Chromosome scale de novo assembly of genomes using chromatin interaction data

Jay Ghurye

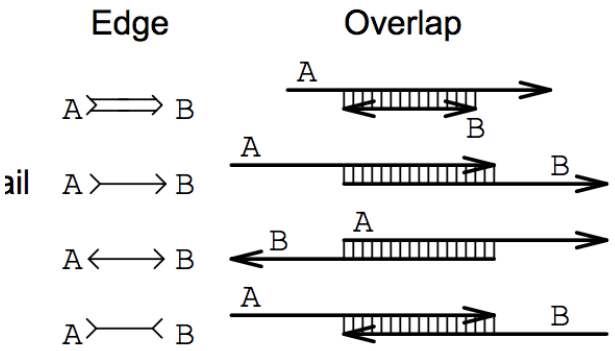
Center for Bioinformatics and Computational Biology

University of Maryland – College Park

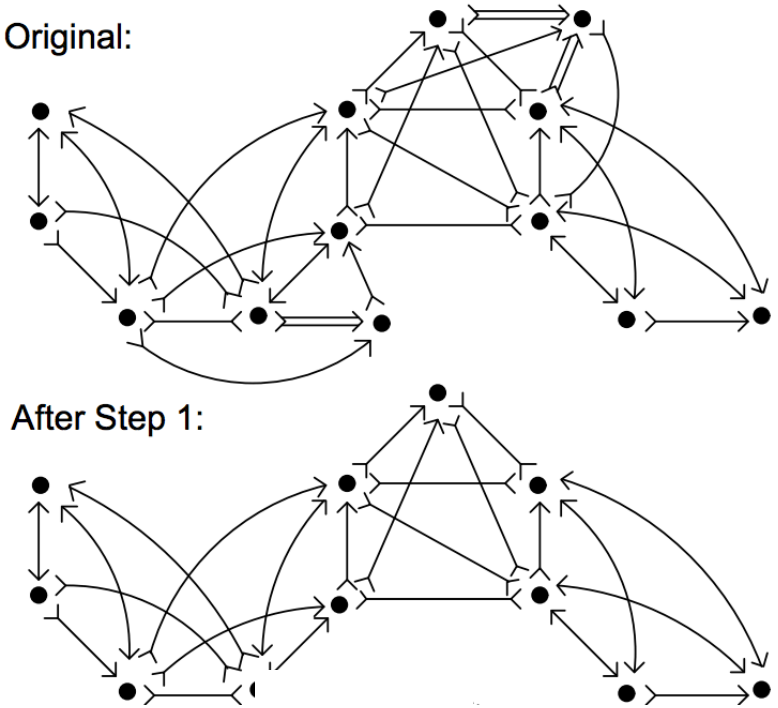
Genome assembly is a big puzzle!

Genome assembly overview

Find overlaps



Construct overlap graph

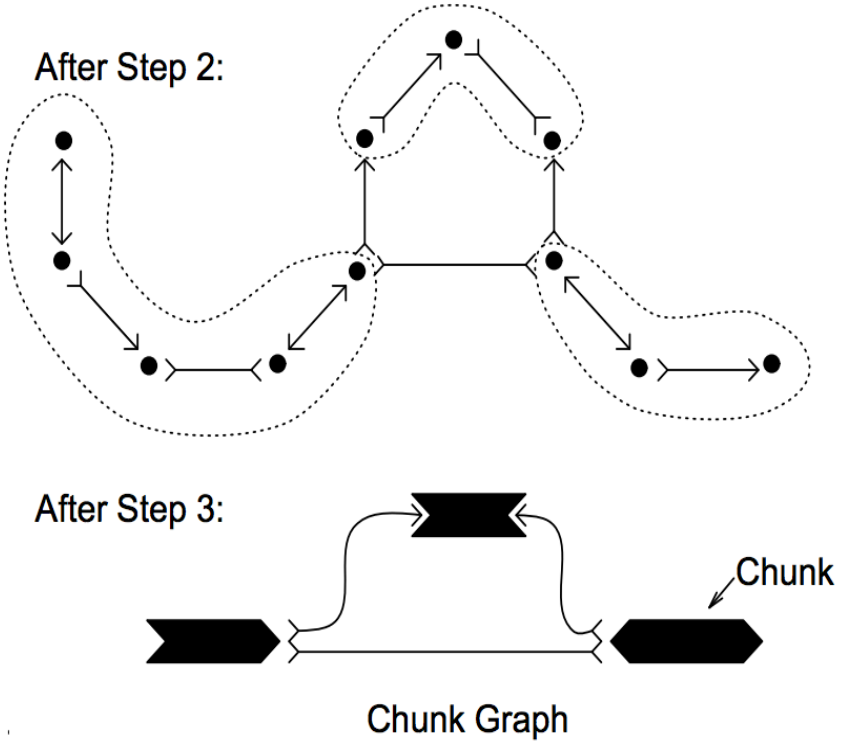


Remove contained edges

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, *2*(2), 275-290.

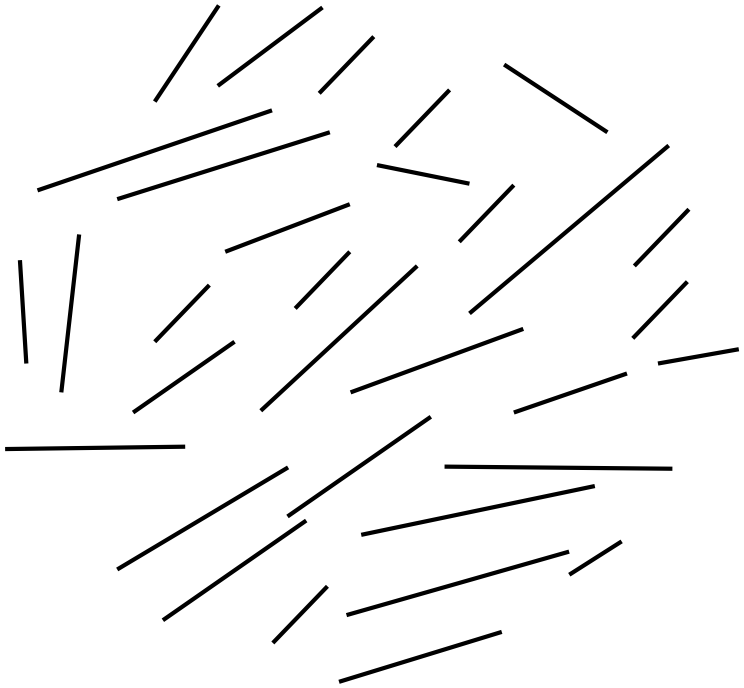
Genome assembly overview

- Remove transitive edges
- Collapse unitigs
- Output unitigs



Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, *2*(2), 275-290.

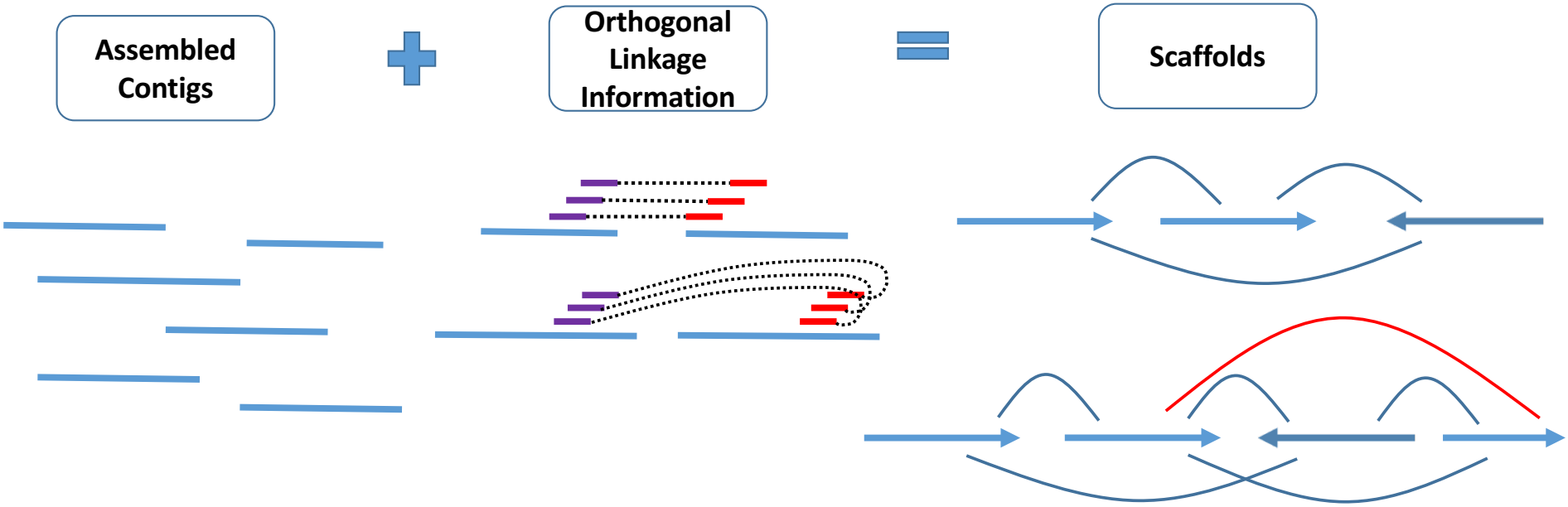
Contigs



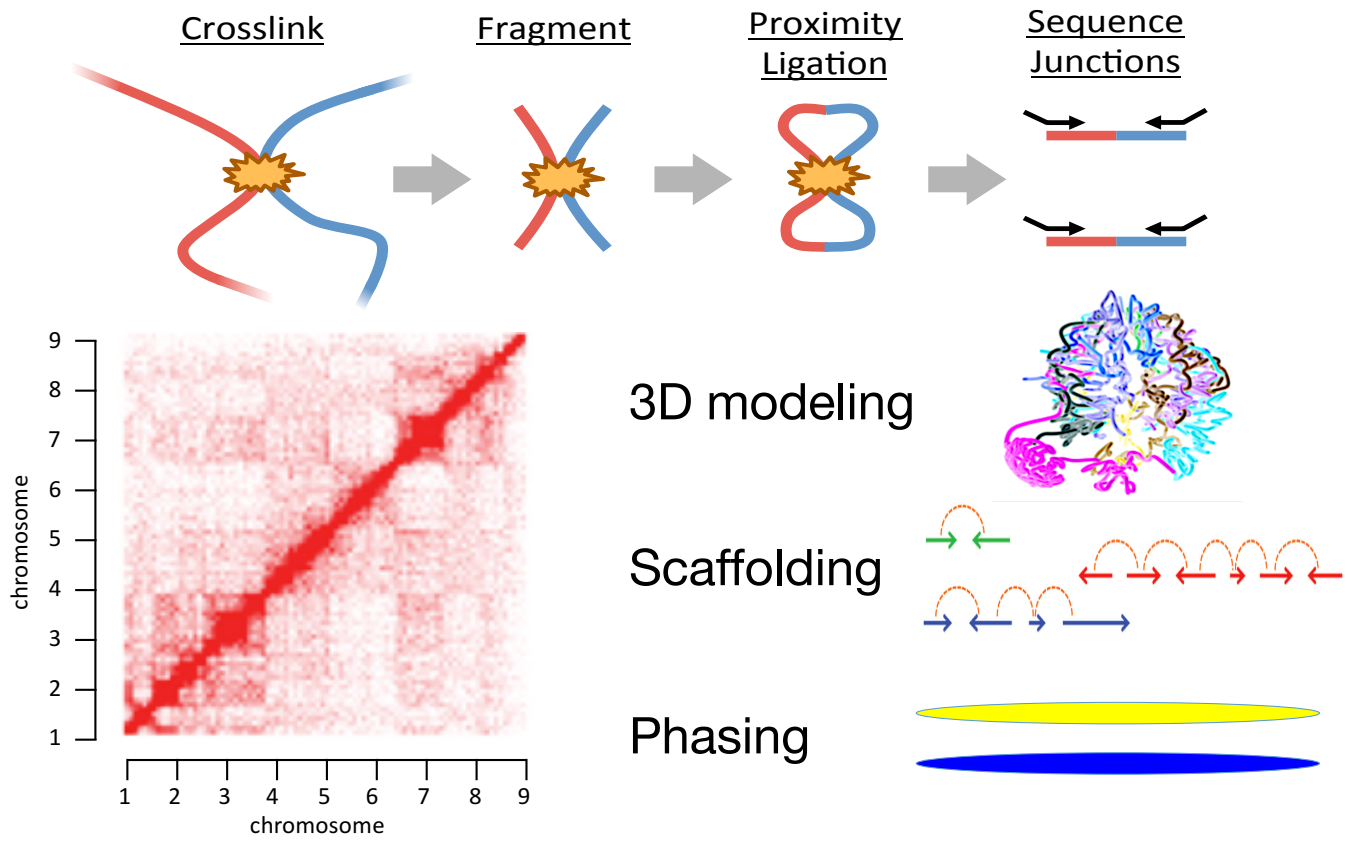
≠

Chromosomes

Genome scaffolding



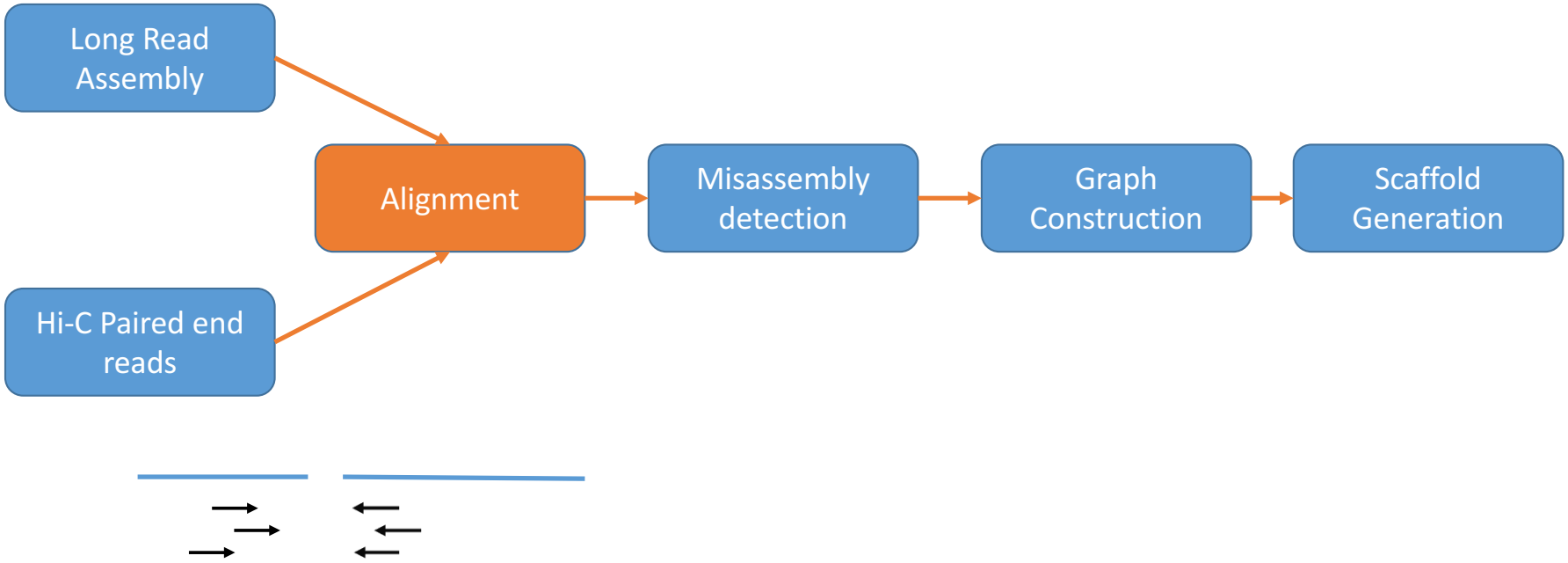
Hi-C chromatin conformation capture



- Duan *Nature* 2010, Burton *Nat Biotech* 2013, Kaplan *Nat Biotech 2013*, Selvaraj *Nat Biotech* 2013

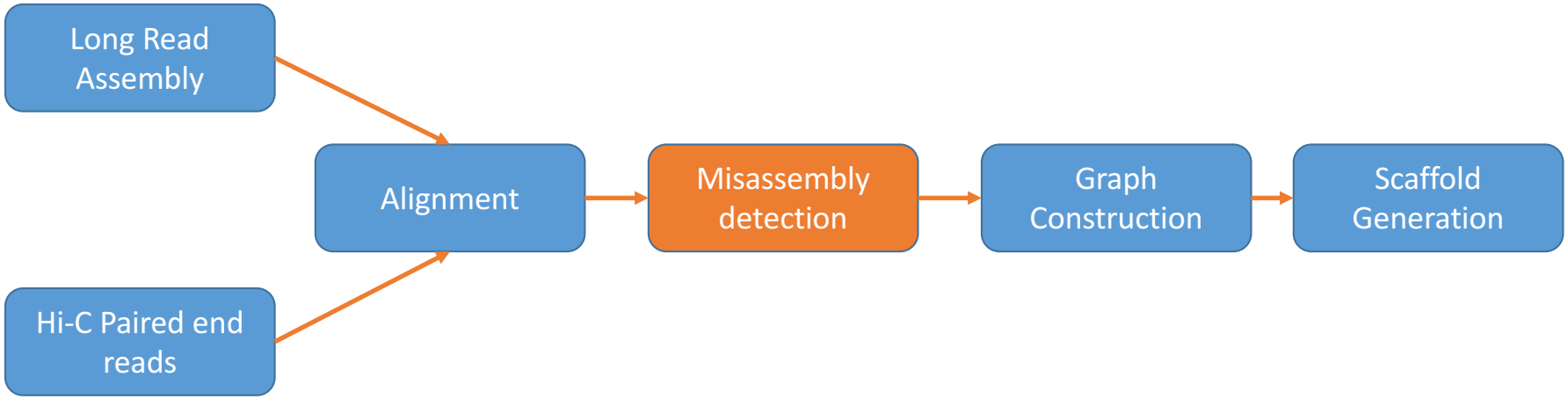
Ivan Liachko, Phase Genomics; Sid Selvaraj, Arima Genomics

SALSA - **S**imple **A**assembL**L**y **S**c**A**ffolder

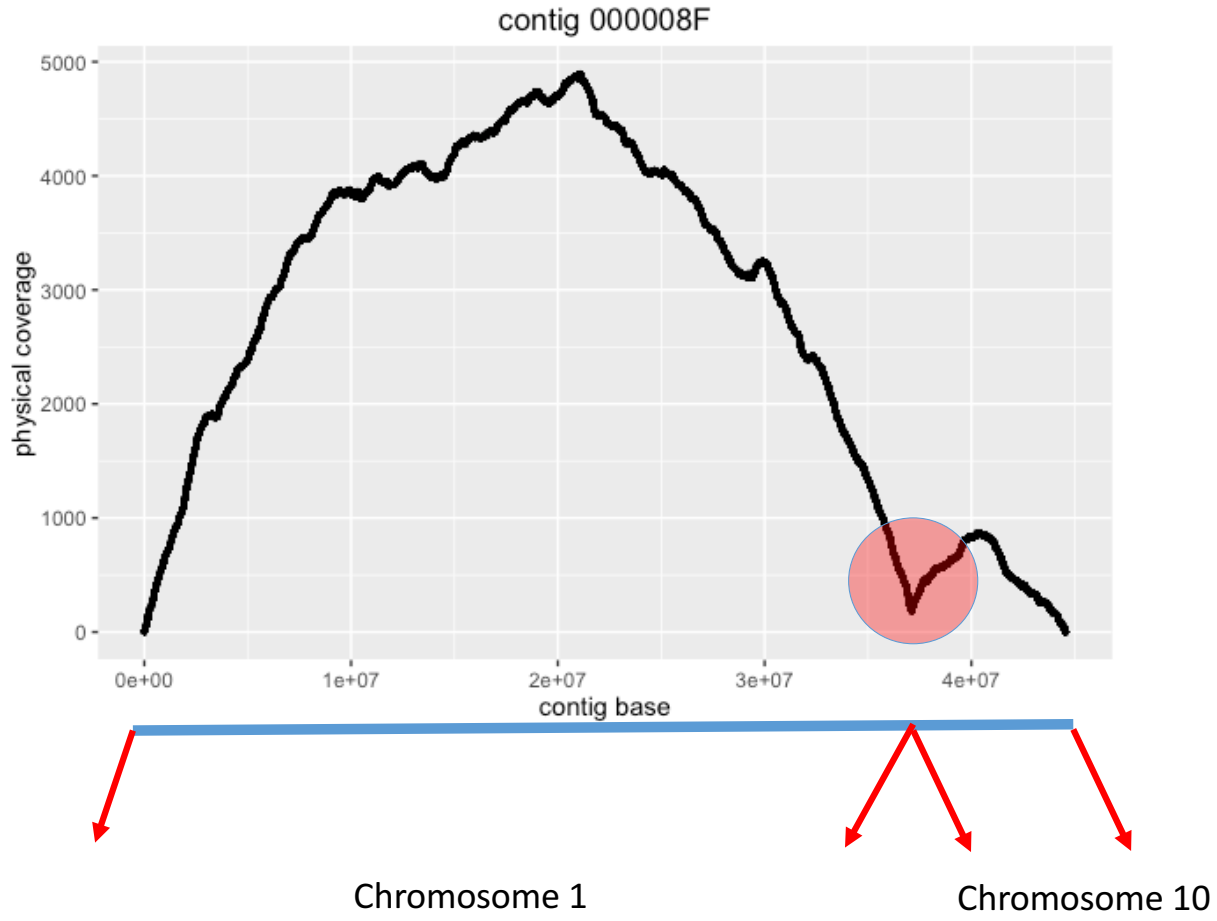


Ghurye, Jay, et al. "Scaffolding of long read assemblies using long range contact information.'' BMC Genomics

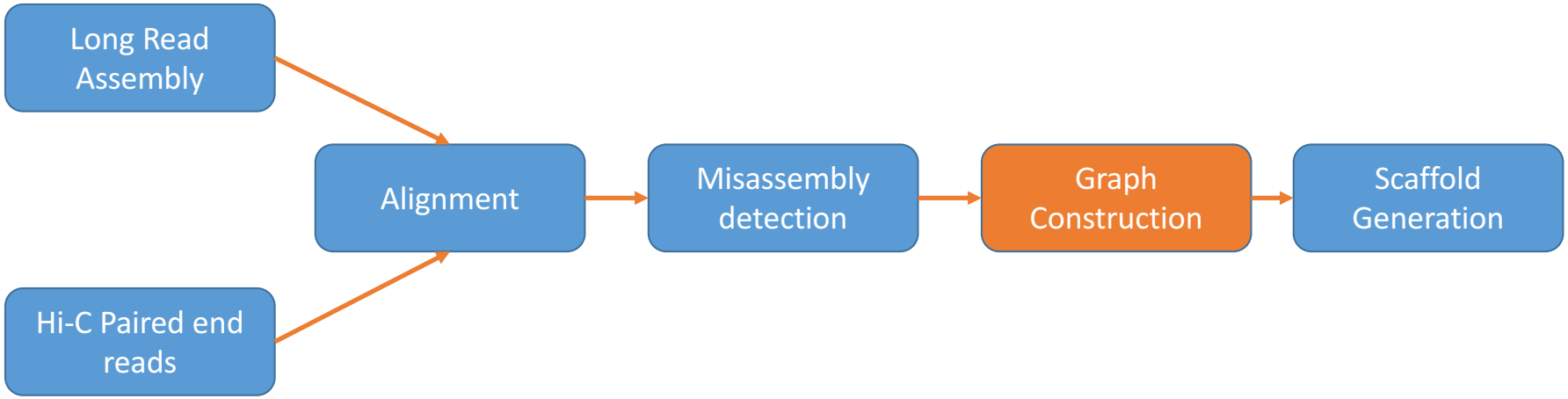
SALSA - **S**imple **A**ssembb**L**y **S**c**A**ffolder



Ghurye, Jay, et al. "Scaffolding of long read assemblies using long range contact information.'' BMC Genomics


contig 000008F
physical coverage
5000
4000
3000
2000
1000
0
0e+00
1e+07
2e+07
3e+07
4e+07
contig base
Chromosome 1
Chromosome 10

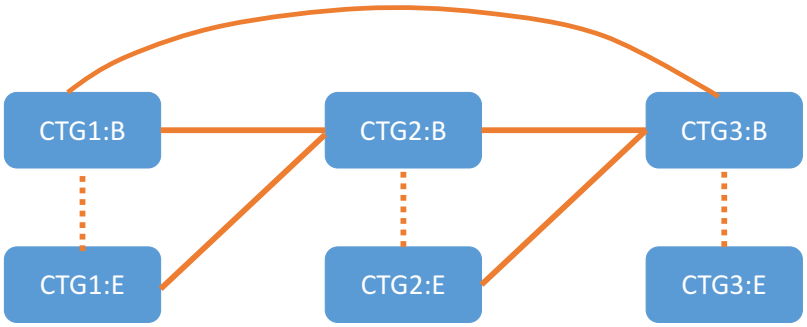
SALSA - **S**imple **A**ssembl**L**y **S**c**A**ffolder

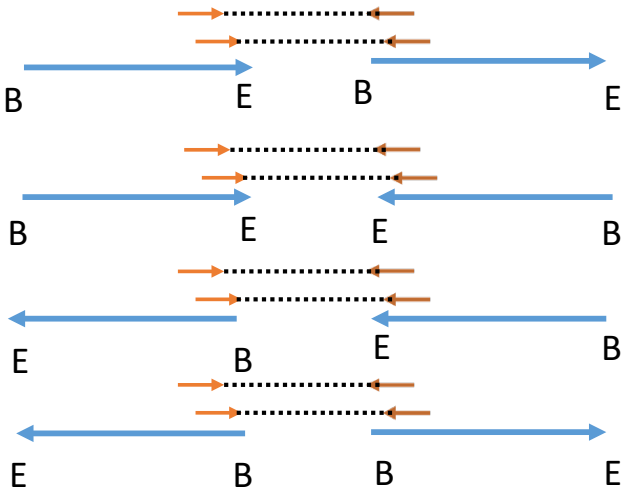


Ghurye, Jay, et al. "Scaffolding of long read assemblies using long range contact information.'' BMC Genomics

Scaffold graph

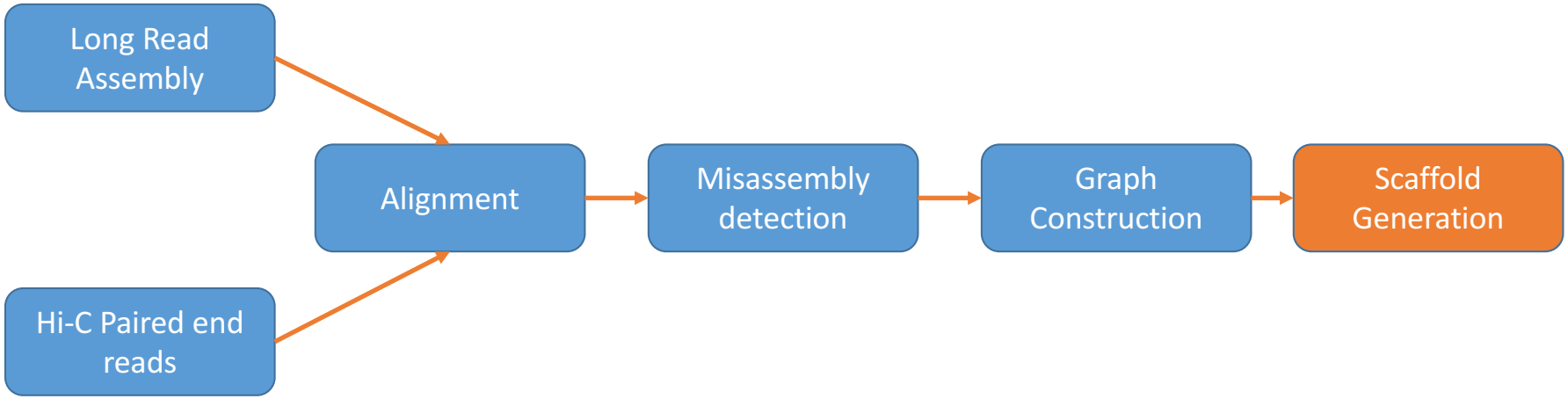
- Nodes are contig ends (Begin and End)
- Edges imply HiC read linkage





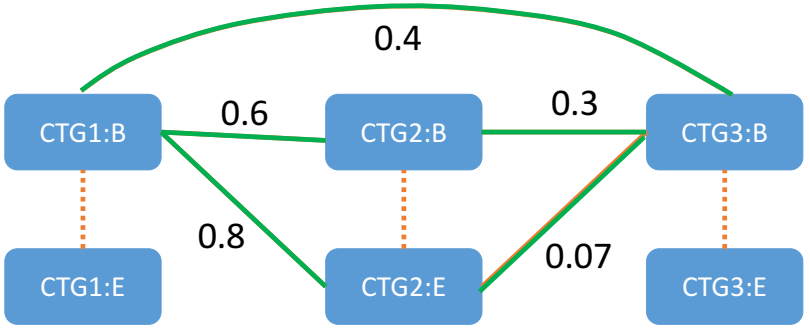
- Number of sites where restriction enzyme reduce length bias
- Find # cut sites for each contig
- Score for particular orientation = # read pairs / #restriction sites

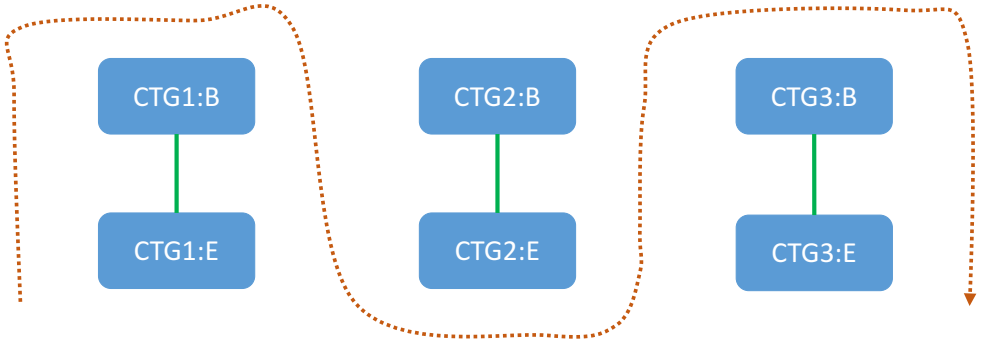
SALSA - **S**imple **A**ssembb**L**y **S**c**A**ffolder



Ghurye, Jay, et al. "Scaffolding of long read assemblies using long range contact information.'' BMC Genomics

Scaffold construction





Consider edges in decreasing order of weights

Traversal of this graph gives scaffold

[CTG1:E, CTG1:B, CTG2:E, CTG2:B, CTG3:B, CTG3:E]

Reverse Reverse Forward

NA12878 assembly evaluation

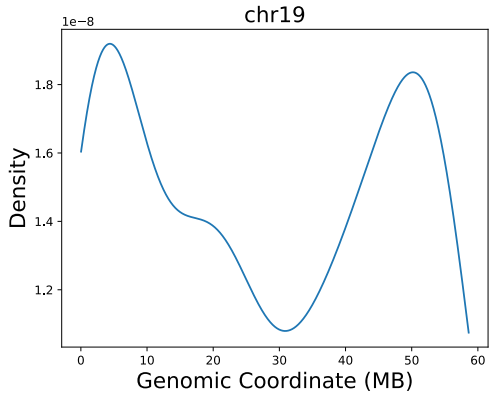
Metric
#Scaffolds
Total Bases
Aligned bases
Breakpoints
Relocations
Translocations
Inversions

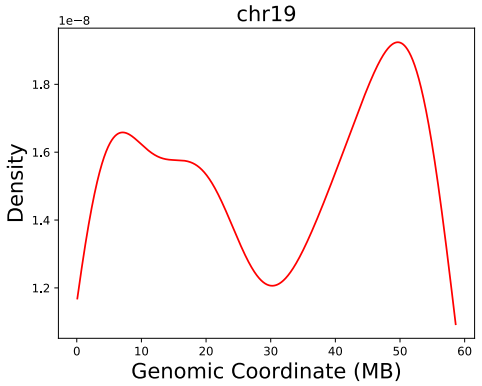
Some errors are common in both scaffolds due to structural variations between GRCh38 and NA12878 Genome

NA12878 Error Distribution

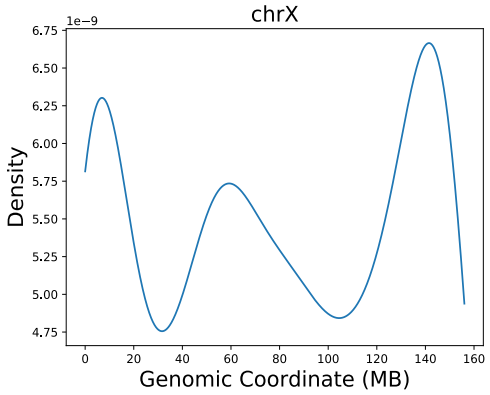
SALSA **LACHESIS**

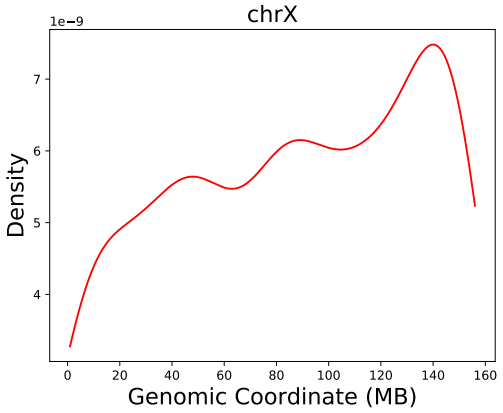
(A)





(B)





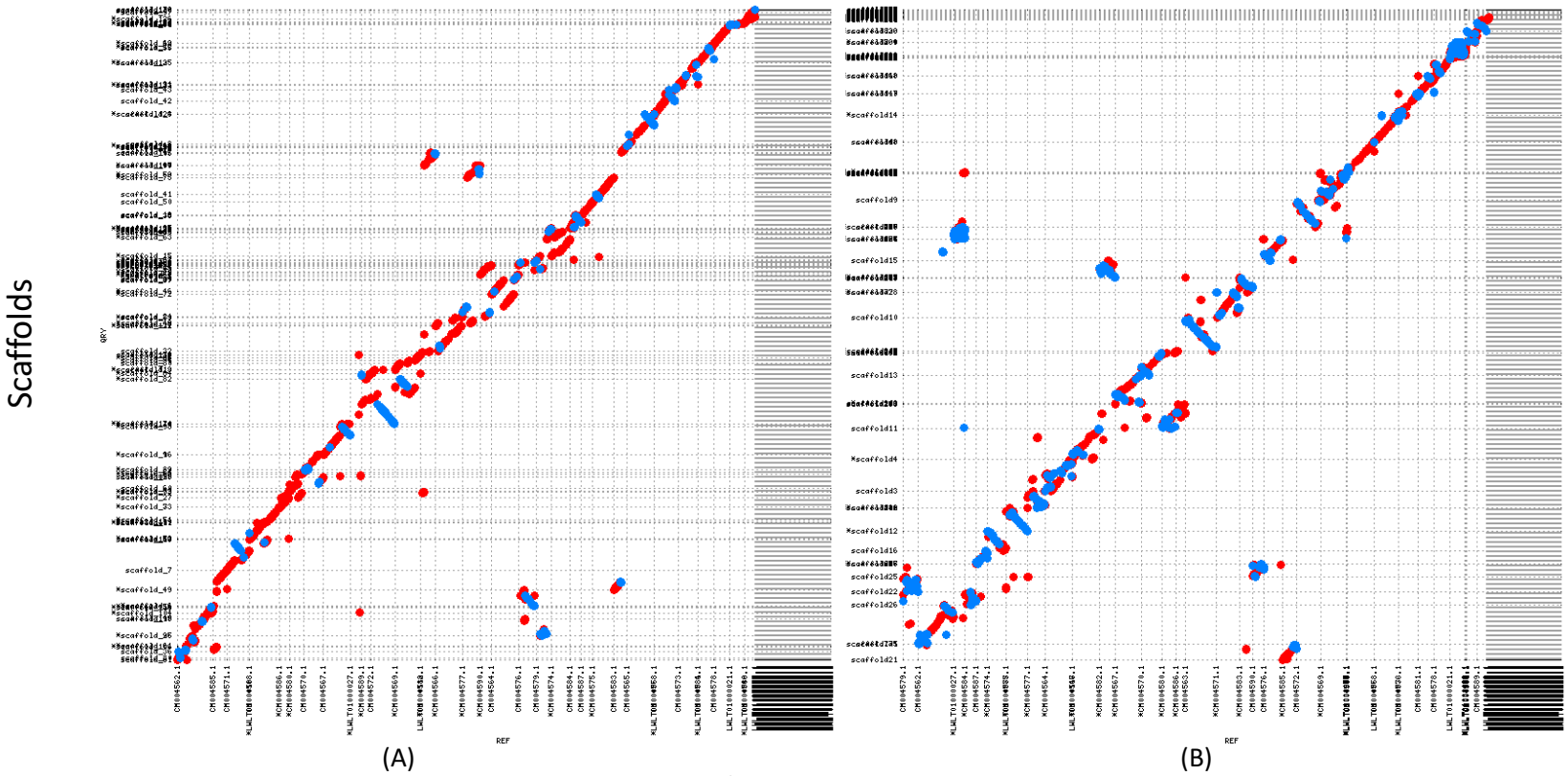
Result – NA19240 : Pacbio + HiC

Feature	Value
#Contigs	3242
Contig NG50	23.98 Mb
# Scaffolds	118
Scaffold NG50	69.49 Mb
Number of Bases	2789088362 (2.78 Gb)
# chr arms (p and q) covered by single scaffold	26
# chr arms covered by 2 scaffolds	14

Result: Goat – Pacbio + HiC

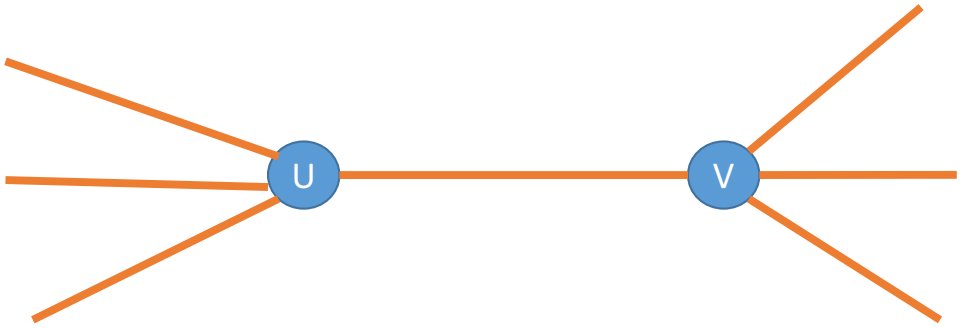
SALSA
365 errors

LACHESIS
1414 errors



SALSA – New Improvements

- Edge weighing Scheme



W(u,v) = RE normalized edge score for edge u,v

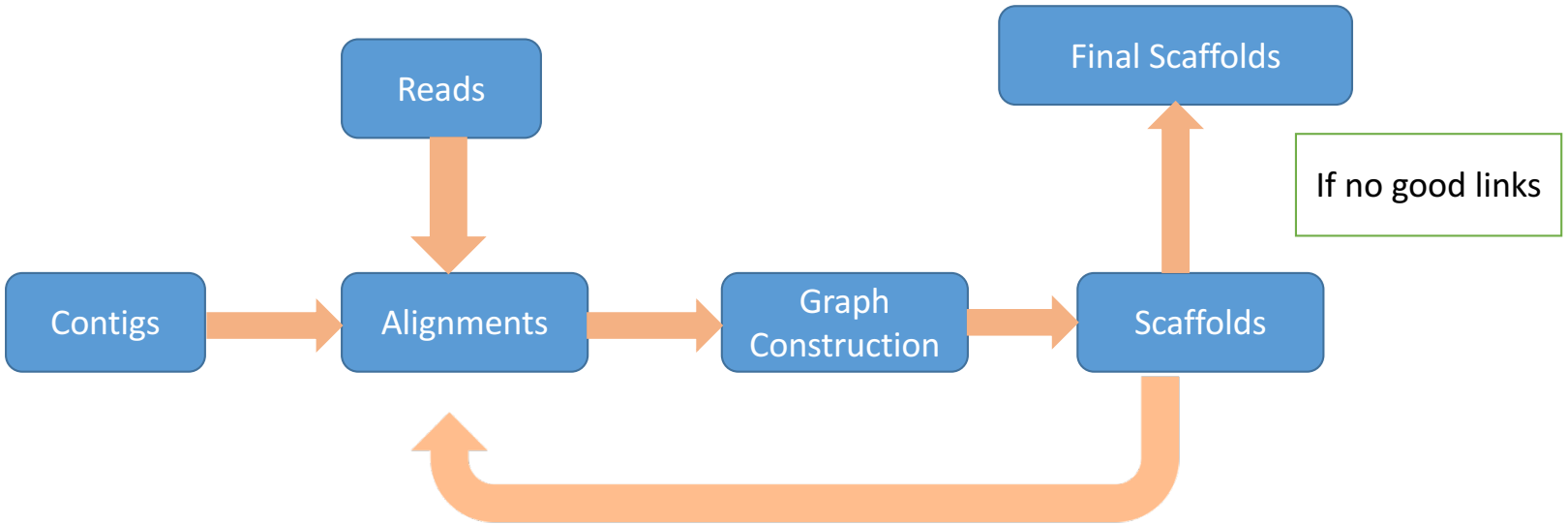
W(u,u1) = max weighted incident edge on u

W(v,v2) = max weighted incident edge on v

W'(u,v) = W(u,v)/ max(W(u,u1), W(v,v2))

SALSA – New Improvements

- Iterative pipeline



Credits: Adam Phillippy

Image attribution, left to right, top to bottom

By André Karwath aka Aka - Own work, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=227170
By Didier Descouens - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=15954199
Public Domain
By Clemson University - USDA Cooperative Extension Slide Series, Bugwood.org, CC BY 3.0 us, https://commons.wikimedia.org/w/index.php?curid=8596665
By Pacific Southwest Region - https://www.flickr.com/photos/usfws_pacificsw/14534381519, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=39739922
https://bigblueblogging.wordpress.com/2015/08/09/common-hermit-crab-pagurus-bernhardus/
http://www.eve.ucdavis.edu/grosberg/research_organisms2.htm
By Thomas Lersch - Own work, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1001910
By Holbrook, John Edwards, 1794-1871 (English Wikipedia) - North American herpetology; or, A description of the reptiles inhabiting the United States. , modified from Biodiversity Heritage Library, Public Domain, https://commons.wikimedia.org/w/index.php?curid=6491262
By The original uploader was Marknesbitt at English Wikipedia (January 2006). Later uploaded by Jeannot12 at fr.wikipedia (December 2006) - (See above.)Transferred from fr.wikipedia to Commons by Crochet.david using CommonsHelper.Originally uploaded 06:55, 18 January 2006 (UTC) by Marknesbitt (talk • contribs) to en:Wikipedia (log)., Public Domain, https://commons.wikimedia.org/w/index.php?curid=6699396
By Michael MacNeil, USDA. Original uploader was Gzuckier at en.wikipedia - Transferred from en.wikipedia(Original text : http://www.genome.gov/pressDisplay.cfm?photoID=67), Public Domain, https://commons.wikimedia.org/w/index.php?curid=11411883
Megan Sayre
By James Gathany/CDC - This media comes from the Centers for Disease Control and Prevention's Public Health Image Library (PHIL), with identification number #4487.Note: Not all PHIL images are public domain; be sure to check copyright status and credit authors and content providers.English | Slovenščina | +/−, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1831559
By Muhammad Mahdi Karim (www.micro2macro.net) Facebook Youtube - Own work, GFDL 1.2, https://commons.wikimedia.org/w/index.php?curid=9556152
By Bjørn Christian Tørrissen - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17340222
By Franz Eugen Köhler, Köhler's Medizinal-Pflanzen - List of Koehler Images, Public Domain, https://commons.wikimedia.org/w/index.php?curid=255643
By User:Zac Wolf (original), en:User:Stefan (cropping) - en:Image:Whale shark Georgia aquarium.jpg, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=1312499
By Mark Peters from Baltimore, USA - Poplar Spring Animal Sanctuary, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=11762434

Acknowledgements

- Jason Chin
- Sergey Koren
- Adam Phillippy
- Arang Rhie
- Derek Bickhart
- Mihai Pop

SALSA: https://github.com/machinegun/hi-c-scaffold