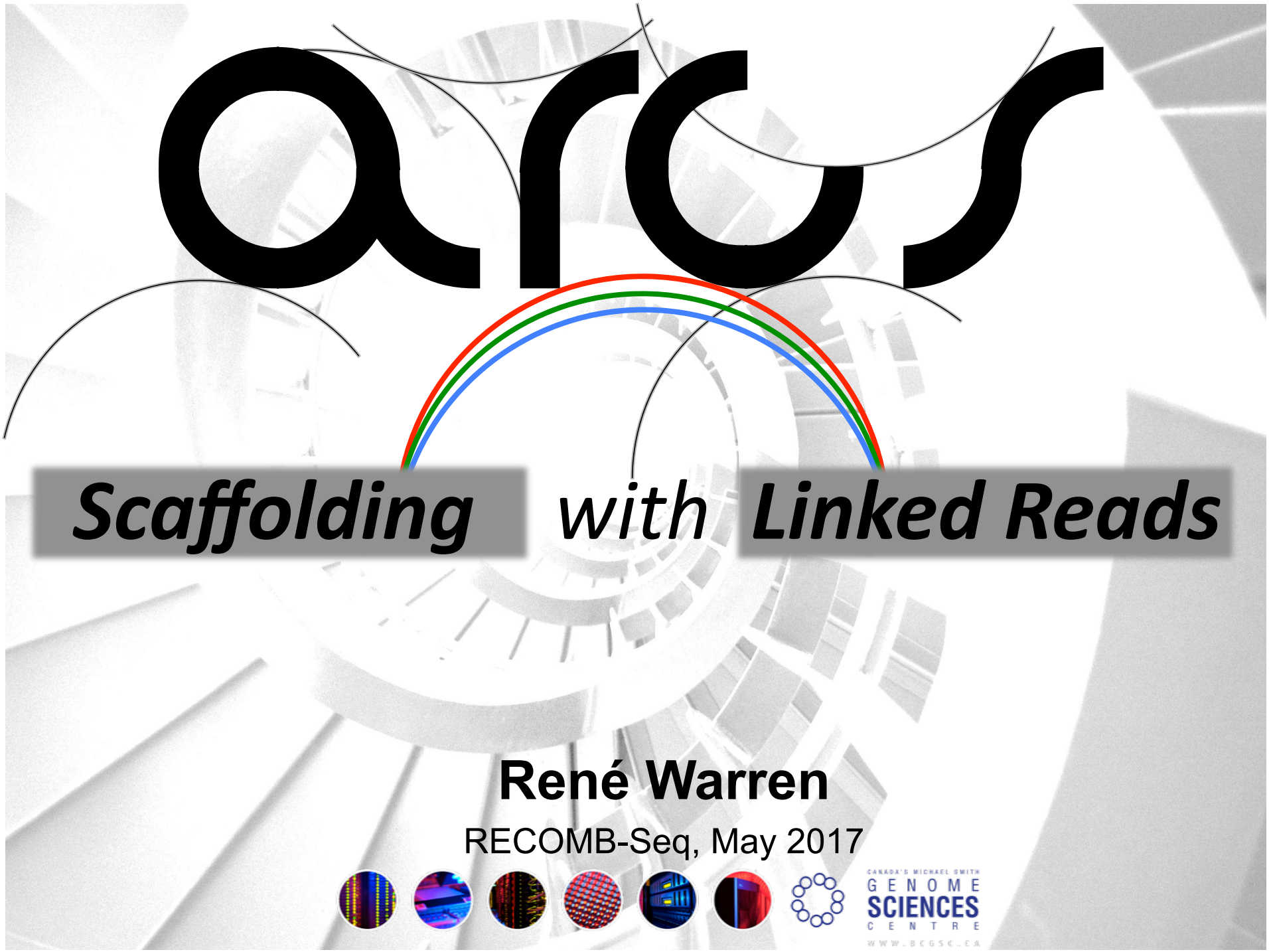


René Warren

RECOMB-Seq, May 2017

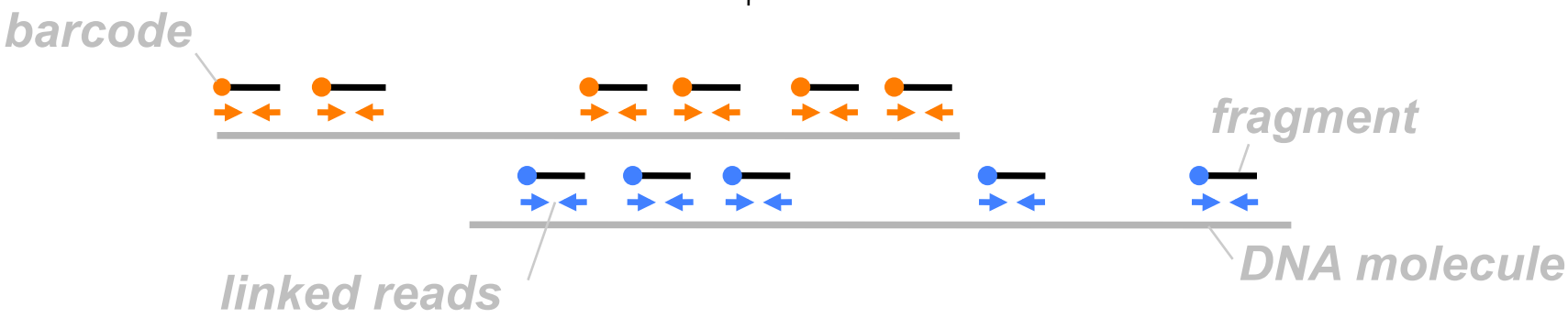
CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE
WWW.BCGSC.CA

Genome scaffolding

- **Order and orient contigs**
  - Separated by gaps or overlaps
  - Goal : one scaffold, one chromosome
  - Recover complete genes
- **10X Genomics' Chromium : co-located linked reads**
  - DNA molecules : ↗ physical, ↘ individual sequence coverage
- **Bioinformatics Technologies**
  - Applicable to linked reads:
    - *fragScaff* HiC data (Adey 2014, Mostovoy 2016)
    - *Architect* Moleculo read cloud (Kuleshov 2016)
    - *Supernova* Chromium (Weisenfeld 2016)

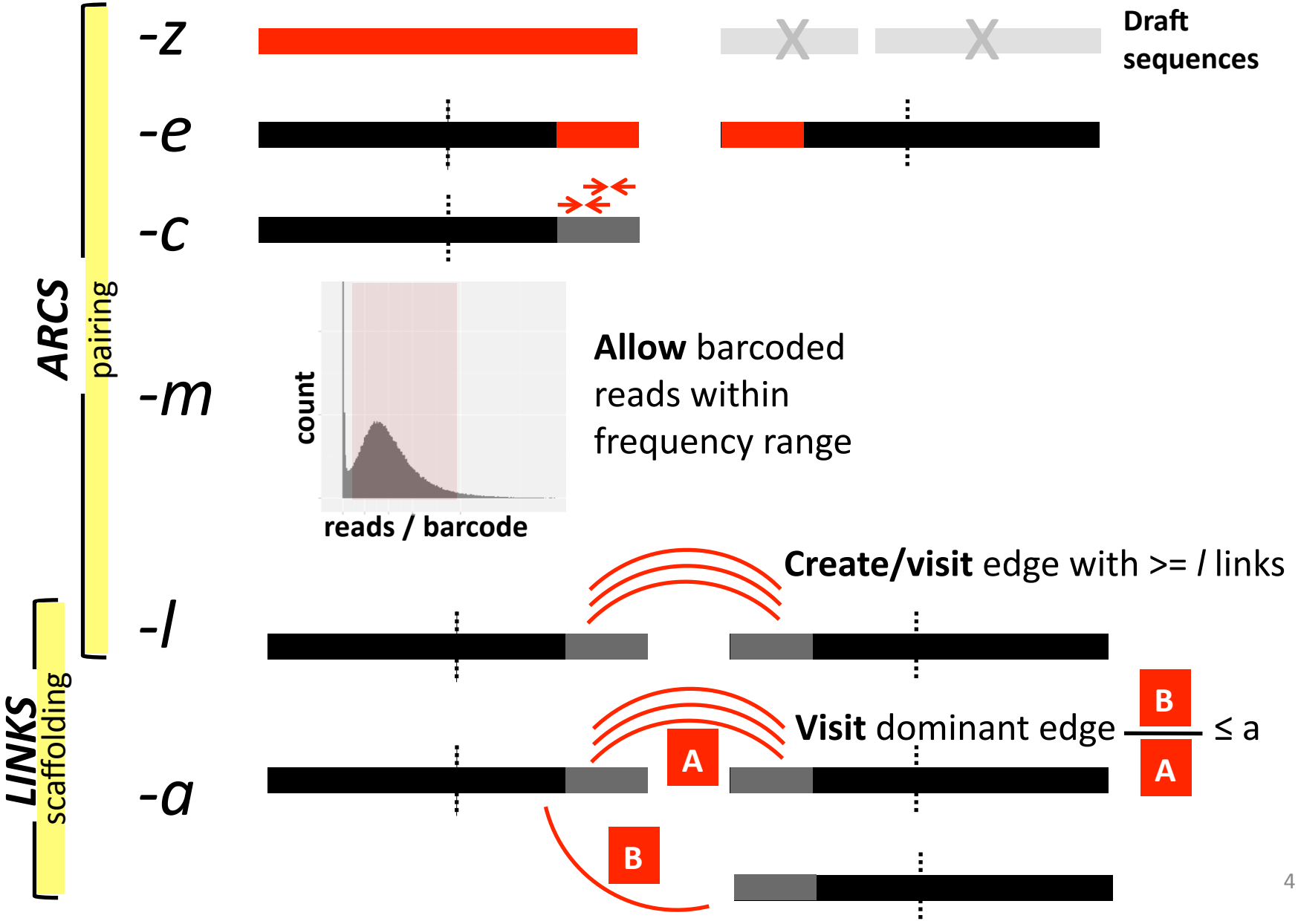
  › Alignment-based *de novo* assembler

Assembly Roundup by Chromium Scaffolding



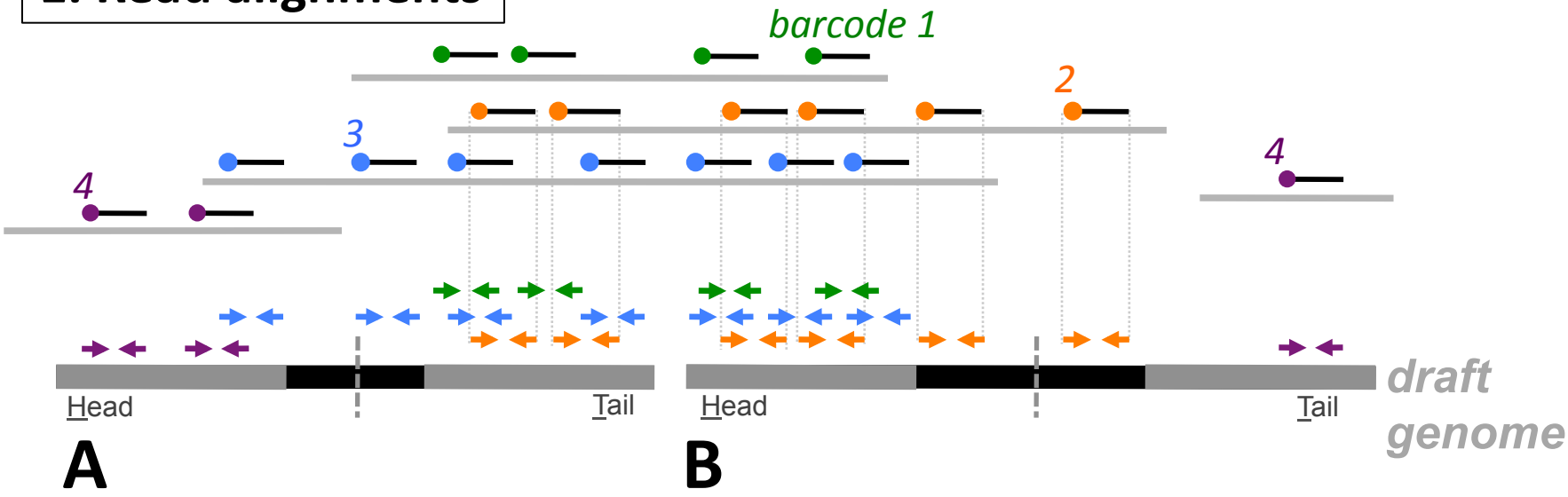
- **Captures** long-range info
- **Identifies** co-located scaffolds in genome draft
- **Improves** *de novo* assembly

Parameters


ARCS
pairing
LINKS
scaffolding
-z
-e
-c
-m
-l
-a
Draft sequences
count
reads / barcode
Allow barcoded reads within frequency range
Create/visit edge with >= l links
Visit dominant edge
B
A
≤ a
A
B

Method

1. Read alignments



Method

1. Read alignments

2. Binomial test & tally

Sequence A barcode	Head	Tail		Sequence B barcode	Head	Tail
1	0	**2**	T-H	1	**2**	0
2	0	**2**	T-H	2	**2**	0
3	1	**2**	T-H	3	**3**	0
4	**2**	0	H-T	4	0	**1**

Is read distribution in 5' (head) or 3' (tail) end different from a uniform distribution?

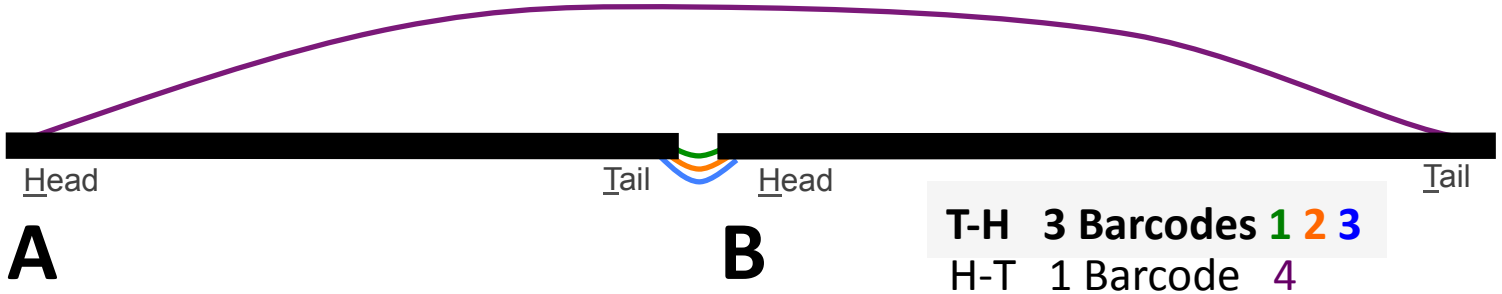
Barcode	(Seq. ID, H/T), count		
1	(A, T): 2	(B, H): 2	
2	(A, T): 2	(B, H): 2	
3	(A, T): 2	(A, H): 1	(B, H): 3
4	(A, H): 2	(B, T): 1	

hash tracks head-tail tally for each barcode

Method

3. Pair sequences

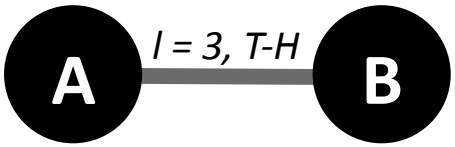
For any two sequences, tally head & tail barcode support



T-H 3 Barcodes 1 2 3
H-T 1 Barcode 4

Seq. ID 1	Seq. ID 2	# HH links	# HT links	# TH links	# TT links
A	*B*	*0*	*1*	*3*	*0*

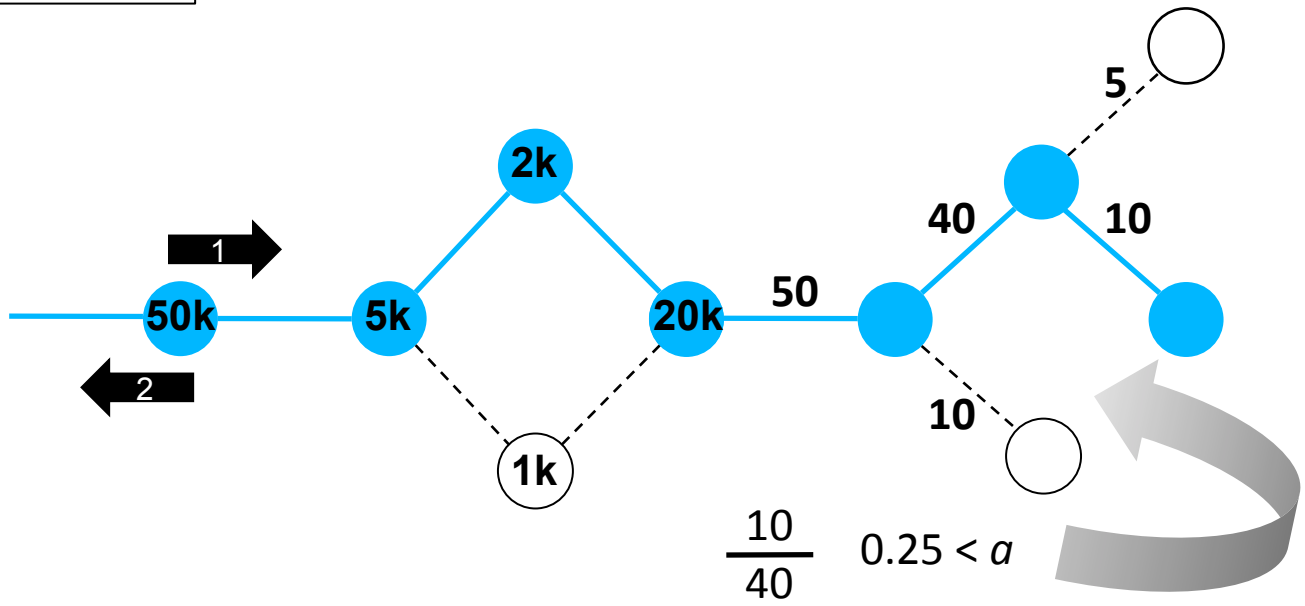
4. Output graph



Apply heuristics favoring most supported H-T combo and build graph

Method

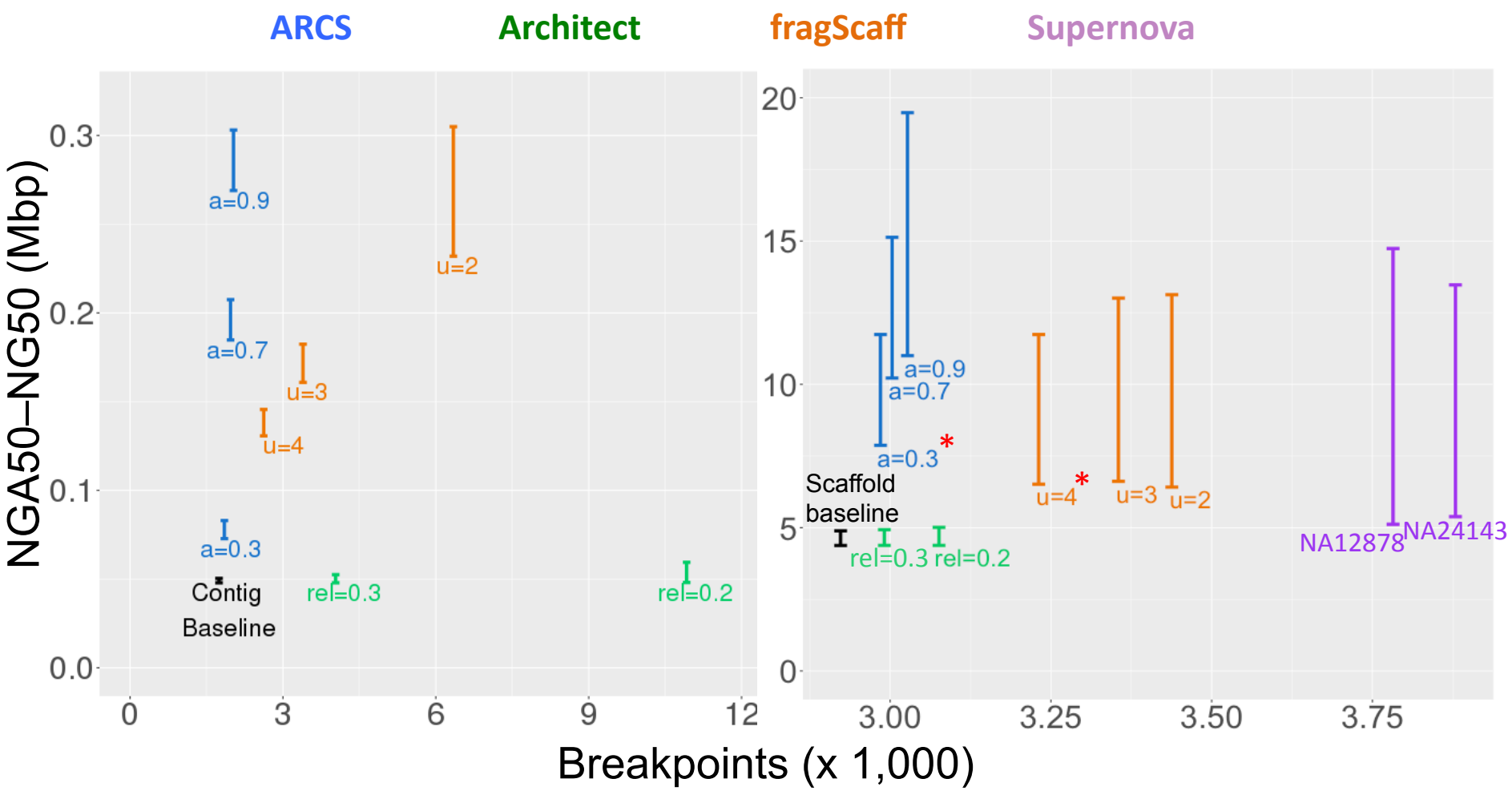
5. Scaffolding



Currently utilizes the *LINKS* (Warren *et al.* 2015) scaffolder

Scaffolding human drafts

comparison to GRCh38 human genome reference



Scaffolding human drafts

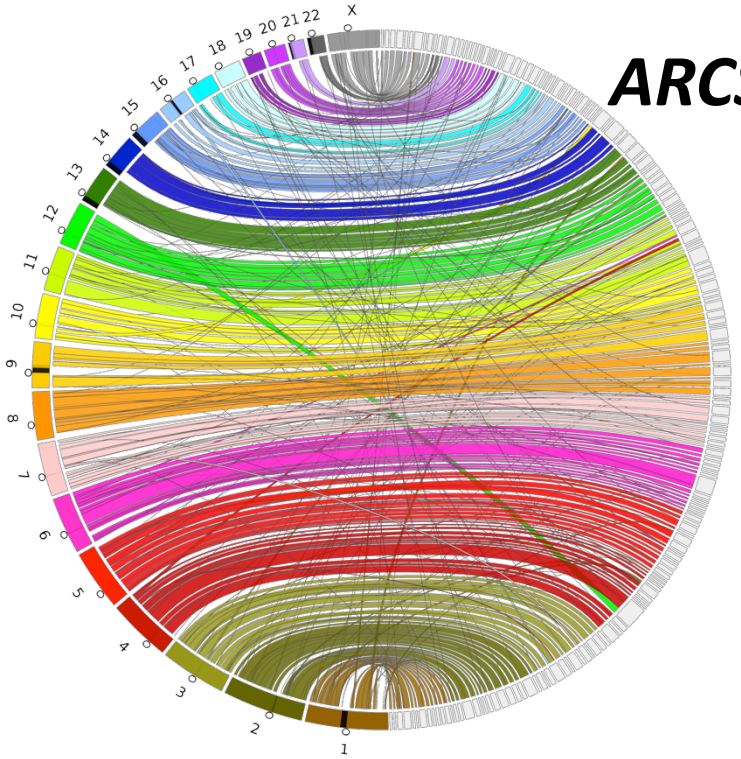
Resources

Tool	Time	Mem
ARCS	**0:55**	**3.4**
fragScaff*	1:59	14.1
Architect	6:07	9.6
*64 threads	h:mm	GB

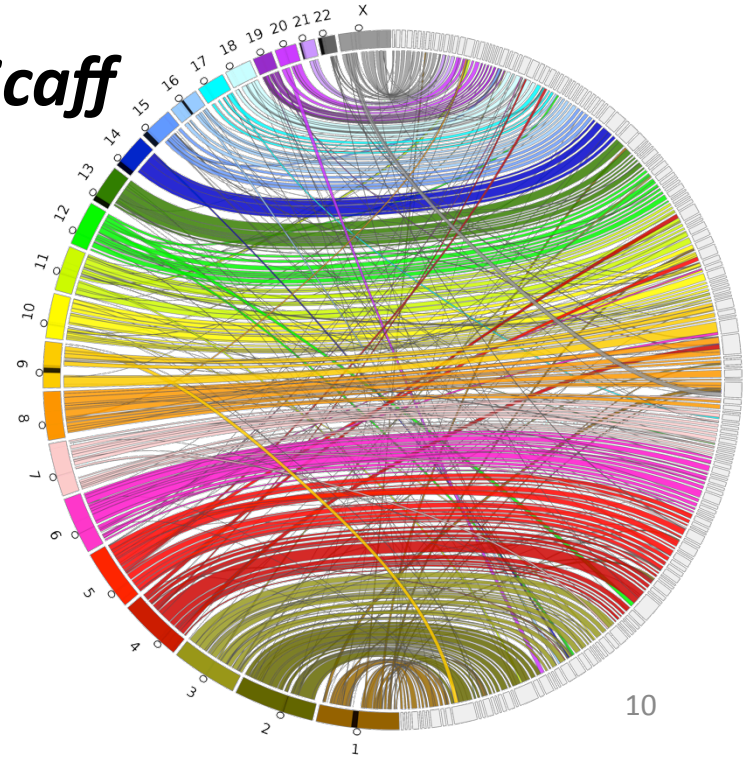
Misassemblies

Assembly	Intra	Inter
Baseline	1090	399
ARCS	**1228**	**434**
fragScaff	1911	510

Inconsistent contig order

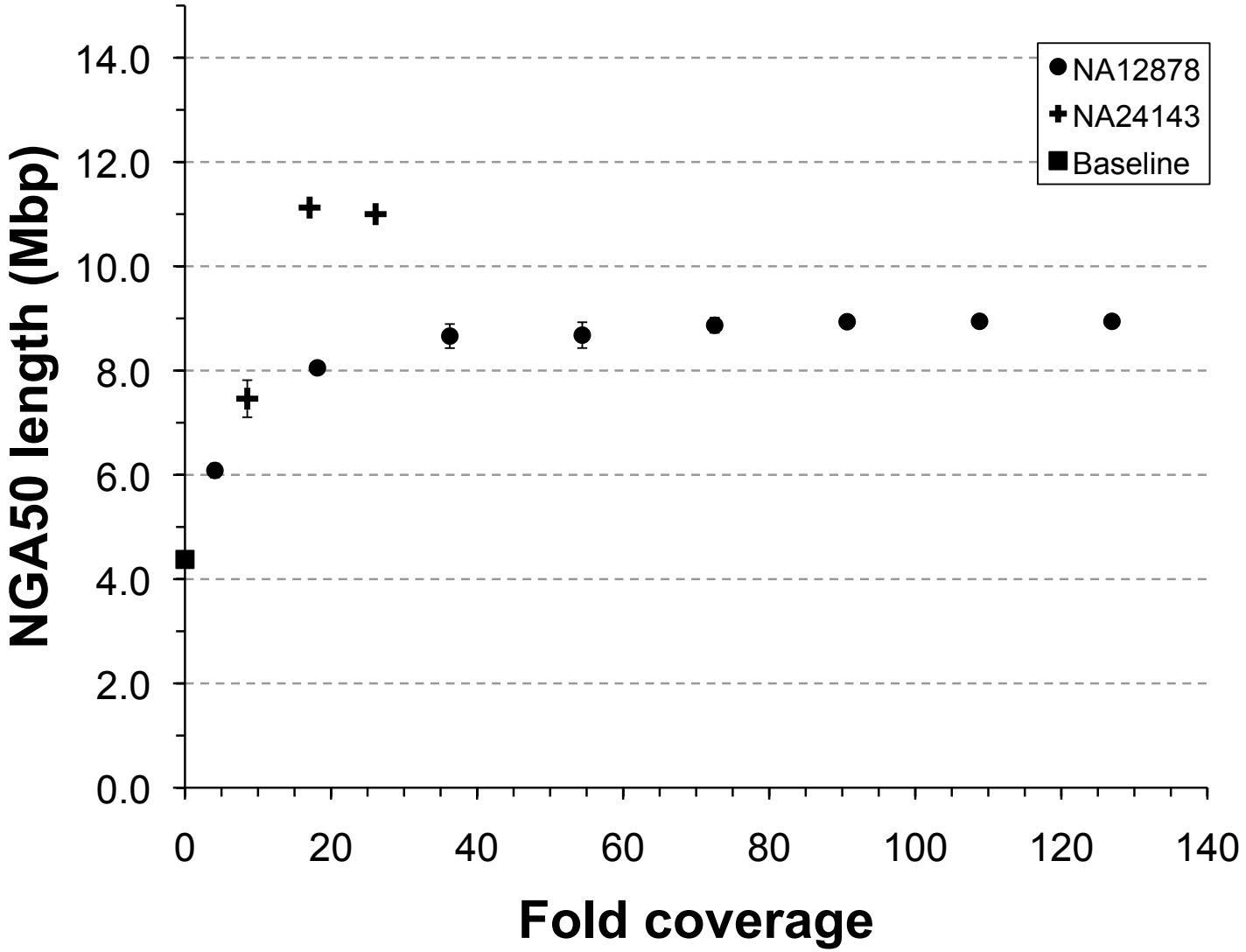


NA24143
1kbp+ GRCh38
alignments



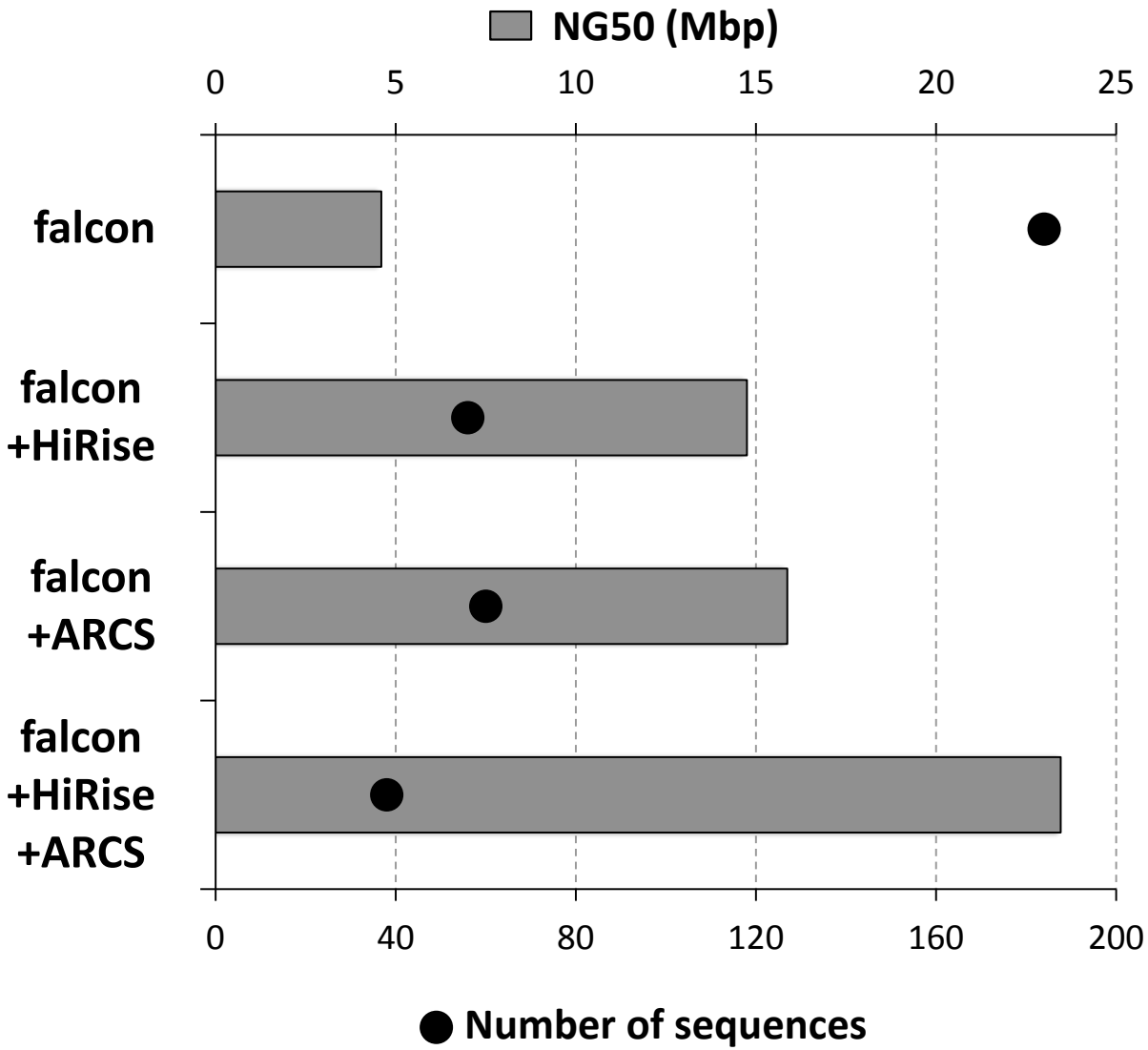
Scaffolding human drafts

works with low (20X) Chromium sequence coverage



Retrospective scaffolding

long-read (pacbio) NA24143 sequencing



Spruce genome scaffolding

Economically important lumber

- **Aims**

  Improve wood quality, pest resistance

- **Challenges**

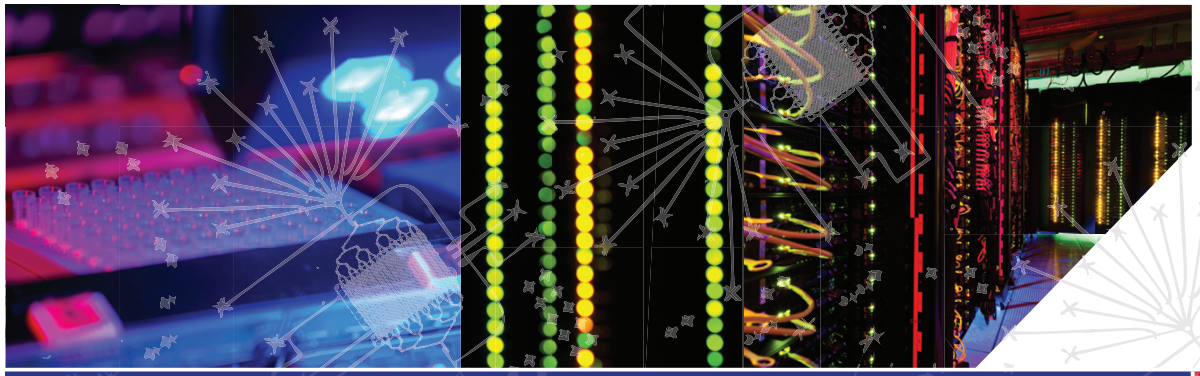
  Large 20 Gbp genome, drafts fragmented

- 8 Chromium lanes (~50-fold coverage)
- 3-fold NG50 increase to 315 kbp (from 87 kbp)
- 9 additional complete genes recovered
  - CEGMA (42%) and BUSCO (31%)

Summary

- Stand-alone, linked-read genome scaffolder
  - retrospective scaffolding
- *Mbp*-range, accurate drafts from short reads
- Fast, scalable, low read coverage

https://github.com/bcgsc/ARCS
https://github.com/warrenlr/LINKS



CANADA'S MICHAEL SMITH

GENOME SCIENCES CENTRE

A leading international centre for genomics and bioinformatics research committed to advancing knowledge of cancer-related diseases, improving human health through disease prevention, diagnosis and therapeutic approaches, and realizing social and economic benefits of genomics research.

54 TERABASES SEQUENCED • A HUMAN GENOME EVERY 17 MINUTES • HIGH-PERFORMANCE COMPUTING

AFFILIATIONS BC Cancer Research Center • BC Cancer Agency • BC Cancer Foundation • Genome BC • Simon Fraser University • University of Britsh Columbia • Genome Sciences Institute

Thank You!

Sarah Yeo* | *Lauren Coombe* | *Justin Chu* | *Ben Vandervalk
Shaun Jackman* | *S. Austin Hammond* | *Inanc Birol

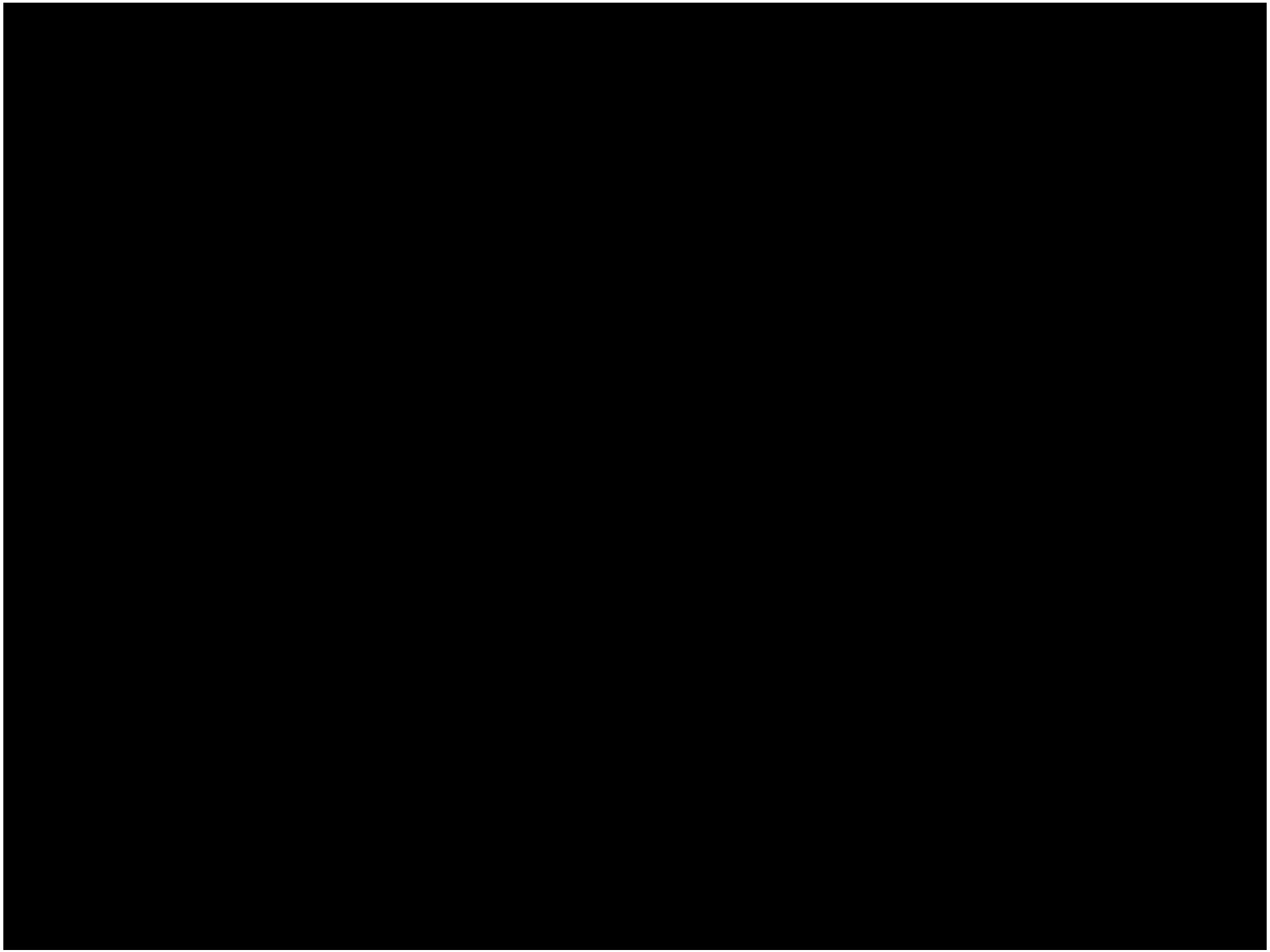
Funding

NIH

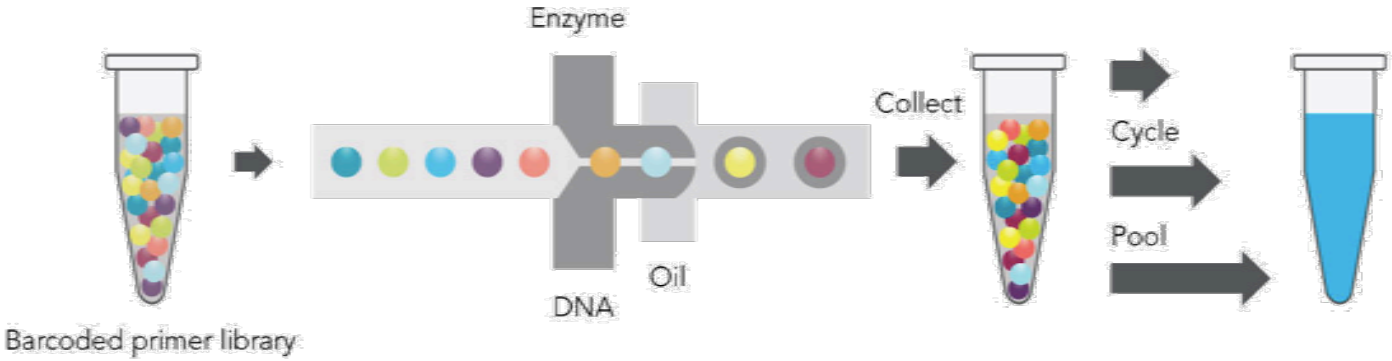
GenomeCanada

BC CANCER FOUNDATION
partners in discovery

BCCA CANCER RESEARCH CENTRE



Overview of 10X Genomics

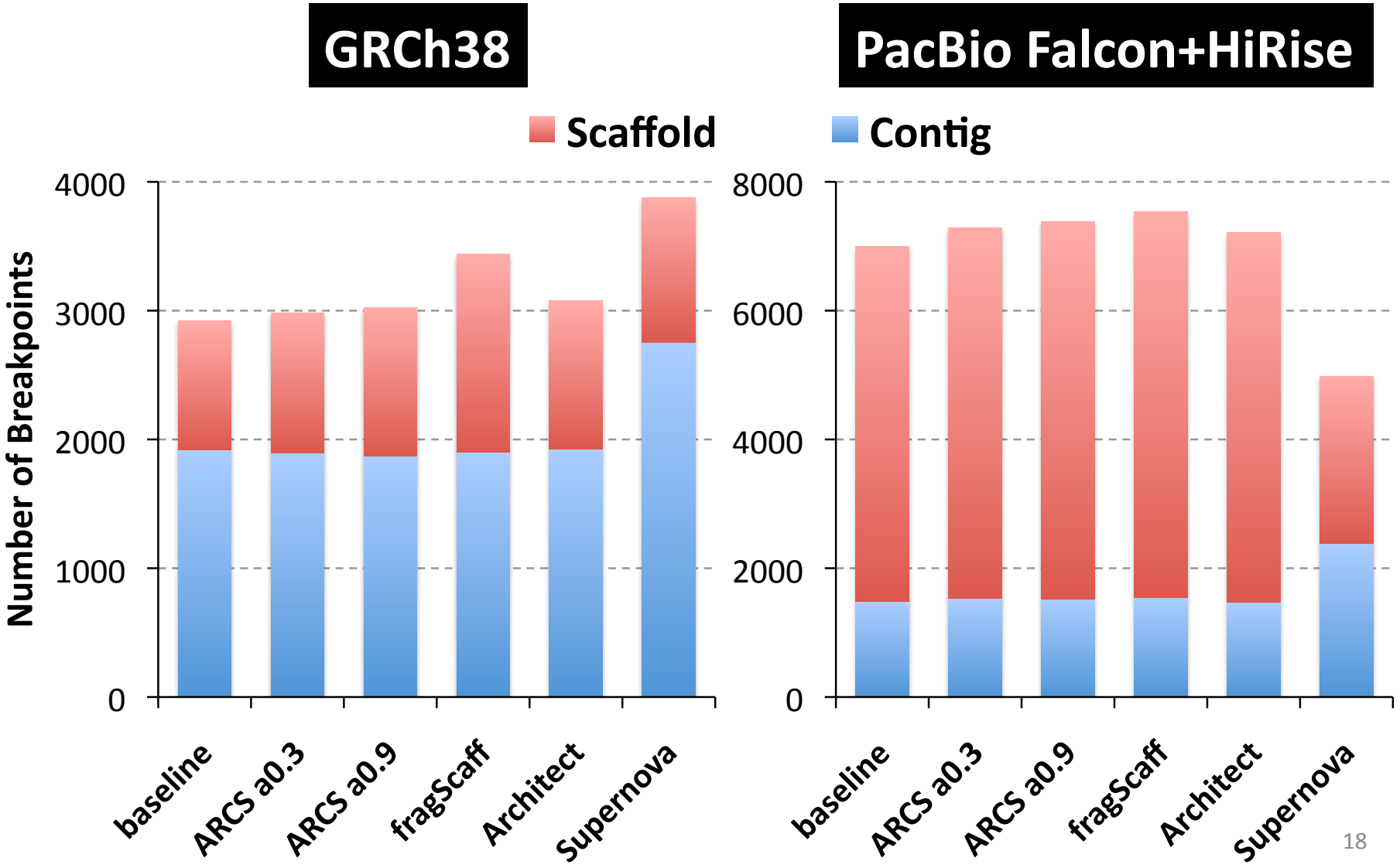


DNA molecules

Picture: http://2.bp.blogspot.com/-9bC-Fg0qMEQ/VPR0yO2y5hI/AAAAAAAABjw/B6M3OLAvqLE/s1600/Screenshot%2B2015-03-02%2B14.36.11.png

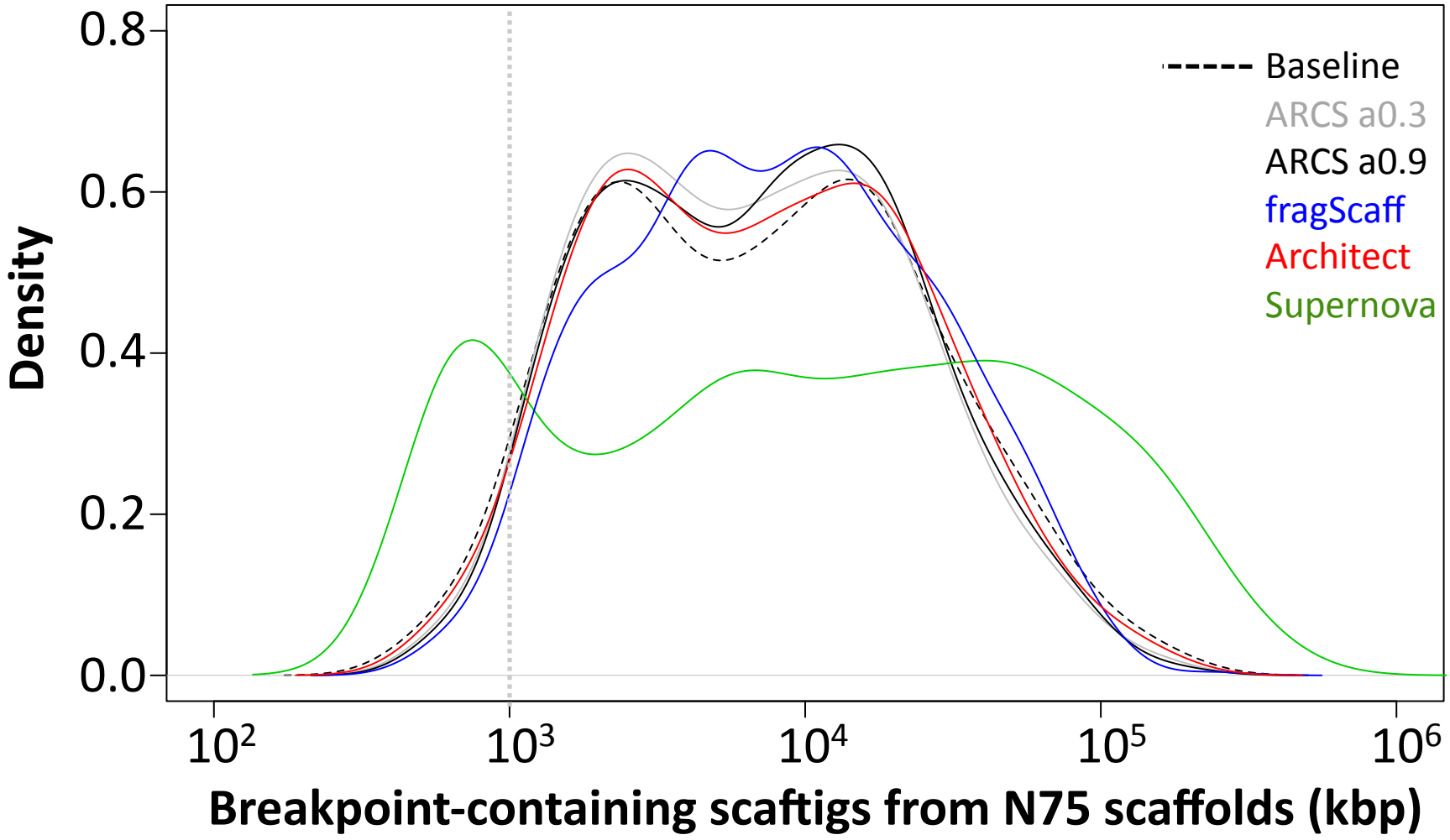
Scaffolding NA24143 assembly drafts

breakpoint analysis



Scaffolding NA24143 assembly drafts

Scaftigs	n:500	N50
Baseline	210,391	33,905
Supernova	69,299	115,158



Scaffolding human drafts

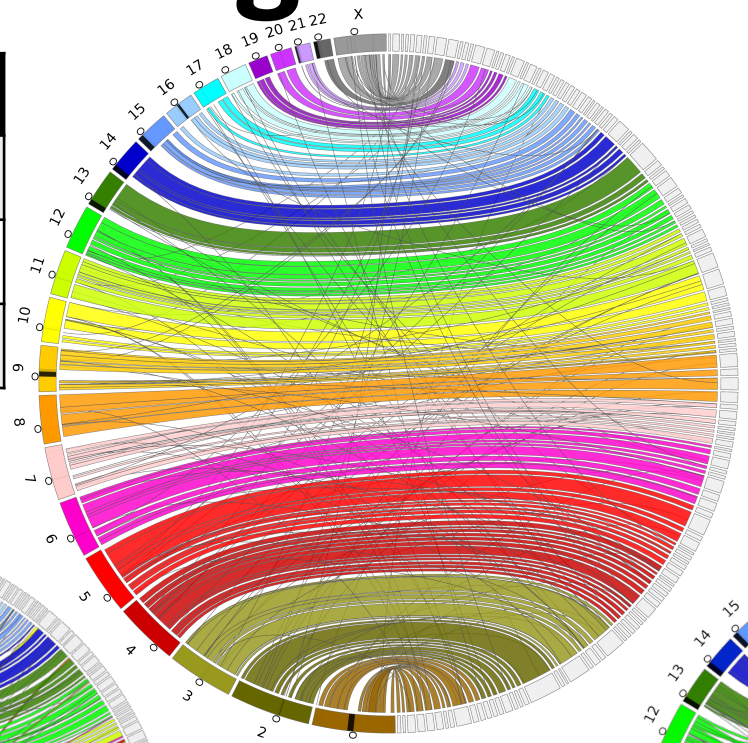
Resources

Tool	Time	Mem
ARCS	**0:55**	**3.4**
Architect	6:07	9.6
fragScaff*	1:59	14.1
*64 threads	h:mm	GB

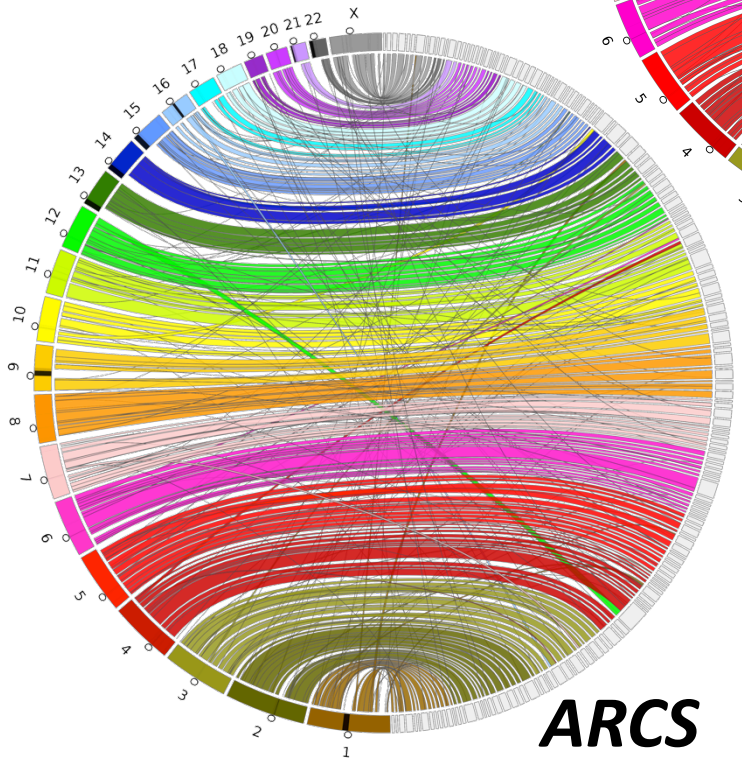
Misassemblies

Assembly	Intra	Inter
Baseline	1090	399
ARCS	1228	**434**
fragScaff	1911	510
Supernova	**1198**	439

Inconsistent contig order

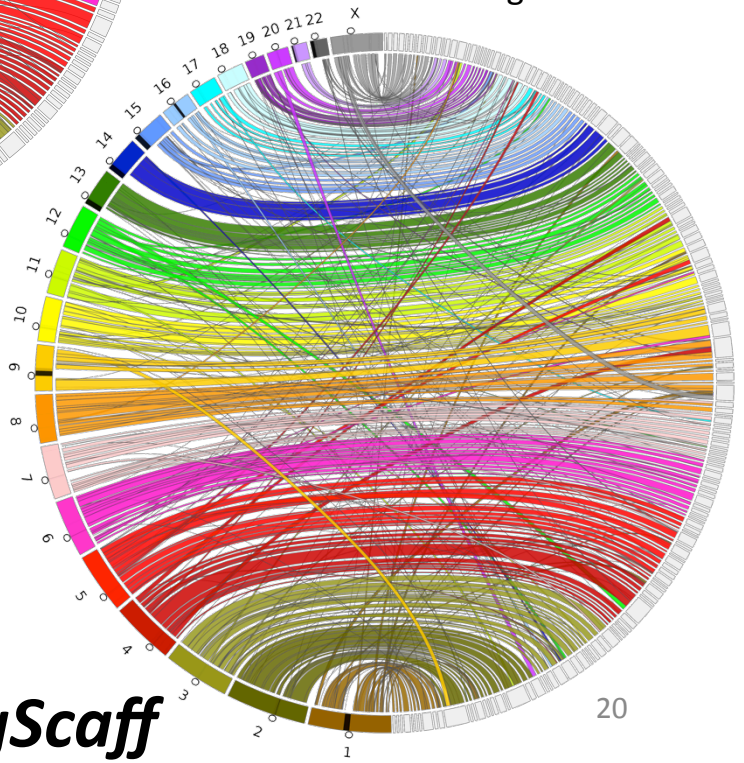


Supernova



ARCS

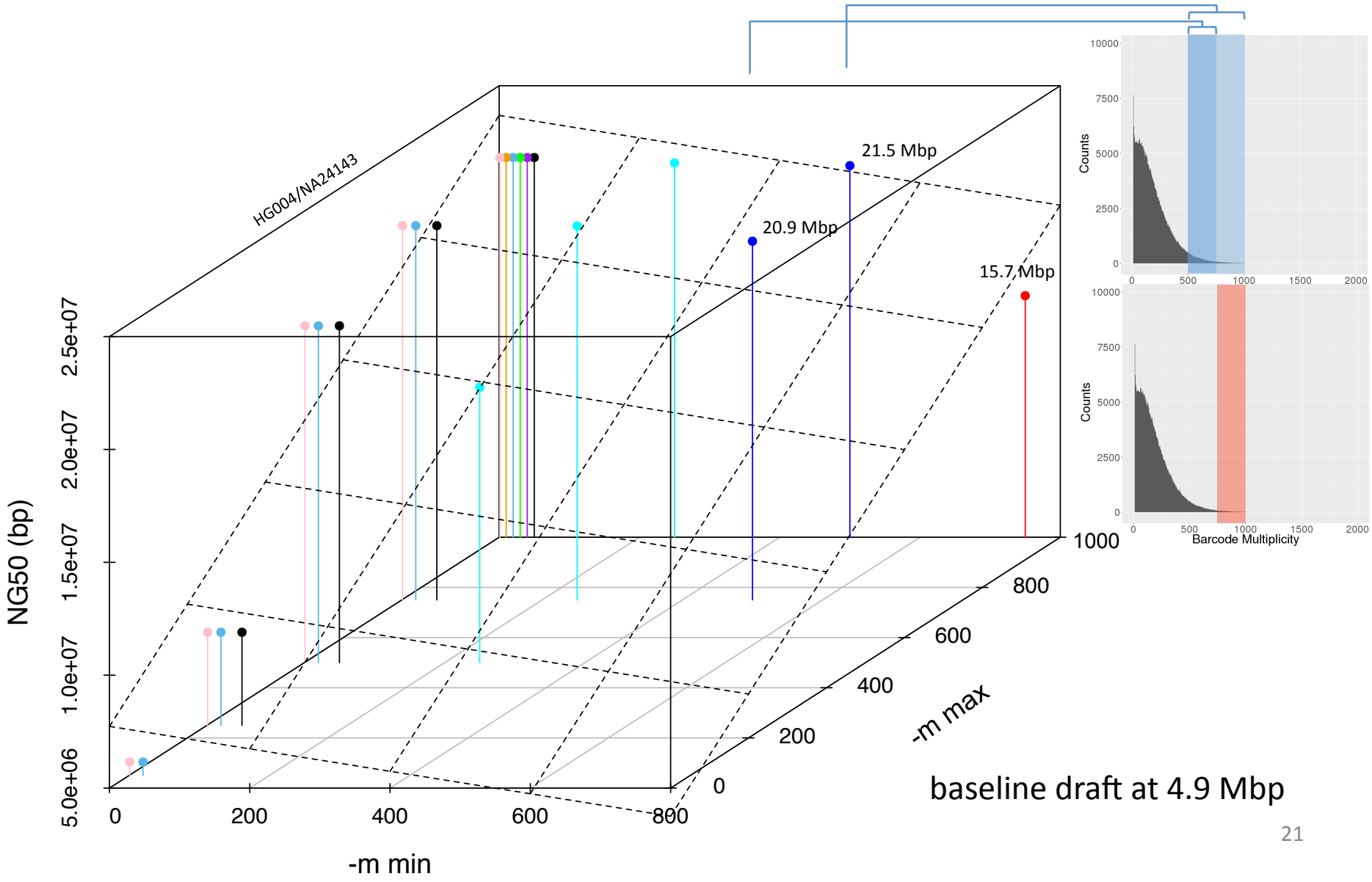
NA24143
1kbp+ GRCh38
alignments



fragScaff

Scaffolding human drafts

Effect of –*m* barcode multiplicity range parameter



baseline draft at 4.9 Mbp

Spruce genome scaffolding

- 8 lanes Chromium ~50-fold
- NG50 : 315 kbp (from 87 kbp)
- Recovers 9 additional CEGs (CEGMA and BUSCO)