HASLR:
Fast Hybrid Assembly of Long Reads

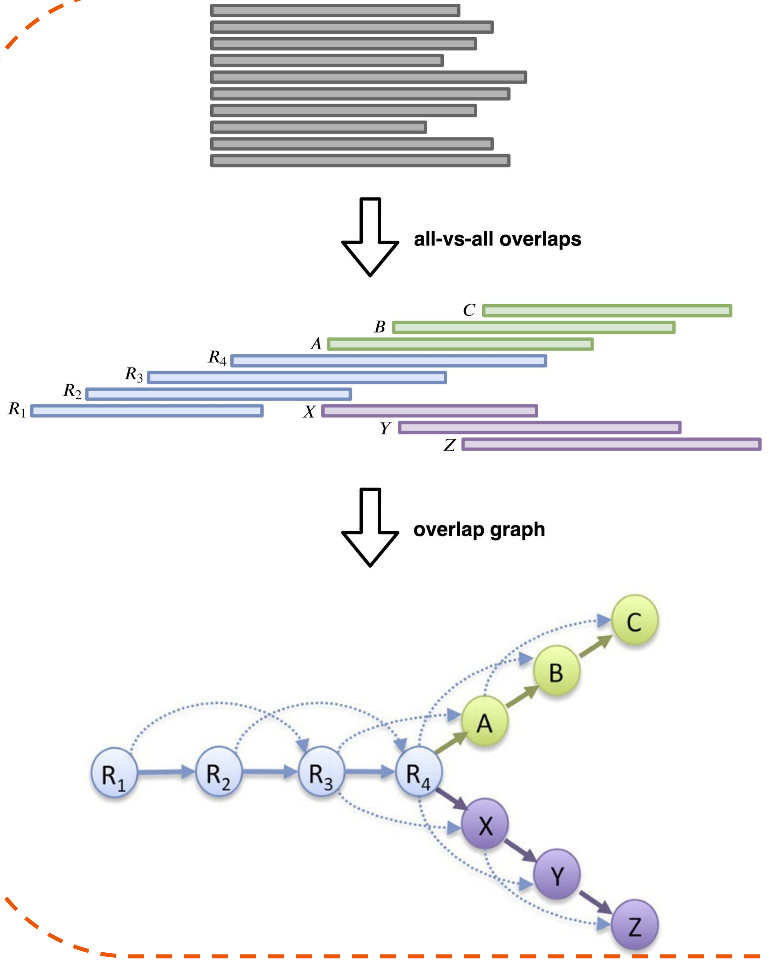
Ehsan Haghshenas, Hossein Asgari, Jens Stoye,
Cedric Chauve, Faraz Hach

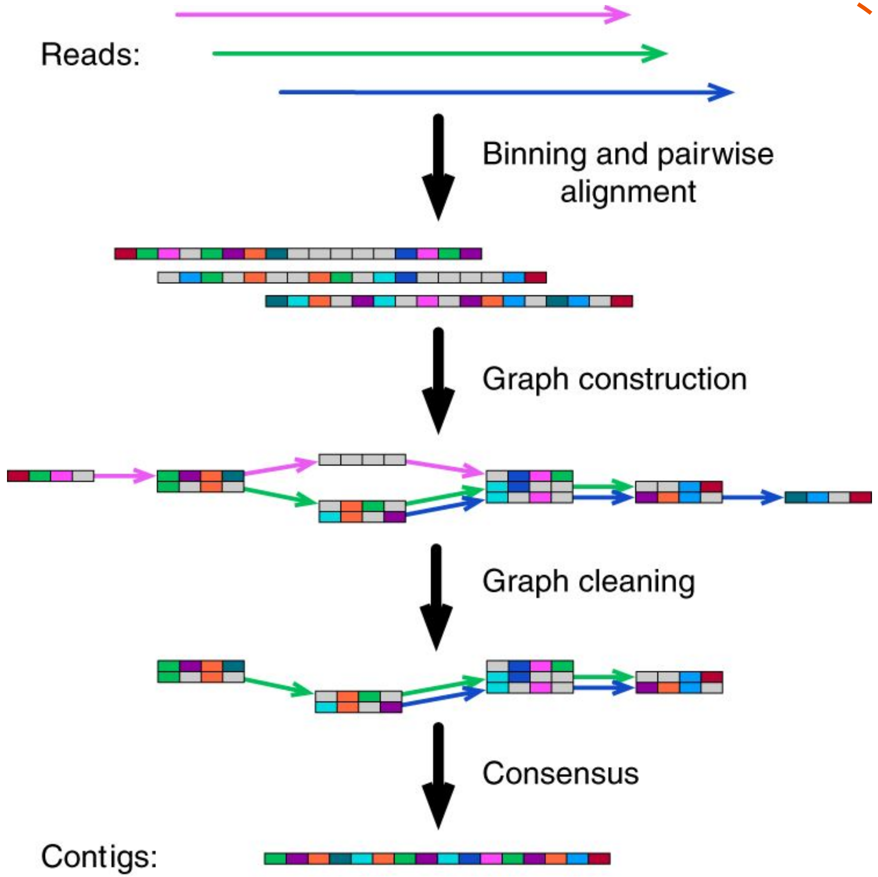
DSB 2020, February 5, 2020.

Summary

- Features of HASLR
    - Simple ideas.
    - Re-use efficient, well-tested, tools.
    - Fast and memory efficient.
    - Low mis-assembly rate.
    - Good contiguity and gene completeness.
    - Base-level accuracy similar to others tools after polishing.

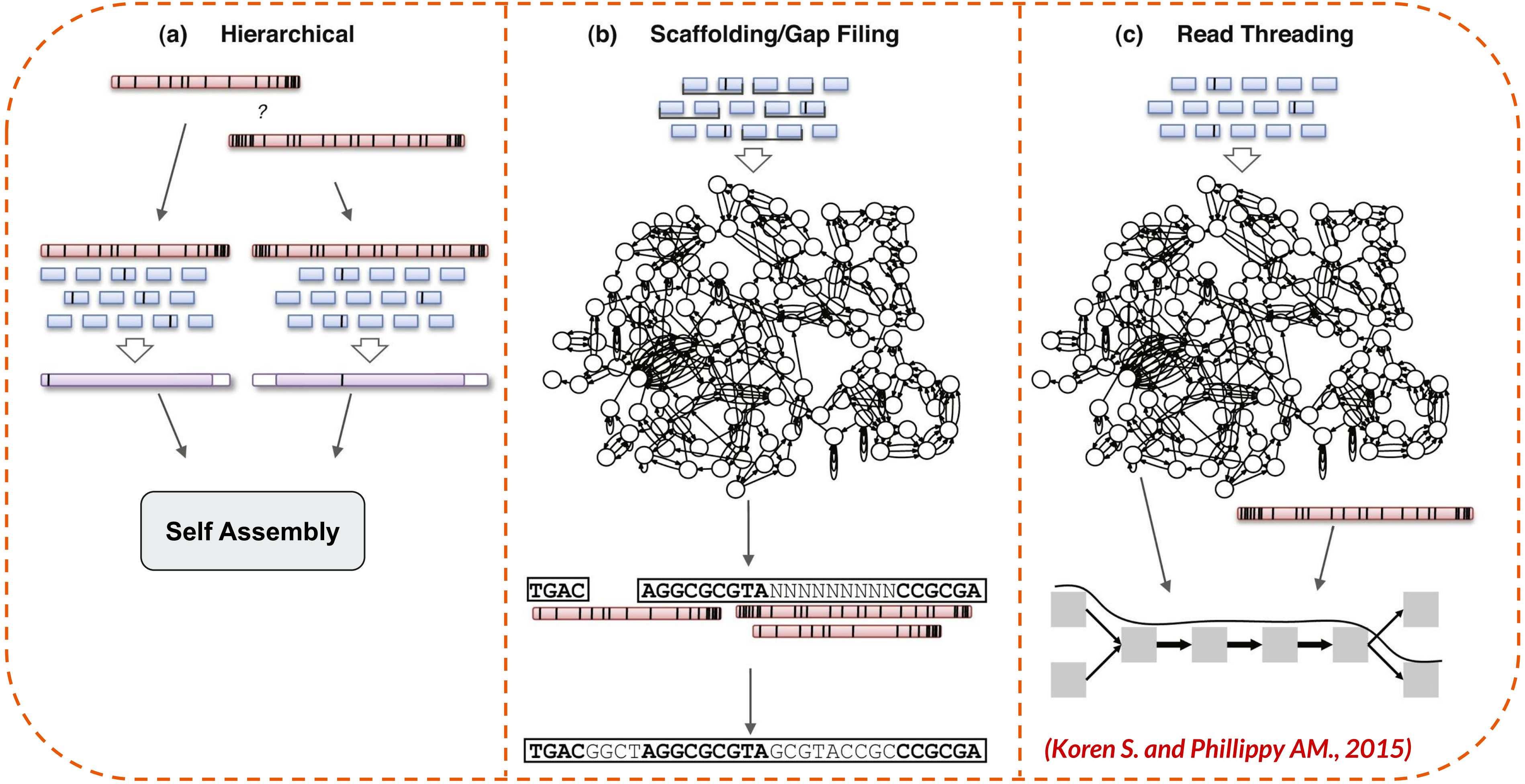
Long read assembly: self assembly





(Ruan J. and Li H., 2019)

Long read assembly: hybrid assembly



(Koren S. and Phillippy AM., 2015)

HASLR’s methodology

Short read assembly

- Build a short read assembly using Minia
  - `-kmer-size 49 -abundance-min 3 -no-ec-removal`
- Identify “unique” short read contigs
  - We assume longer contigs are more likely to come from unique regions of the genome
  - Let f_{avg} and f_{std} be average and standard deviation of “mean k-mer frequency” of the longest 30 short read contigs
  - Every short read contig whose mean k-mer frequency is below $f_{avg} + 3 f_{std}$ is considered to be unique

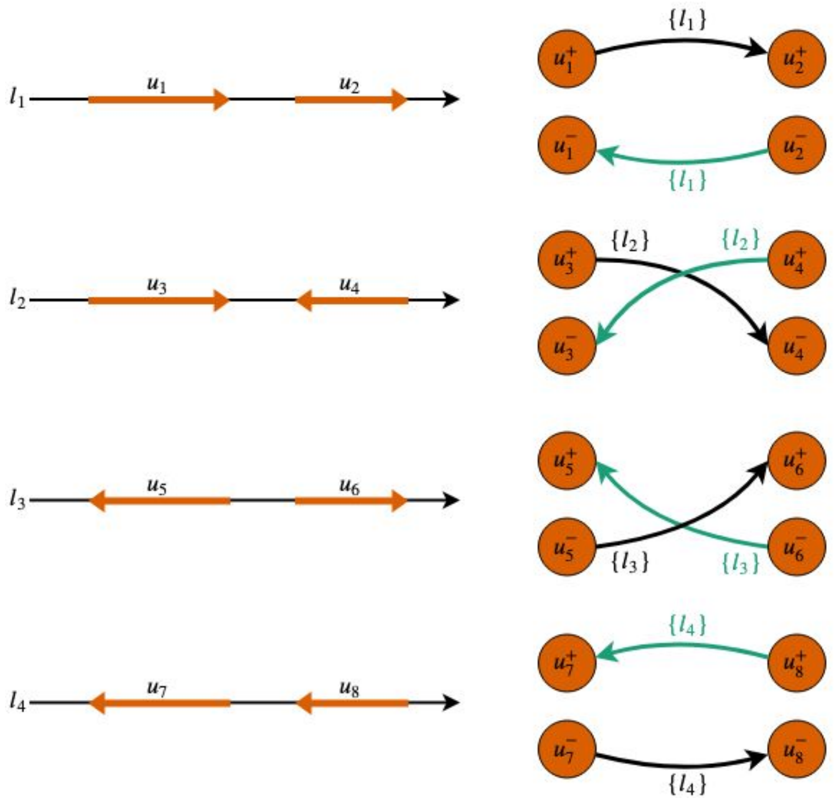
Aligning unique contis to long reads

- Align unique contigs against longest **25x coverage** of long reads
    - Using minimap2
    - Coverage is calculated based on the estimated genome size
- For each long read, select a subset of non-overlapping unique contigs alignments whose total identity score is maximal

$$S(j) = \mathbf{max}\{S(j-1),\ \underbrace{S(prev(j))}_{\text{largest index } z<j \text{ such that } a_j \text{ and } a_z \text{ are non-overlapping}} + \underbrace{a_j[nmatch]}_{\text{number of matches in } j\text{-th alignment}}\}$$

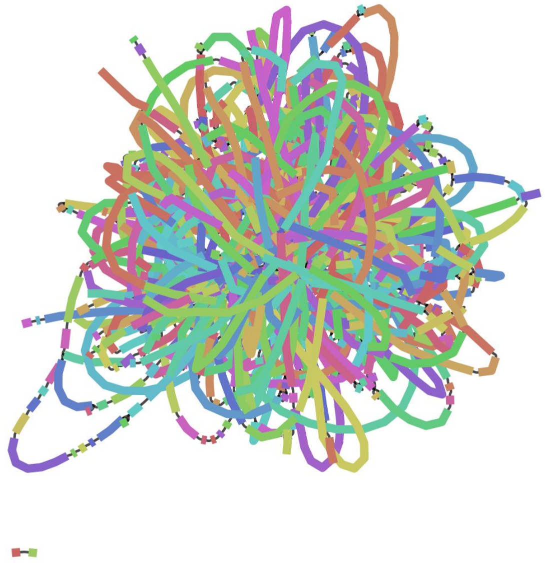
Backbone graph

- Two nodes for each unique contig
  - representing forward and reverse strand
- Edges are added between nodes if their corresponding unique contigs align to some long reads **consecutively**
  - one edge for forward and another for reverse strand



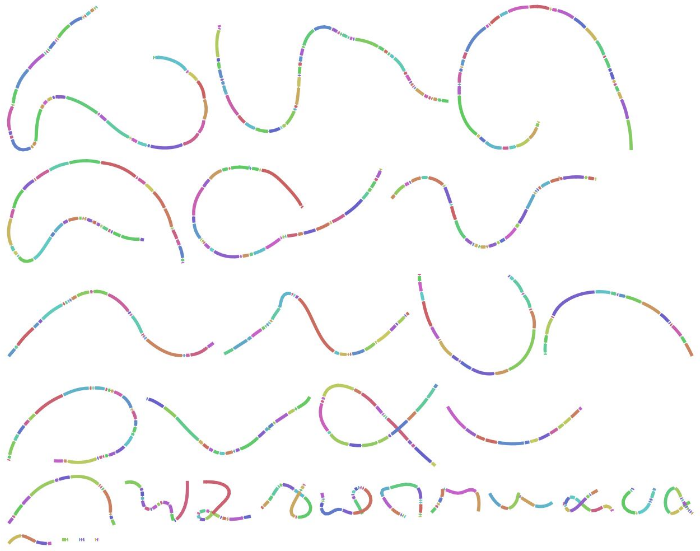
Mis-mappings

- Wrong alignment of unique contigs onto long reads cause wrong edges


Yeast PacBio dataset

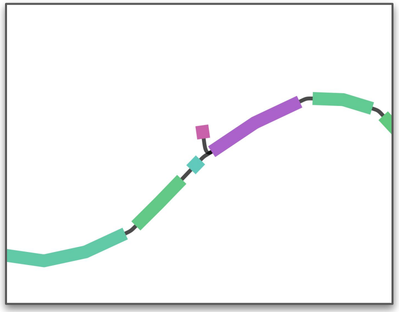
Mis-mappings

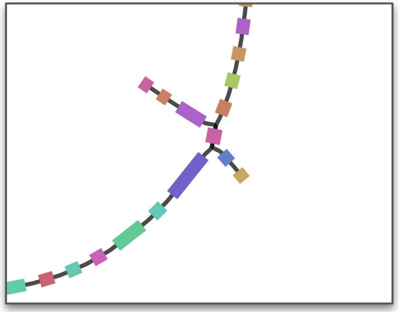
- Wrong alignment of unique contigs onto long reads cause wrong edges
- Remove low support edges
  - Less than 3 long reads
- Still there are some artifacts in the graph structure


Yeast PacBio dataset

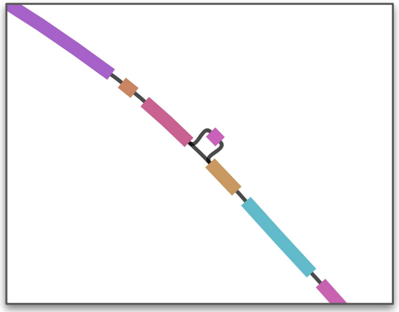
Graph cleaning

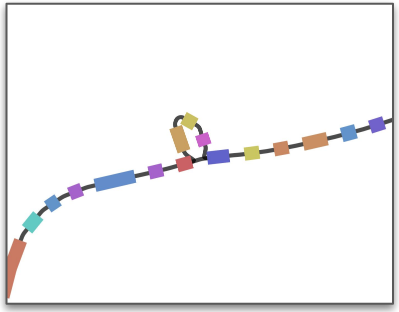
Tip



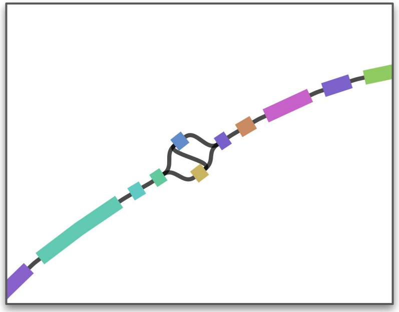


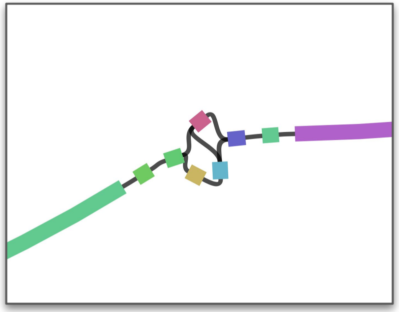
Simple bubble





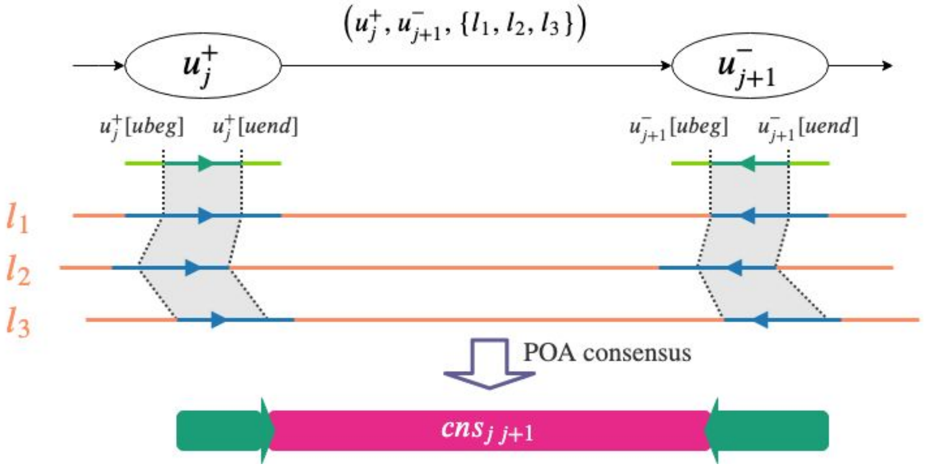
Super bubble





Consensus calling

- Find the region of unique contigs that is shared by all supporting long reads
- Calculate consensus using partial order alignment
  - SPOA in global alignment mode
- Can be done for each edge independently
  - Easy to parallelize



Generating the final assembly

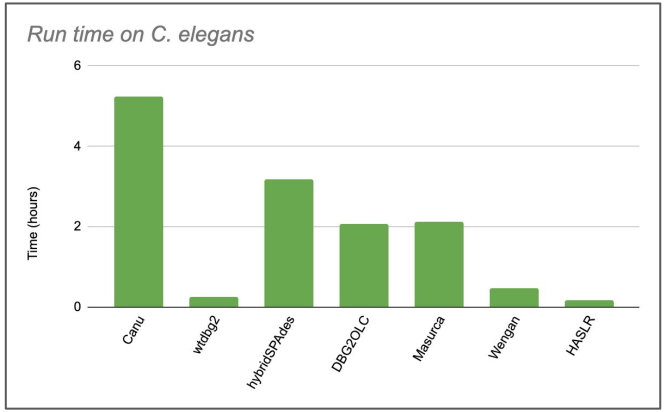
- Generate one contig per simple path (unitig) in the graph
- For each simple path, concatenate the sequence of the unique short read contigs and the consensus sequences.

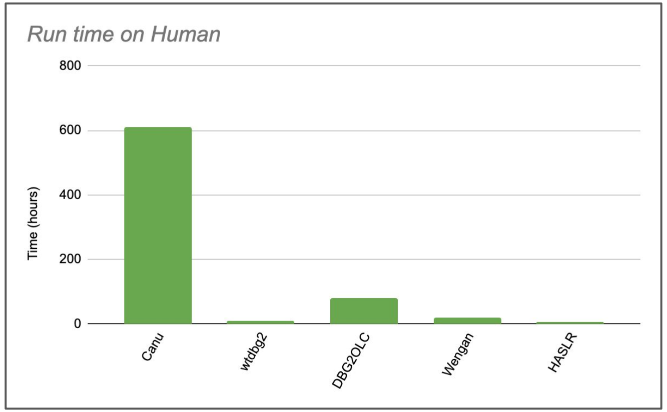
Results

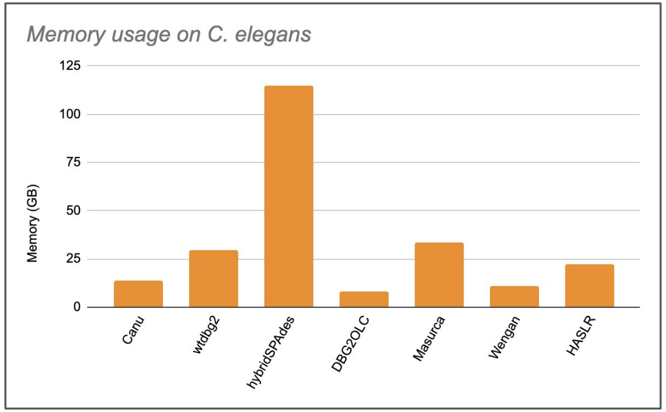
Simulated dataset

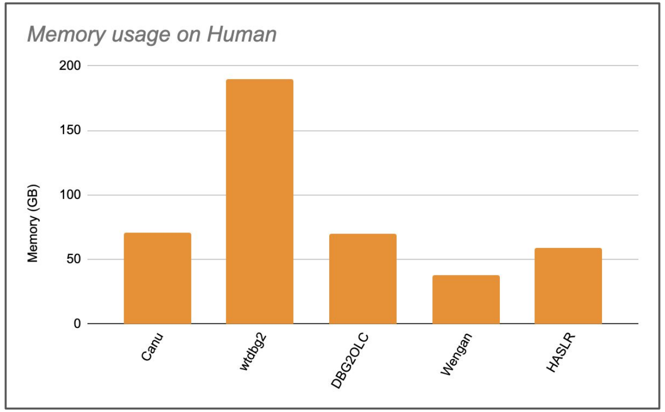
Genome	Assembler	Genome fraction	NGA50	Extensive misassembly	Local misassembly	Mismatch rate	Indel rate
C.elegans	Canu	99.847	13,775,238	3	1	5.88	67.73
	wtdbg2	95.468	81,074	194	506	246.33	657.89
	hybridSPAdes	98.643	924,797	67	197	73.26	9.14
	Unicycler	NA					
	DBG2OLC	99.692	6,732,354	10	7	8.55	174.21
	Masurca	99.609	4,614,507	34	123	14.89	4.56
	Wengan	98.917	2,042,350	53	20	7.26	59.81
	HASLR	99.182	6,455,832	0	0	14.74	230.58
Human	Canu	97.279	15,045,226	854	99	37.7	196.78
	wtdbg2	92.735	87,595	3,436	13,041	224.02	598.87
	hybridSPAdes	NA					
	Unicycler	NA					
	DBG2OLC	91.013	14,385,033	221	246	8.43	201.56
	Masurca	NA					
	Wengan	94.617	11,216,374	185	70	3.84	33.5
	HASLR	91.213	17,025,446	2	5	11.32	207.88

Simulated dataset





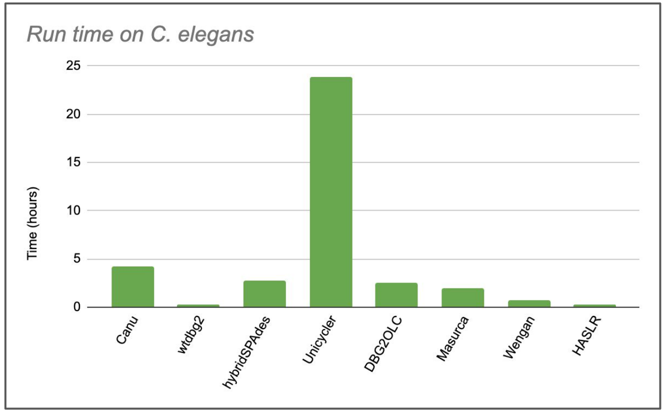


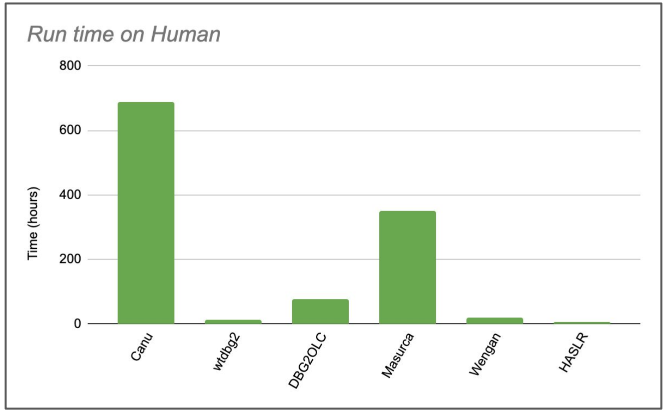


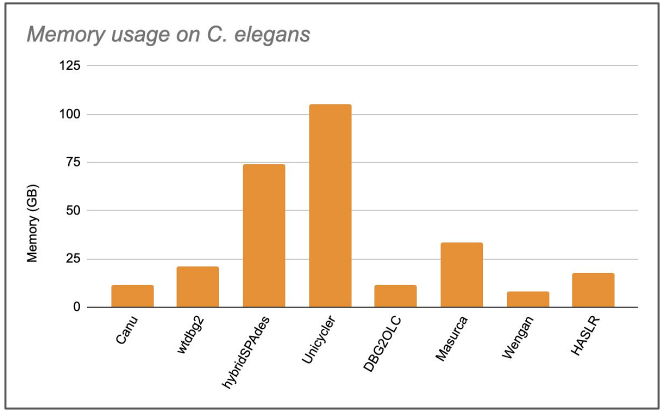
Real dataset

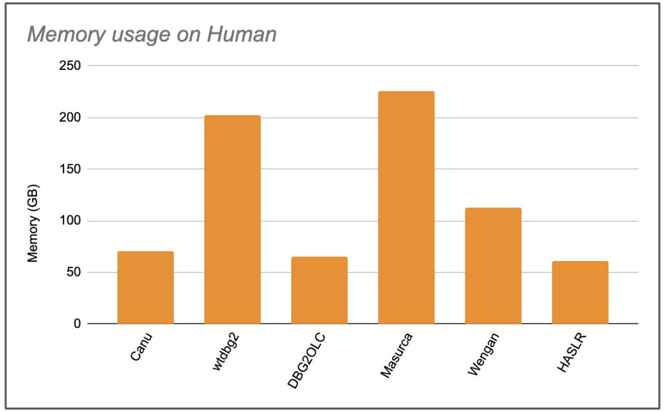
Genome	Assembler	Genome fraction	NGA50	Extensive misassembly	Local misassembly	Mismatch rate	Indel rate
C.elegans (PacBio)	Canu	99.665	561,201	723	596	65.28	58.82
	wtdbg2	98.994	561,292	329	596	26.82	79.72
	hybridSPAdes	96.720	84,003	633	638	108.04	15.96
	Unicycler	97.102	139,992	940	692	58.36	45.47
	DBG2OLC	99.100	421,196	546	383	44.75	80.61
	Masurca	97.013	471,366	368	504	49.20	23.50
	Wengan	93.341	341,861	308	336	35.75	121.11
	HASLR	97.431	453,631	259	331	26.08	140.40
CHM1 (PacBio)	Canu	96.084	2,329,909	6,715	7,048	145.81	120.69
	wtdbg2	92.896	2,081,842	3,535	6,286	118.45	72.54
	hybridSPAdes	NA					
	Unicycler	NA					
	DBG2OLC	95.547	1,599,466	3,718	8,690	116.81	116.89
	Masurca	93.782	1,761,291	4,984	7,491	180.83	57.53
	Wengan	88.948	875,489	2,771	7,577	115.65	160.71
	HASLR	92.664	1,699,092	2,097	7,661	113.06	281.74

Real dataset

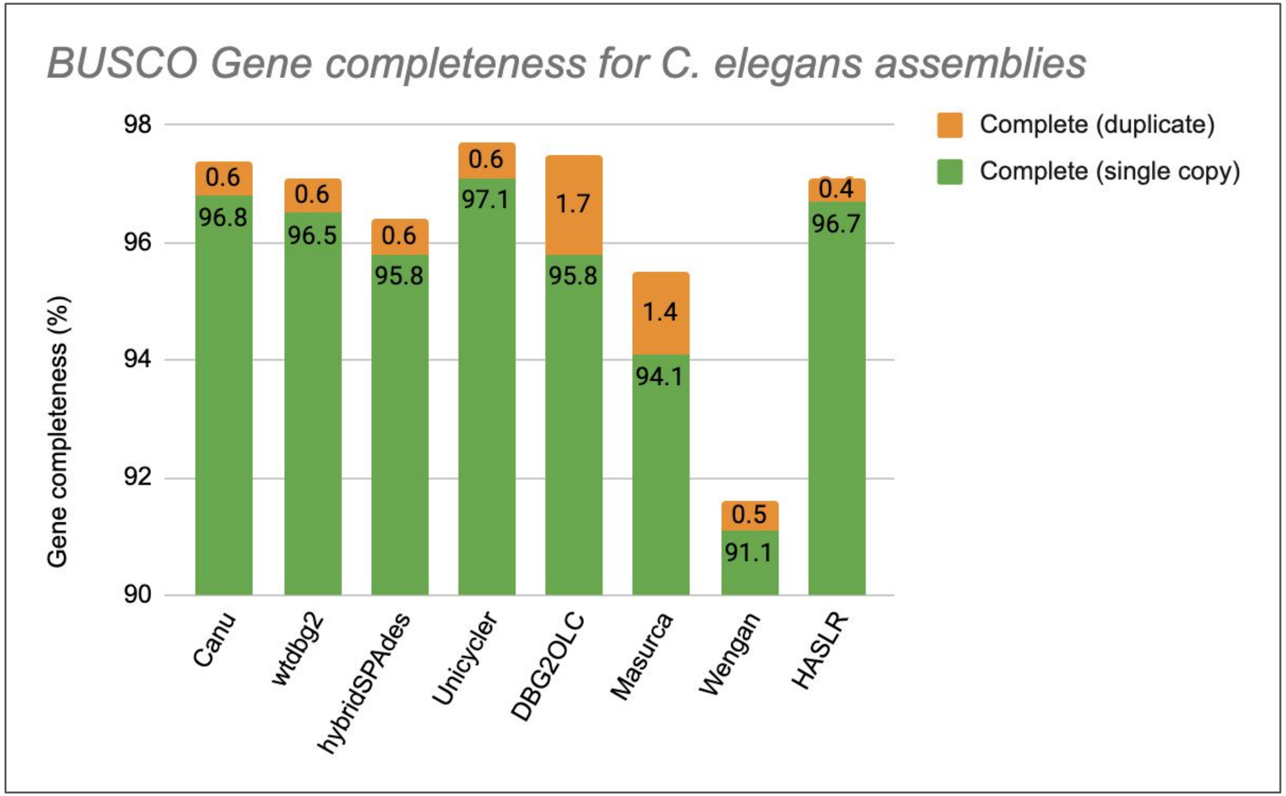








Gene completeness



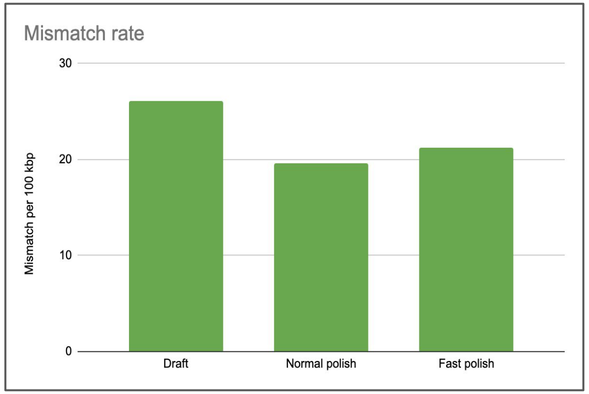
Effect of polishing

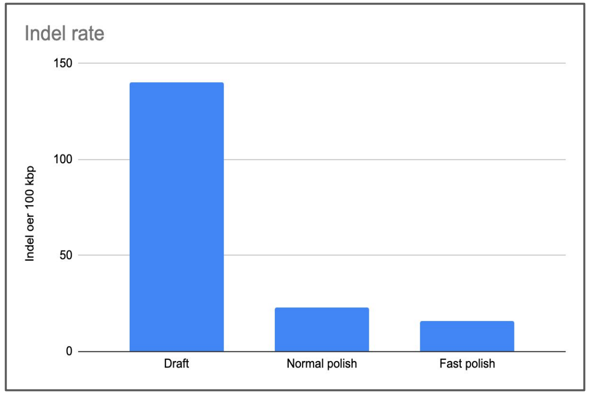
		Mismtach rate		Indel rate	
Dataset	Assembler	draft	polished	draft	polished
C.elegans (PacBio)	Canu	65.28	65.88	58.82	29.71
	wtdbg2	26.82	25.90	79.72	27.11
	hybridSPAdes	108.04	27.88	15.96	45.43
	Unicycler	58.36	36.97	45.47	32.08
	DBG2OLC	44.75	46.50	80.61	43.52
	Masurca	49.20	30.90	23.50	31.97
	Wengan	35.75	21.13	121.11	22.82
	HASLR	26.08	19.61	140.40	22.92

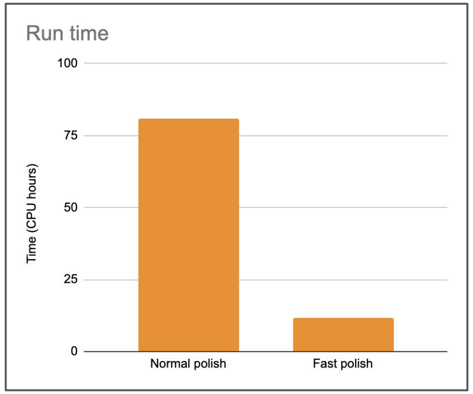
Polishing is done using arrow (https://github.com/PacificBiosciences/GenomicConsensus)

Faster polishing?

- What if we only polish regions between unique contigs?







- Not integrated with HASLR yet

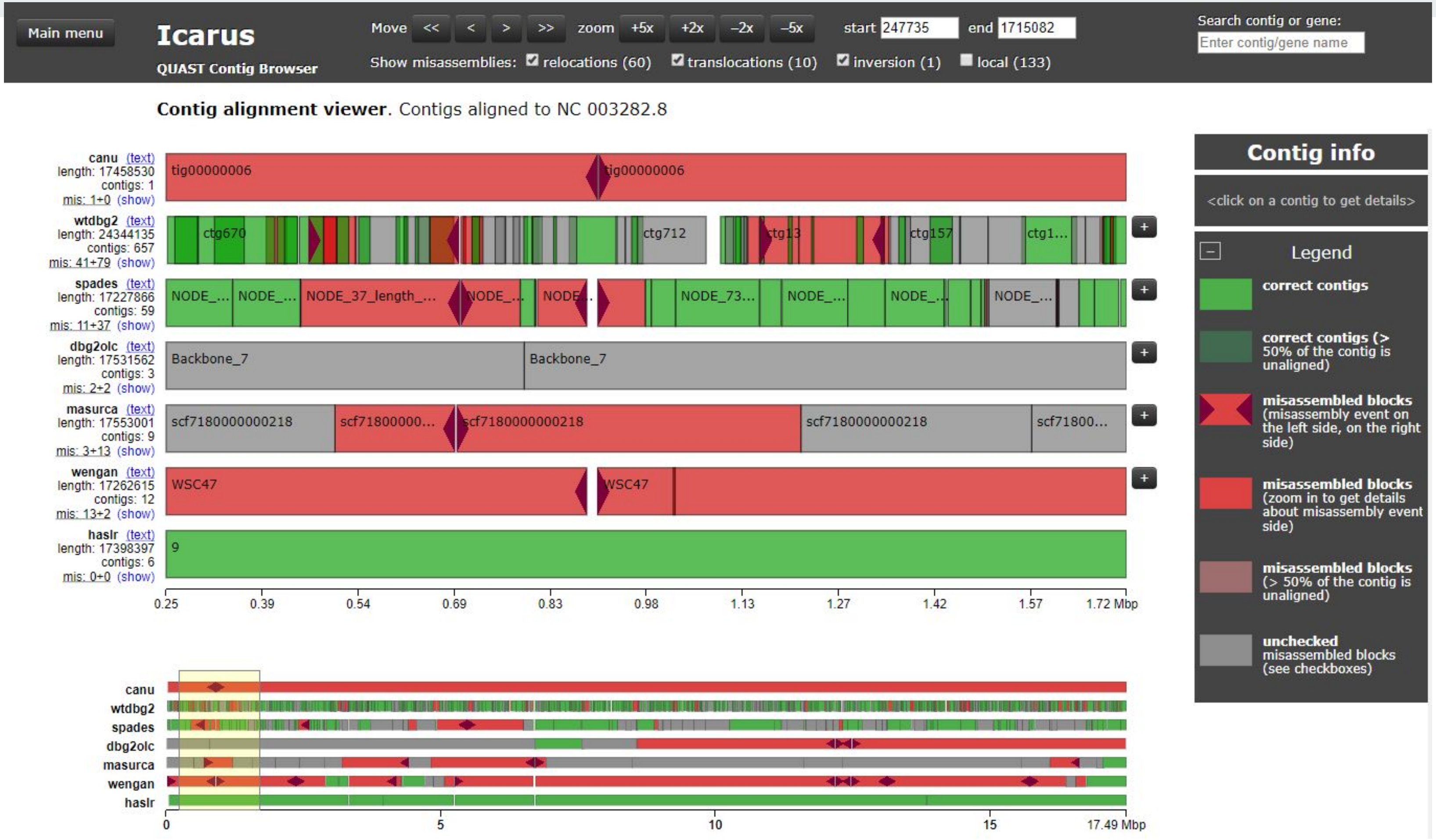
Summary

- HASLR is a fast and memory efficient assembly pipeline.
- It relies on a combination of simple ideas and well-tested assembly tools.
- It generates a conservative assembly, characterized by a low rate of mis-assemblies at the expense of a lower genome fraction.
- Its main innovation is the introduction of the backbone graph for scaffolding and gap filling.
- Available on bioconda and github
  - https://github.com/vpc-ccg/haslr

Future directions

- Advanced bubble/tip cleaning algorithm.
- Integrating fast polishing module.
- Support for ultra-long nanopore reads.
- Improving genome coverage.
    - Using an OLC approach on unused long reads
- Diploid genome assembly.
    - Clustering long read subsequences into two groups before consensus calling

Thank you!



Icarus

QUAST Contig Browser

start 247735 end 1715082

Show misassemblies: [✓ relocations (60)] [✓ translocations (10)] [✓ inversion (1)] local (133)

Search contig or gene: Enter contig/gene name

Contig alignment viewer. Contigs aligned to NC 003282.8

canu (text)
length: 17458530
contigs: 1
mis: 1+0 (show)

wtdbg2 (text)
length: 24344135
contigs: 657
mis: 41+79 (show)

spades (text)
length: 17227866
contigs: 59
mis: 11+37 (show)

dbg2olc (text)
length: 17531562
contigs: 3
mis: 2+2 (show)

masurca (text)
length: 17553001
contigs: 9
mis: 3+13 (show)

wengan (text)
length: 17262615
contigs: 12
mis: 13+2 (show)

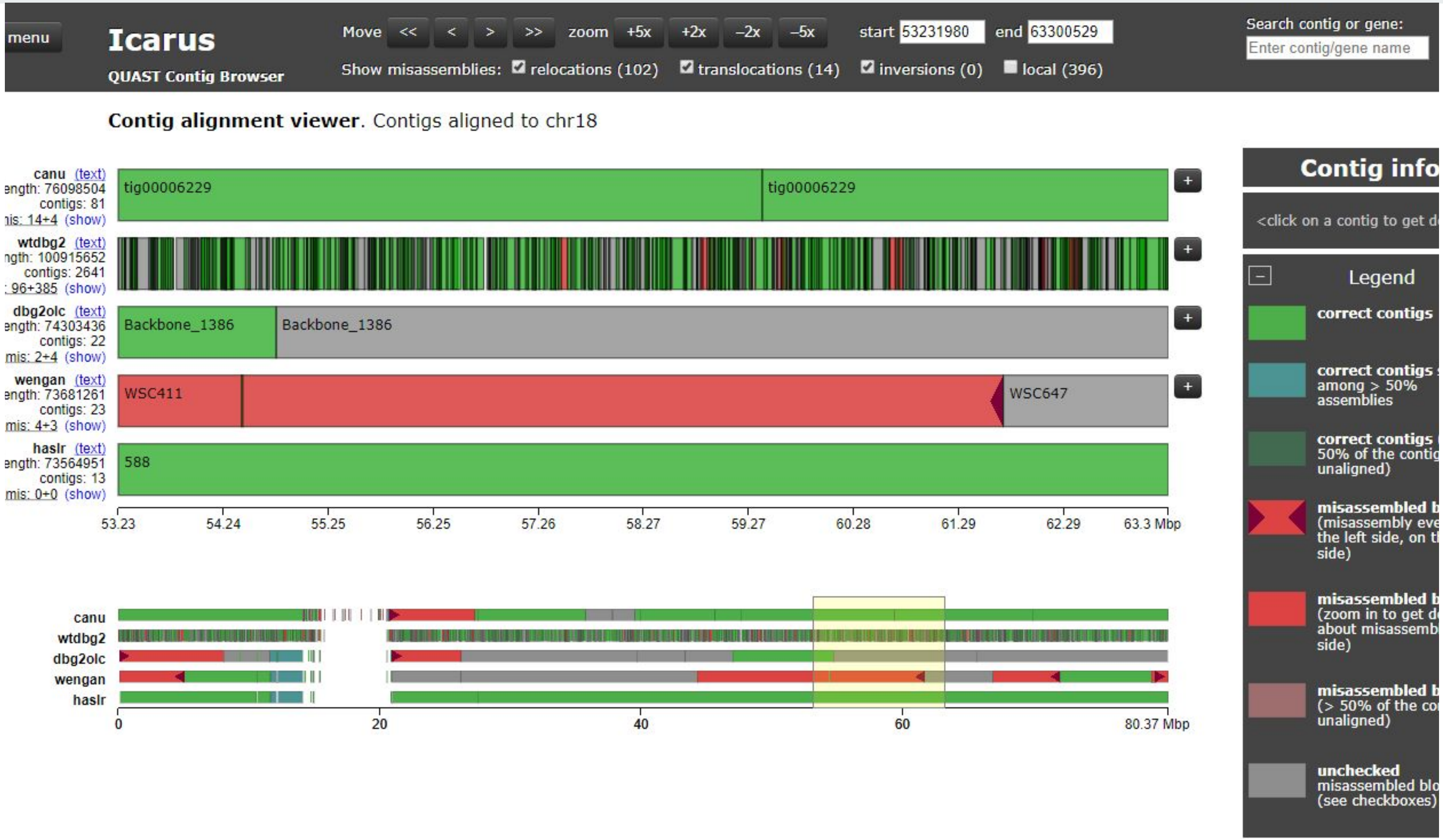
haslr (text)
length: 17398397
contigs: 6
mis: 0+0 (show)

Contig info

<click on a contig to get details>

Legend

🟩	**correct contigs**
🟩	**correct contigs** (> 50% of the contig is unaligned)
🟥	**misassembled blocks** (misassembly event on the left side, on the right side)
🟥	**misassembled blocks** (zoom in to get details about misassembly event side)
🟥	**misassembled blocks** (> 50% of the contig is unaligned)
⬜	**unchecked** misassembled blocks (see checkboxes)


menu
Icarus
QUAST Contig Browser
Move << < > >>
zoom +5x +2x –2x –5x
Search contig or gene:
Enter contig/gene name
Contig info
<click on a contig to get d
Legend
correct contigs
correct contigs among > 50% assemblies
correct contigs 50% of the contig unaligned)
misassembled b (misassembly eve the left side, on t side)
misassembled b (zoom in to get d about misassemb side)
misassembled b (> 50% of the co unaligned)
unchecked misassembled blo (see checkboxes)
tig00006229
tig00006229
Backbone_1386
Backbone_1386
WSC411
WSC647
588
53.23
54.24
55.25
56.25
57.26
58.27
59.27
60.28
61.29
62.29
63.3 Mbp
canu
wtdbg2
dbg2olc
wengan
haslr
0
20
40
60
80.37 Mbp

start 53231980 end 63300529

Show misassemblies: [✓ relocations (102)] [✓ translocations (14)] [✓ inversions (0)] local (396)

Contig alignment viewer. Contigs aligned to chr18

canu (text)
ength: 76098504
contigs: 81
nis: 14+4 (show)

wtdbg2 (text)
ngth: 100915652
contigs: 2641
: 96+385 (show)

dbg2olc (text)
ength: 74303436
contigs: 22
mis: 2+4 (show)

wengan (text)
ength: 73681261
contigs: 23
mis: 4+3 (show)

haslr (text)
ength: 73564951
contigs: 13
mis: 0+0 (show)