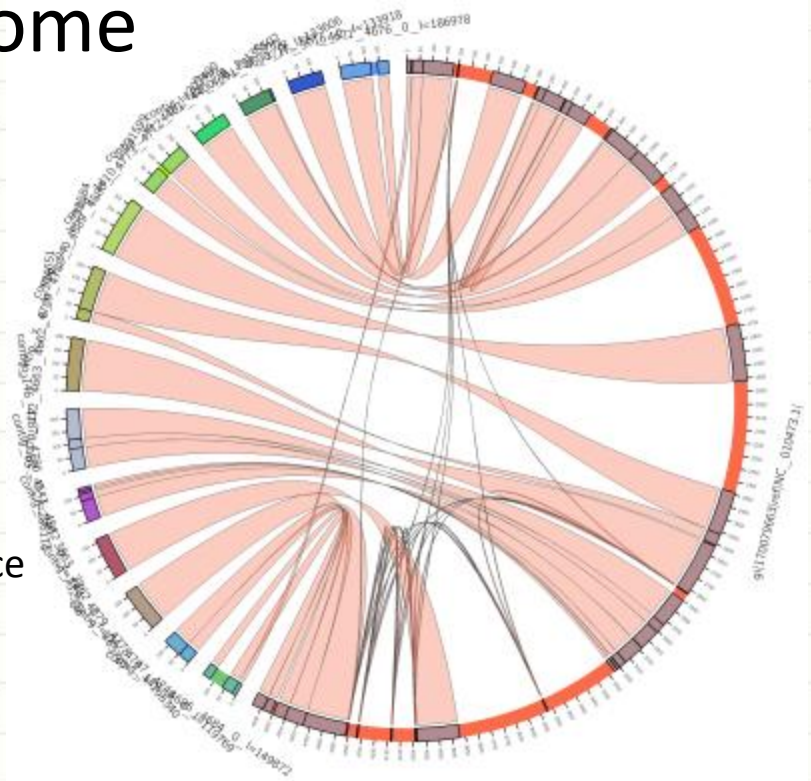
COURSEWORK
REFERENCE ASSISTED ASSEMBLY

Dmitry Kuzminov

13.12.2012

Finding synteny blocks with Sibelia

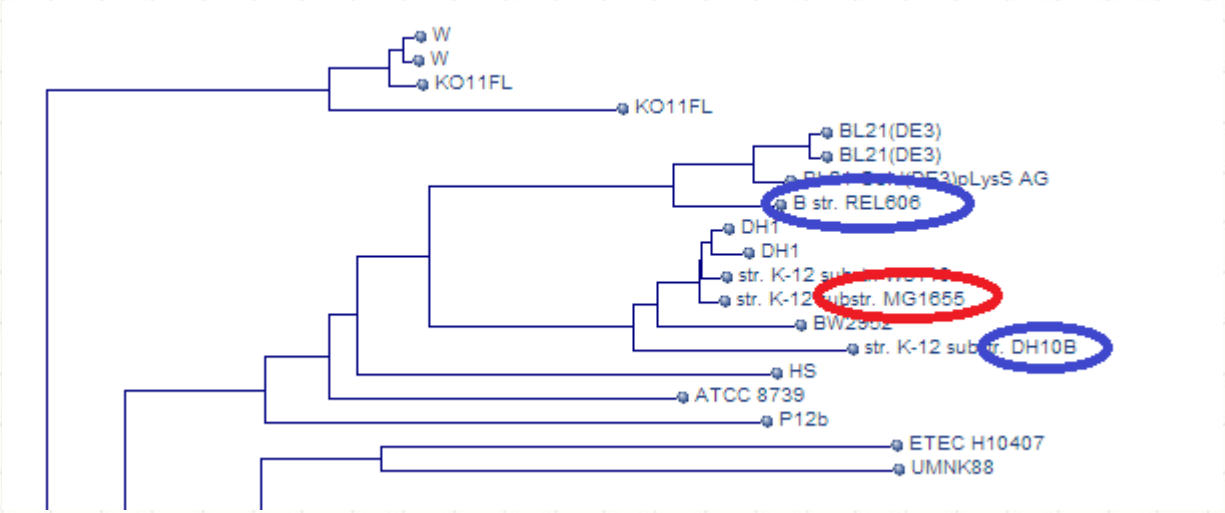
- Looks for synteny blocks
- Looks between different chromosomes and within a single chromosome
- Provides circle diagrams



This picture: mapping 14 largest contigs to a reference

Sample data

- E.coli genome
  - Assembly for K-12 substr. MG1655
  - Reference genome B str. REL606
  - Reference genome K-12 substr. DH10B



Assembly parameters to regard

- Evaluate with Quast software
  - # misassemblies
  - N50

Rectangles assembly

QUAST β

QUality ASsesment Tool for Genome Assemblies by Algorithmic Biology Lab

E. coli K12.MG1655

30 November 2012, Friday, 21:33:56

Contigs shorter than 200 bp were skipped

Extended report

Basic statistics	**Rectangles_contigs**
# contigs	145
Largest contig	236 111
Total length	4 644 119
NG50	119 769
Misassemblies	
# misassemblies	1
Misassembled contigs length	19 528
Genome statistics	
Genome fraction (%)	99.99
# genes	4287 + 35 part
# operons	854 + 29 part
# mismatches per 100 kbp	2.990
# indels per 100 kbp	0.500
# N's per 100 kbp	0

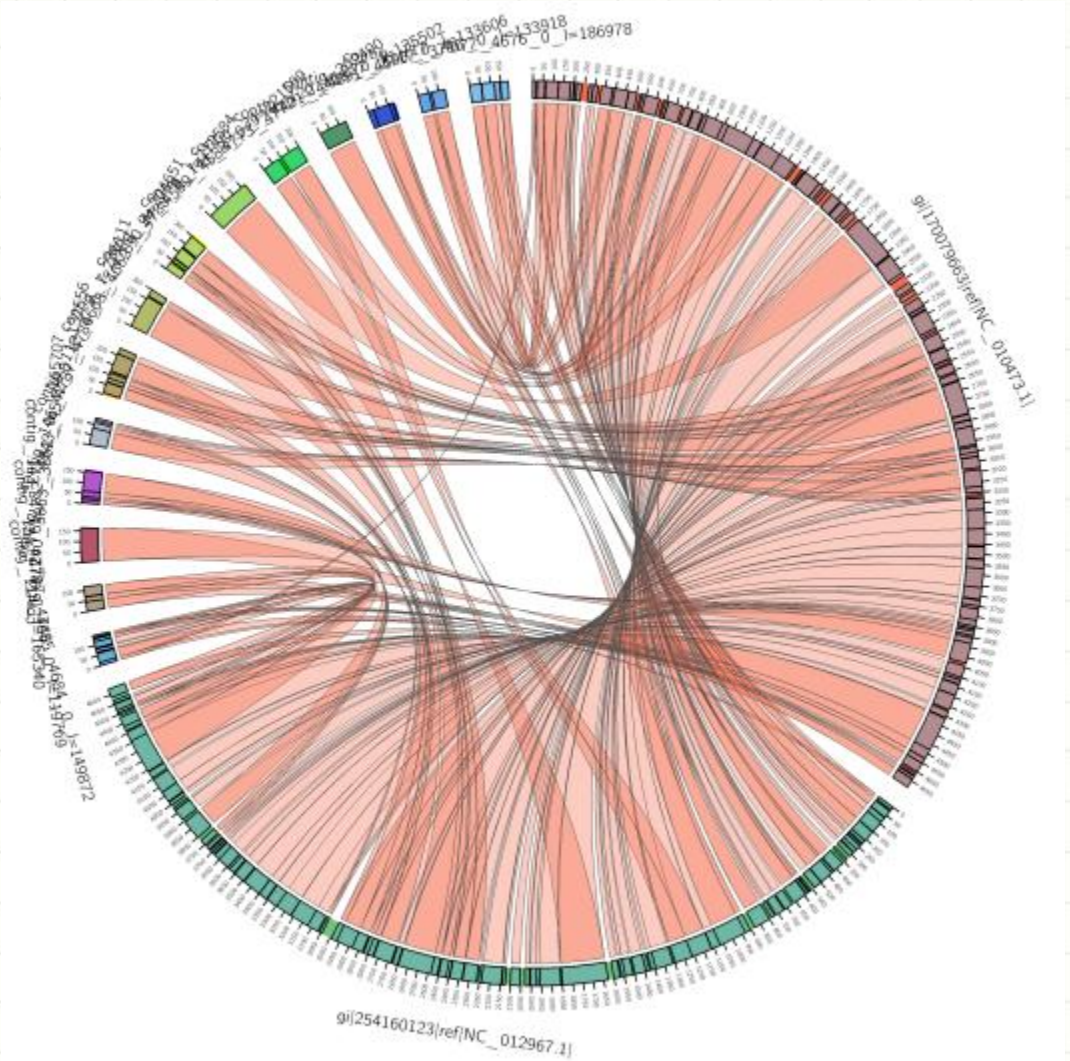
misassemblies = 1

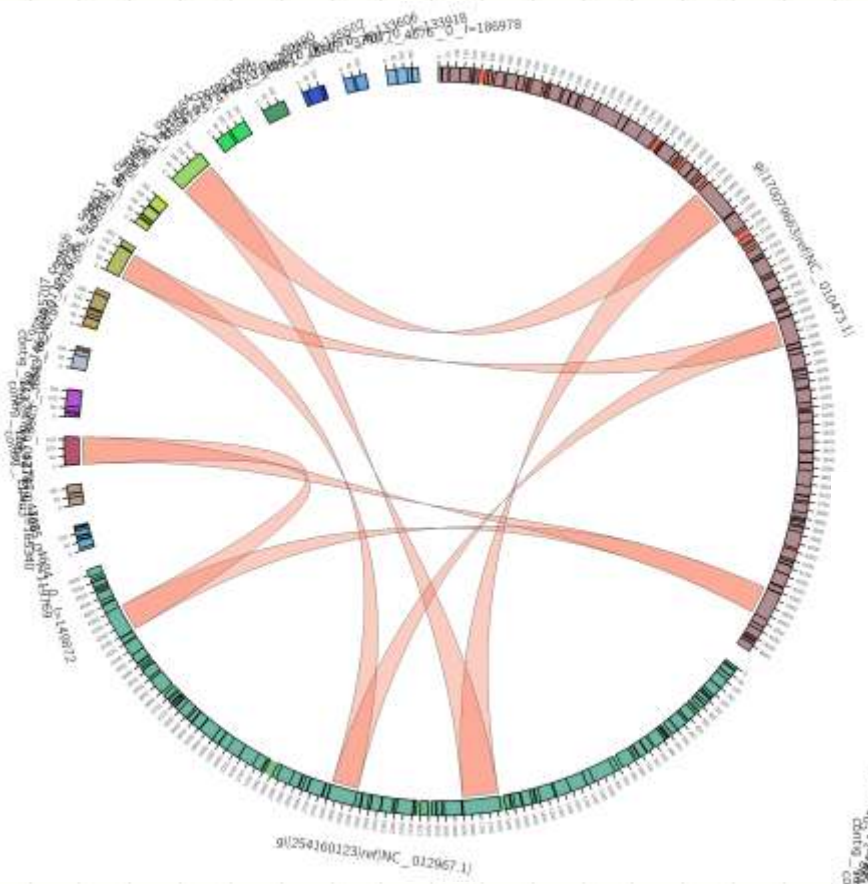
N50 = 119769

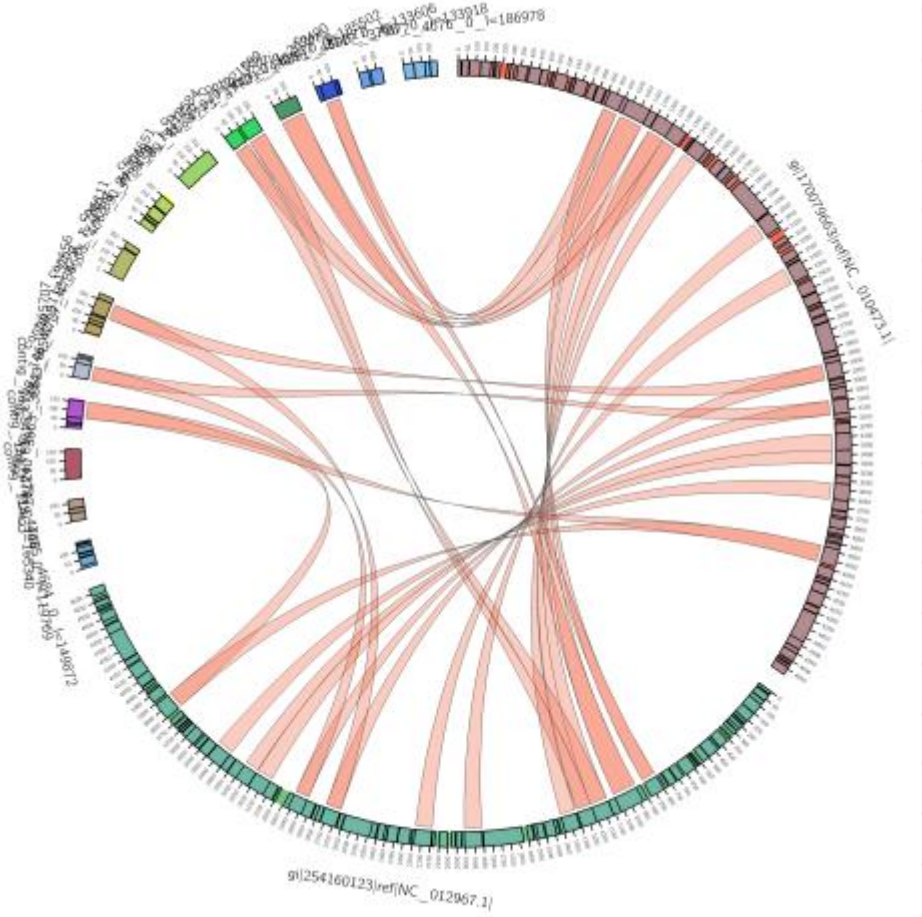
contigs = 145

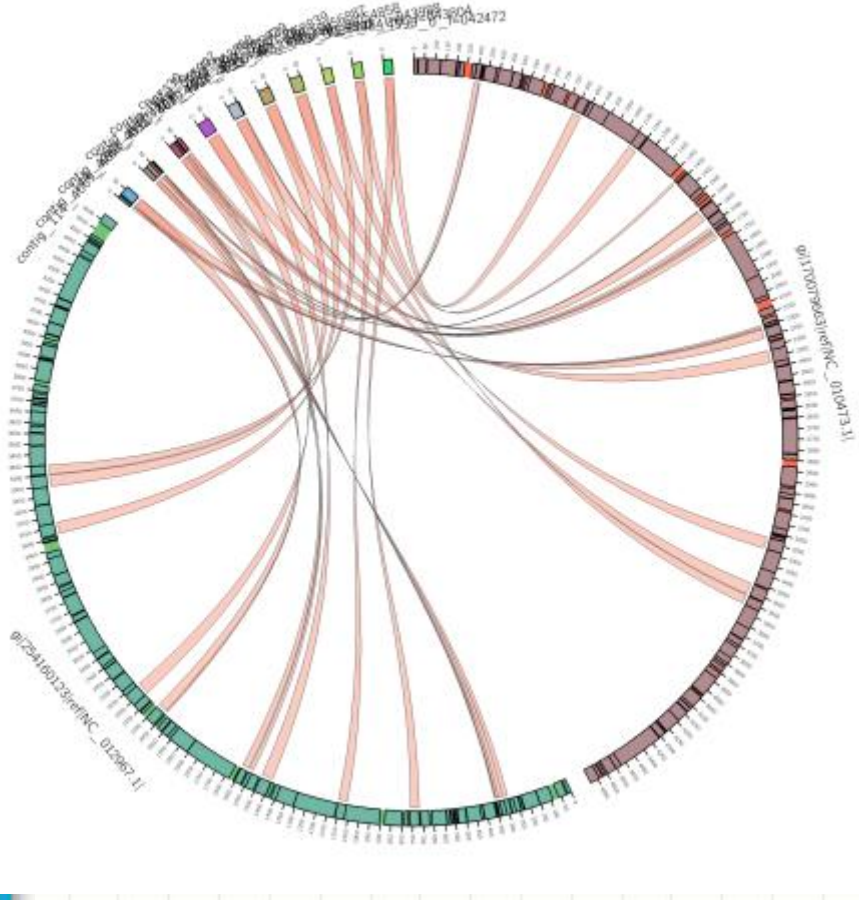
Genome fraction = 99.99%

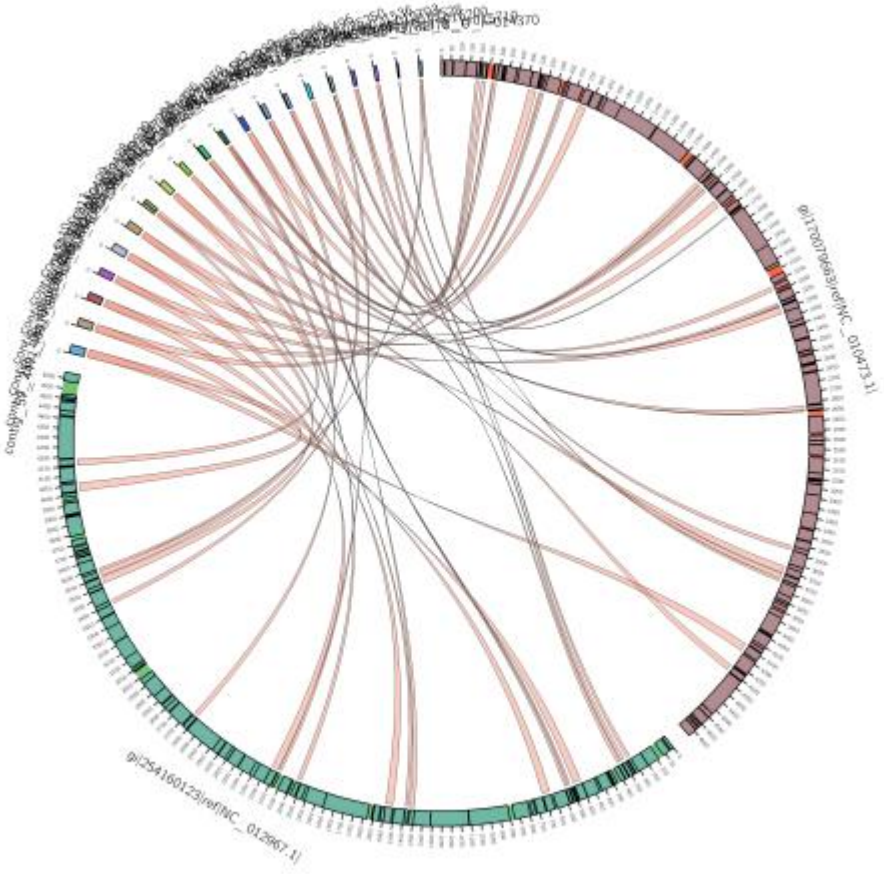
14 contigs > N50 (119769)




gi|254160123|ref|NC_012967.1|


gi|254160123|ref|NC_012967.1|


gi|170079663|ref|NC_010473.1|
gi|254160123|ref|NC_012967.1|


gi|170079663|ref|NC_010473.1|
gi|254160123|ref|NC_012967.1|

Results

Extended report

Basic statistics	Rectangles_contigs_New	
# contigs	140	145
Largest contig	268 904	236 111
Total length	4 638 845	4 644 119
NG50	132 556	119 769
Misassemblies		
# misassemblies	1	1
Misassembled contigs length	19 528	19 528
Genome statistics		
Genome fraction (%)	99.90	99.99
# genes	4284 + 35 part	4287 + 35 part
# operons	853 + 30 part	854 + 29 part
# mismatches per 100 kbp	2.870	2.990
# indels per 100 kbp	0.470	0.500
# N's per 100 kbp	0	0

Results: is it possible to improve?

- N50 x10
  - N50 = 1197690 = 25%. Seems to be impossible
- N50 x5
  - N50 = 119769 * 5. Unlikely
- N50 x2
  - This should be the target

QUESTIONS?

Initial Proposals

- Article: http://genomebiology.com/2009/10/8/R88
- General principle described in the article
- The algorithm is implemented in Arachne software

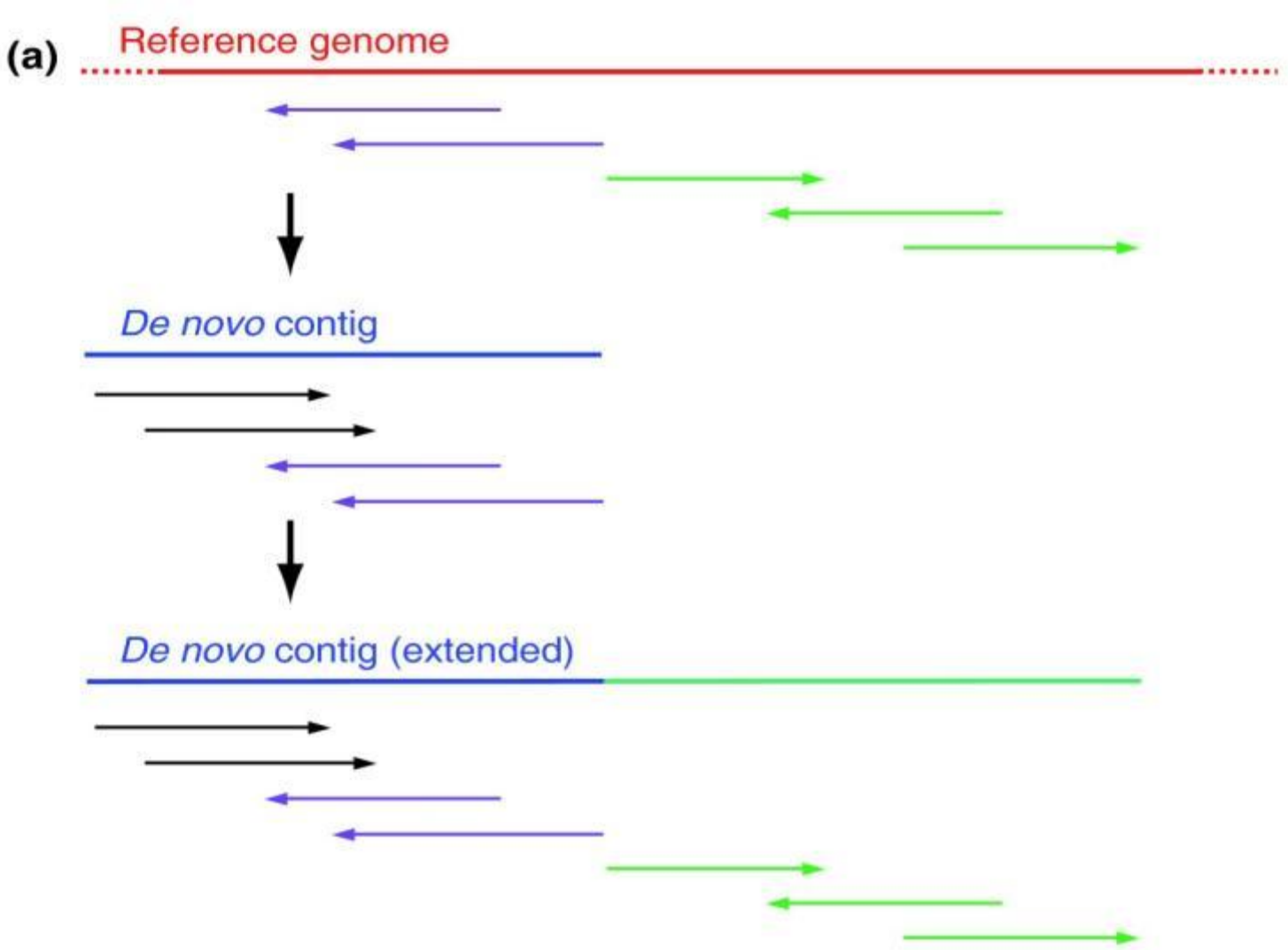
General Idea

- We have as an input:
  - Contigs
    - These contigs are assembled with another algorithm
  - Reference genome of another specie
    - Assembled with other tools in advance
- With the reference we could assemble contigs better than just *de novo*

Extending Contigs

- We could extend contigs by gluing them together if the mapping to the reference shows that these contigs are neighbours.

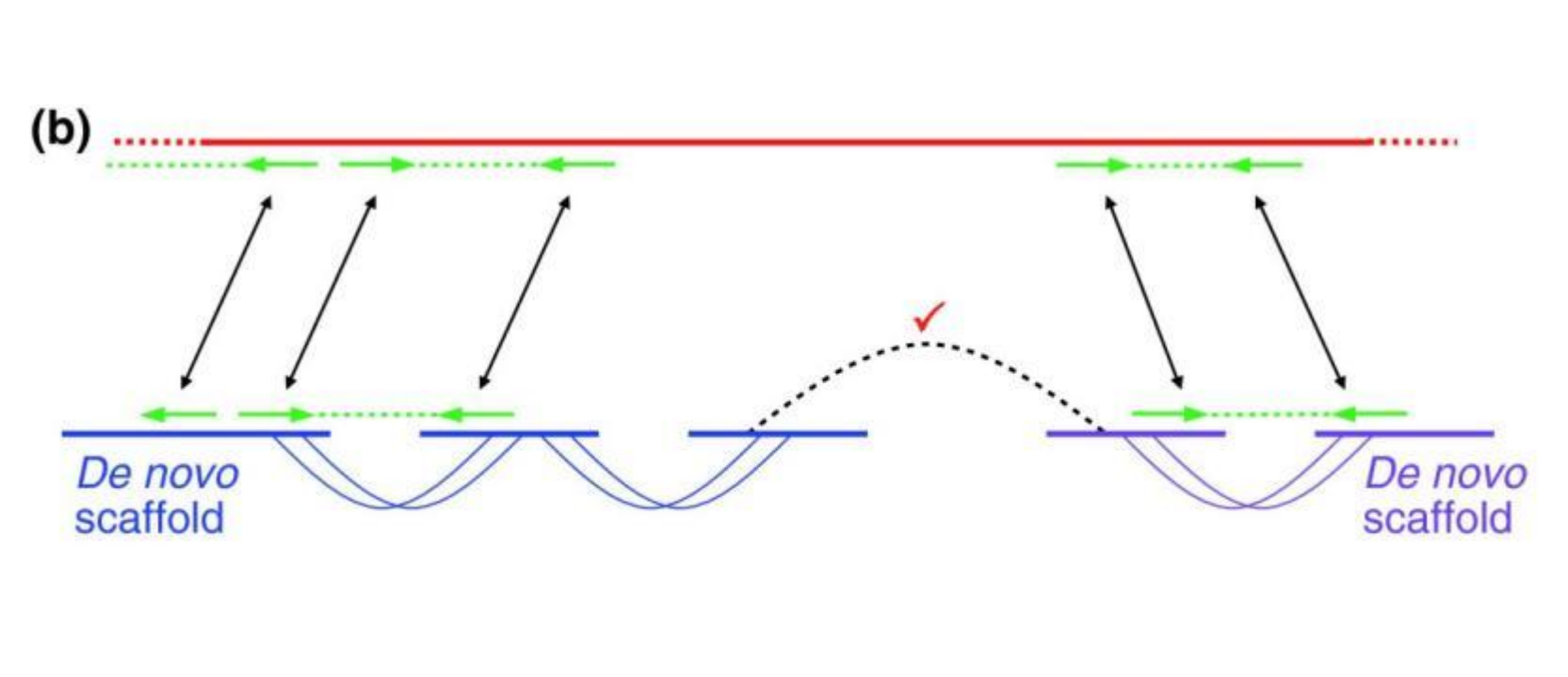
Extending Contigs



Connecting Scaffolds

- We could connect scaffolds if the mapping to the reference shows that these scaffolds are close to each other.

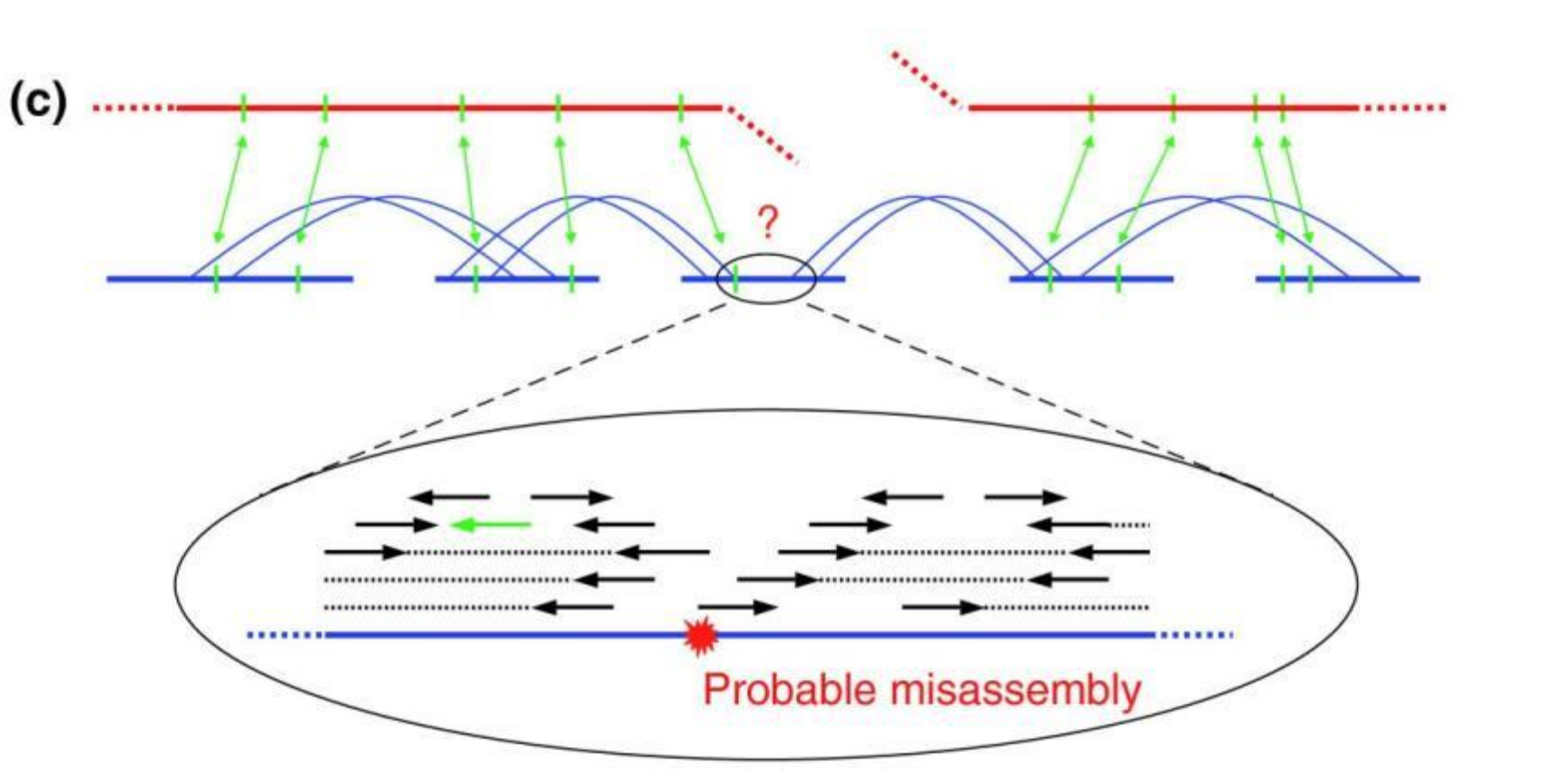
Connecting Scaffolds



Error Correction

- We could correct errors. If the reference shows that two parts of a contig should be far away from each other we could correct this erroneous link.

Error Correction



IDBA_hybrid

Problems

- No description of how this tool works. Even no description of how to treat the output
- Works good (at least produces the result) on test data but throws an exception on real data

IDBA_hybrid Tool Description

(Taken from the download page as is)

IDBA-Hybrid is an iterative De Bruijn Graph De Novo Assembler for hybrid sequencing. It is an extension of IDBA-UD algorithm. It aims at using a closed related reference genome to help de novo assembly, especially when sequencing depth is low. IDBA-Hybrid does alignment between reads and reference first to extract similar regions in the reference genome, and then it correct the similar regions based on the alignment results and apply local assembly technique to resolve potential structure virations. Finally, it groups all the reads and the contigs got from those similar regions to do de novo assembly. The expriments showed it outperforms all existing de novo or hybrid assembly algorithms, especilly when the sequencing depth is low and the reference genome is similar to the target genome.

Algorithm Analysis

So does this description reflect what this tool actually does?

As the author answered:

[I am sorry that our paper about IDBA-Hybrid is under preparation. I will let you, once we finish it. If you are not comfortable about that, you can use IDBA-UD to do assembly.]

So lets try to analyze the source code…

Configuration Parameters

Below are the parameters, the default, min and max values:

- K-value: default = 30, min = 20, max = 100
- Step: default = 20
- Similarity: default = 0.95

Algorithm (with default values)

1. Align reads to reference with k = 30
2. Modify **reference** to make it closer to reads
3. For k := 20 to 100 step 20
    1. Arrange
    2. Correct reads (to improve consistency)
    3. Save as “reads_k”

2. Scaffold the alignment for “reads_100”

Conclusion

- Idba_hybrid does only mapping to reference
- idba_hybrid doesn’t perform complex corrections, just modifies single nucleotides
- Idba_hybrid can be used for assembly against the reference of the same specie
- idba_hybrid cannot be used for genome assembly of another specie