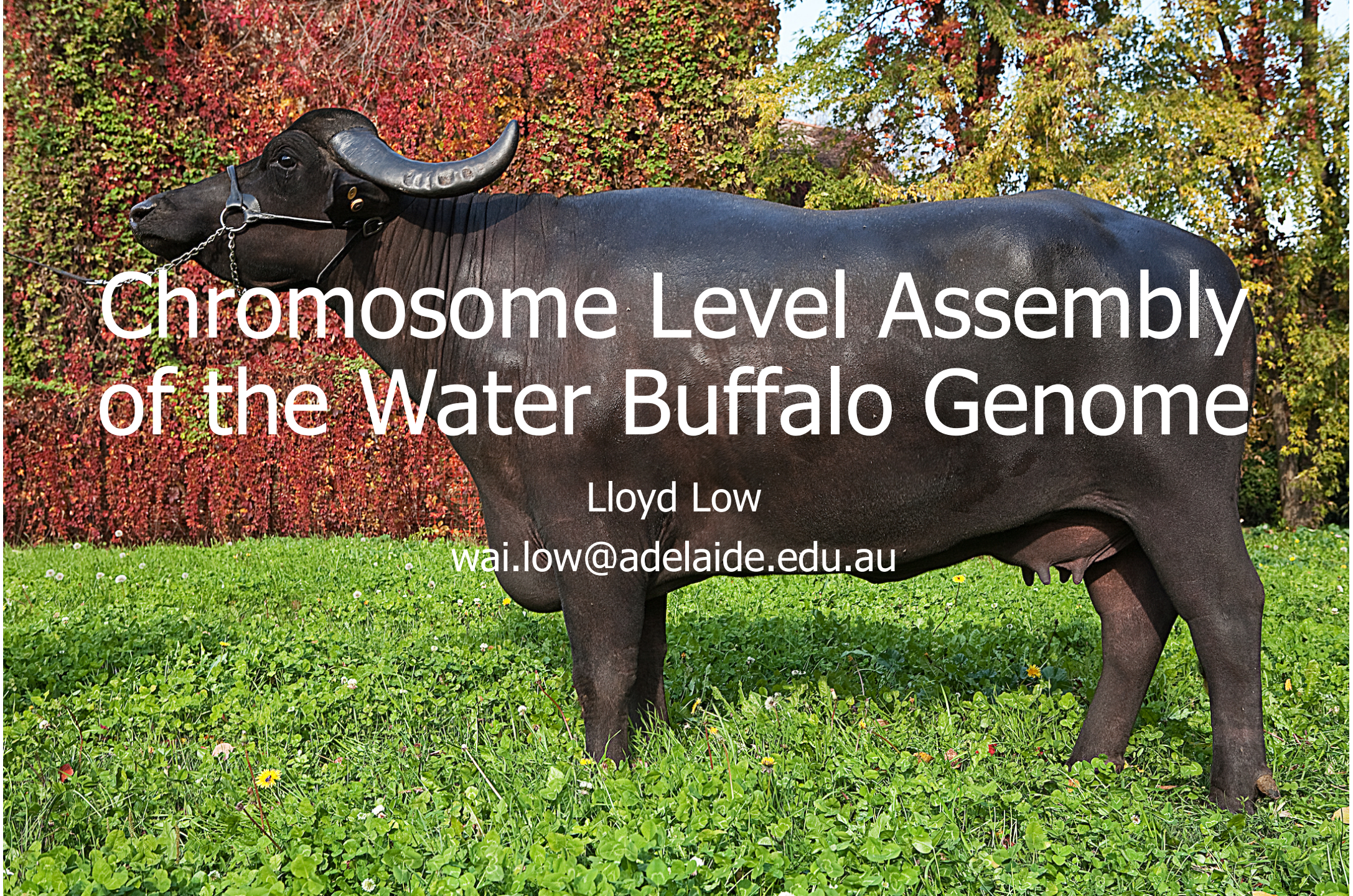
Davies Research Centre

THE UNIVERSITY of ADELAIDE

Davies Research Centre
...excellence in ruminant science



Chromosome Level Assembly of the Water Buffalo Genome

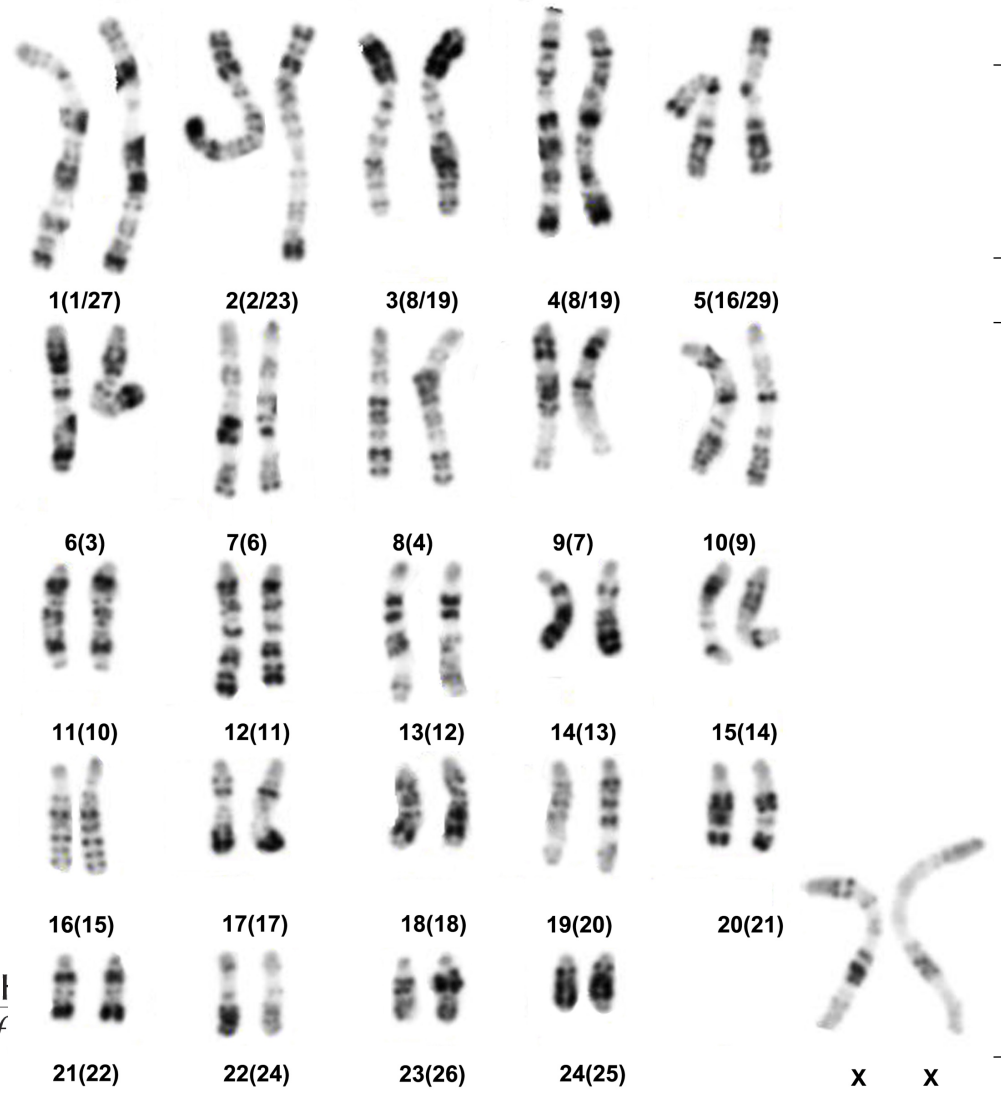
Lloyd Low

wai.low@adelaide.edu.au

Reference buffalo genome – why?

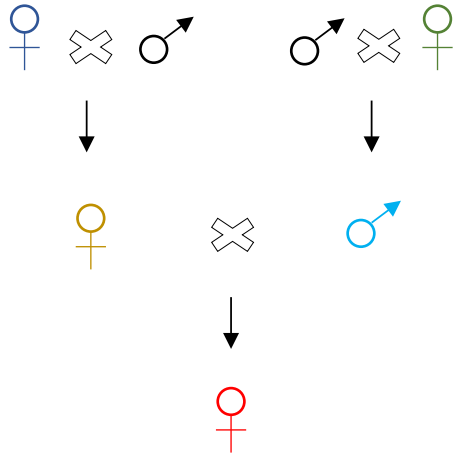
- Apply genome-based selection method for genetic improvement
  - Dairy
  - Meat
  - Draught animal
- Manage genetic diversity
- Uncover interesting biology of the species
- Comparative genomics (mammal and ~96 MYA divergence with human)

Karyotype and pedigree

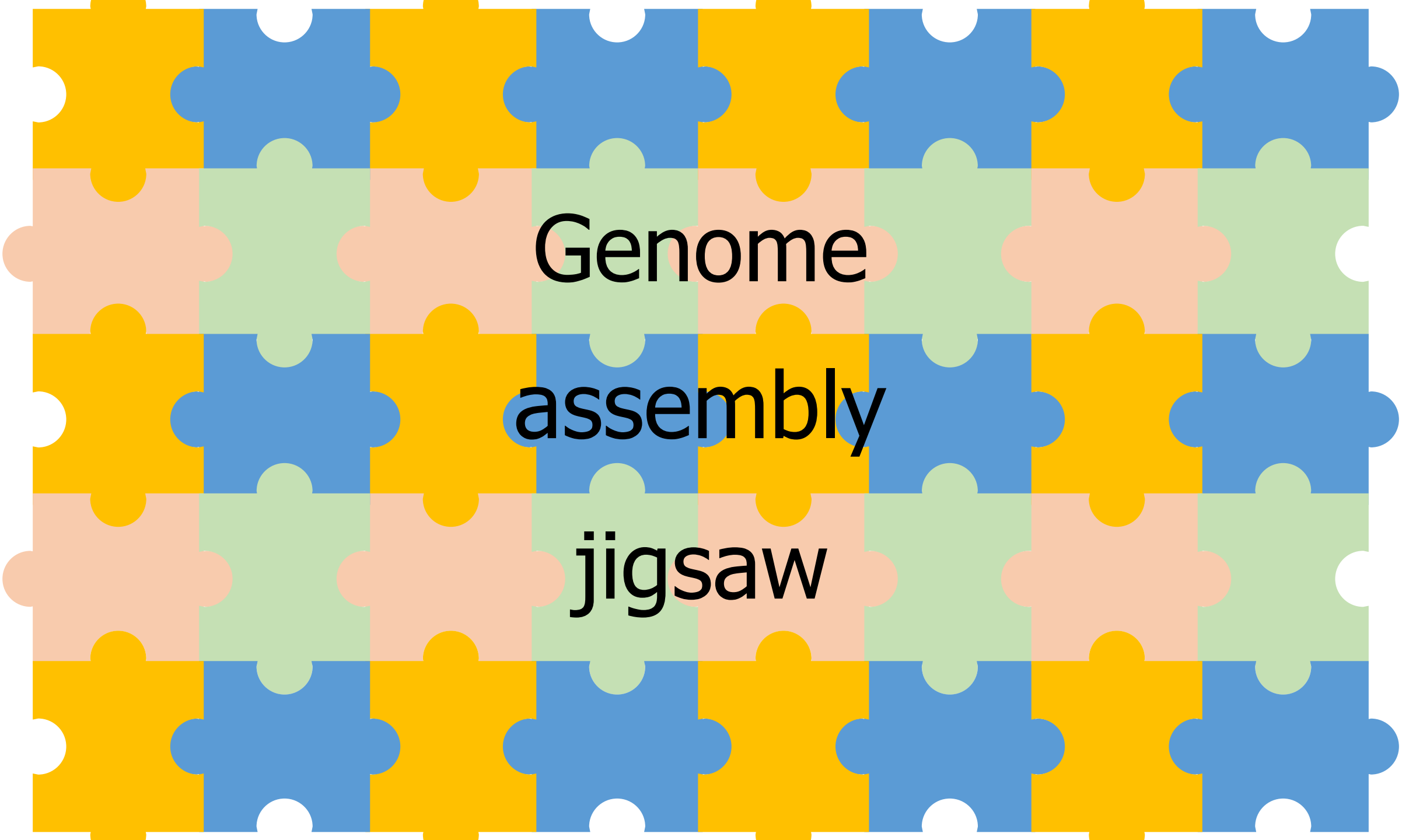


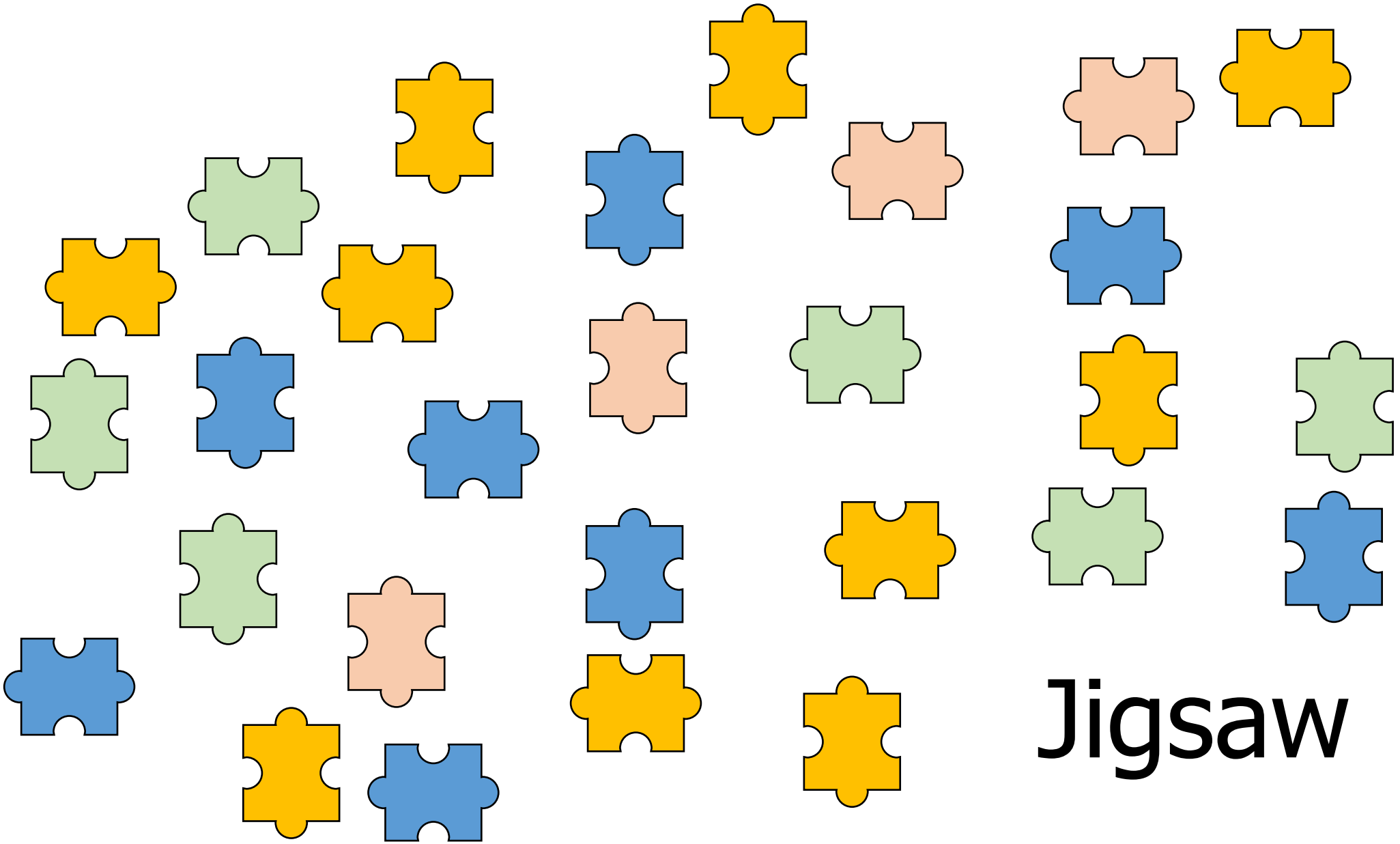
submetacentric

acrocentric

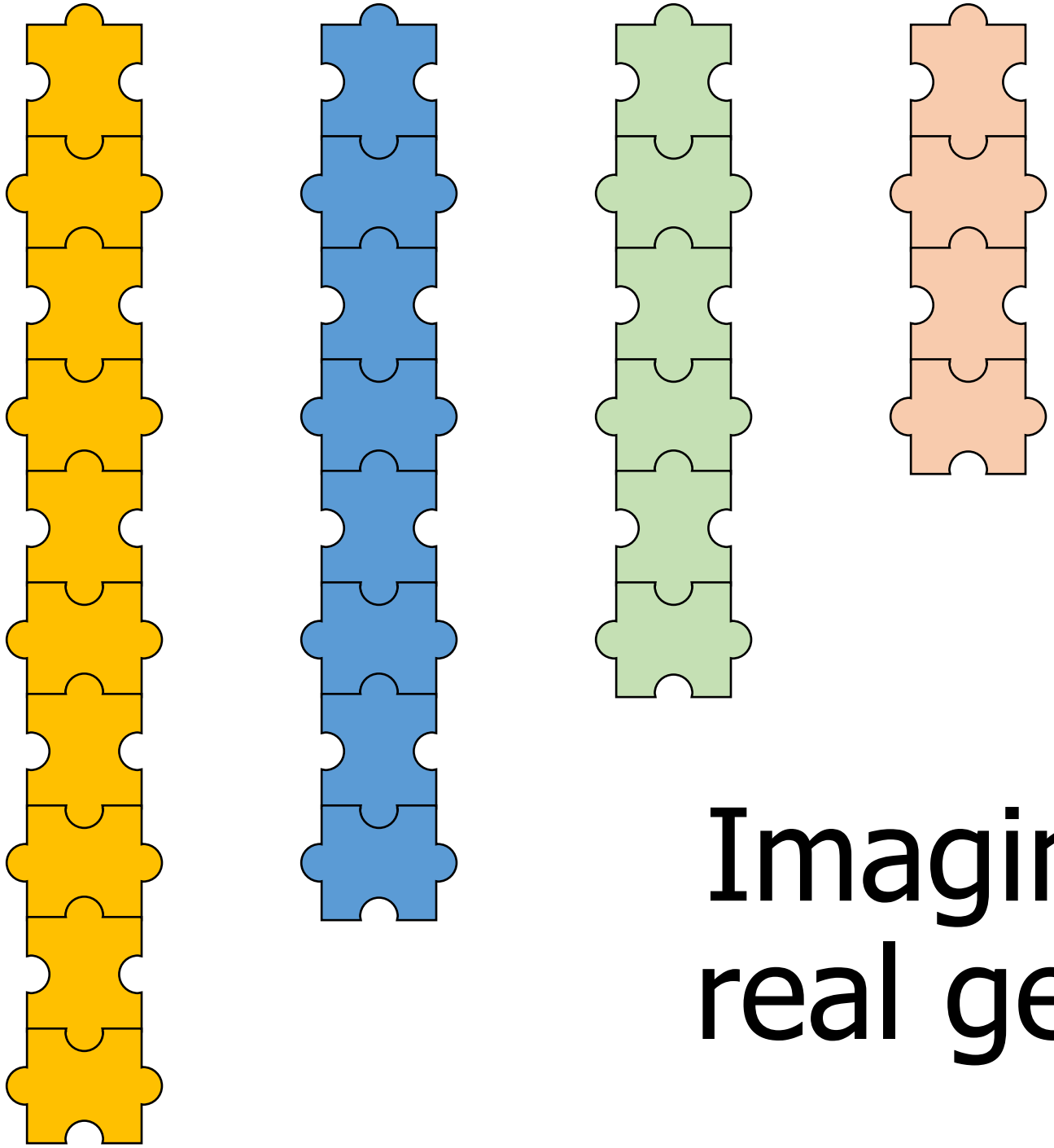


Olimpia

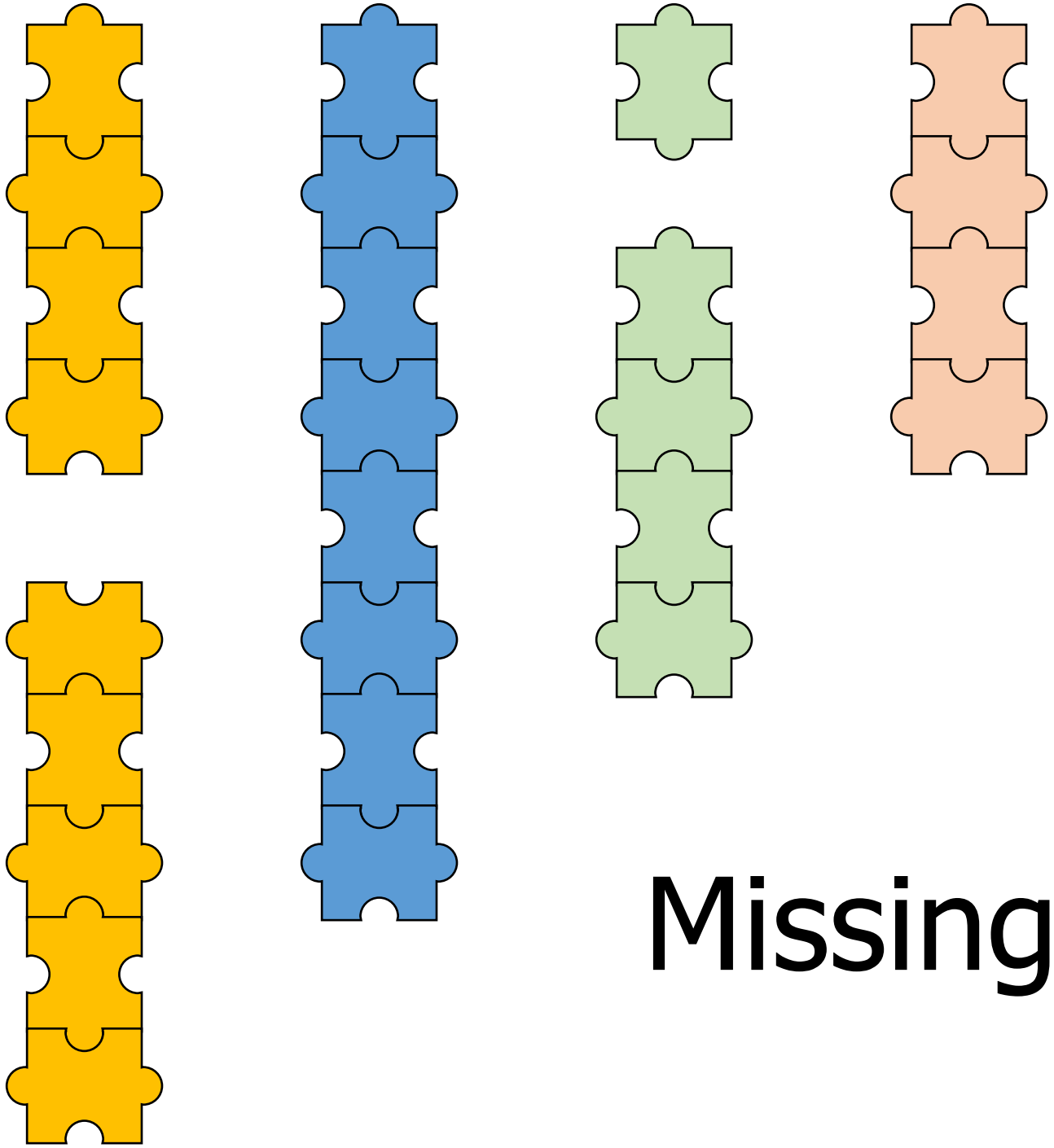

Genome
assembly
jigsaw



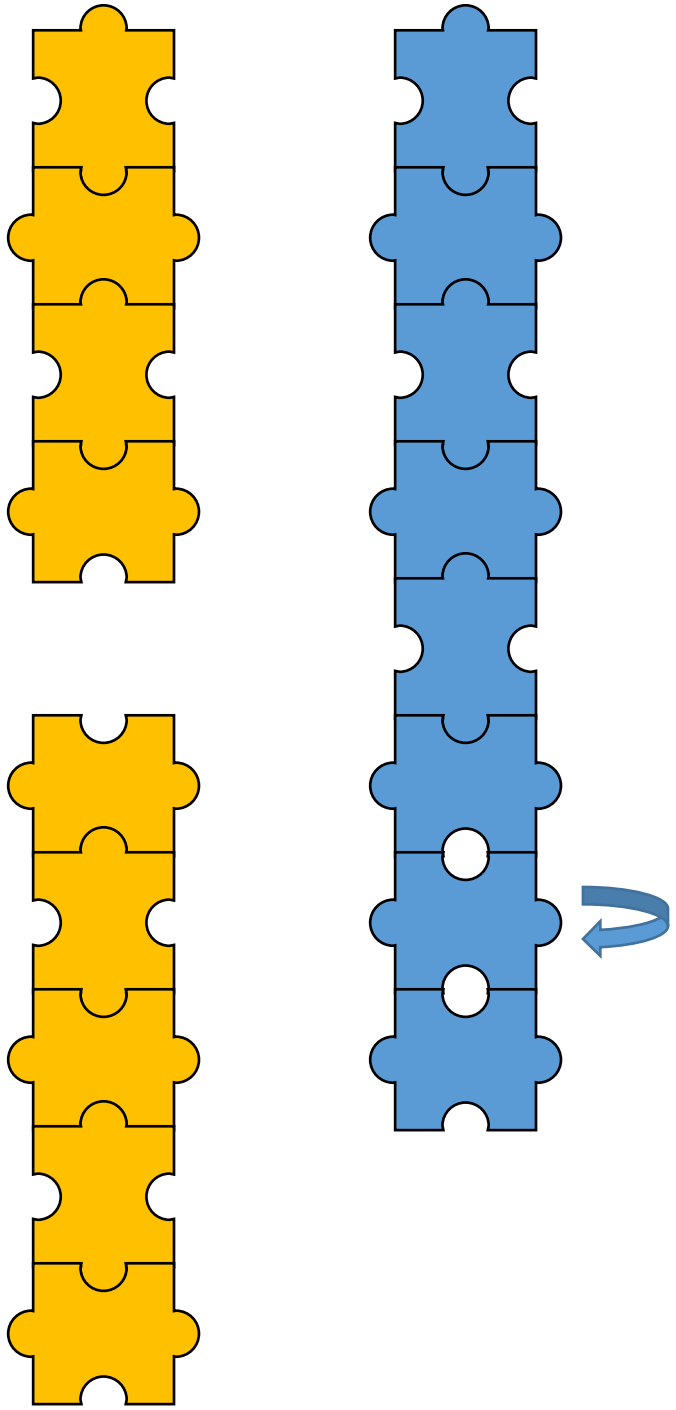
Jigsaw



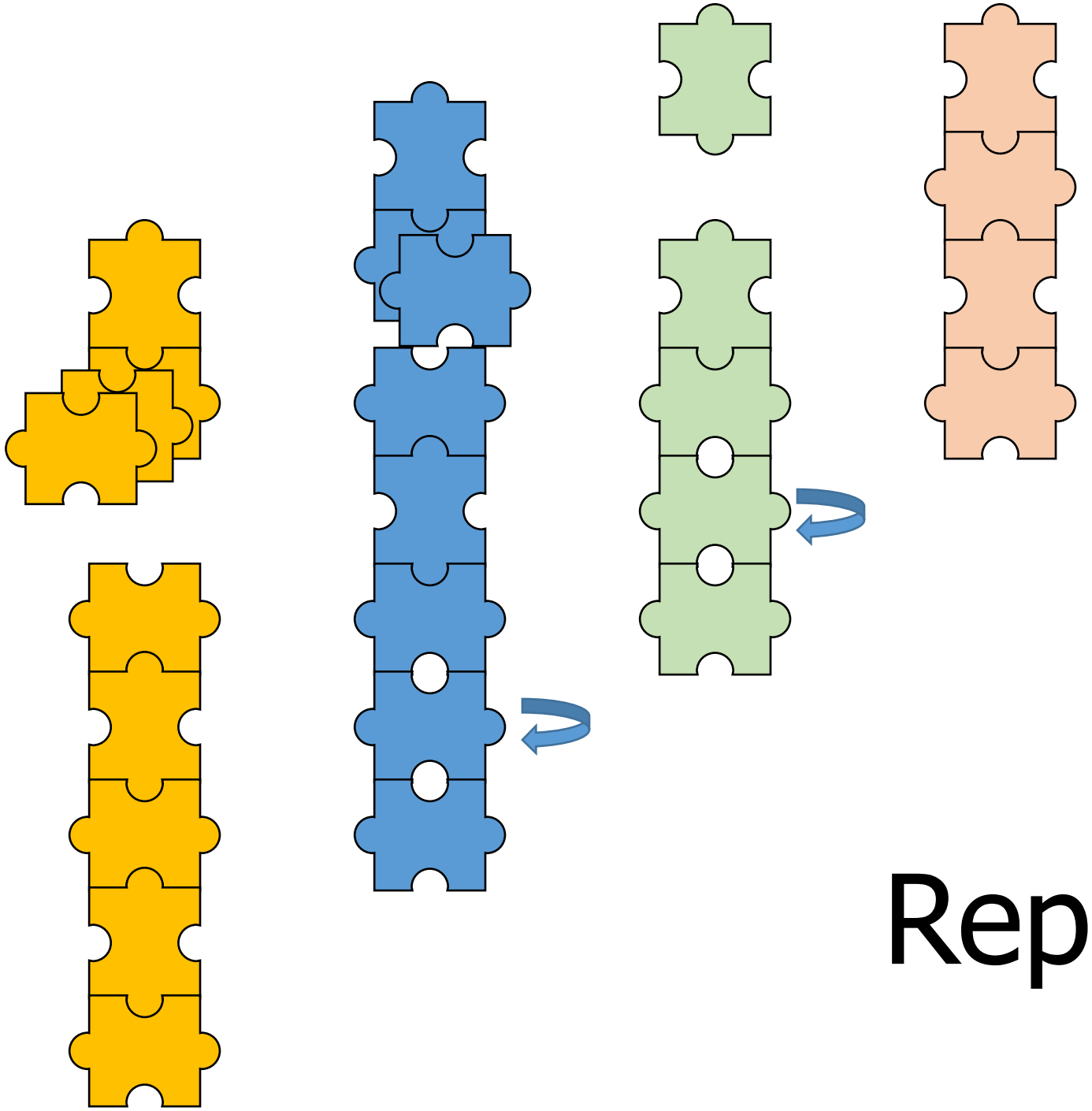
Imagine the real genome



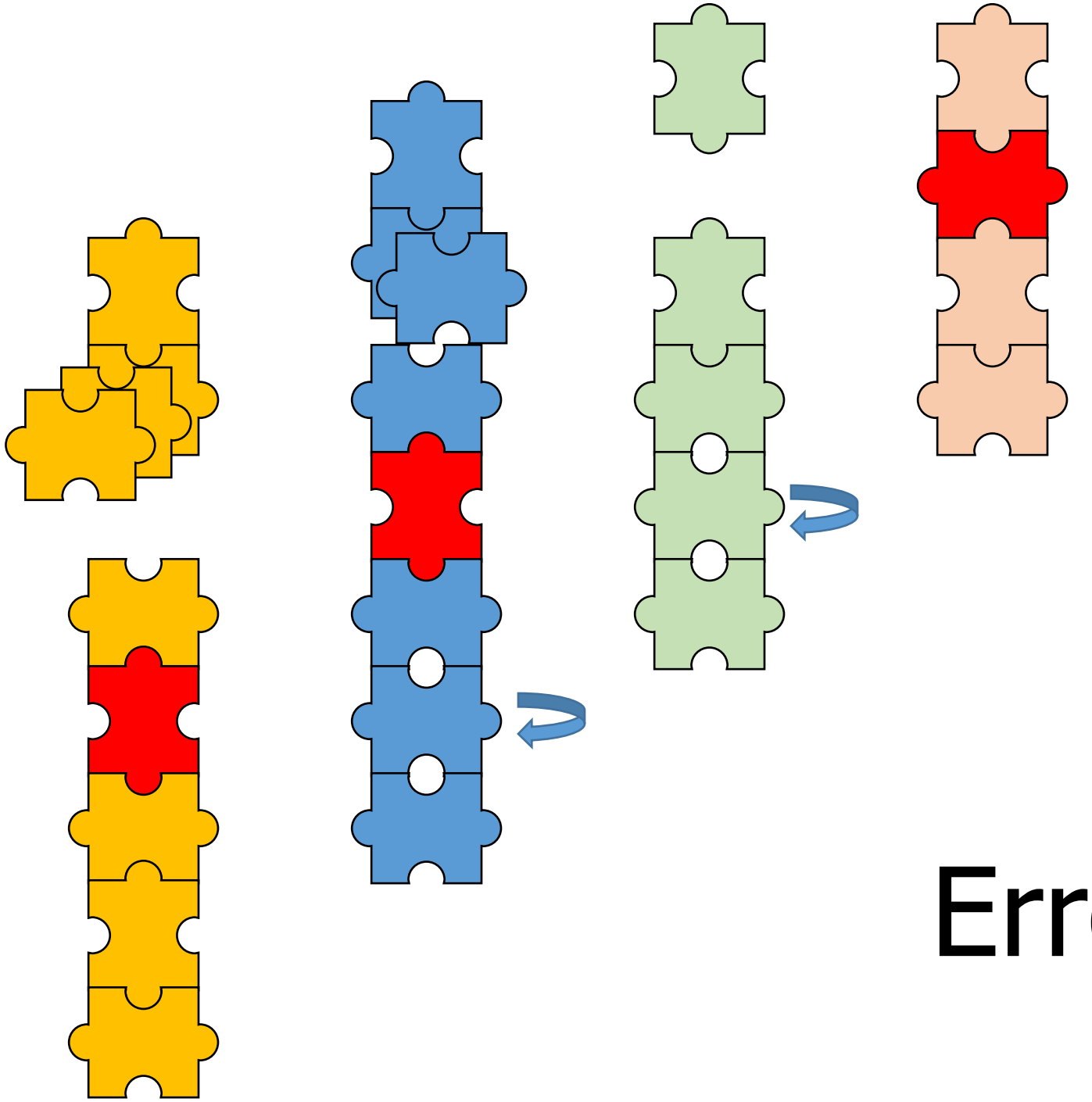
Missing pieces



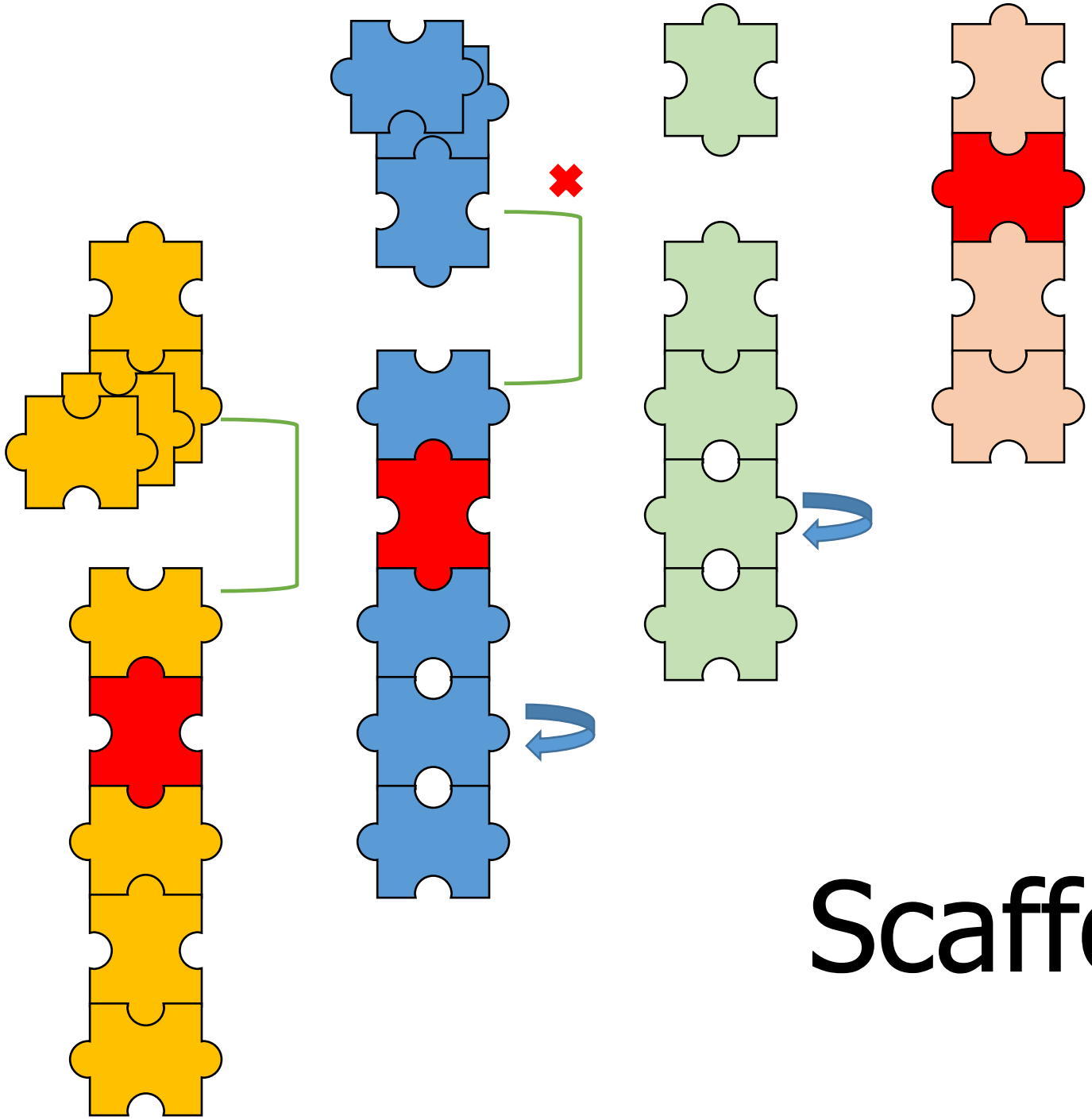
Wrong orientation



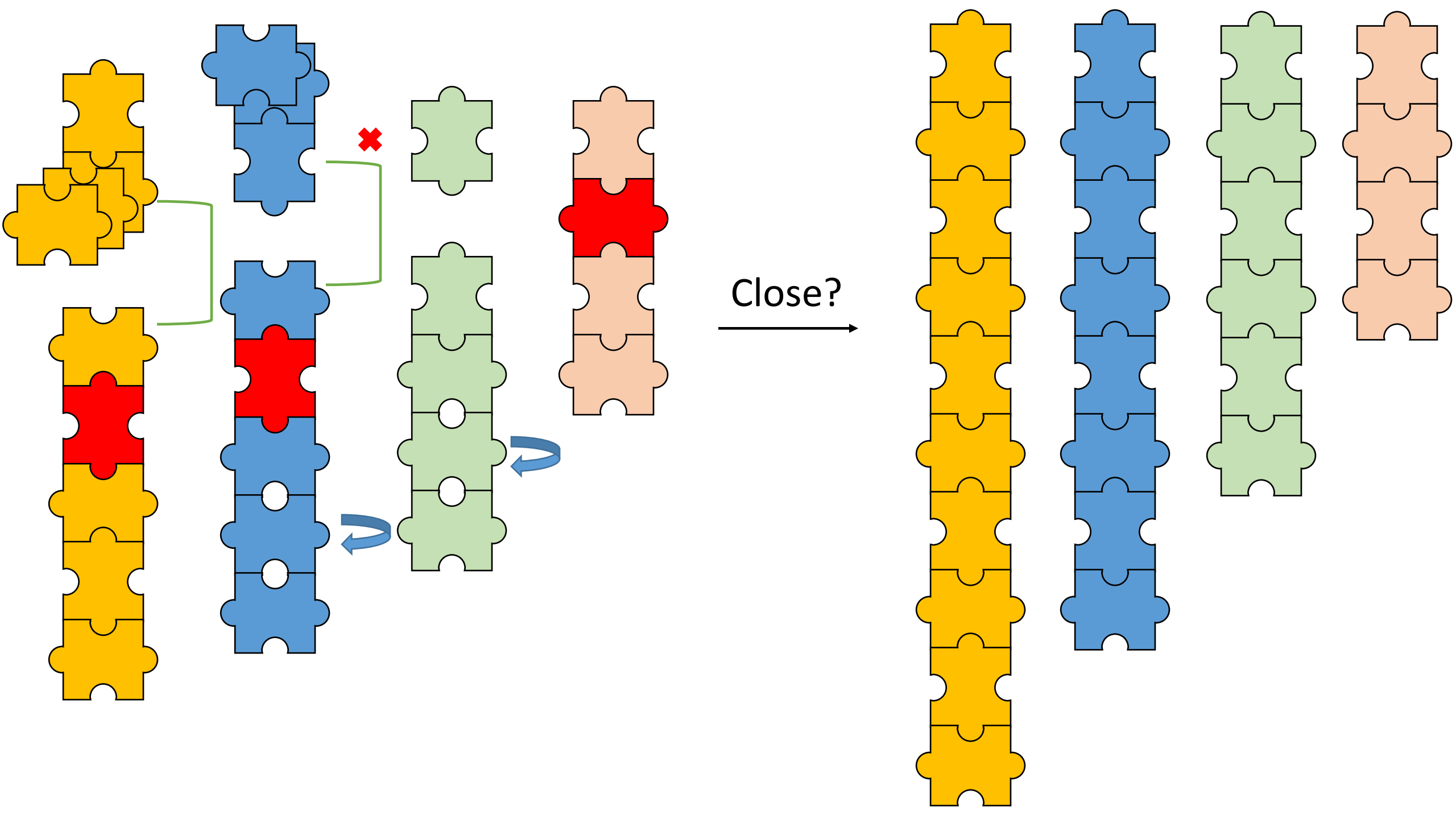
Repeats



Errors



Scaffolding


Close?

Published water buffalo assembly

Description	Published assembly
Total sequence length (bp)	2,836,166,969
Total assembly gap length (bp)	74,388,041
Number of contigs	630,368
Contig N50 (bp)	21,938
Contig L50	35,881
Number of scaffolds	366,983
Scaffold N50 (bp)	1,412,388
Scaffold L50	581

Genome assembly and transcriptome resource for river buffalo, *Bubalus bubalis* (2*n* = 50)

John L Williams, Daniela Iamartino, Kim D Pruitt, Tad Sonstegard, Timothy P L Smith, Wai Yee Low, Tommaso Biagini, Lorenzo Bomba, Stefano Capomaccio, Bianca Castiglioni ... Show more

GigaScience, Volume 6, Issue 10, 1 October 2017, Pages 1–6, https://doi.org/10.1093/gigascience/gix088

Published: 01 September 2017 **Article history**

THE UNIVERSITY of ADELAIDE

Davies Research Centre
...excellence in ruminant science

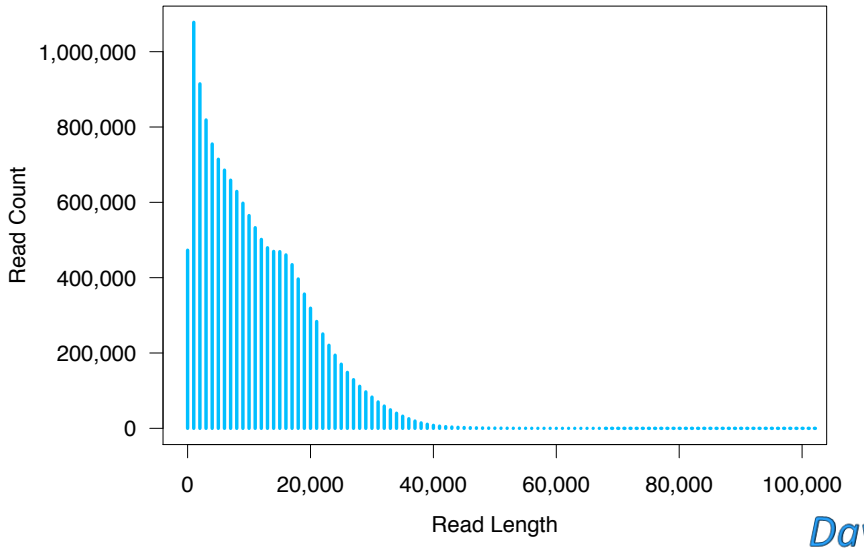
Sequenced data

- PacBio (~69X)
- Chicago (~22X)
- HiC (~53X)
- Illumina PE (~80X)

PacBio sequences

SEQUENCING DATA	
Libraries	7
Sequel Cells	57
RS II Cells	8
Sequel Yield	191 Gb
RSII Yield	8.0 Gb
Total Yield	**199 Gb**

	Raw Reads	Raw Bases	Mean Read L	Read N50
Sequel Data	14,350,446	164 Gb	11.5 kb	17 kb
RS II Data	1,421,854	8 Gb	5.8 kb	16 kb
All Data	14,870,495	171 Gb	11.5 kb	17 kb



Acknowledgements:
Tim Smith, USDA-ARS
Sarah Kingan,
Pacific Biosciences

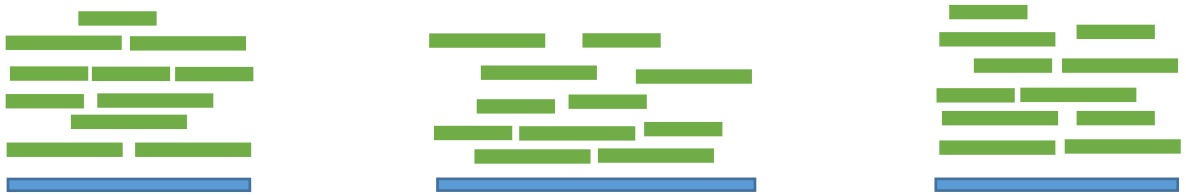
THE UNIVERSITY of ADELAIDE

Davies Research Centre
...excellence in ruminant science

Assembly of contigs

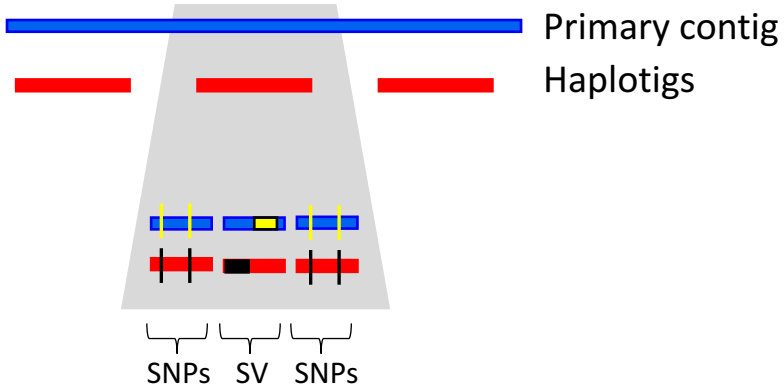
Raw PacBio reads

Create pre-assembled reads

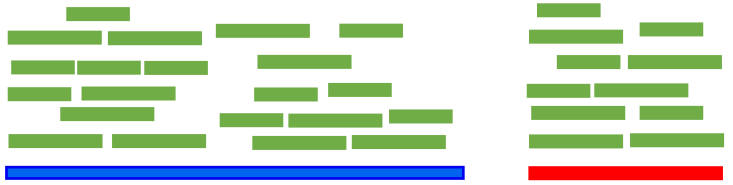


Assemble pre-assembled reads

Falcon-Unzip haplotype resolved assembly



Polish with BLASR/Arrow



Chin et al (2016)

THE UNIVERSITY of ADELAIDE

Davies Research Centre
...excellence in ruminant science

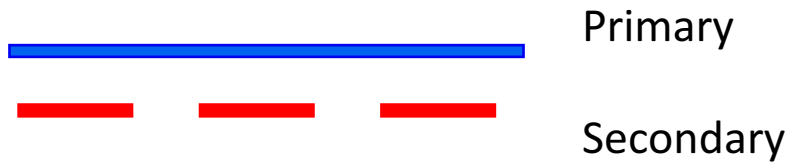
Contigs

FALCON ASSEMBLY

CONTIG TYPE	NUMBER	LENGTH	N50	LONGEST
Primary	1694	2.66 Gb	18.7 Mb	88.5 Mb
Secondary (i.e. Associate)	5205	0.218 Gb	0.044 Mb	0.402 Mb

FALCON-UNZIP ASSEMBLY

CONTIG TYPE	NUMBER	LENGTH	N50	LONGEST
Primary	953	2.65 Gb	18.8 Mb	88.9 Mb
Secondary (i.e. Haplotigs)	7956	1.53 Gb	0.394 Mb	2.77 Mb



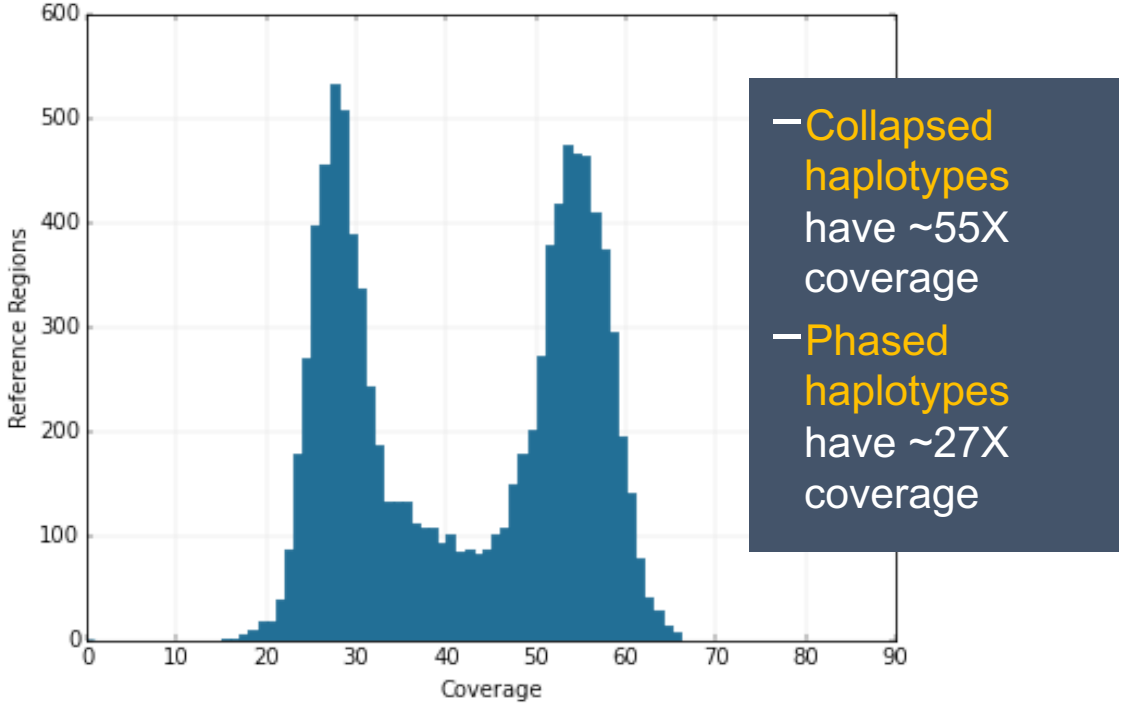
Comparison of FALCON and FALCON-UNZIP

CONTIG TYPE	FALCON	FALCON-UNZIP
Primary length	2.66 Gb	2.65 Gb
Primary N50	18.7 Mb	18.8 Mb
Secondary length	0.218 Gb	1.53 Gb
Proportion phased	8.2%	58%

Polishing - contigs

- ROUND ONE
  - polishing with phased reads within Unzip Module
- ROUND TWO
  - polishing with all reads mapped to combined reference (primary contigs plus haplotigs)
  - resequencing pipeline on SMRTlink

RAW READ COVERAGE HISTOGRAM

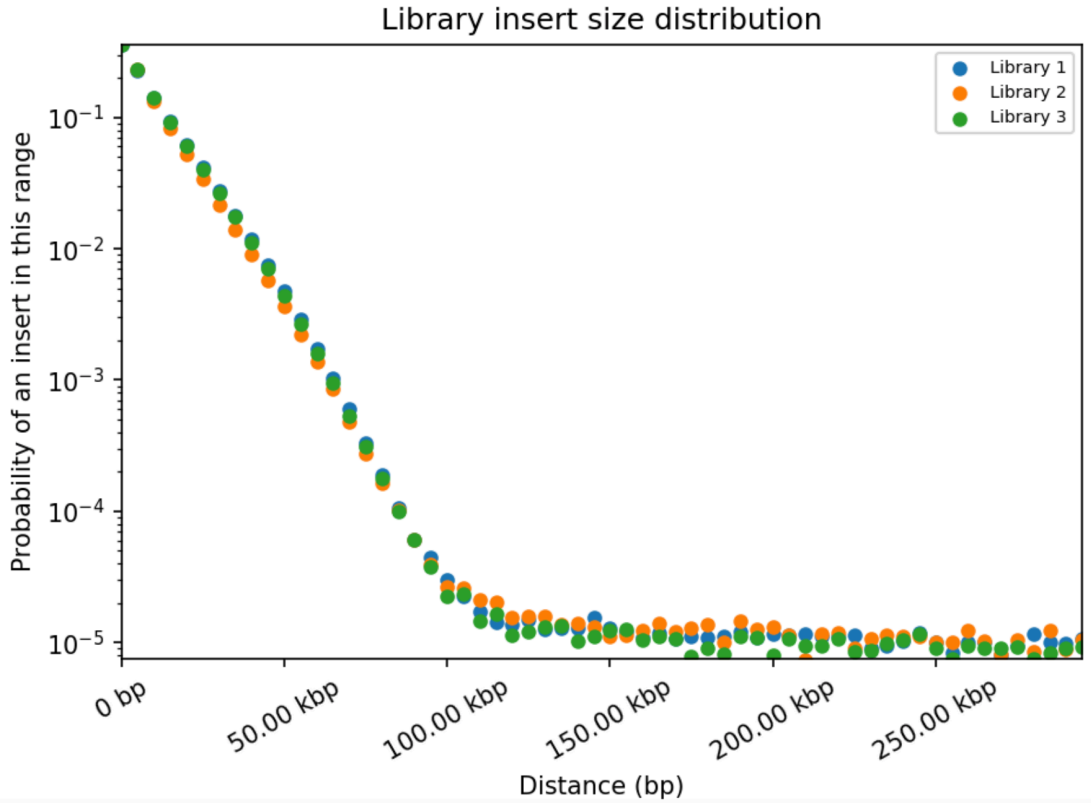


- Collapsed haplotypes have ~55X coverage
- Phased haplotypes have ~27X coverage

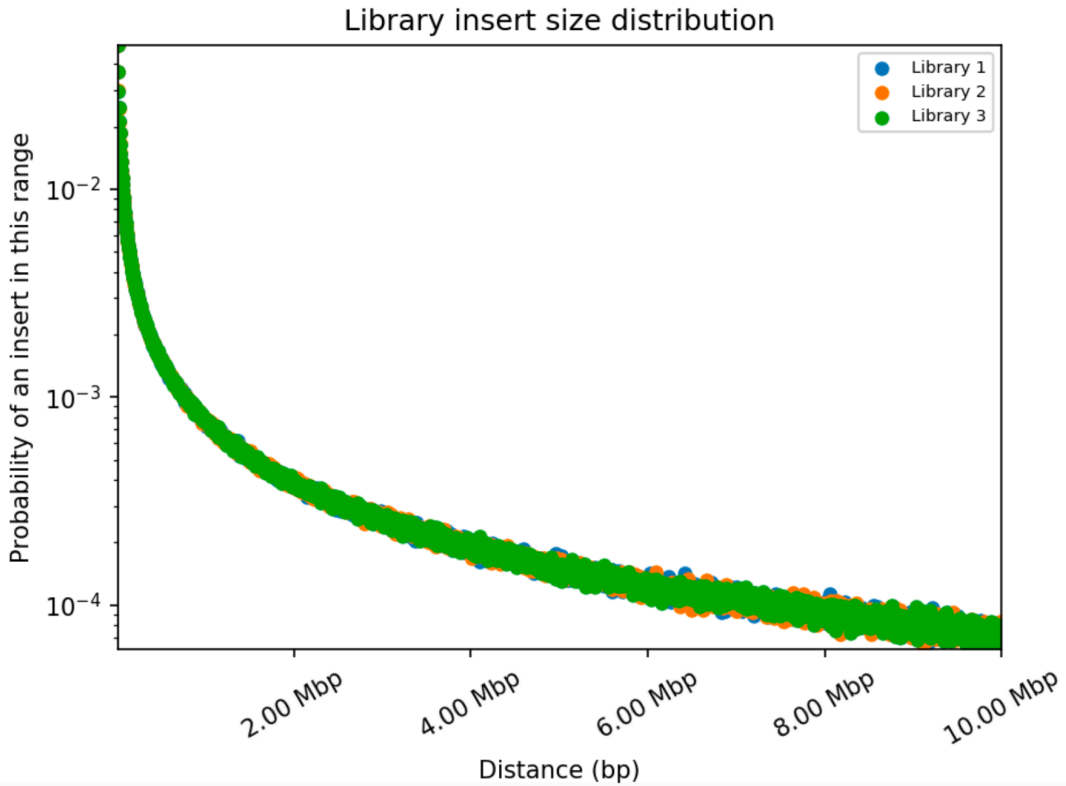
Acknowledgements:
Sarah Kingan,
Pacific Biosciences

Scaffolding

Chicago



HiC



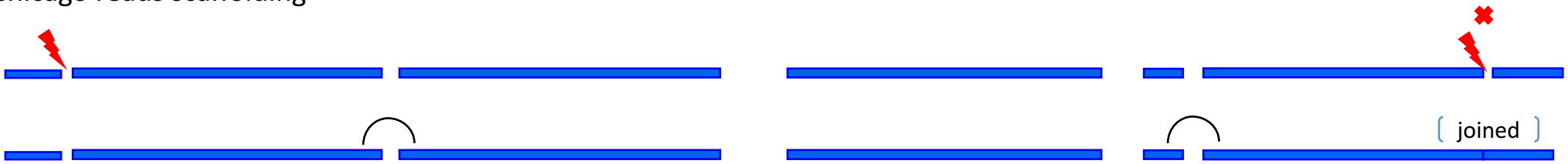
THE UNIVERSITY of ADELAIDE

Acknowledgements:
Thomas Swale, Dovetail Genomics

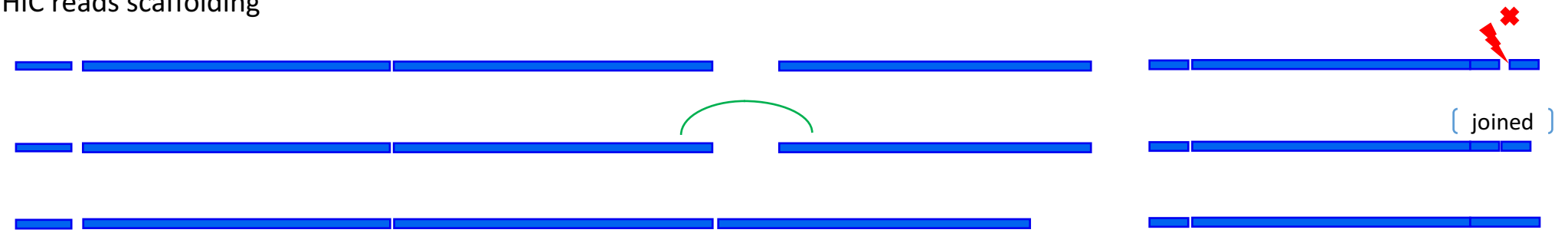
Davies Research Centre
...excellence in ruminant science

Scaffolding

Chicago reads scaffolding

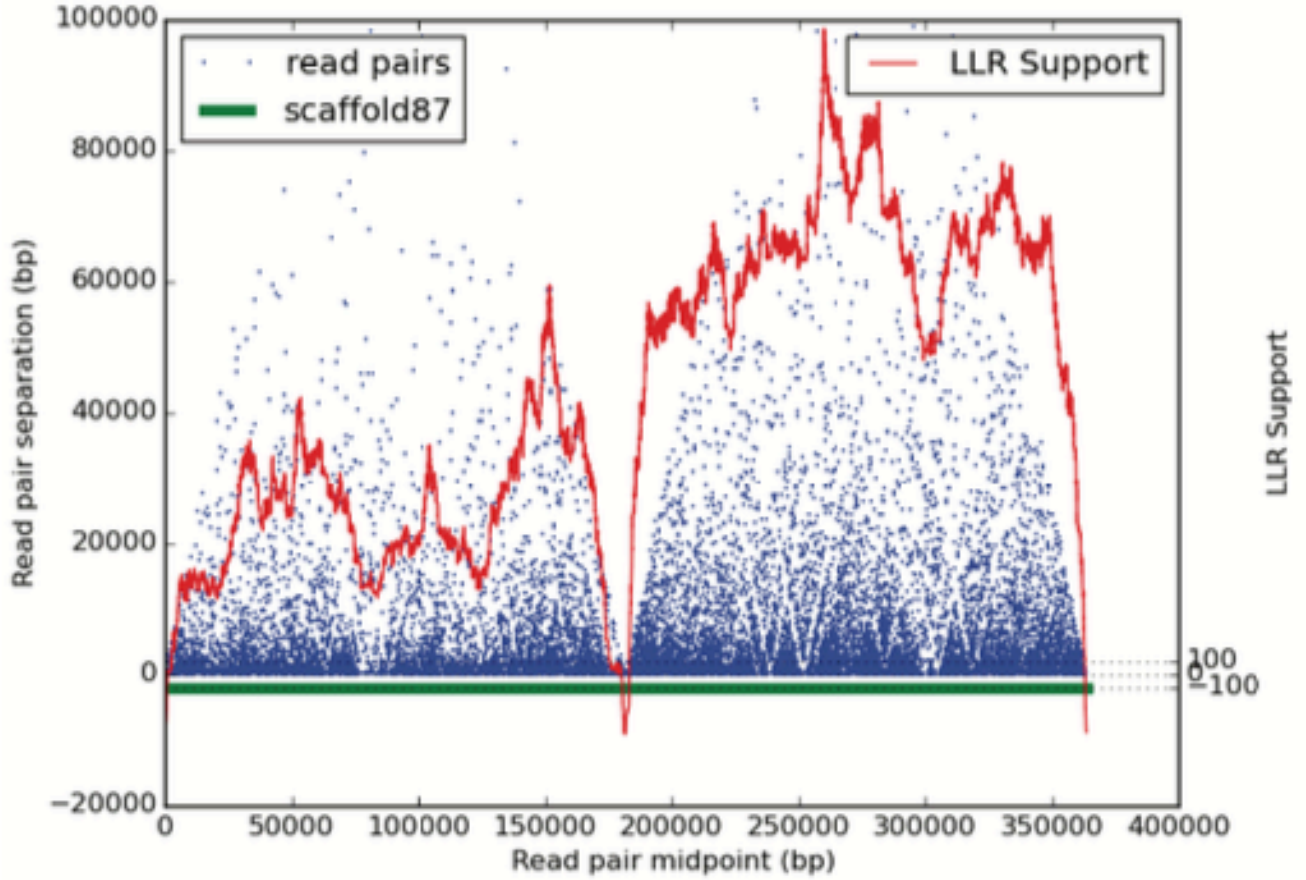


HiC reads scaffolding



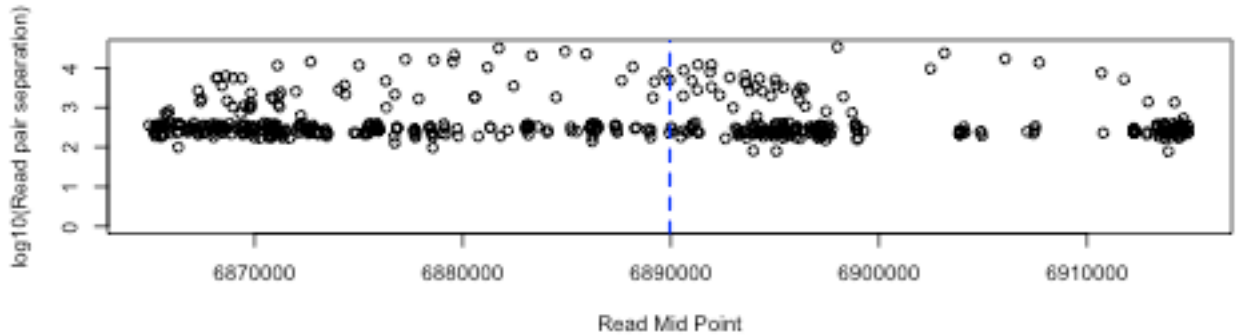
Range of chicago: 1-100 kb
Range of HiC: 10-10,000 kb

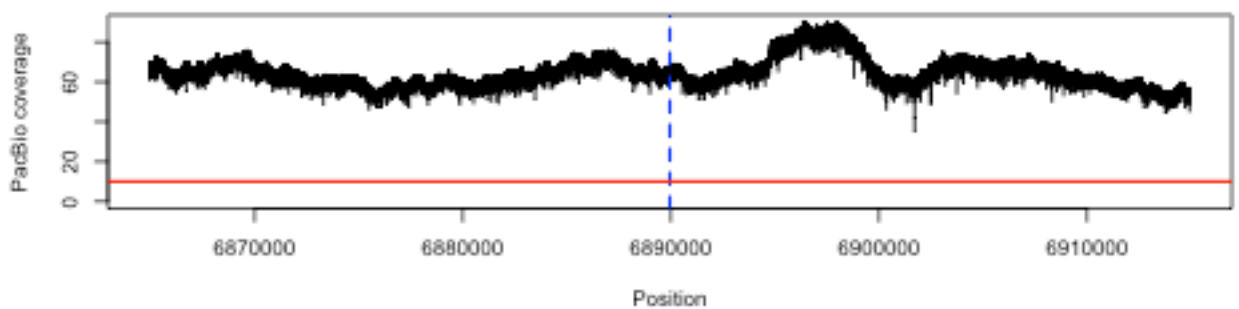
HiRise breaks

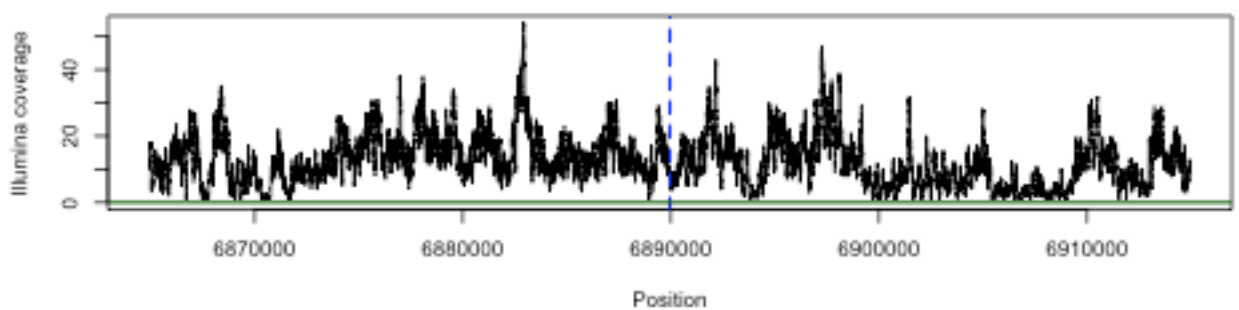


Putnam et al 2016 Genome Research

Case 1: PacBio coverage seems good but a breakpoint is given by HiRise



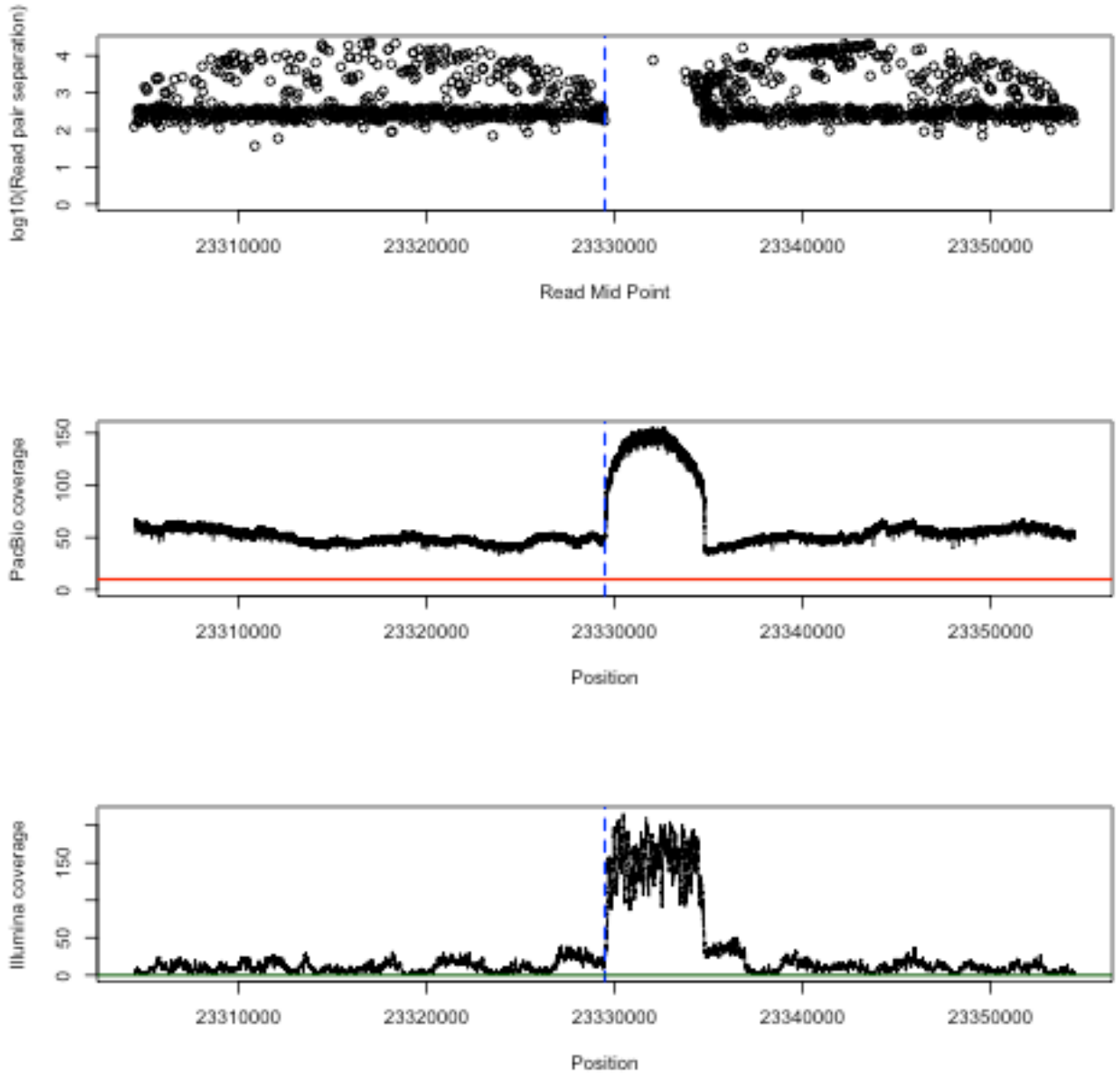




THE UNIVERSITY of ADELAIDE

Davies Research Centre
…excellence in ruminant science

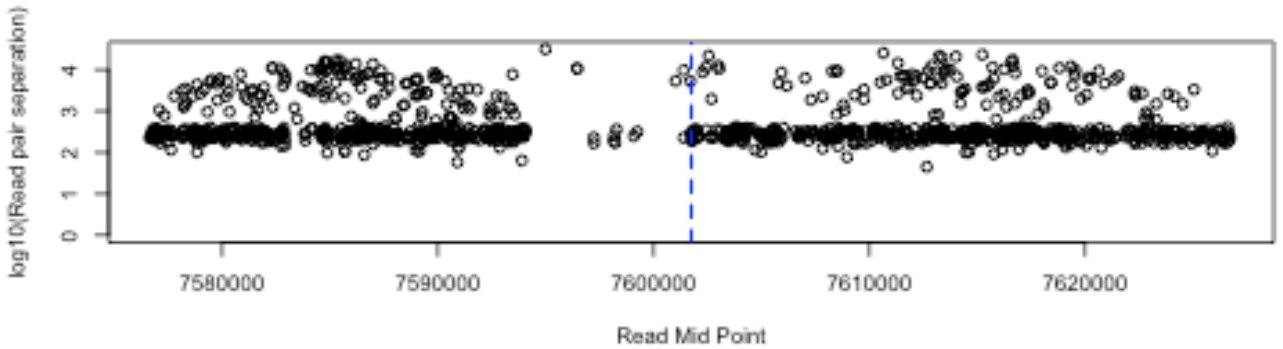
Case 2: Breakpoint in unusually high coverage region by PacBio reads

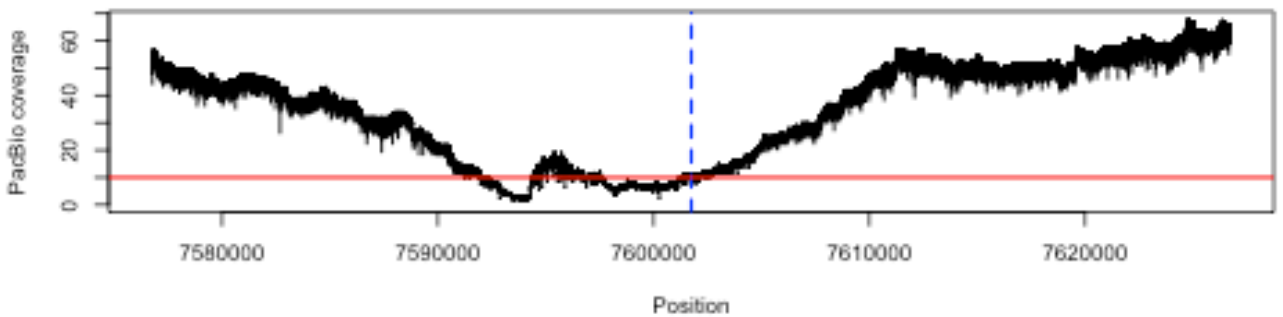


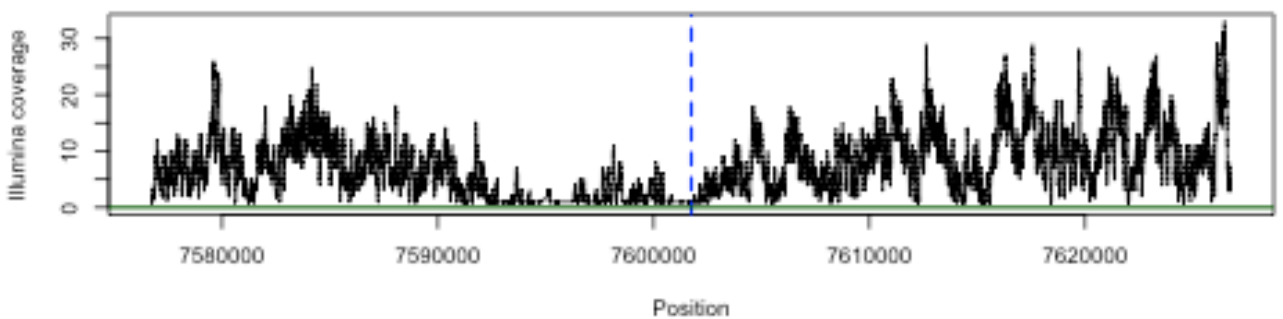
THE UNIVERSITY of ADELAIDE

Davies Research Centre
...excellence in ruminant science

Case 3: HiRise breaks at low PacBio coverage region





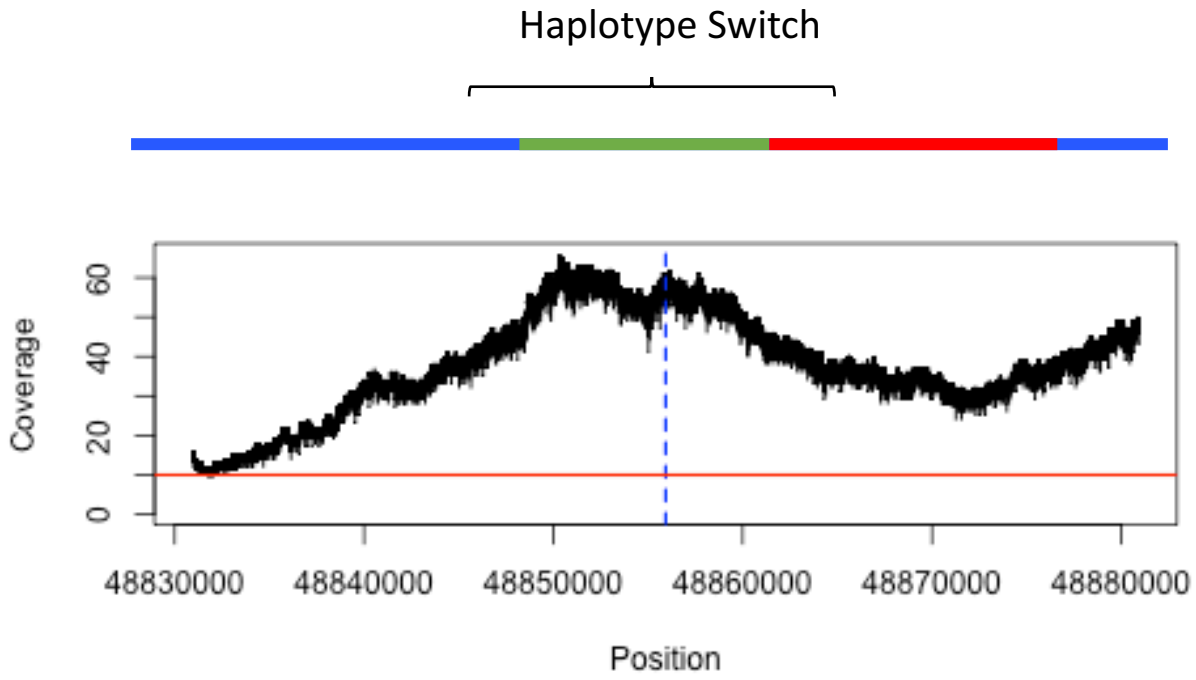


THE UNIVERSITY of ADELAIDE

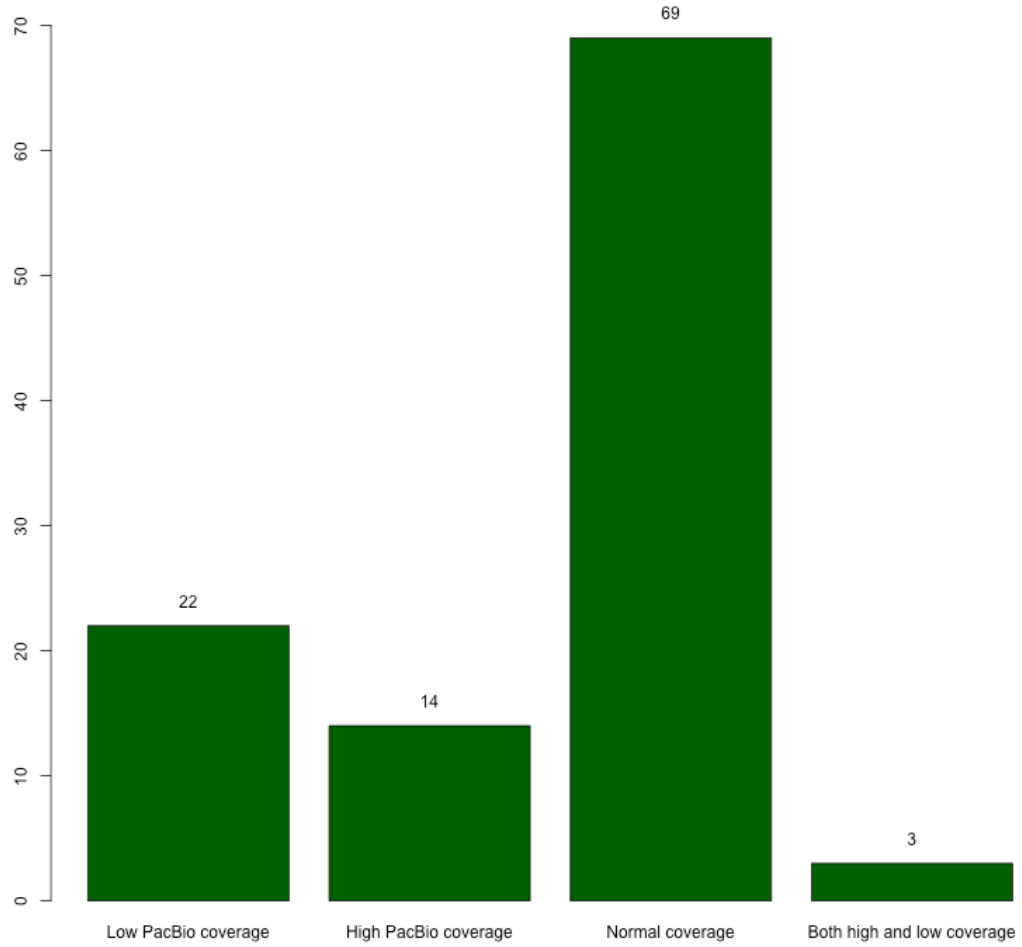
Davies Research Centre
...excellence in ruminant science

Issues with false breaks

- Total 108 HiRise breaks on the primary contigs
- One explanation is haplotype switches
- Paired-end read mapping has no alternate haplotypes as targets

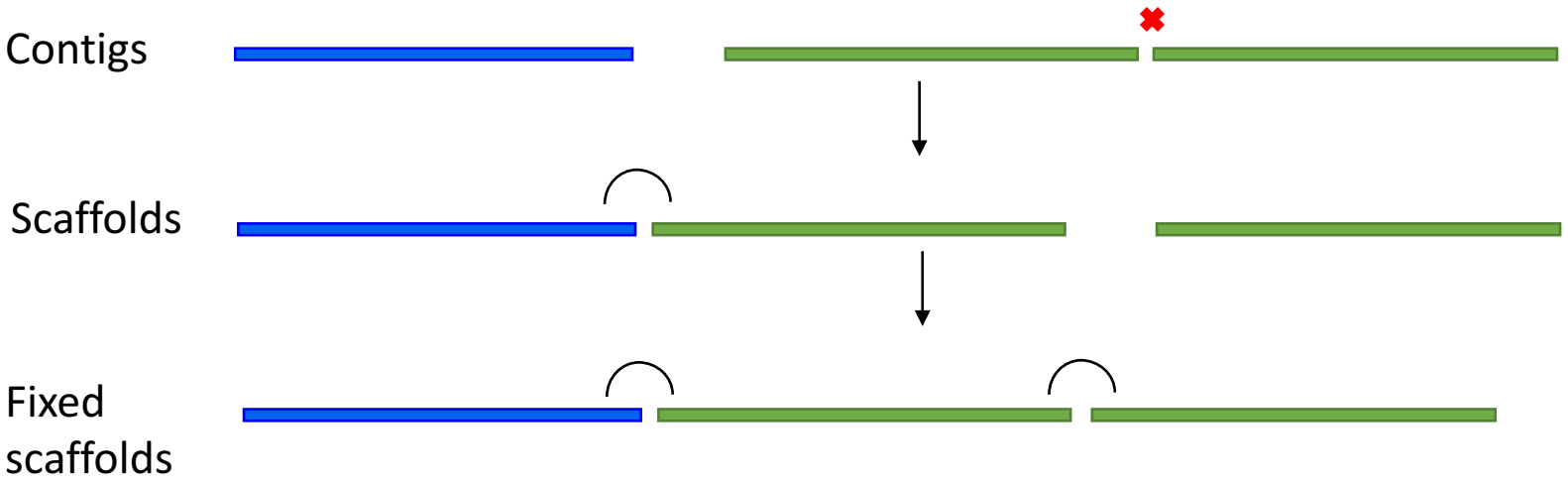


Issues with false breaks



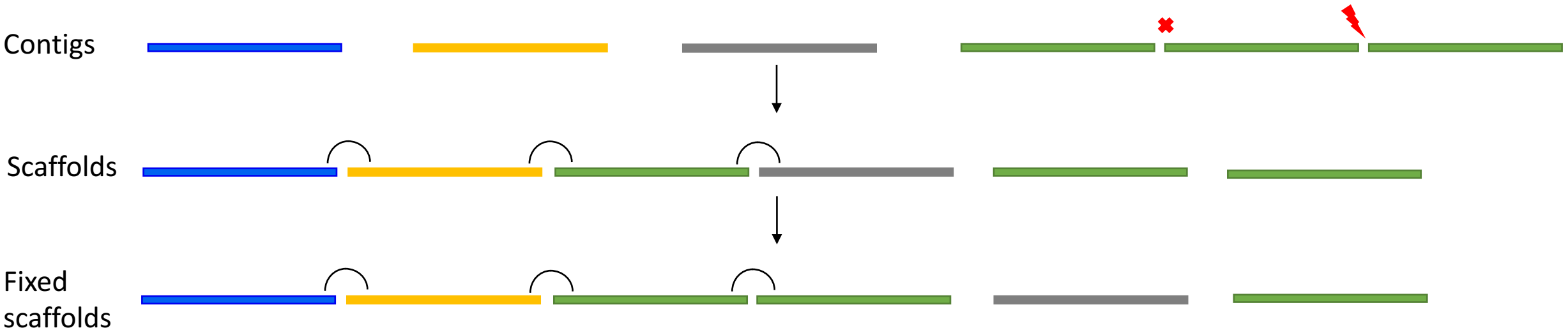
Remove false breaks

Simple scenario



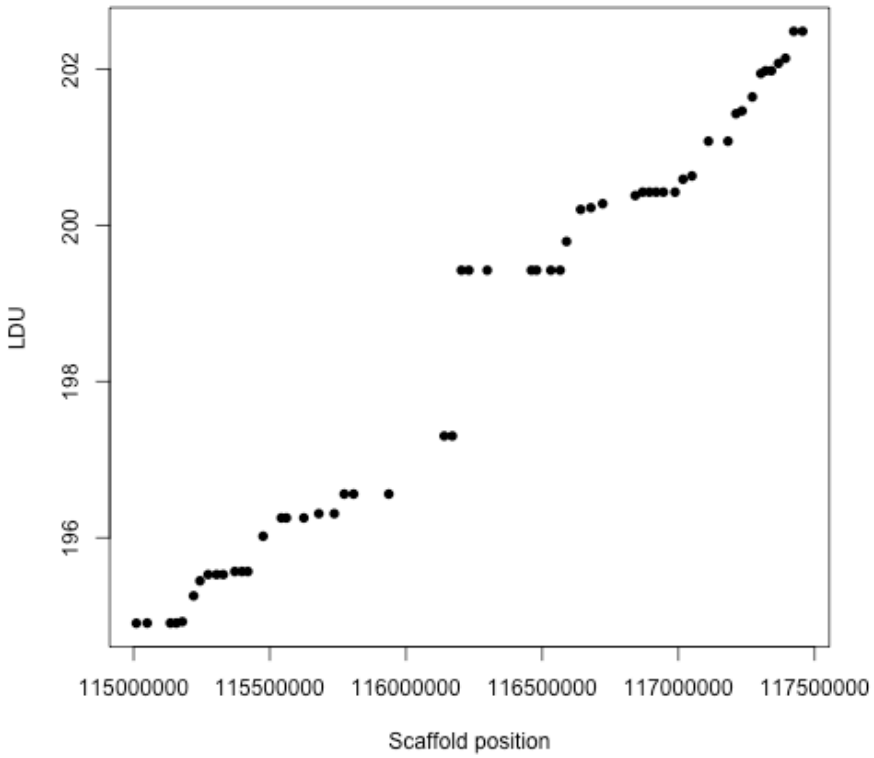
Remove false breaks

Slightly more complicated scenario

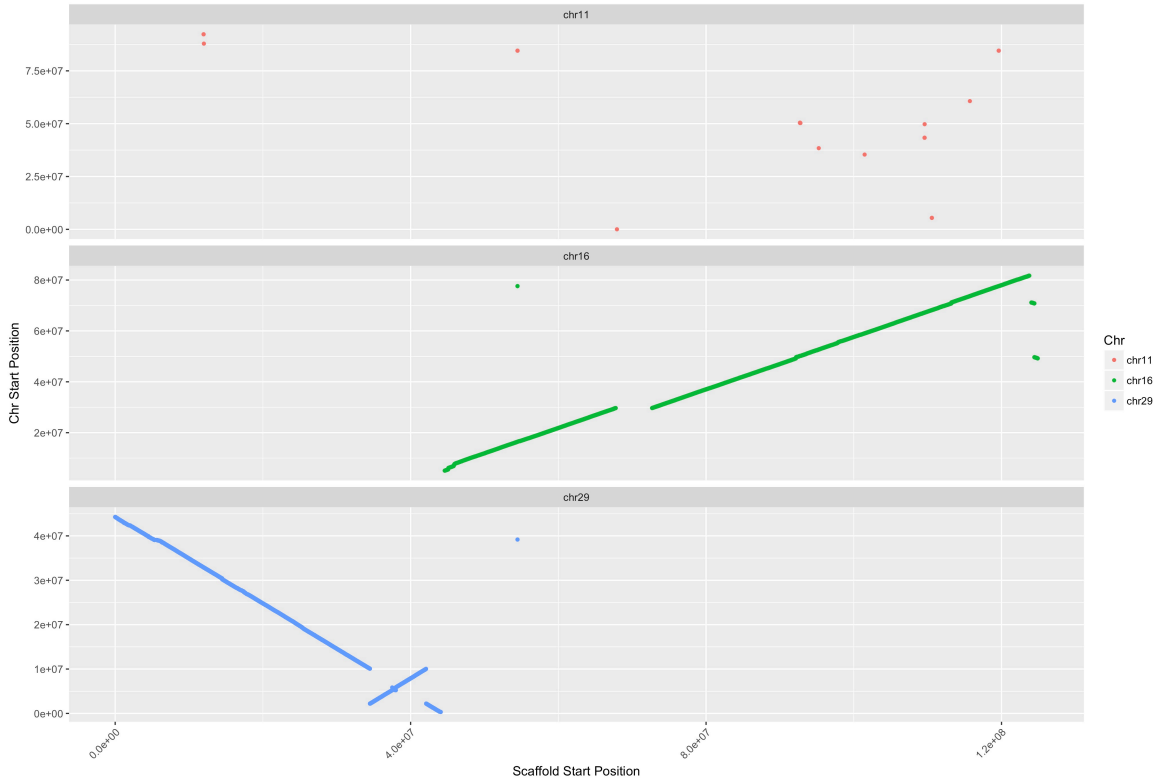


Scaffold conflict resolution

Linkage disequilibrium map



Conservation of synteny

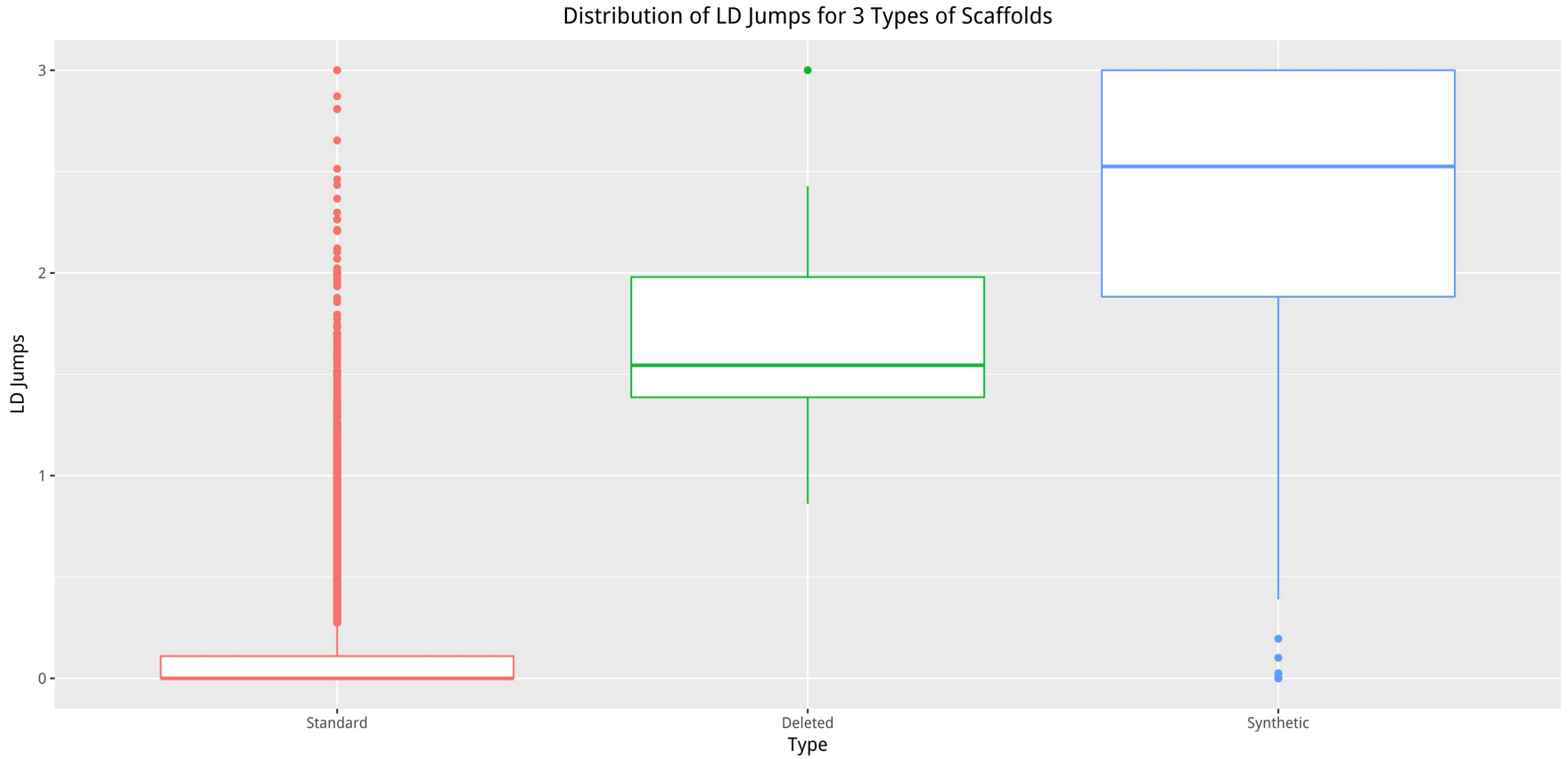


THE UNIVERSITY of ADELAIDE

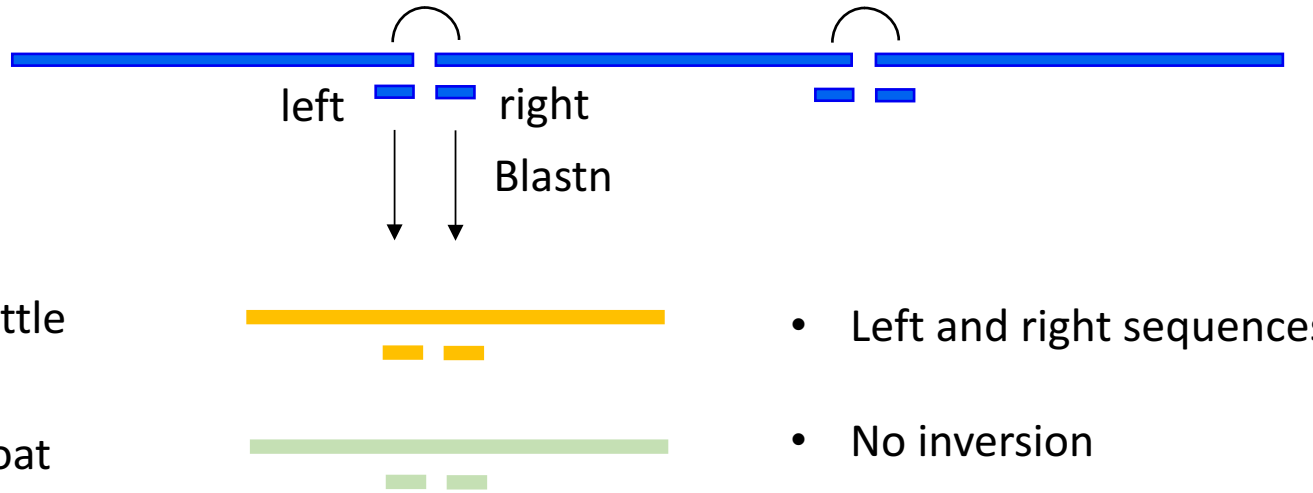
Acknowledgements:
Rick Tearle, Davies Research Centre

Davies Research Centre
…excellence in ruminant science

Scaffold conflict resolution - LD



Scaffold conflict resolution - synteny



- Left and right sequences match the same subject chromosome
- No inversion
- Left and right sequences are within 1 Mb

Scaffold conflict resolution

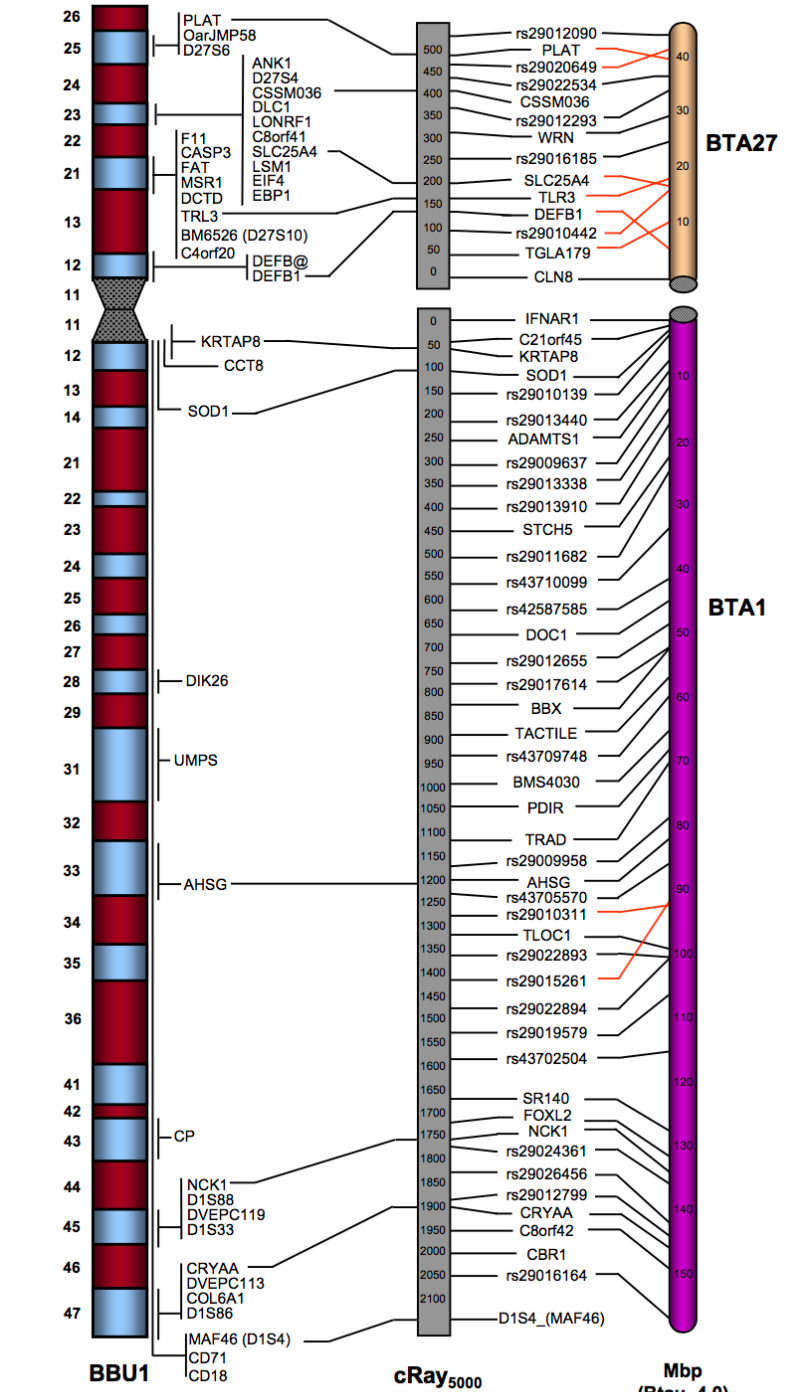
- Total 484 gaps, major scaffolds contains 457 gaps

Conservation of synteny with cattle	LD jump		
	False	True	Not available
False	158	153	16
True	81	42	7

Conservation of synteny with goat	
False	143
True	26

- 3 pairs of scaffolds are further joined based on synteny and LD data
- 8 scaffolds are further corrected on suspicious contigs joins

RH Map

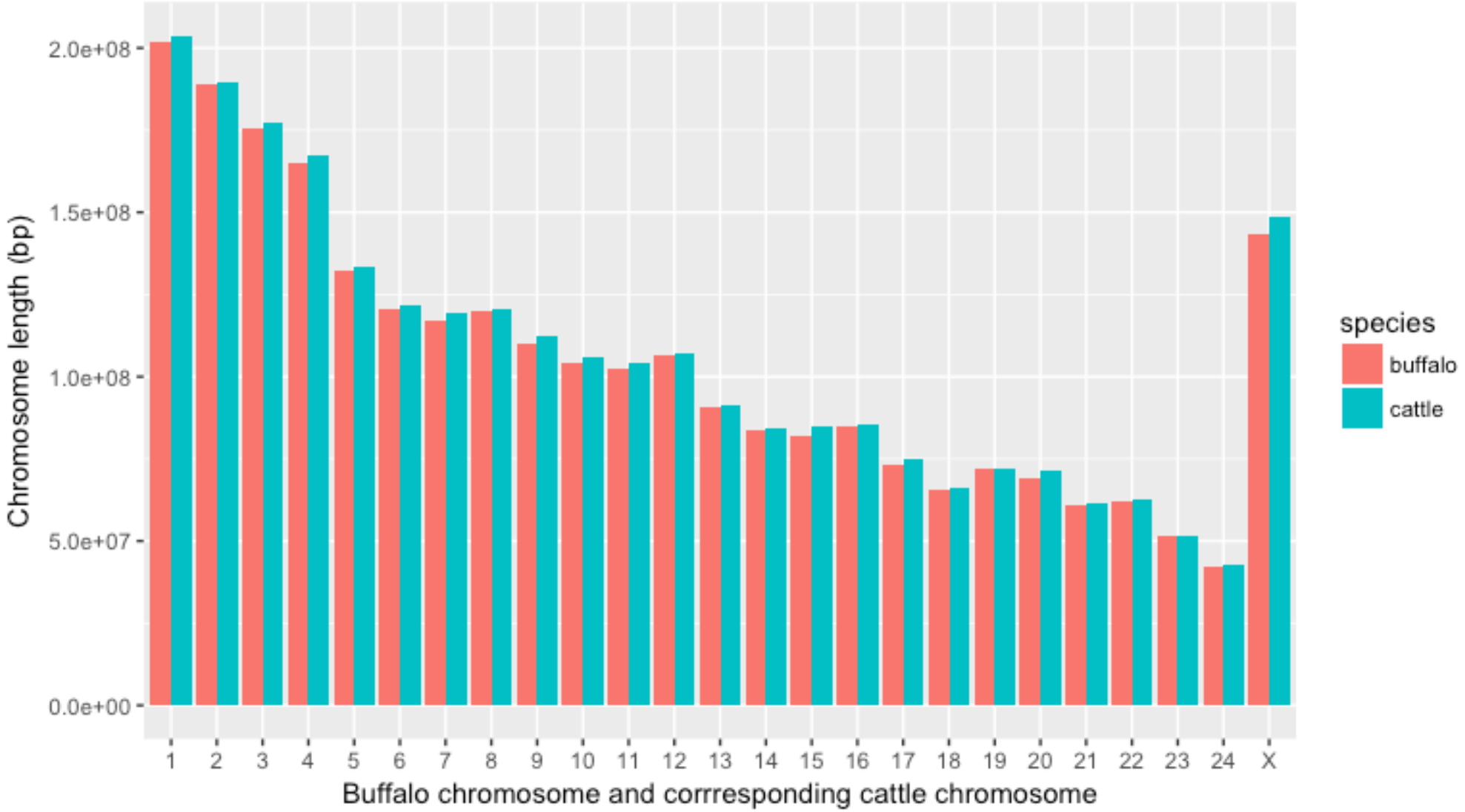


Amaral et al 2008

THE UNIVERSITY of ADELAIDE

Davies Research Centre
...excellence in ruminant science

Comparison with cattle

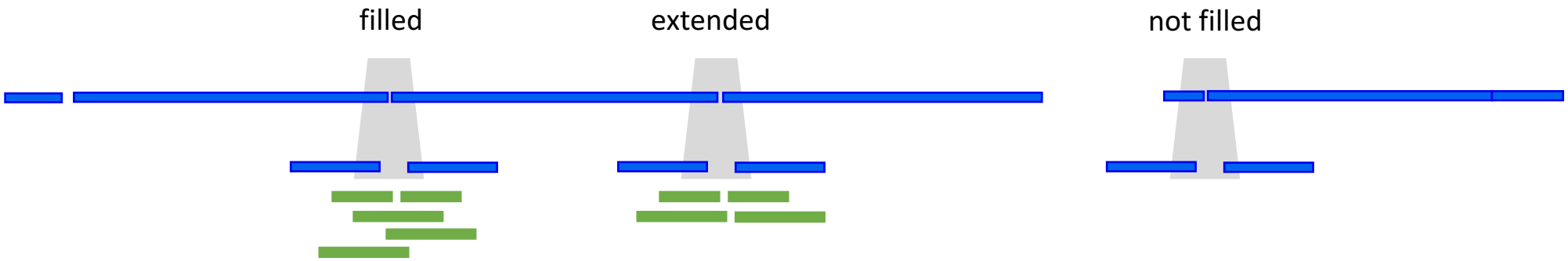


Buffalo chromosome	Cattle chromosome	Post filter aligned cattle sequences (bp)	Aligned buffalo chromosome sequences (bp)	Proportion of aligned buffalo chromosome in corresponding cattle chromosome
chr1	chr27 + chr1	79,411,655	78,424,865	0.9876
chr2	chr23 + chr2	80,454,600	78,916,878	0.9809
chr3	chr19 + chr8	78,479,786	76,730,536	0.9777
chr4	chr28 + chr5	69,215,103	67,515,827	0.9754
chr5	chr29 + chr16	55,206,967	53,479,067	0.9687
chr6	chr3	51,521,713	50,784,560	0.9857
chr7	chr6	41,661,618	41,119,539	0.987
chr8	chr4	49,376,403	48,676,612	0.9858
chr9	chr7	46,778,549	46,139,263	0.9863
chr10	chr9	41,281,480	40,445,643	0.9798
chr11	chr10	45,577,025	44,587,309	0.9783
chr12	chr11	48,089,325	47,281,056	0.9832
chr13	chr12	34,705,169	33,351,805	0.961
chr14	chr13	37,138,042	36,163,781	0.9738
chr15	chr14	35,469,423	34,443,725	0.9711
chr16	chr15	34,652,165	33,783,533	0.9749
chr17	chr17	30,473,860	29,988,091	0.9841
chr18	chr18	31,490,276	30,944,184	0.9827
chr19	chr20	29,265,525	28,912,705	0.9879
chr20	chr21	31,028,704	29,017,488	0.9352 ← min
chr21	chr22	29,563,981	29,035,737	0.9821
chr22	chr24	27,688,622	26,847,733	0.9696
chr23	chr26	23,404,542	22,731,039	0.9712
chr24	chr25	21,212,092	20,796,988	0.9804
chrX	chrX	42,150,076	40,174,442	0.9531

Masked repeats
Aligned length > 100bp
Percent identity > 90%

Gap fill

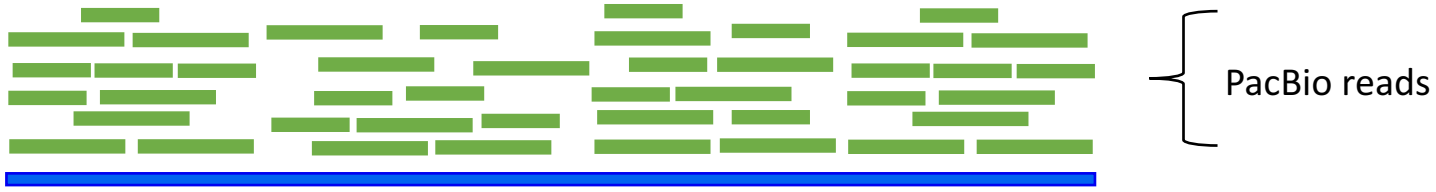
Gap fill with PBJelly



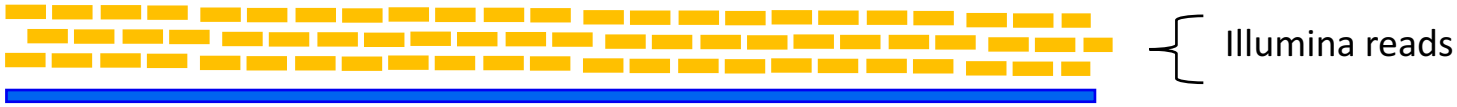
GAP STATUS	NUMBER
Over filled	195
Filled	162
Minimum read failed	63
Single extended	16
Double extended	14
Not filled	38

Polishing - scaffolds

Polish with BLASR/Arrow



Error correction with PILON



Improvement over published assembly

Description	Published assembly	Current assembly	Improvement
Total sequence length (bp)	2,836,166,969	2,654,063,837	
Total assembly gap length (bp)	74,388,041	484,000	
Number of contigs	630,368	953	
Contig N50 (bp)	21,938	18,784,635	+856 fold
Contig L50	35,881	42	-854 fold
Number of scaffolds	366,983	510	
Scaffold N50 (bp)	1,412,388	117,187,264	+83 fold
Scaffold L50	581	9	-65 fold

Top ranked mammalian assemblies

Description	Human	Mouse	Goat	Water buffalo
Total sequence length (bp)	3,253,848,404	2,818,974,548	2,922,813,246	2,654,063,837
Total assembly gap length (bp)	161,368,351	79,435,572	38,187	484,000
Number of contigs	1,519	885	30,399	953
Contig N50 (bp)	56,413,054	32,273,079	26,244,591	18,784,635
Contig L50	19	26	32	42
Number of scaffolds	858	336	29,907	510
Scaffold N50 (bp)	59,364,414	52,589,046	87,277,232	117,187,264
Scaffold L50	17	18	13	9

Acknowledgements

The Davies Research Centre

Rick Tearle

John Williams

ARS USDA

Derek Bickhart

Benjamin Rosen

Timothy Smith

Pacific Biosciences

Sarah Kingan

Dovetail Genomics

Thomas Swale

Università Cattolica

Paolo Ajmone-Marsan

7th International Symposium on Animal Functional Genomics

In association with the Functional Annotation of Animal Genomes Workshop

National Wine Centre, Adelaide, South Australia

Save the date 12-14th November 2018

Details and registration **www.ISAFG2018.com**

International Symposium on

Animal Functional Genomics

Adelaide 2018

In conjunction with

FAANG

Functional Annotation of Animal Genomes

Major Sponsors

GOLD SPONSOR

GENESEEK AUSTRALASIA a Neogen Company

NEOGEN

DovetailGenomics

ThermoFisher SCIENTIFIC

illumina

THE UNIVERSITY of ADELAIDE

THE UNIVERSITY of ADELAIDE

Davies Research Centre

...excellence in ruminant science