RAGOUT

Enlarge your contigs

Ragout – a reference-assisted assembly tool for bacterial genomes

Mikhail Kolmogorov[1], Brian Raney[2], Benedict Paten[2] and Son Pham[3]

[1]St. Petersburg University of the Russian Academy of Sciences,
[2]University of California Santa Cruz, [3]University of California San Diego

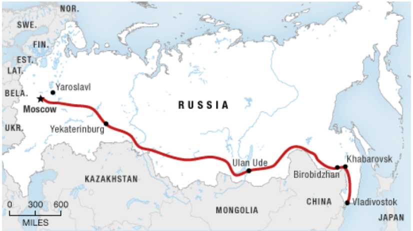
ISMB 2014, Boston

Outline

1. Introduction
2. Quick Overview
3. Algorithm Description
4. Results
5. Further plans

Trans-Siberian Railway

- The longest railroad in the world
- 9248 km
- $\sim$ 15 000 000 railroad ties



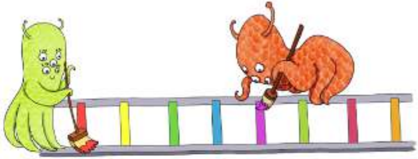
A Secret Story

- Once Aliens came:

A Secret Story

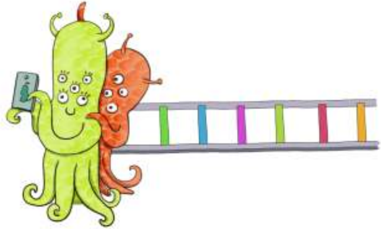
- Once Aliens came:

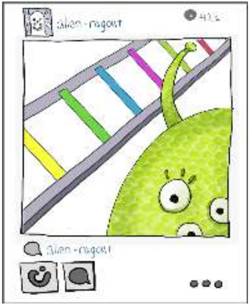
- And they have painted the ties in different colors:



A Secret Story II

- After, they took a lot of pictures:



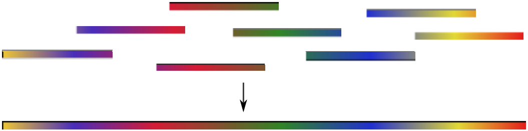


A Secret Story III

- And after they had been gone, rain has wanished all dyes from the railroad :(

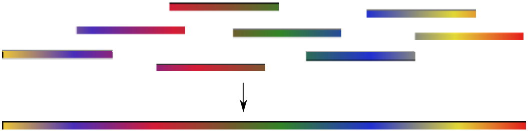
A Secret Story III

- And after they had been gone, rain has wanished all dyes from the railroad :(
- Can we now reconstruct the original coloring using those pictures?



A Secret Story III

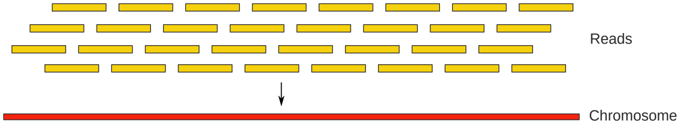
- And after they had been gone, rain has wanished all dyes from the railroad :(
- Can we now reconstruct the original coloring using those pictures?



- This is exactly a problem that genome assemblers solve!
  - SPAdes
  - ABySS
  - Velvet
  - SOAPdenovo
  - SGA
  - ...

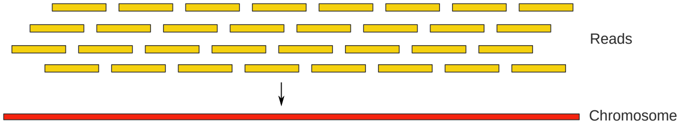
Genome Assembly

- Join short overlapping reads into chromosomes
- Expectation:

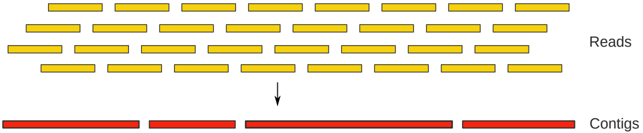


Genome Assembly

- Join short overlapping reads into chromosomes
- Expectation:

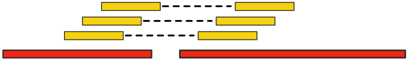


- Reality:



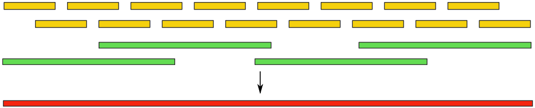
Complete Sequence?

- Jumping libraries:



illumina

- Long reads:

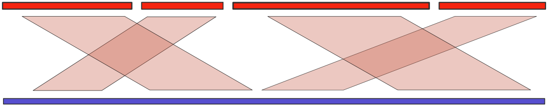


PACIFIC BIOSCIENCES

- Still expensive and not as reliable as short reads
- Is there any alternative?

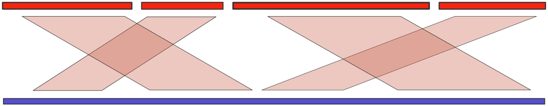
Reference-assisted Assembly

- Using a complete genome of another closely-related organism
- Contigs are being aligned on that *reference* genome



Reference-assisted Assembly

- Using a complete genome of another closely-related organism
- Contigs are being aligned on that *reference* genome



- Structural variations?

Rearrangement Approaches

- Gaul and Blanchette. "Ordering Partially Assembled Genomes Using Gene Arranements", *Springer, 2006*
  - Tries to minimize number of structural variations between two genomes
- Kim et. al. "Reference-assisted Chromosome Assembly", *PNAS, 2013*
  - First attempt to use multipe genomes simultaneously
  - One *reference* and multiple *outgroups*
  - Still heavily rely on that reference
- Both approaches may introduce errors

Rearrangement Approaches

- Gaul and Blanchette. "Ordering Partially Assembled Genomes Using Gene Arranements", *Springer, 2006*
  - Tries to minimize number of structural variations between two genomes
- Kim et. al. "Reference-assisted Chromosome Assembly", *PNAS, 2013*
  - First attempt to use multipe genomes simultaneously
  - One *reference* and multiple *outgroups*
  - Still heavily rely on that reference
- Both approaches may introduce errors
- So maybe we need multiple references?

Outline

1 Introduction

2 Quick Overview

3 Algorithm Description

4 Results

5 Further plans

Ragout Recipe

- Ragout – Reference-Assisted Genome Ordering UTility
- Written in Python/C++
- Ingredients:
  - Multiple references (in FASTA format)
  - Contigs/scaffolds from short-read assembly
  - Phylogenetic tree
- Output: scaffolds

Outline

1 Introduction

2 Quick Overview

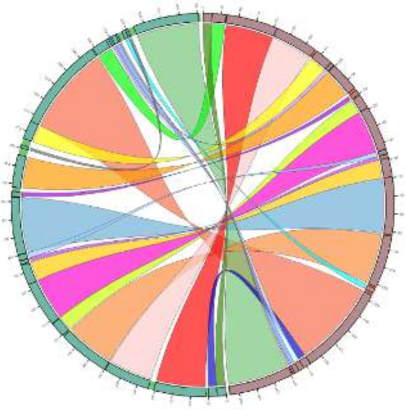
3 Algorithm Description

4 Results

5 Further plans

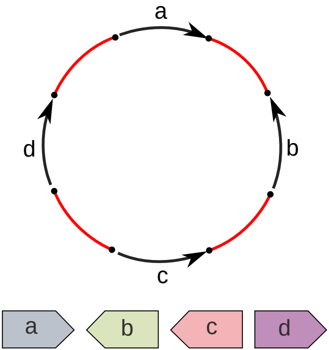
Genome Representation

- Comparing nucleotie by nucleotide is expensive
- Extract conserved segements (synteny blocks)
- Assumption: each block is represented exactly once in each genome



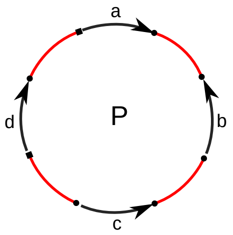
SIBELIA

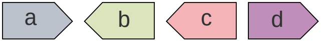
Genome as Synteny Blocks and Adjacencies

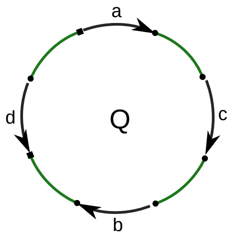


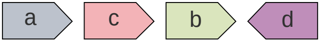
- Chromosome is represented as an aleterating cycle of **directed black** and **undirected red** edges
- **Black** edges correspond to synteny blocks
- **Red** edges connect ends of adjacent synteny blocks

Breakpoint Graphs Are Simple!

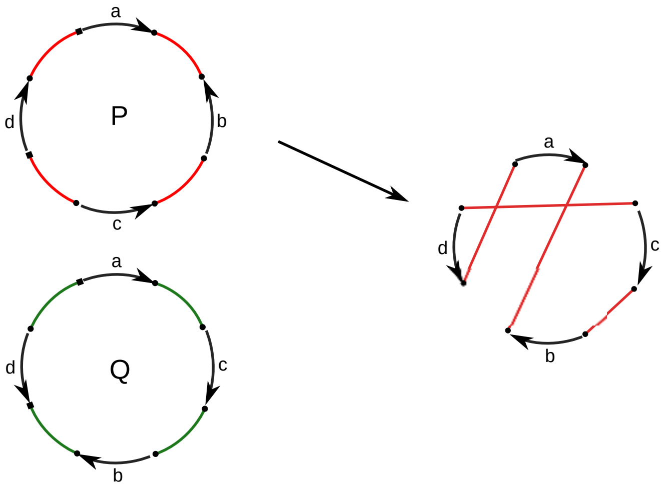




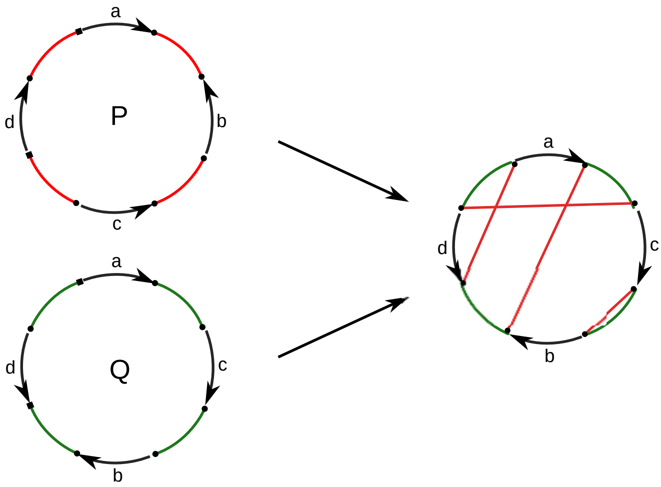




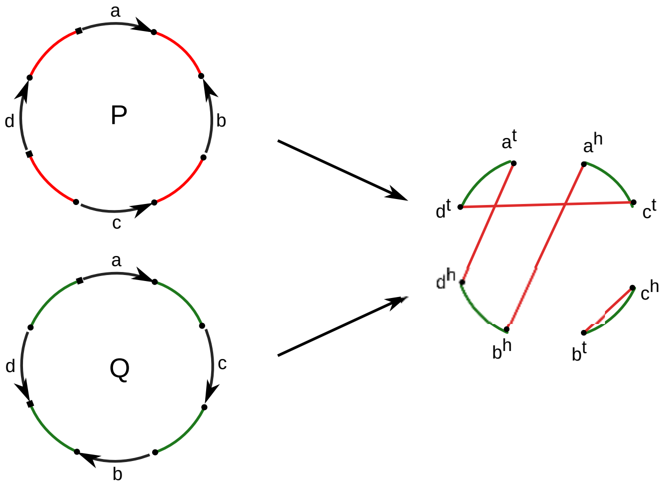
Breakpoint Graphs Are Simple!



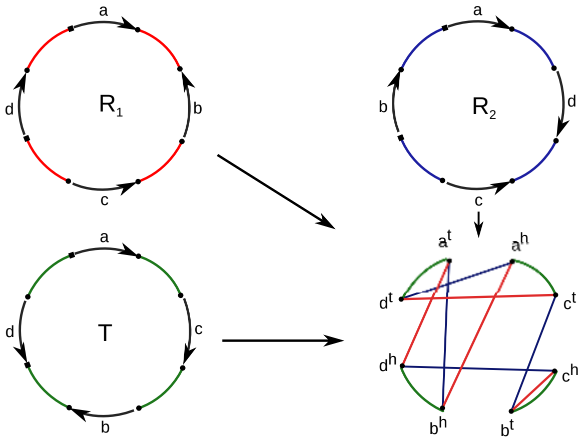
Breakpoint Graphs Are Simple!



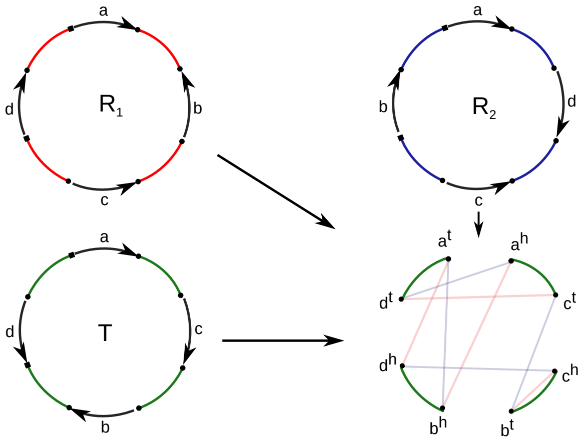
Breakpoint Graphs Are Simple!



Breakpoint Graphs Are Simple!

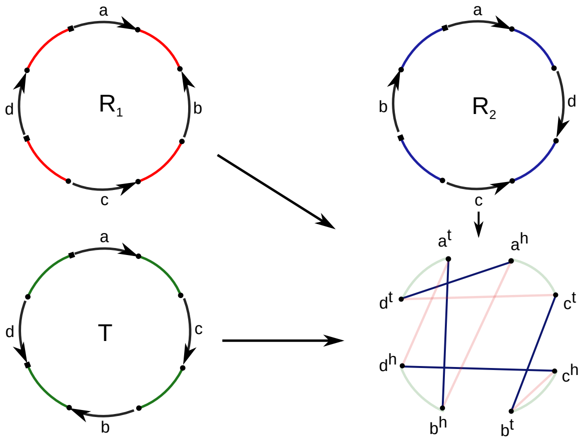


Breakpoint Graphs Are Simple!



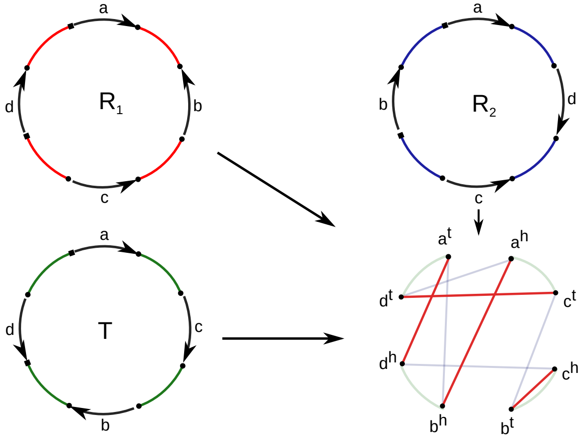
- Each color defines a perfect matching

Breakpoint Graphs Are Simple!



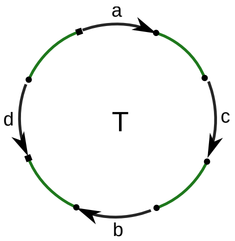
- Each color defines a perfect matching

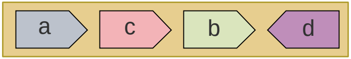
Breakpoint Graphs Are Simple!



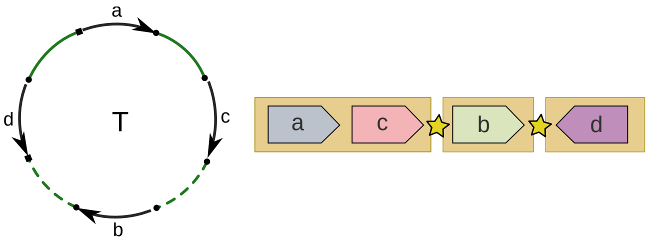
- Each color defines a perfect matching

Incomplete Breakpoint Graph



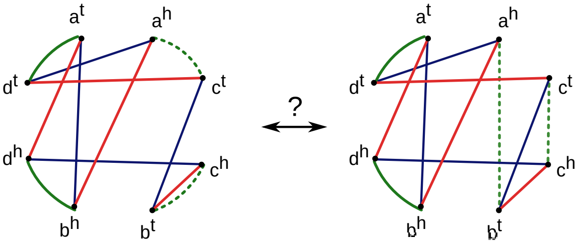


Incomplete Breakpoint Graph



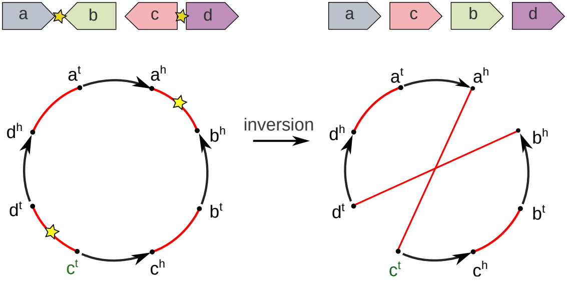
- Some adjacencies are missing

Incomplete Breakpoint Graph II



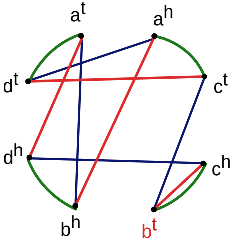
- Find missing edges
- **= Recover perfect matching**
- There are multiple variants of such matching
- How to find the correct one?

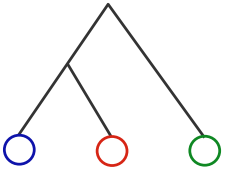
States of Adjacencies



- *State* = adjacent vertex
- *State* of c^t: $d^t \rightarrow a^h$
- Rearrangements change *states* of adjacencies

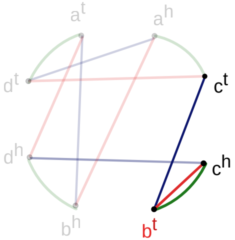
Objective Function

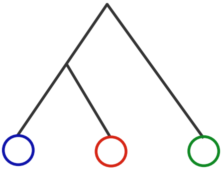




- Choose an arbitrary perfect matching

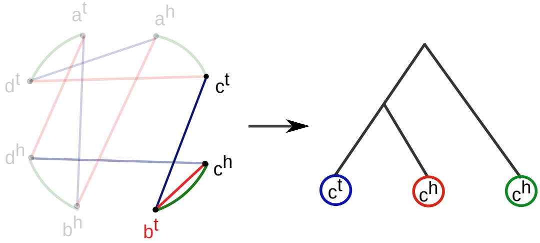
Objective Function





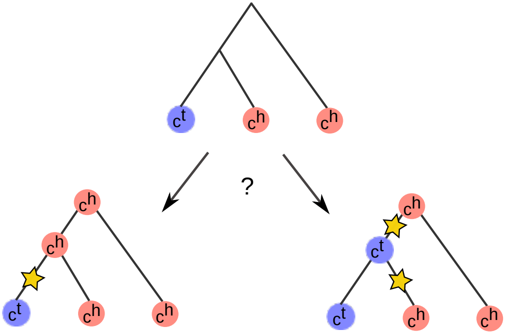
- Choose an arbitrary perfect matching
- Pick a vertex from the graph

Objective Function



- Choose an arbitrary perfect matching
- Pick a vertex from the graph
- Label tree nodes as *states* of chosen vertex in genomes
- The tree represents evolution of breakpoint states

Parsimony Procedure



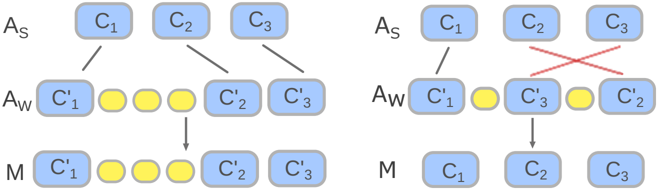
- Find scenario with minimum number of changes
- Associated cost for graph vertex u and tree T:

$$P(u, T) = \sum_{branch\ (i,\ j), i \neq j} W(branchlength)$$

Optimal Contigs Order

- Cost for a complete graph G: $\sum_{u \in G} P(u, T)$
- Want a prefect matching which minimizes this cost
- An efficient solution:
  - Node weight $\rightarrow$ edge weight
  - Find minimum weight perfect matching
  - Blossom algorithm in $\mathrm{O}(n^4)$

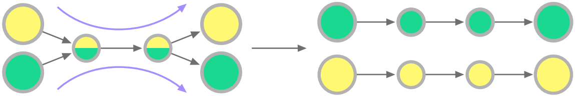
Iterative Assembly



(a) Locally consistent (b) Locally inconsistent

- Solve the dilemma about choice of synteny block size
- Merge scaffolds with different precision into one assembly

Refinement with Assembly Graph



- Incorporate very small/repetitive contigs
- Analogiously to repeat resolution in short-read assembly

Outline

1. Introduction
2. Quick Overview
3. Algorithm Description
4. Results
5. Further plans

Toy Test – One *E. Coli* Reference

	Ragout	MCM	OSLay
Scaffolds	**1**	**1**	8
Contigs (coverage)	**129 (97.9%)**	77 (97.6%)	80 (96.7%)
Miss-ordered	**0**	**0**	1

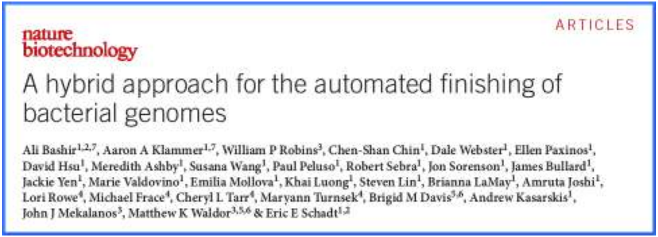
- One *E. Coli* reference without rearrangements
- #Contigs – 156 (98.18%)

Assembly with Rearrangements – Four *H. Pylori* References

#References	Scaffolds	Contigs (cov.)	Miss-ordered
		Ragout	
1	2	91 (97.7%)	6
2	2	**95 (97.8%)**	1
3	**1**	**95 (97.8%)**	1
4	**1**	93 (97.6%)	**0**
		RACA	
2	3	35 (83.6%)	2
3	2	35 (83.6%)	1
4	2	35 (83.8%)	1

- Four *H. Pylori* references with rearrangements
- #Contigs – 183 (98.57%)

Long Reads or ...



ARTICLES

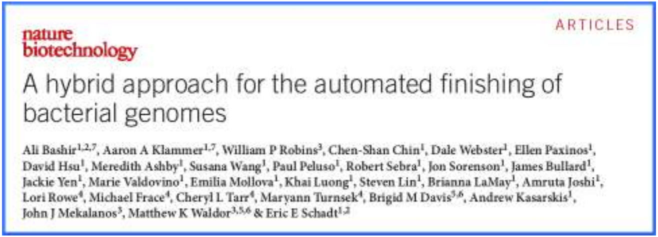
nature biotechnology

A hybrid approach for the automated finishing of bacterial genomes

Ali Bashir[1,2,7], Aaron A Klammer[1,7], William P Robins[3], Chen-Shan Chin[1], Dale Webster[1], Ellen Paxinos[1], David Hsu[1], Meredith Ashby[1], Susana Wang[1], Paul Peluso[1], Robert Sebra[1], Jon Sorenson[1], James Bullard[1], Jackie Yen[1], Marie Valdovino[1], Emilia Mollova[1], Khai Luong[1], Steven Lin[1], Brianna LaMay[1], Amruta Joshi[1], Lori Rowe[4], Michael Frace[4], Cheryl L Tarr[4], Maryann Turnsek[4], Brigid M Davis[3,6], Andrew Kasarskis[1], John J Mekalanos[3], Matthew K Waldor[3,5,6] & Eric E Schadt[1,2]

- 40 bp non-paired Illumina reads
- Roche 454 reads
- PacBio reads

Long Reads or ...



nature biotechnology

ARTICLES

A hybrid approach for the automated finishing of bacterial genomes

Ali Bashir[1,2,7], Aaron A Klammer[1,7], William P Robins[3], Chen-Shan Chin[1], Dale Webster[1], Ellen Paxinos[1], David Hsu[1], Meredith Ashby[1], Susana Wang[1], Paul Peluso[1], Robert Sebra[1], Jon Sorenson[1], James Bullard[1], Jackie Yen[1], Marie Valdovino[1], Emilia Mollova[1], Khai Luong[1], Steven Lin[1], Brianna LaMay[1], Amruta Joshi[1], Lori Rowe[4], Michael Frace[4], Cheryl L Tarr[4], Maryann Turnsek[4], Brigid M Davis[3,6], Andrew Kasarskis[1], John J Mekalanos[3], Matthew K Waldor[3,5,6] & Eric E Schadt[1,2]

- 40 bp non-paired Illumina reads
- Roche 454 reads?
- PacBio reads?
- Can we replace long reads with Ragout here?

Long Reads or Reference-assisted Assembly?

#References	Scaffolds	Contigs (cov.)	Miss-ordered
		Ragout	
1	3	**185 (94.8%)**	3
2	**2**	179 (94.7%)	4
3	**2**	174 (94.7%)	**0**
		RACA	
2	6	124 (85.8%)	0
3	3	127 (90.0%)	**0**

- Three *V. Cholerae* references with rearrangements
- #Contigs – 1407 (96.89%)
- Results are shown without refinement (poor assembly quality)

Outline

1. Introduction

2. Quick Overview

3. Algorithm Description

4. Results

5. Further plans

New results & further plans

- Assembly of *Drosophila yakuba* with three other *Drosophila* species:

Scaffolds	10
Contigs	1538 (94.92%)
Miss-ordered	26
Contigs N50	162 216
Scaffolds N50	30 316 814

- Assembly of multiple mouse lines
- Capturing rearrangements with assembly graph
- Illumina BaseSpace integration

BaseSpace®
Genomics Cloud Computing

Acknowledgements

Son Pham

Pavel Avdeyev

Dmitry Meleshko

Nikolay Vyahhi

Brian Raney Benedict Paten

Tamara Panesh

Anna Arthuykhova

- Travel funding was generously provided by Akamai Technologies

http://fenderglass.github.io/Ragout