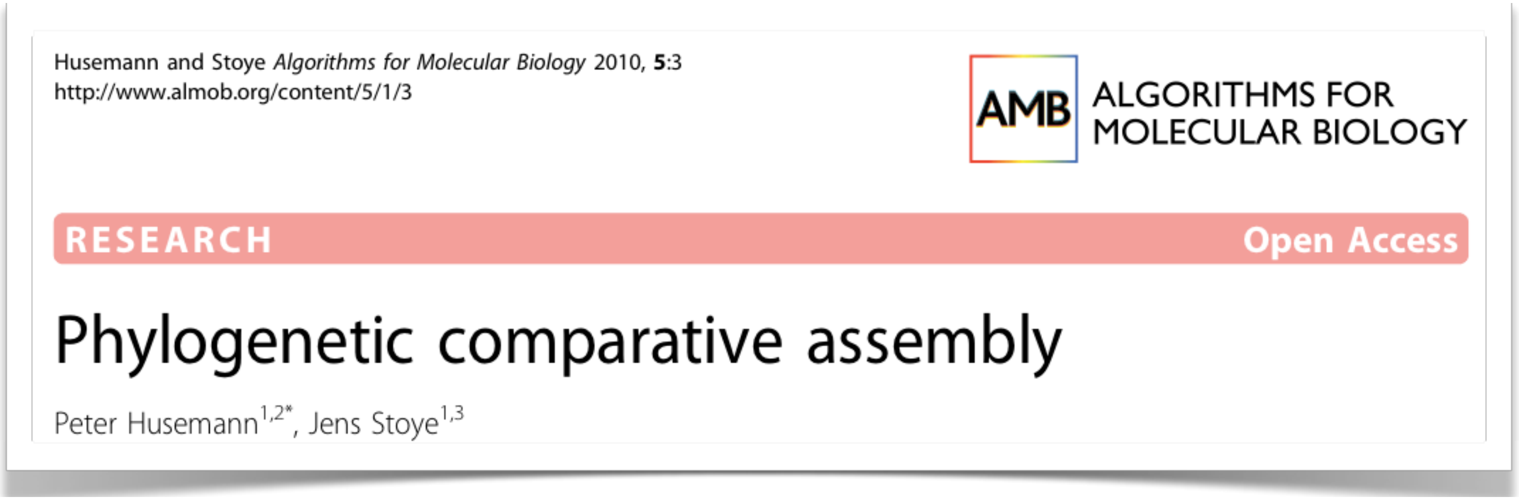
Weight of edge between two contigs

All pairs of matches between {v_i,v_j} x r

score depends on dist. between matches and phylogenetic distance to ref. genome







Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH **Open Access**

Phylogenetic comparative assembly

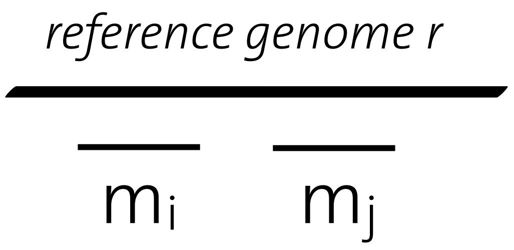
Peter Husemann[1,2*], Jens Stoye[1,3]

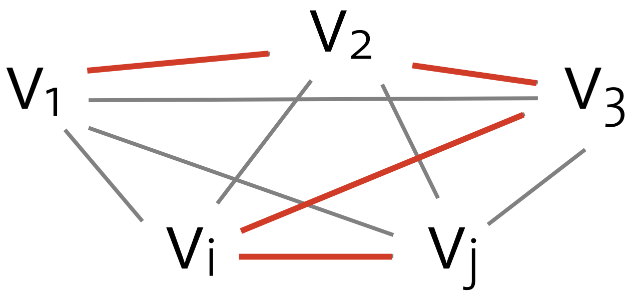
$$w_r(v_i,v_j) = \sum_{m_i^r \in \mathcal{M}_i^r,\, m_j^r \in \mathcal{M}_j^r} s\left(d(\pi(m_i^r),\pi(m_j^r)), d_{\mathcal{T}}\right) \cdot \mathrm{qhits}(m_i^r) \cdot \mathrm{qhits}(m_j^r)$$

Weight of edge between two contigs

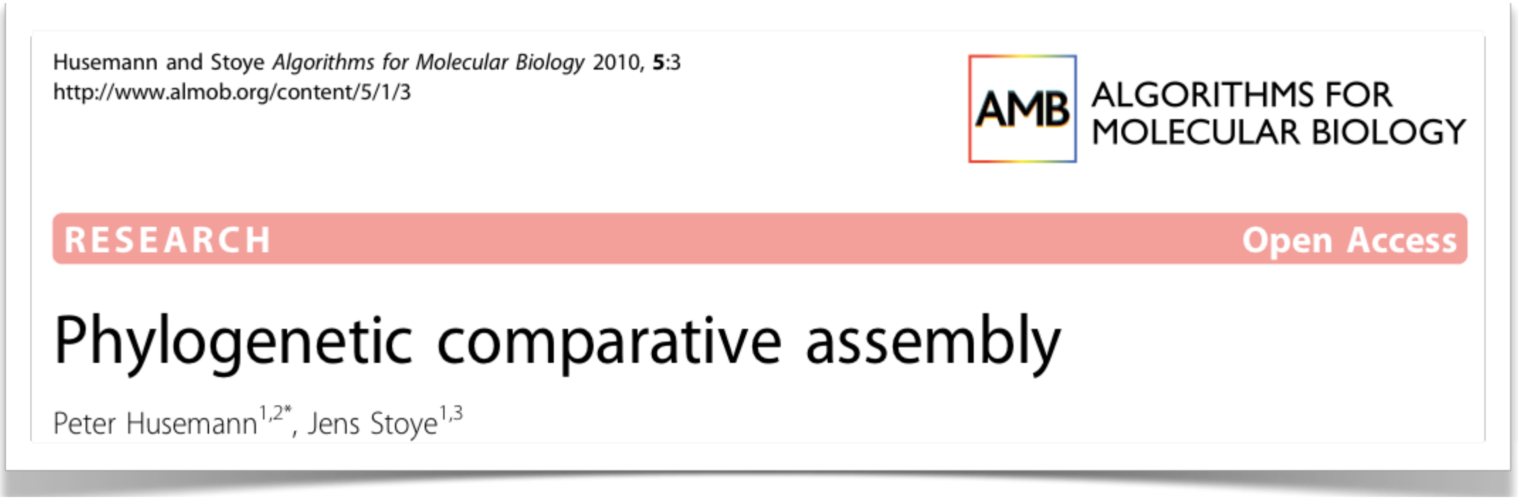
All pairs of matches between {v_i,v_j} x r

score depends on dist. between matches and phylogenetic distance to ref. genome





$$s(d,d_{\mathcal{T}}) := \frac{1}{d_{\mathcal{T}} \cdot \sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d}{d_{\mathcal{T}} \cdot \sigma}\right)^2}$$



Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Phylogenetic comparative assembly

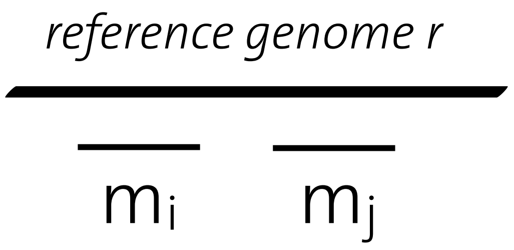
Peter Husemann[1,2*], Jens Stoye[1,3]

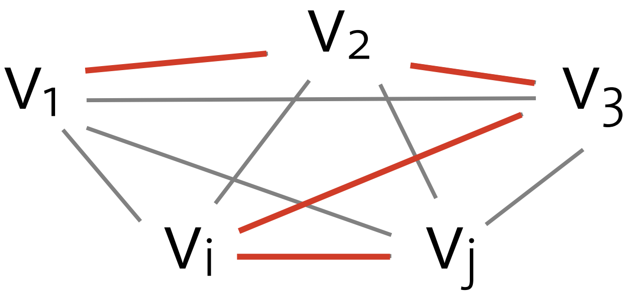
$$w_r(v_i, v_j) = \sum_{m_i^r \in \mathcal{M}_i^r, m_j^r \in \mathcal{M}_j^r} s\left(d(\pi(m_i^r), \pi(m_j^r)), d_{\mathcal{T}} \right) \cdot \mathrm{qhits}(m_i^r) \cdot \mathrm{qhits}(m_j^r)$$

Weight of edge between two contigs

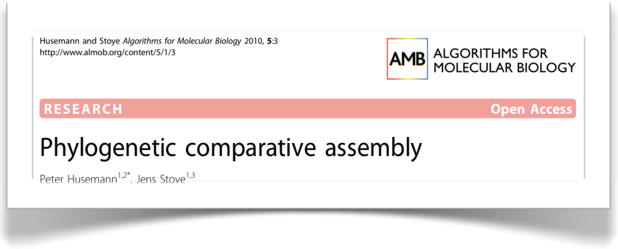
All pairs of matches between {v_i,v_j} x r

score depends on dist. between matches and phylogenetic distance to ref. genome





$$s(d, d_{\mathcal{T}}) := \frac{(1-\varphi)}{d \cdot d_{\mathcal{T}} \cdot \sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d}{d_{\mathcal{T}} \cdot \sigma_1}\right)^2} + \frac{\varphi}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d-\mu}{\sigma_2}\right)^2}$$



Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

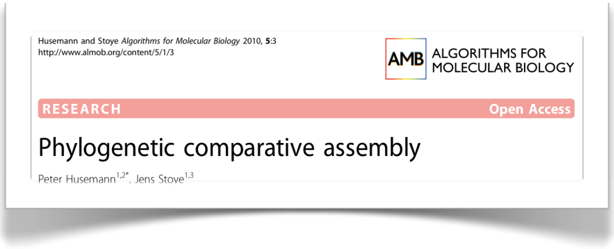
RESEARCH Open Access

Phylogenetic comparative assembly

Peter Husemann[1,2*], Jens Stoye[1,3]

Closest species as reference

Organism	Closest Reference	OSLay		Projector2	
		TP	FP	TP	FP
C. aurimucosum	*C. glutamicum*	0	1	10	20
C. kroppenstedtii	*C. jeikeium*	0	0	1	2
C. urealyticum	*C. jeikeium*	6	6	8	18



Husemann and Stoye *Algorithms for Molecular Biology* 2010, **5**:3
http://www.almob.org/content/5/1/3

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Phylogenetic comparative assembly

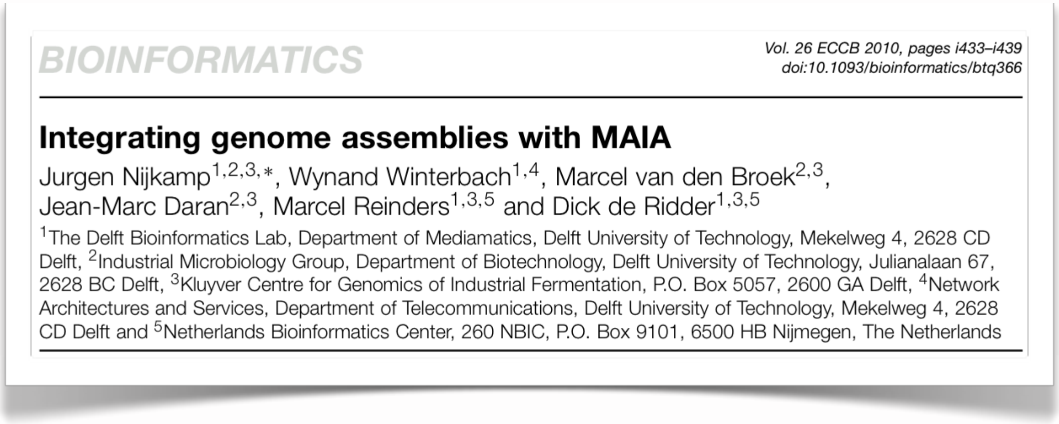
Peter Husemann[1,2*], Jens Stoye[1,3]

Closest species as reference

Organism	Closest Reference	OSLay		Projector2	
		TP	FP	TP	FP
C. aurimucosum	*C. glutamicum*	0	1	10	20
C. kroppenstedtii	*C. jeikeium*	0	0	1	2
C. urealyticum	*C. jeikeium*	6	6	8	18

PGA		*treecat*	
TP	FP	TP	FP
14.5 (16)	66.5 (70)	**17**	**66**
2.0 (2)	**4.0 (4)**	**3**	6
20.9 (25)	72.5 (76)	**27**	**70**

Multiple reference species



BIOINFORMATICS

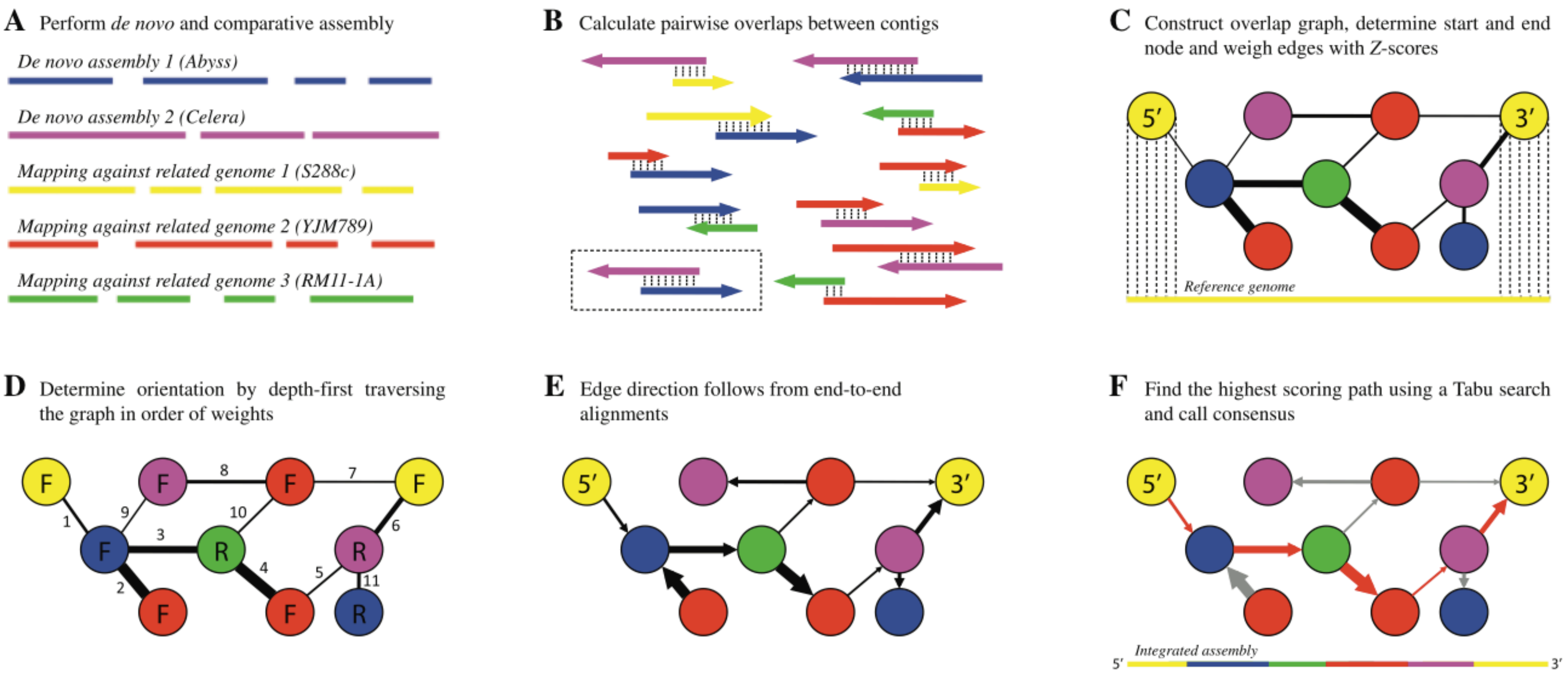
Vol. 26 ECCB 2010, pages i433–i439
doi:10.1093/bioinformatics/btq366

Integrating genome assemblies with MAIA

Jurgen Nijkamp[1,2,3,*], Wynand Winterbach[1,4], Marcel van den Broek[2,3], Jean-Marc Daran[2,3], Marcel Reinders[1,3,5] and Dick de Ridder[1,3,5]

[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Industrial Microbiology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, [4]Network Architectures and Services, Department of Telecommunications, Delft University of Technology, Mekelweg 4, 2628 CD Delft and [5]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

A “meta” assembler



A Perform *de novo* and comparative assembly

B Calculate pairwise overlaps between contigs

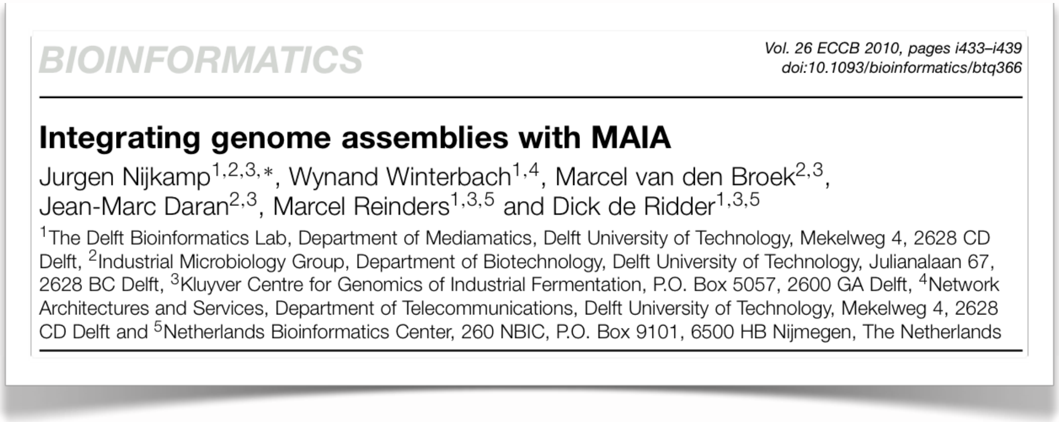
C Construct overlap graph, determine start and end node and weigh edges with *Z*-scores

D Determine orientation by depth-first traversing the graph in order of weights

E Edge direction follows from end-to-end alignments

F Find the highest scoring path using a Tabu search and call consensus

De novo assembly 1 (Abyss)
De novo assembly 2 (Celera)
Mapping against related genome 1 (S288c)
Mapping against related genome 2 (YJM789)
Mapping against related genome 3 (RM11-1A)
5′
3′
Reference genome
F
F
F
F
F
R
R
F
F
R
1
2
3
4
5
6
7
8
9
10
11
5′
3′
5′
3′
Integrated assembly
5′
3′



BIOINFORMATICS

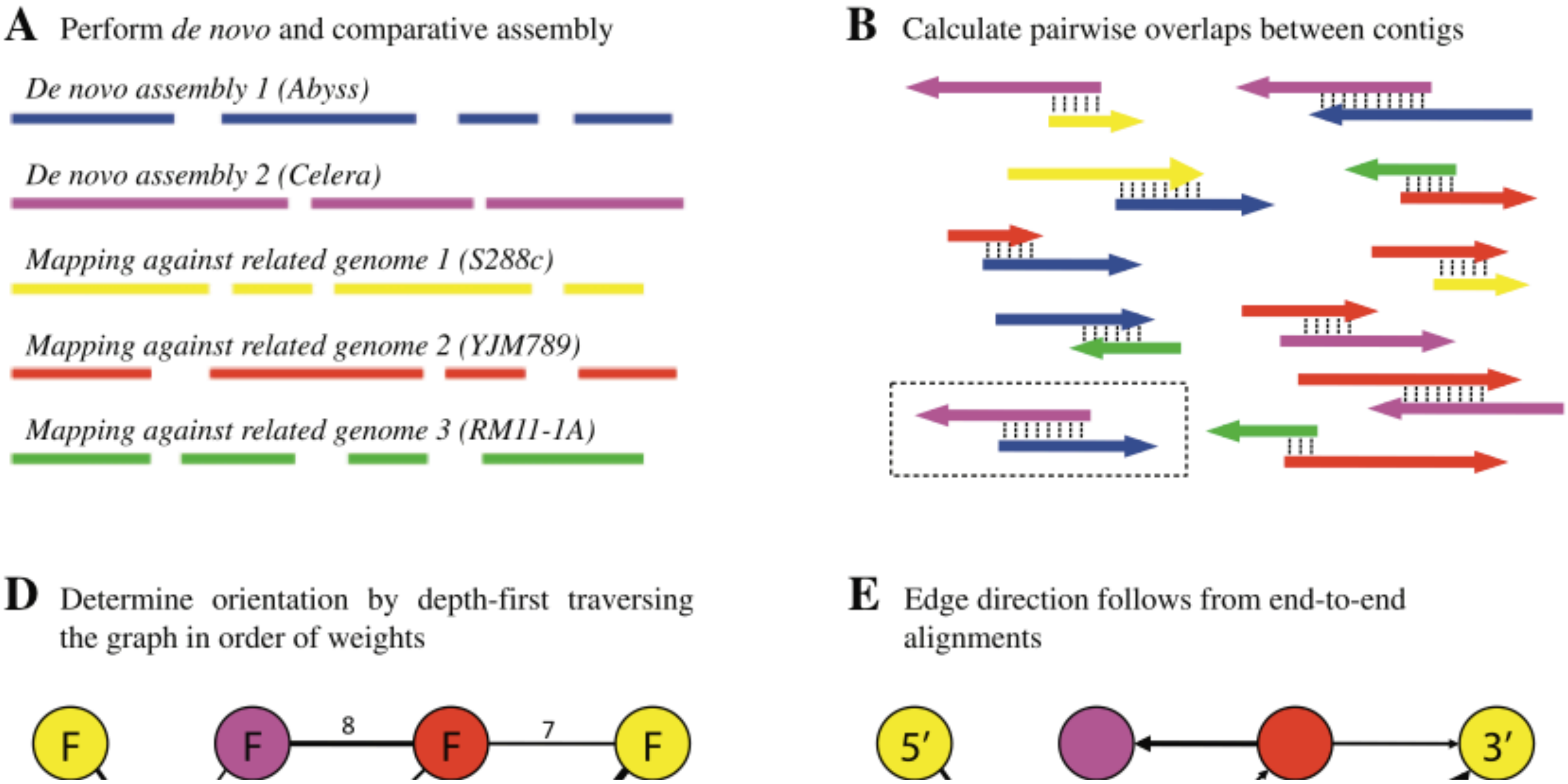
Vol. 26 ECCB 2010, pages i433–i439
doi:10.1093/bioinformatics/btq366

Integrating genome assemblies with MAIA

Jurgen Nijkamp[1,2,3,*], Wynand Winterbach[1,4], Marcel van den Broek[2,3], Jean-Marc Daran[2,3], Marcel Reinders[1,3,5] and Dick de Ridder[1,3,5]

[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Industrial Microbiology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, [4]Network Architectures and Services, Department of Telecommunications, Delft University of Technology, Mekelweg 4, 2628 CD Delft and [5]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

A “meta” assembler



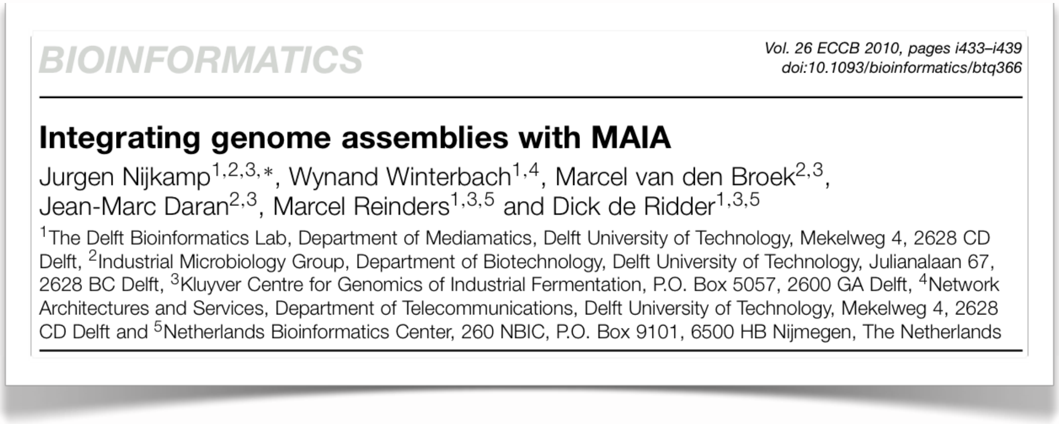
A Perform *de novo* and comparative assembly

B Calculate pairwise overlaps between contigs

D Determine orientation by depth-first traversing the graph in order of weights

E Edge direction follows from end-to-end alignments

De novo assembly 1 (Abyss)
De novo assembly 2 (Celera)
Mapping against related genome 1 (S288c)
Mapping against related genome 2 (YJM789)
Mapping against related genome 3 (RM11-1A)
F
F
8
F
7
F
5′
3′



BIOINFORMATICS

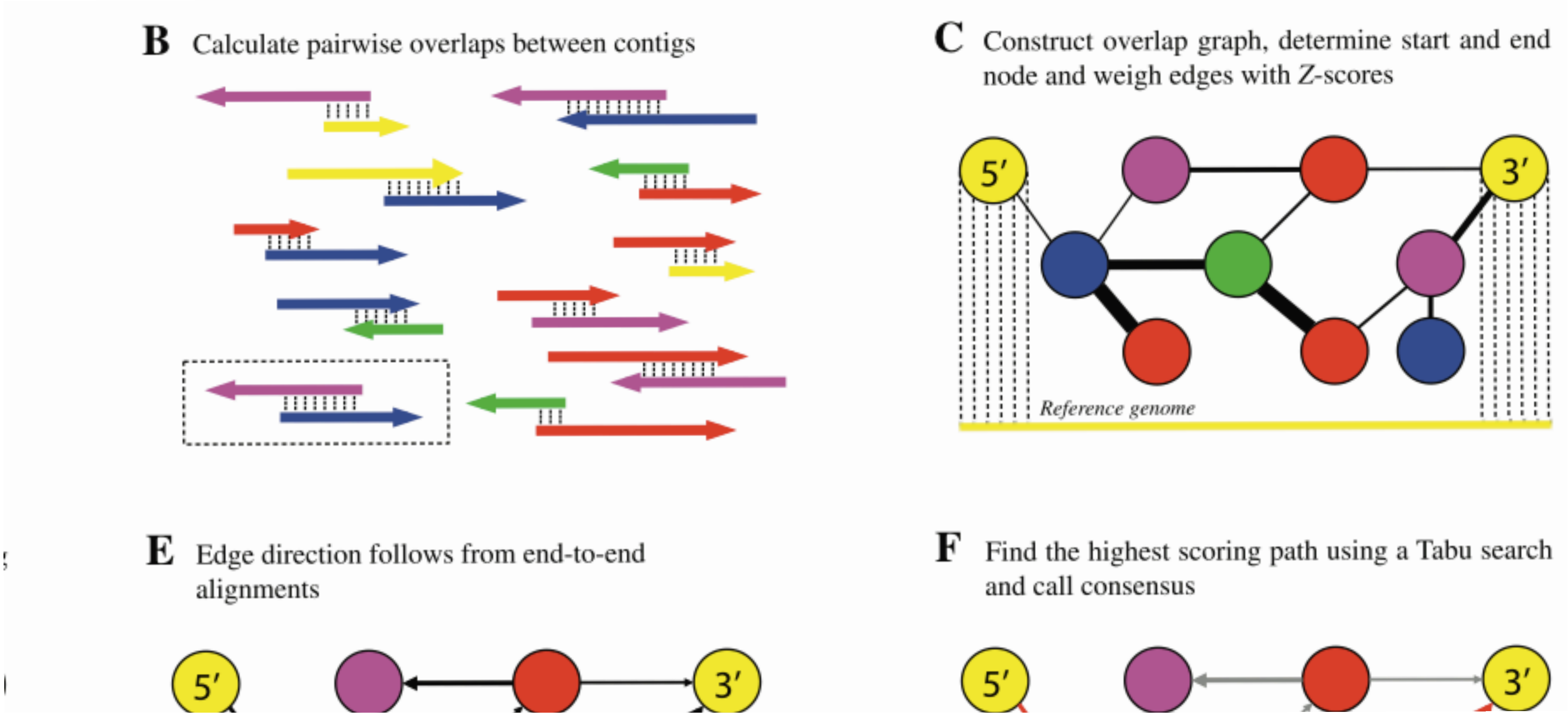
Vol. 26 ECCB 2010, pages i433–i439
doi:10.1093/bioinformatics/btq366

Integrating genome assemblies with MAIA

Jurgen Nijkamp[1,2,3,*], Wynand Winterbach[1,4], Marcel van den Broek[2,3], Jean-Marc Daran[2,3], Marcel Reinders[1,3,5] and Dick de Ridder[1,3,5]

[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Industrial Microbiology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, [4]Network Architectures and Services, Department of Telecommunications, Delft University of Technology, Mekelweg 4, 2628 CD Delft and [5]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

A “meta” assembler



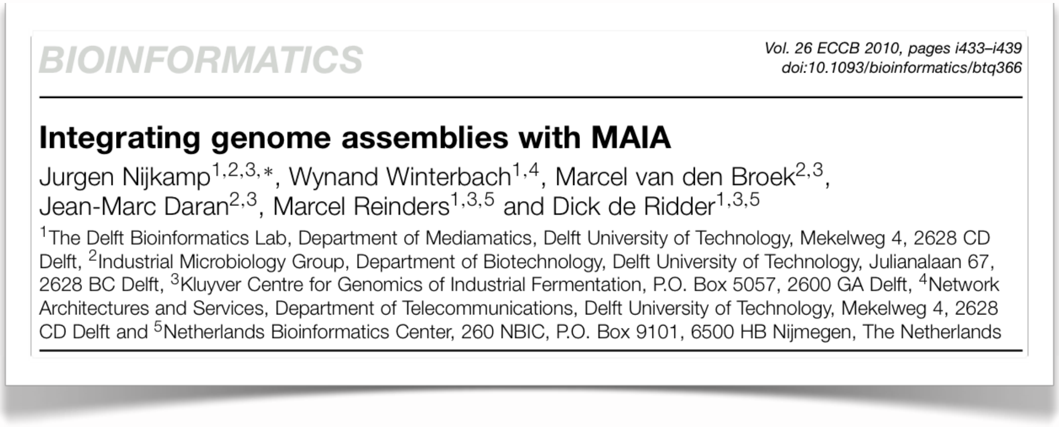
B Calculate pairwise overlaps between contigs

C Construct overlap graph, determine start and end node and weigh edges with *Z*-scores

E Edge direction follows from end-to-end alignments

F Find the highest scoring path using a Tabu search and call consensus

A “meta” assembler



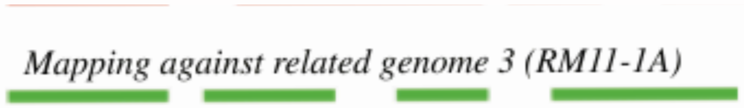
BIOINFORMATICS

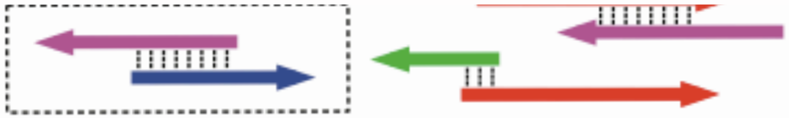
Vol. 26 ECCB 2010, pages i433–i439
doi:10.1093/bioinformatics/btq366

Integrating genome assemblies with MAIA

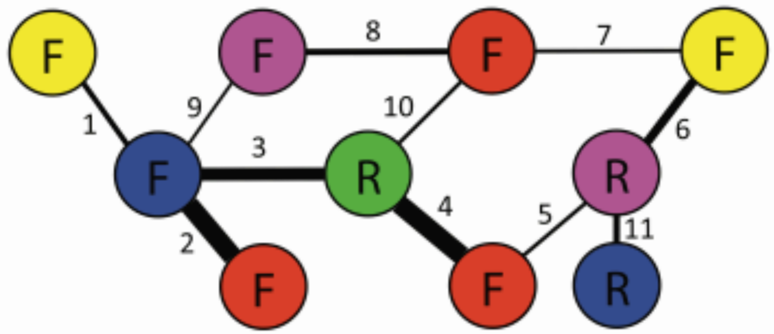
Jurgen Nijkamp[1,2,3,*], Wynand Winterbach[1,4], Marcel van den Broek[2,3], Jean-Marc Daran[2,3], Marcel Reinders[1,3,5] and Dick de Ridder[1,3,5]

[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Industrial Microbiology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, [4]Network Architectures and Services, Department of Telecommunications, Delft University of Technology, Mekelweg 4, 2628 CD Delft and [5]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

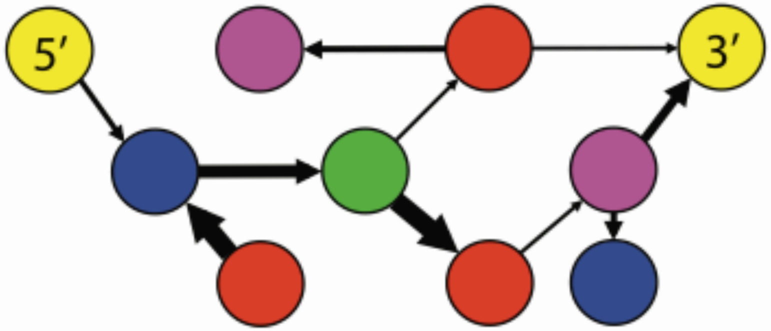


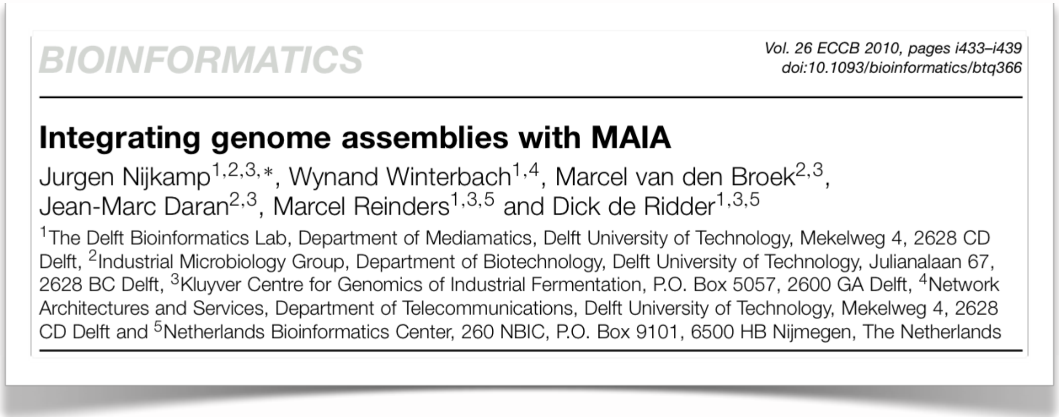


D Determine orientation by depth-first traversing the graph in order of weights



E Edge direction follows from end-to-end alignments





BIOINFORMATICS

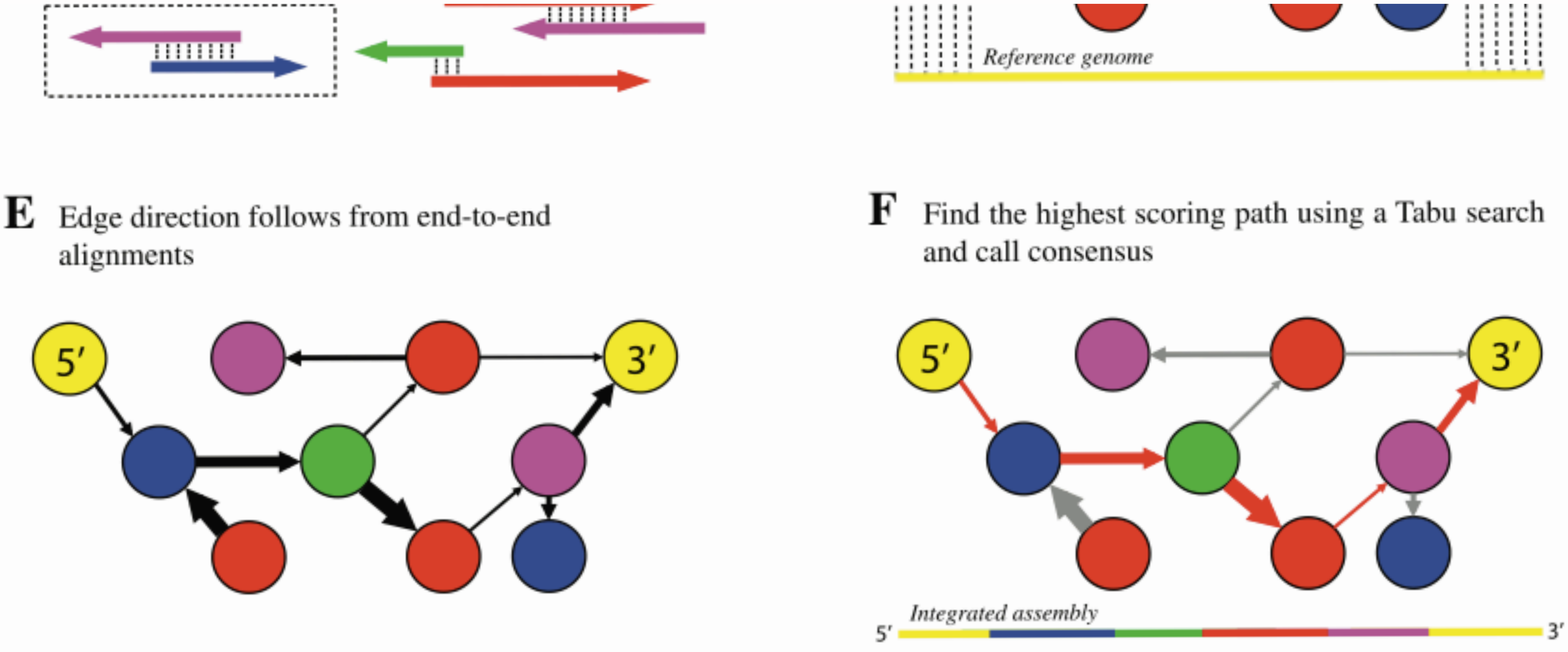
Vol. 26 ECCB 2010, pages i433–i439
doi:10.1093/bioinformatics/btq366

Integrating genome assemblies with MAIA

Jurgen Nijkamp[1,2,3,*], Wynand Winterbach[1,4], Marcel van den Broek[2,3], Jean-Marc Daran[2,3], Marcel Reinders[1,3,5] and Dick de Ridder[1,3,5]

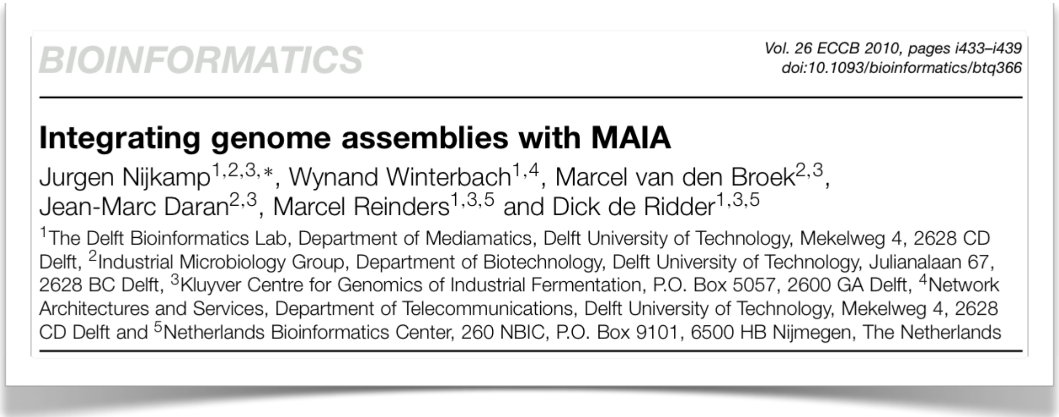
[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Industrial Microbiology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, [4]Network Architectures and Services, Department of Telecommunications, Delft University of Technology, Mekelweg 4, 2628 CD Delft and [5]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

A “meta” assembler



E Edge direction follows from end-to-end alignments

F Find the highest scoring path using a Tabu search and call consensus



BIOINFORMATICS

Vol. 26 ECCB 2010, pages i433–i439
doi:10.1093/bioinformatics/btq366

Integrating genome assemblies with MAIA

Jurgen Nijkamp[1,2,3,*], Wynand Winterbach[1,4], Marcel van den Broek[2,3], Jean-Marc Daran[2,3], Marcel Reinders[1,3,5] and Dick de Ridder[1,3,5]

[1]The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Industrial Microbiology Group, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, [4]Network Architectures and Services, Department of Telecommunications, Delft University of Technology, Mekelweg 4, 2628 CD Delft and [5]Netherlands Bioinformatics Center, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

Strategy	Assembly	Package	# contigs	Total size (Mb)	N50 (kb)	Mapped reads (%)
Single input	*De novo*	Abyss	1223	11.64	20	84.8
	De novo	Celera	4148	9.03	3	62.8
	Comparative (S288c)	Maq	375	12.06	162	96.9
	Comparative (YJM789)	Maq	907	11.77	44	90.8
	Comparative (RM11-1A)	Maq	795	11.54	41	78.2
Hybrid	*De novo*	Velvet	654	11.40	72	75.5
	De novo + comparative	Minimus	71	12.21	290	92.1
	De novo + comparative	MAIA	29	12.01	918	96.5

Comparative genomics approaches

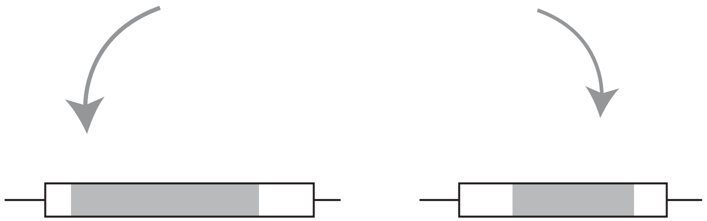
- Attempt to bridge contigs after assembly/annotations
- Ensembl Compara (unpublished)
- ESPRIT (Dessimoz *et al.*, in review)

ESPRIT

*“**E**stablishing **S**plit **P**rotein **R**egions **I**n **T**entative genomes”*

Parts of the same gene misannotated as separate genes

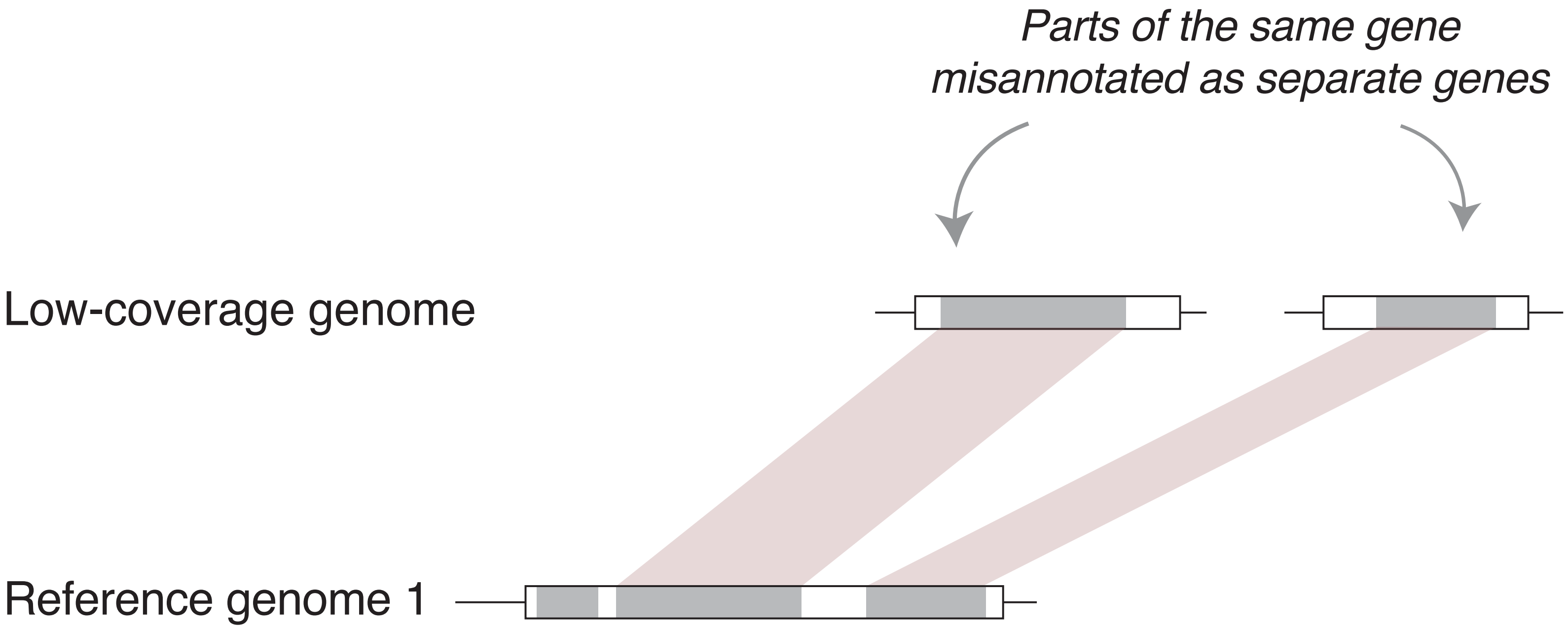
Low-coverage genome



Christophe Dessimoz, Stefan Zoller, Tereza Manousaki, Huan Qiu, Axel Meyer, and Shigehiro Kuraku, *Comparative genomics approach to detecting split coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes)*, Briefings in Bioinformatics, in review

ESPRIT

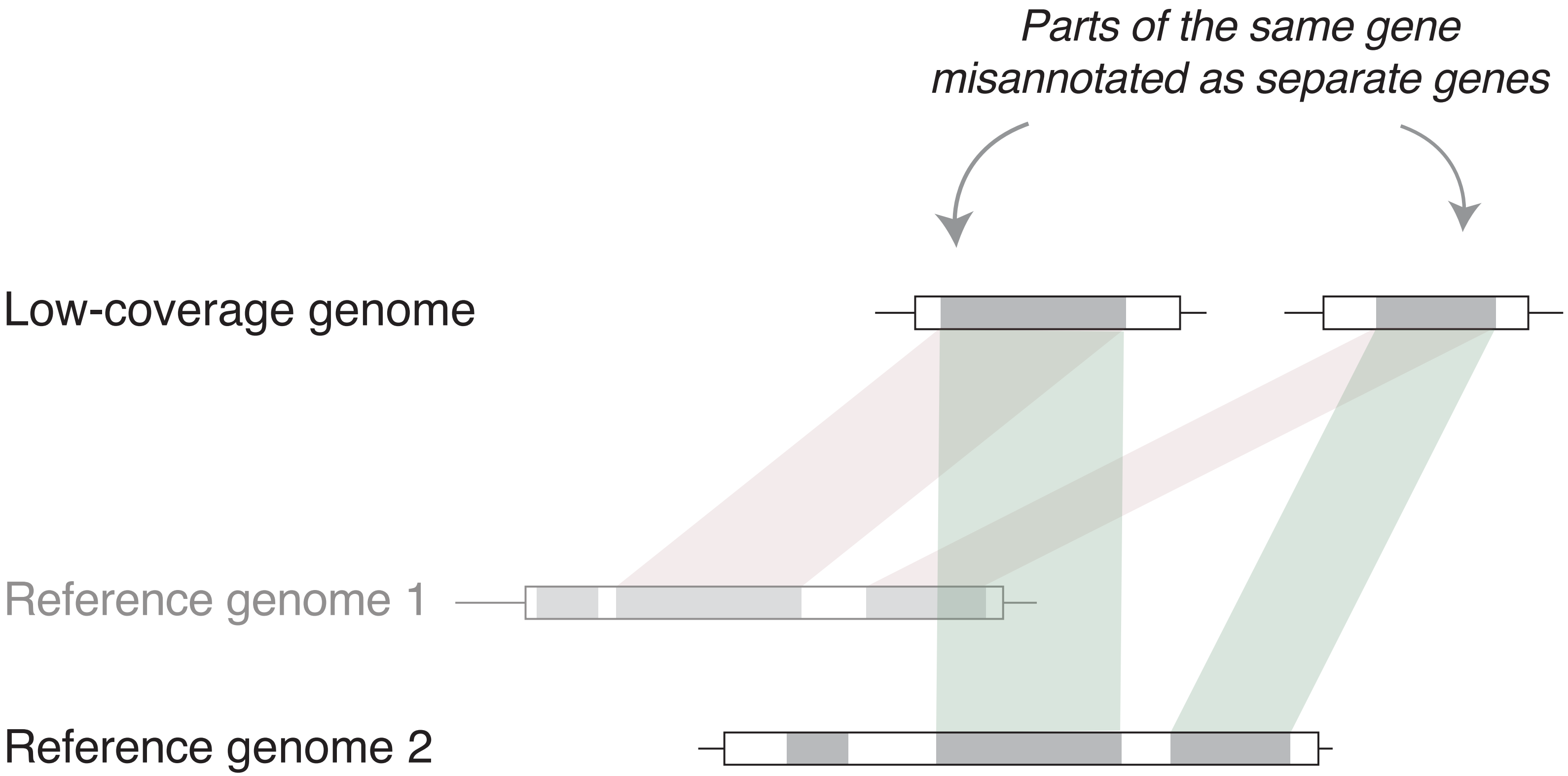
*“**E**stablishing **S**plit **P**rotein **R**egions **I**n **T**entative genomes”*



Christophe Dessimoz, Stefan Zoller, Tereza Manousaki, Huan Qiu, Axel Meyer, and Shigehiro Kuraku, *Comparative genomics approach to detecting split coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes)*, Briefings in Bioinformatics, in review

ESPRIT

*“**E**stablishing **S**plit **P**rotein **R**egions **I**n **T**entative genomes”*



Parts of the same gene misannotated as separate genes

Low-coverage genome

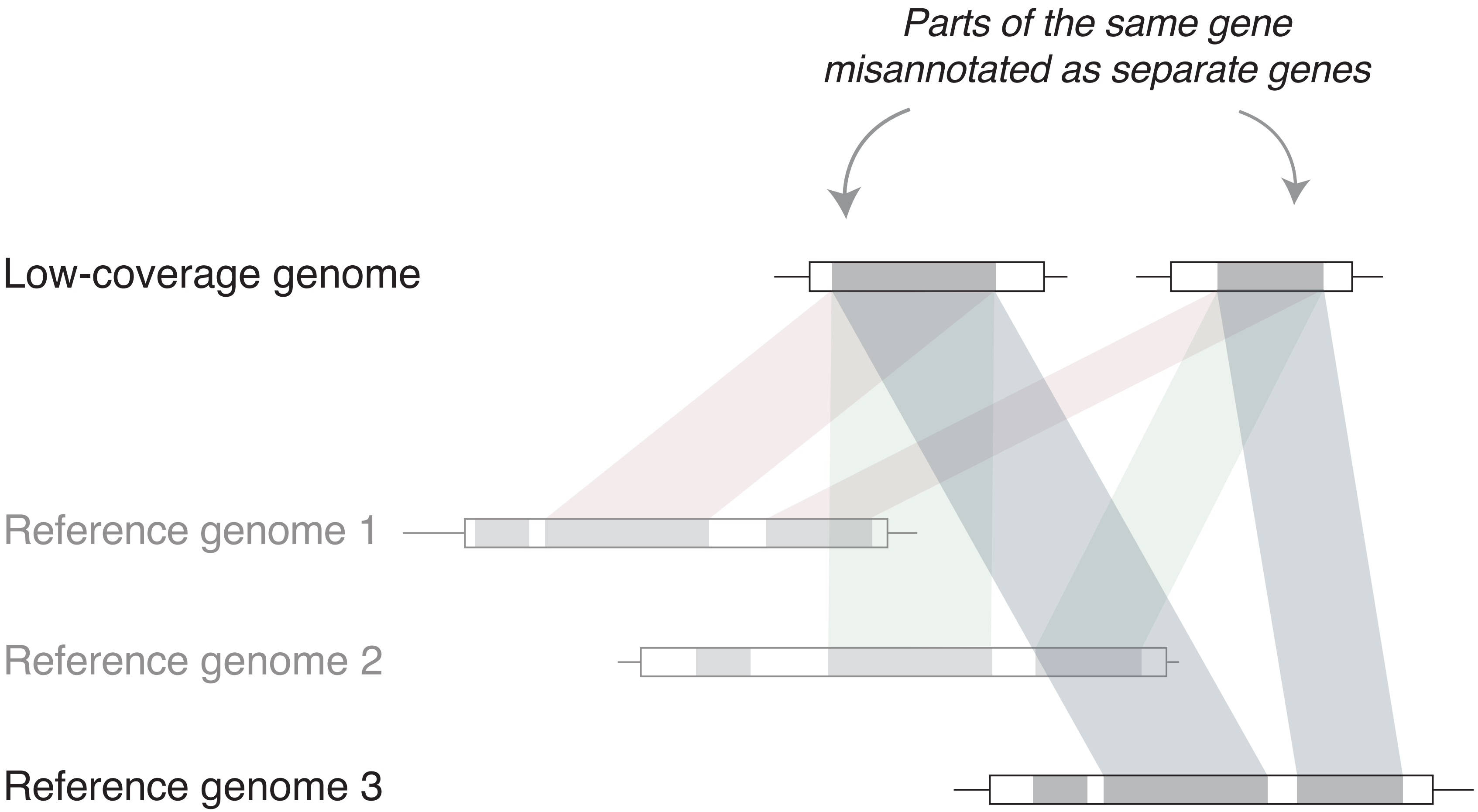
Reference genome 1

Reference genome 2

Christophe Dessimoz, Stefan Zoller, Tereza Manousaki, Huan Qiu, Axel Meyer, and Shigehiro Kuraku, *Comparative genomics approach to detecting split coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes)*, Briefings in Bioinformatics, in review

ESPRIT

*"**E**stablishing **S**plit **P**rotein **R**egions **I**n **T**entative genomes"*



Christophe Dessimoz, Stefan Zoller, Tereza Manousaki, Huan Qiu, Axel Meyer, and Shigehiro Kuraku, *Comparative genomics approach to detecting split coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes)*, Briefings in Bioinformatics, in review

Open Challenges

- How to select & weight appropriate reference genomes.
- Duplications/repetitive sequences remain a challenge with these methods.

Conclusions

- Recently, a new assembly approach has emerged: phylogeny-based assembly.
- It is complementary to *de novo* assembly and assembly based on a single reference alignment.
- It can be done as part of the assembly process itself (4 published methods reviewed) or after assembly/annotation (Ensembl compara, ESPRIT)