PACIFIC BIOSCIENCES®

Improved Phased Assembly using HiFi Data

Ivan Sović, Ph.D., PacBio Assembly Tech Lead

Zev Kronenberg, Christopher Dunn, Derek Barnett, Sarah Kingan, James Drake, Jonas Korlach

@IvanSovic

UC Davis Workshop, 2020

IMPROVED PHASED ASSEMBLY USING HIFI DATA

James Drake Derek Barnett Ivan Sović

Zev Kronenberg Christopher Dunn

Sarah Kingan

Jonas Korlach

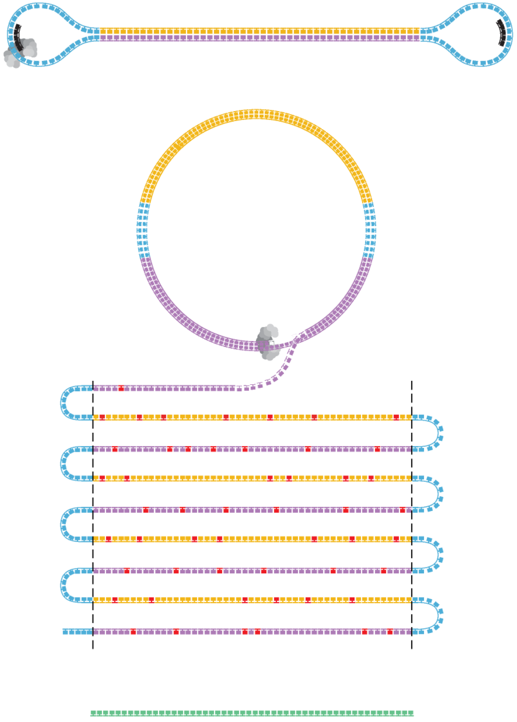
WHAT ARE HIFI READS?

- **They are long**
  - Up to 25 kb

- **They are accurate**
  - Long reads with ≥Q20 (99%) accuracy

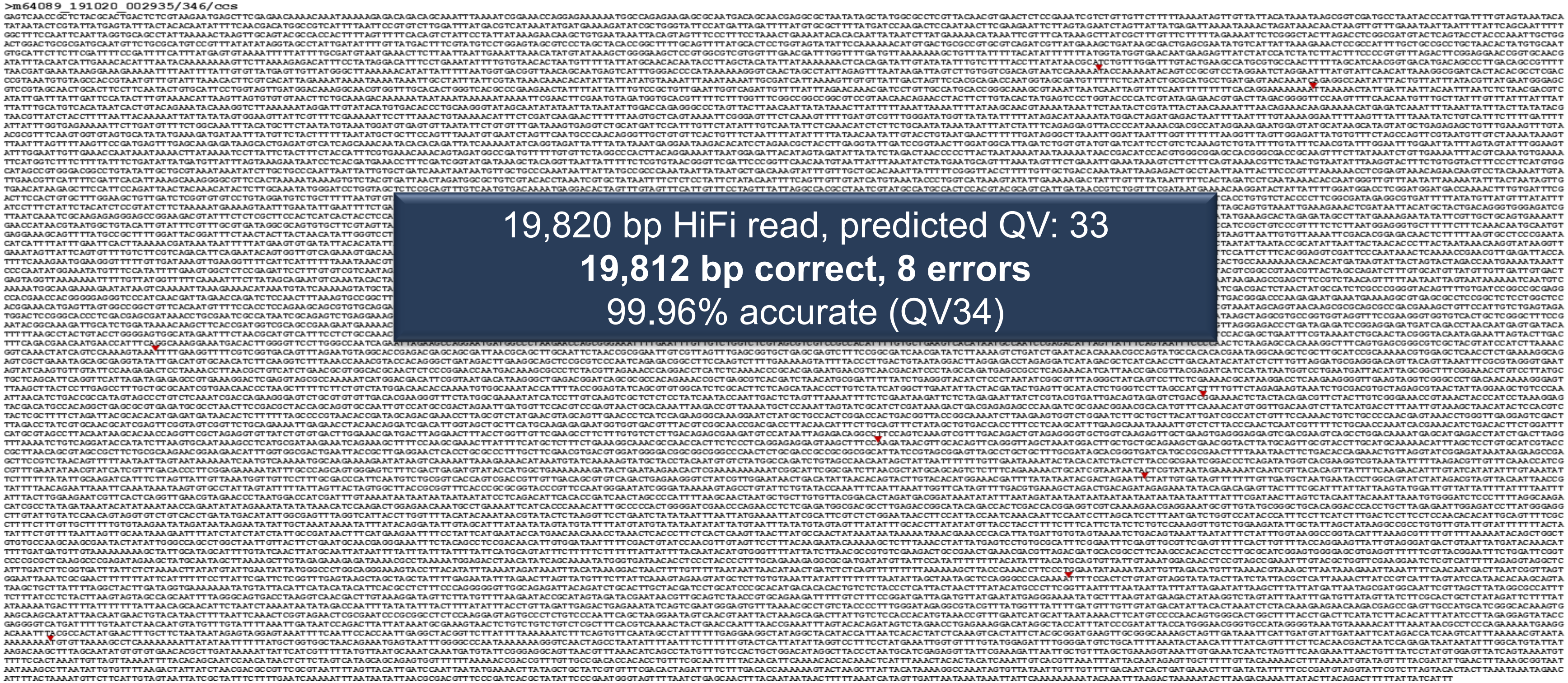
- **They have single-molecule resolution**
  - Sequence DNA or RNA

- **They have little bias**
  - No DNA amplification, least GC content and sequence complexity bias



HiFi READ
(>99% Accuracy)

HOW ACCURATE ARE HIFI READS?



```
>m64089_191020_002935/346/ccs
```

19,820 bp HiFi read, predicted QV: 33

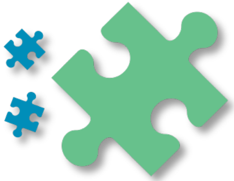
19,812 bp correct, 8 errors

99.96% accurate (QV34)

HIFI READS FOR IMPROVED ASSEMBLY

Contiguity

- Resolve Repetitive Regions
- High Contig N50



Correctness

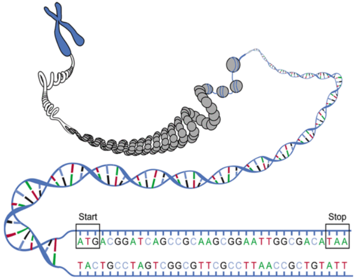
- Base QV
- Phasing accuracy

AGTTTCGATAGA

AGTT-CGAAAGA

Completeness

- Gene Space
- Repetitive Regions

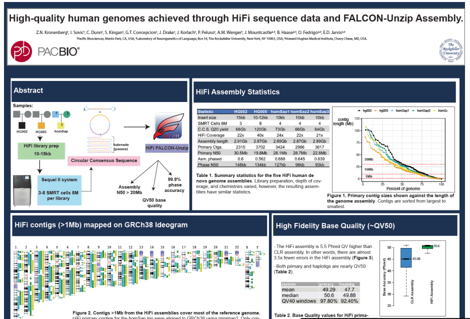

Start
Stop
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Compute

- CPU / Wall Time
- RAM
- Disk Storage

ASSEMBLY METHODS FOR HIFI READS

FALCON-Unzip



High-quality human genomes achieved through HiFi sequence data and FALCON-Unzip Assembly.

Z.N. Kronenberg[1], I. Sovic[1], C. Dunn[1], S. Kingan[1], G.T. Concepcion[1], J. Drake[1], J. Korlach[1], P. Peluso[1], A.M. Wenger[1], J. Mountcastle[2,3], B. Haase[2,3], O. Fedrigo[2,3], E.D. Jarvis[2,3]

[1]Pacific Biosciences, Menlo Park, CA, USA; [2]Laboratory of Neurogenetics of Language, Box 54, The Rockefeller University, New York, NY 10065, USA; [3]Howard Hughes Medical Institute, Chevy Chase, MD, USA.

PACBIO®

Abstract

HiFi Assembly Statistics

Statistic	HG002	HG005	homSap1	homSap2	homSap3
Insert size	15kb	10-12kb	10kb	10kb	10kb
SMRT Cells 8M	3	8	4	4	4
C.C.S. Q20 yield	68Gb	120Gb	73Gb	66Gb	64Gb
HiFi Coverage	22x	40x	24x	22x	21x
Assembly length	2.91Gb	2.87Gb	2.89Gb	2.87Gb	2.89Gb
Primary Ctgs.	2310	3702	3424	2986	3617
Primary N50	30.5Mb	19.8Mb	26.1Mb	28.7Mb	22.8Mb
Asm. phased	0.6	0.562	0.688	0.645	0.639
Phase N50	148kb	134kb	127kb	98kb	93kb

Table 1. Summary statistics for the five HiFi human de novo genome assemblies. Library preparation, depth of coverage, and chemistries varied, however, the resulting assemblies have similar statistics.

Figure 1. Primary contig sizes shown against the length of the genome assembly. Contigs are sorted from largest to smallest.

HiFi contigs (>1Mb) mapped on GRCh38 Ideogram

Figure 2. Contigs >1Mb from the HiFi assemblies cover most of the reference genome. HiFi primary contigs for the homSap trio were aligned to GRCh38 using minimap2. Only contigs greater than 1Mb are shown. Breaks in the alignments do not always correspond to contig boundaries.

High Fidelity Base Quality (~QV50)

- The HiFi assembly is 5.5 Phred QV higher than CLR assembly. In other words, there are almost 3.5x fewer errors in the HiFi assembly (**Figure 3**).
- Both primary and haplotigs are nearly QV50 (**Table 2**).

statistic	primary	haplotig
mean	49.29	47.7
median	50.6	49.88
QV40 windows	97.80%	92.40%

Table 2. Base Quality values for HiFi primary/haplotig number of 10 are shown.

Highly accurate phasing achieved through FALCON-Unzip

- Falcon-Unzip phased between 56.2-68.8% of the genomes we assembled (**Table1**).
- Average haplotig phasing accuracy was ~ 99.9% (**Figure 4**).
- Haplotigs (phase blocks) contiguity, measured as N50, ranged from 93kbp-150kbp and is positively correlated with HiFi insert size.
- Phase blocks can span many genes. For example, in **Figure 5**, several genes are phased, including those with clinically relevant variants (OMIM; ClinVar)

Figure 4. Haplotig Phasing Accuracy. Each point represents a haplotig, the size of the point denotes haplotig size. Points along the vertical and horizontal axis are enriched for either maternal or paternal kmers. The overall accuracy is listed within each pane.

FALCON-Unzip Phase Blocks
200Kbp phased region spanning disease associated variants phased in all individuals
Haplotype transmission in HomSap Trio

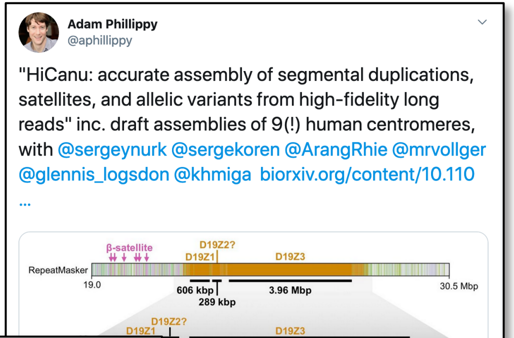
Figure 5. Example of haplotigs phasing clinically relevant genes. The Top panel of genome browser shows the haplotigs/phased regions aligned to GRCh38. Genes tracks and clinical variants are shown below.

Sequen

Conclu

FALC
HiFi c
five hu
-FALC
-Falco
-FALC
ences
For fu
Weng
prove
gy (20
Vollge
huma
63503

HiCanu

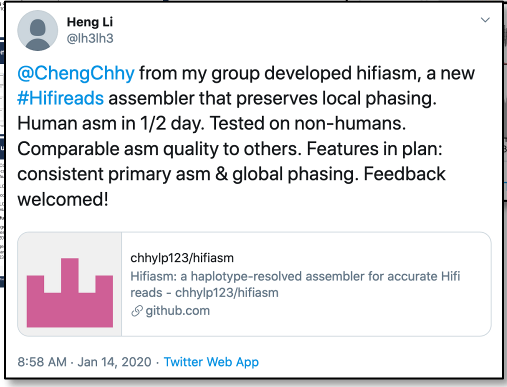


Adam Phillippy
@aphillippy

"HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads" inc. draft assemblies of 9(!) human centromeres, with @sergeynurk @sergekoren @ArangRhie @mrvollger @glennis_logsdon @khmiga biorxiv.org/content/10.110 ...

Twitter Web App

hifiasm



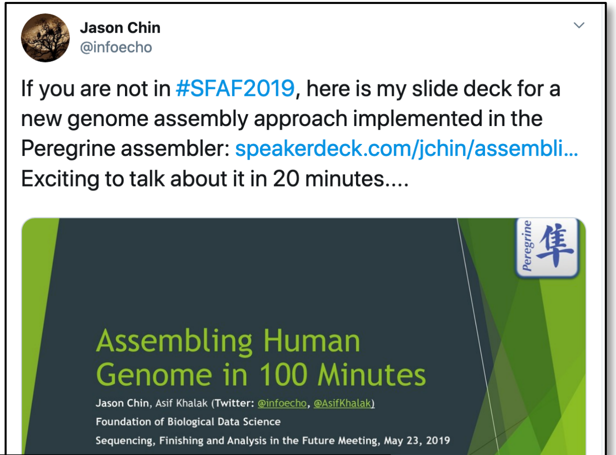
Heng Li
@lh3lh3

@ChengChhy from my group developed hifiasm, a new #Hifireads assembler that preserves local phasing. Human asm in 1/2 day. Tested on non-humans. Comparable asm quality to others. Features in plan: consistent primary asm & global phasing. Feedback welcomed!

chhylp123/hifiasm
Hifiasm: a haplotype-resolved assembler for accurate Hifi reads - chhylp123/hifiasm
github.com

8:58 AM · Jan 14, 2020 · Twitter Web App

Peregrine



Jason Chin
@infoecho

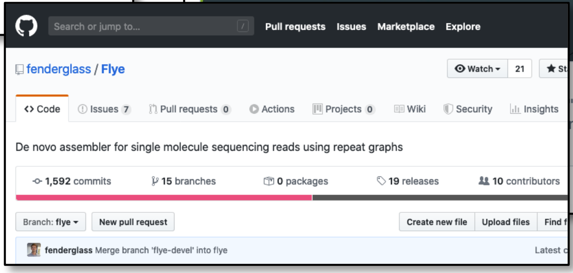
If you are not in #SFAF2019, here is my slide deck for a new genome assembly approach implemented in the Peregrine assembler: speakerdeck.com/jchin/assembli... Exciting to talk about it in 20 minutes....

Assembling Human Genome in 100 Minutes

Jason Chin, Asif Khalak (Twitter: @infoecho, @AsifKhalak)
Foundation of Biological Data Science
Sequencing, Finishing and Analysis in the Future Meeting, May 23, 2019

"SHIMMER" indexing
most unbiased way to ...

Flye



fenderglass / Flye

De novo assembler for single molecule sequencing reads using repeat graphs

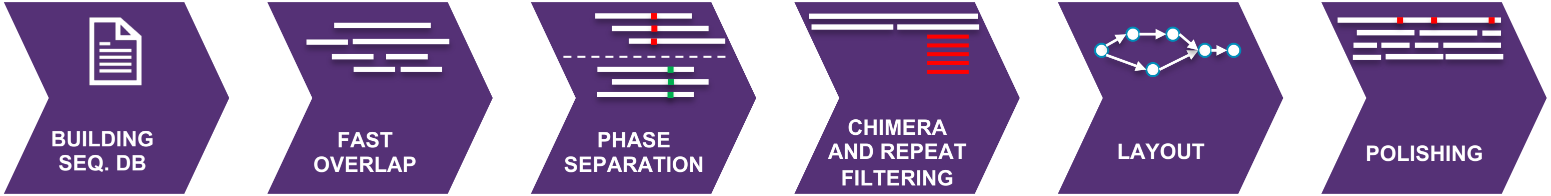
1,592 commits · 15 branches · 0 packages · 19 releases · 10 contributors

fenderglass Merge branch 'flye-devel' into flye

https://github.com/cschin/Peregrine
https://github.com/chhylp123/hifiasm
https://github.com/fenderglass/Flye
https://www.biorxiv.org/content/10.1101/2020.03.14.992248v3
https://www.pacb.com/proceedings/high-quality-human-genomes-achieved-through-hifi-sequence-data-and-falcon-unzip-assembly/

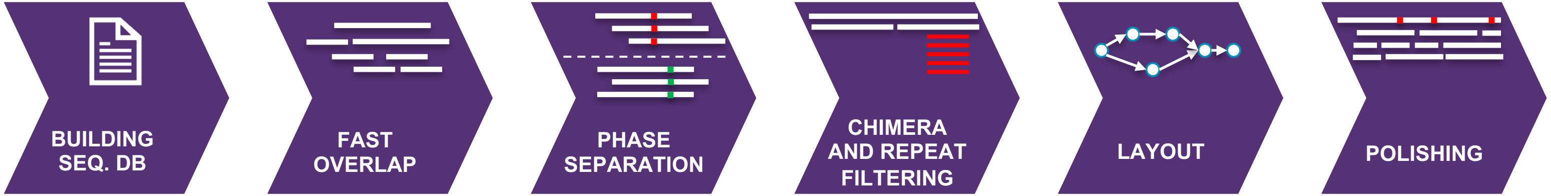
IMPROVED AND PHASED ASSEMBLY (IPA)

IMPROVED AND PHASED ASSEMBLY (IPA)



- **Goals:**
  1. **Fast assembly and quick turnaround time**
  2. **High contiguity**
  3. **Fully phased haplotigs**
  4. **High per-base quality of polished assemblies**
  5. **Ease of use**

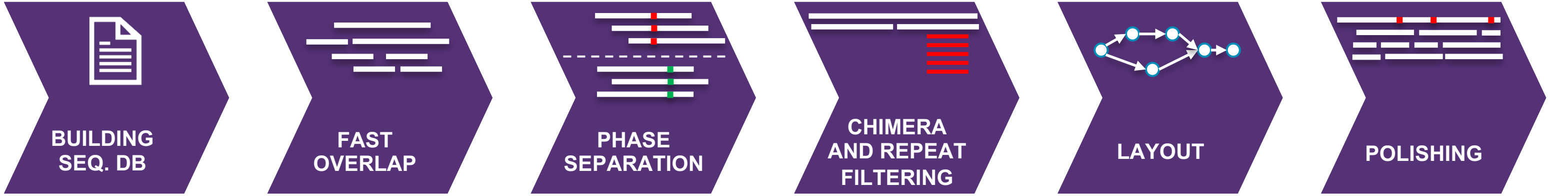
IMPROVED AND PHASED ASSEMBLY (IPA)



- Work in progress
  - Currently in Beta
  - Rapidly being updated

IPA METHODS

IPA WORKFLOW



BUILDING SEQ. DB

FAST OVERLAP

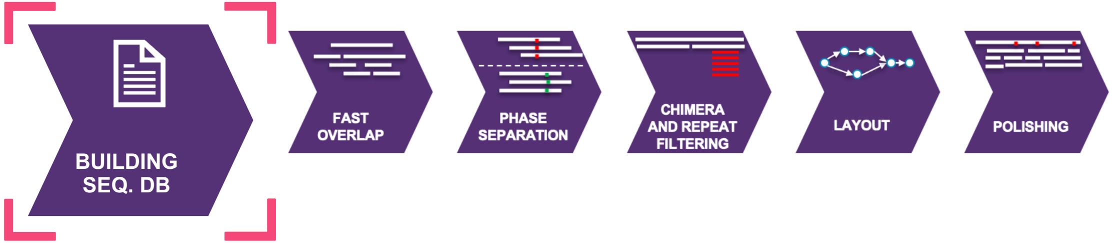
PHASE SEPARATION

CHIMERA AND REPEAT FILTERING

LAYOUT

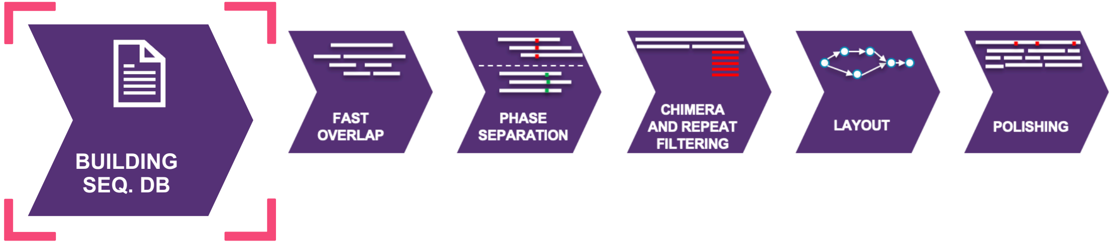
POLISHING

IPA WORKFLOW



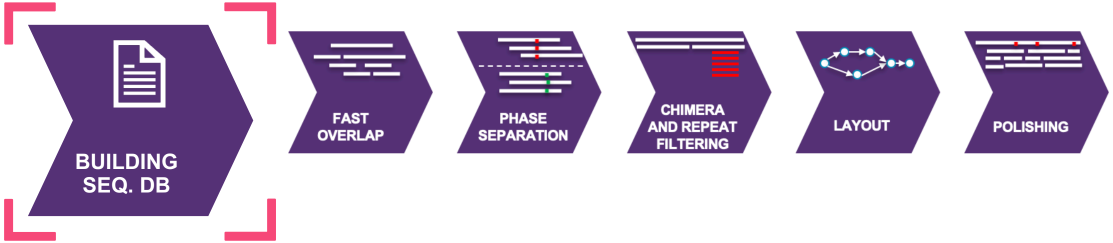
- Sequence Database

IPA WORKFLOW



- **Sequence Database**
  - Converting one or more input files into a unified database format
  - **SeqDB** – database of all input reads, for fast random access
  - **SeedDB** – database of seeds (e.g. minimizers) precomputed from the DB

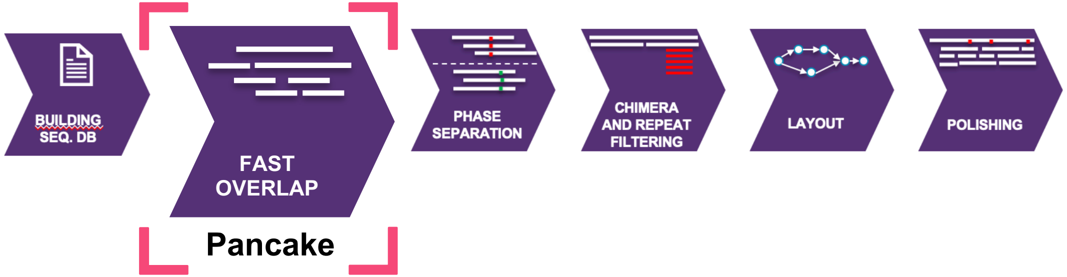
IPA WORKFLOW



Sequence Database

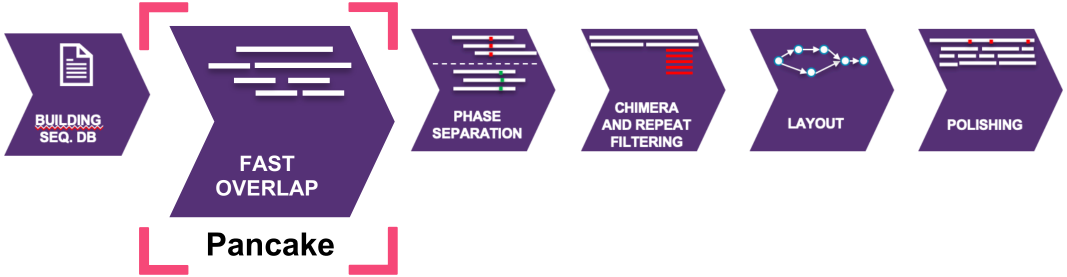
- Supported formats: FASTA, FASTQ, BAM, XML and FOFN (including gzipped FASTA and FASTQ)
- Compression
- Arbitrary method for seed generation in SeedDB
  - Minimizers
  - Full set of dense k-mers
  - **Spaced seeds**
  - Other approaches are trivial to add
- Other features:
  - Fetching sequences in original or homopolymer compressed space
  - Converting headers to IDs

IPA WORKFLOW



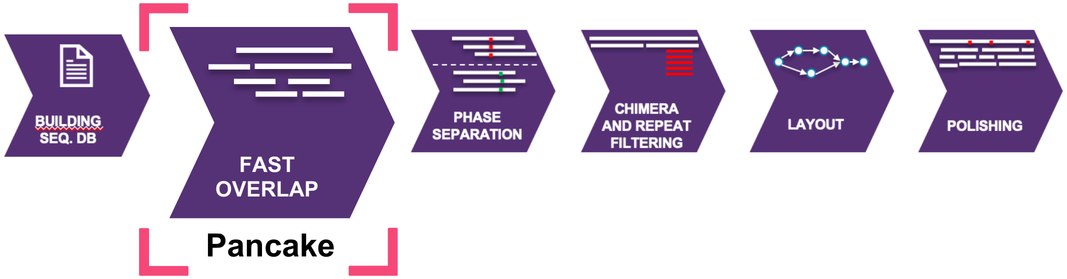
- **Pancake**

IPA WORKFLOW



- **Pancake**
  - New overlapper
  - Extremely fast and accurate
  - Can overlap a 30x NA19240 (18kb) dataset in **20 CPU hrs**

IPA WORKFLOW



BUILDING SEQ. DB

FAST OVERLAP

PHASE SEPARATION

CHIMERA AND REPEAT FILTERING

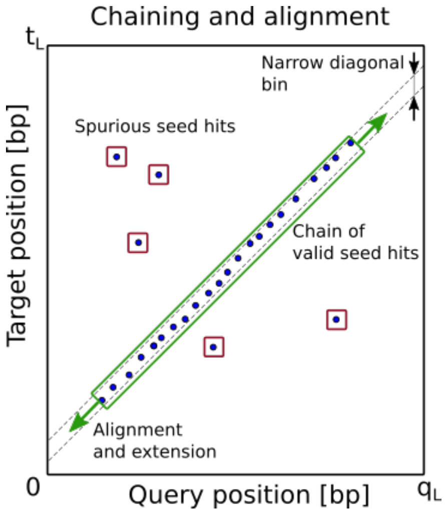
LAYOUT

POLISHING

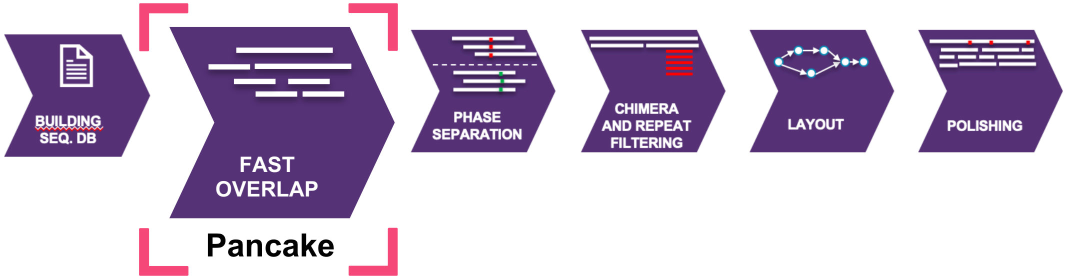
Pancake

Pancake

- Runs on a pair of blocks from the SeqDB (query and target)
- Algorithm:
  - For each read in the query block collect all seed hits in the target block
  - Sort and bin the seed hits in narrow diagonal bins
    - Initial seeding of potential local alignments
  - For each candidate diagonal, perform fast alignment computation and alignment extension to form dovetail overlaps
  - Filter low quality overlaps

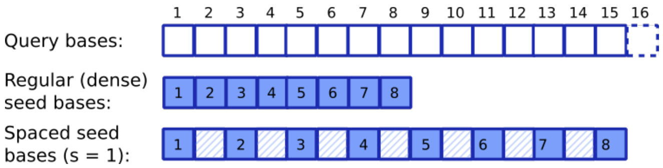


IPA WORKFLOW



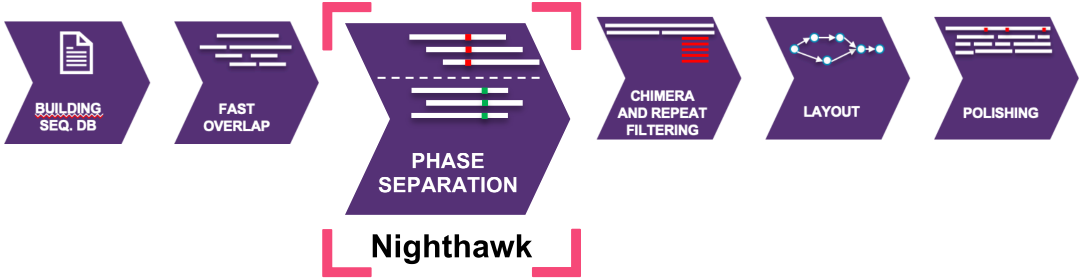
Pancake

- Spaced seeds (minimizers)
  - Novel adaptation in combination with minimizers
  - Seeds are constructed by skipping zero or more bases after every inclusive base
  - Efficient to compute
  - Seeds cover larger regions
  - Increases specificity
  - Minimizer approach applied on spaced seeds



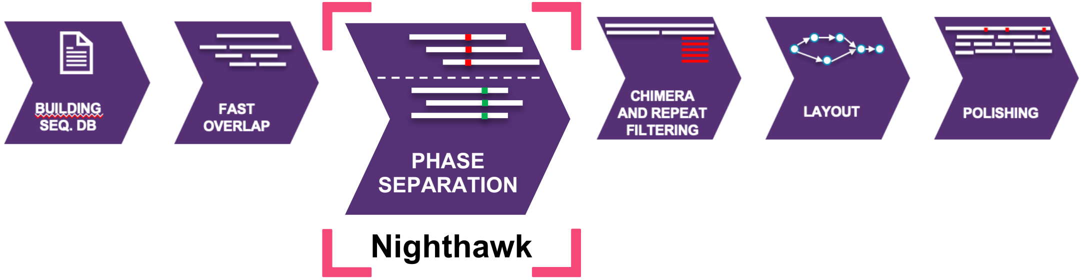
- By default, spacing of 1 is used

IPA WORKFLOW



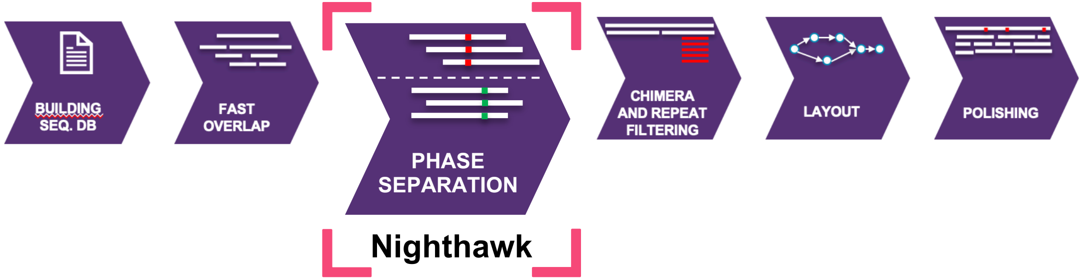
- Nighthawk

IPA WORKFLOW



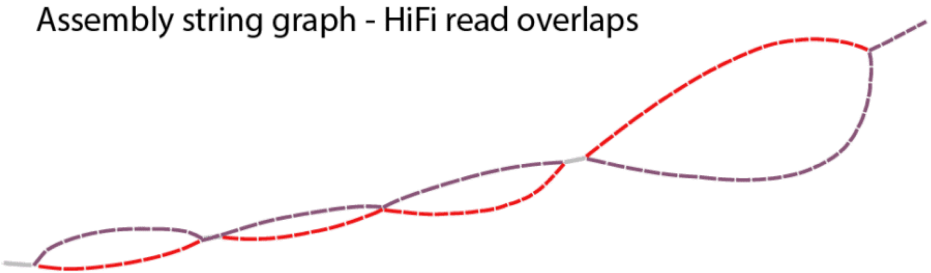
- **Nighthawk**
  - New phasing tool
  - Novel approach based on the de Bruijn graph!
  - Works directly on overlap piles!

IPA WORKFLOW



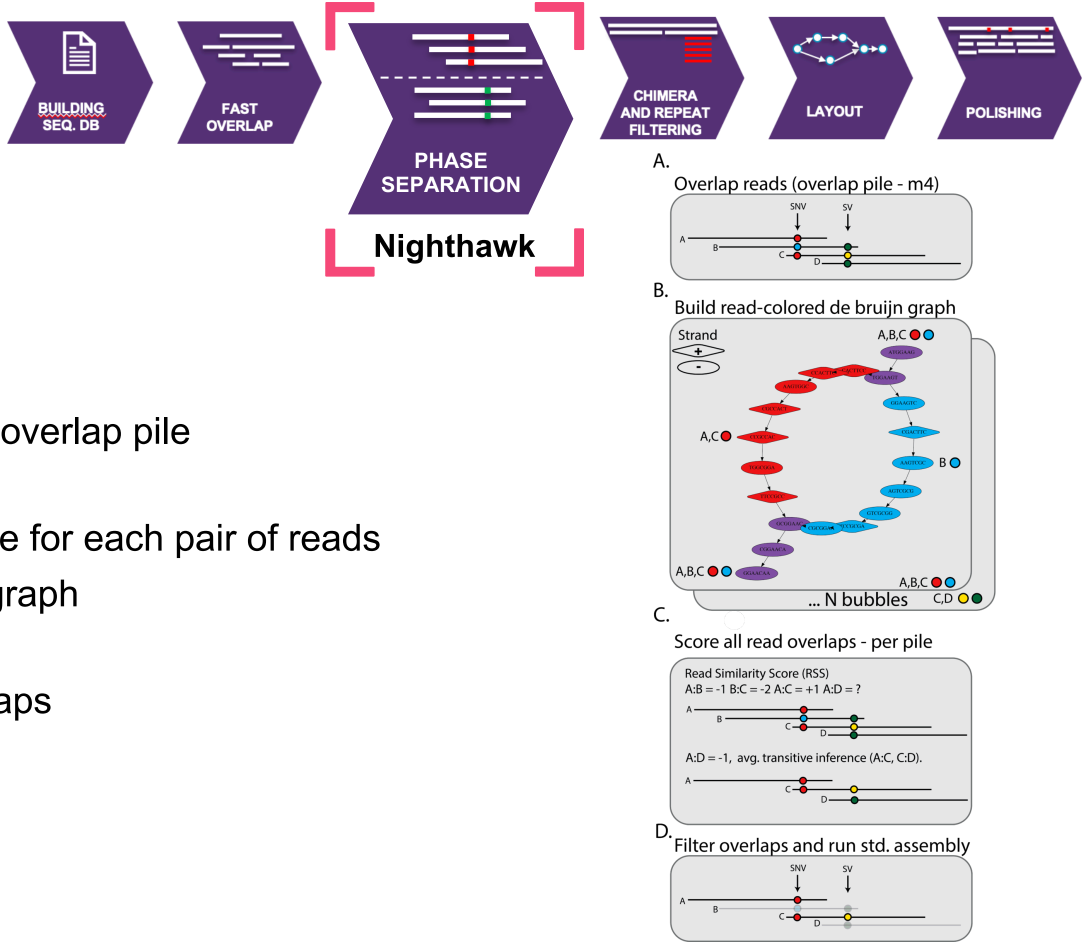
Nighthawk

- Idea:
  - Discover and remove overlaps between reads coming from different haplotypes
  - **Phasing before layout – unlike FALCON-Unzip**
  - Goal: Natural phase separation at the layout stage



Assembly string graph - HiFi read overlaps

IPA WORKFLOW



A.
Overlap reads (overlap pile - m4)

B.
Build read-colored de bruijn graph

C.
Score all read overlaps - per pile

Read Similarity Score (RSS)
A:B = -1 B:C = -2 A:C = +1 A:D = ?

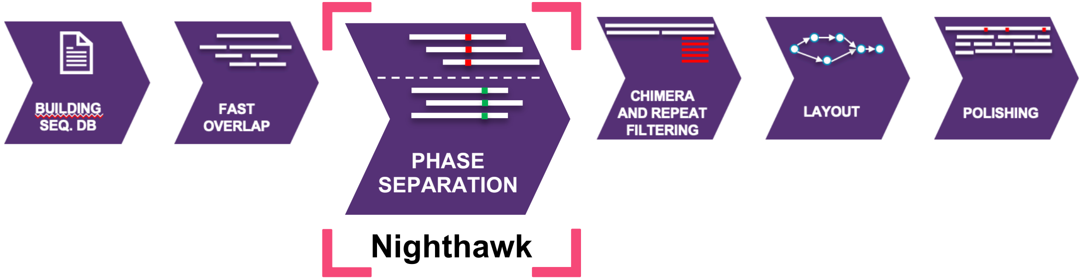
A:D = -1, avg. transitive inference (A:C, C:D).

D.
Filter overlaps and run std. assembly

Nighthawk

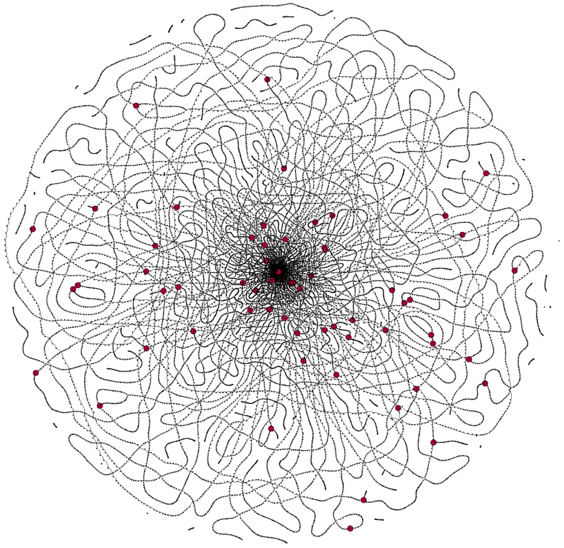
- Algorithm:
  - Builds a de Bruijn graph for each overlap pile
  - Analyzes the bubbles
  - Computes a Read Similarity Score for each pair of reads
  - Phases bubbles in the de Bruijn graph
  - Performs transitive inference
  - Finally – filters cross-phase overlaps

IPA WORKFLOW

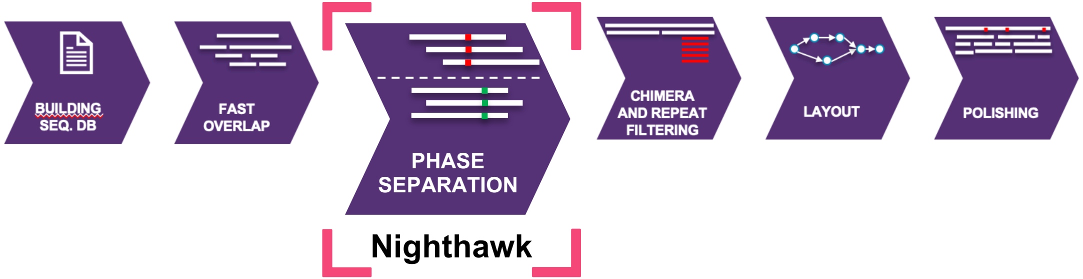


Nighthawk

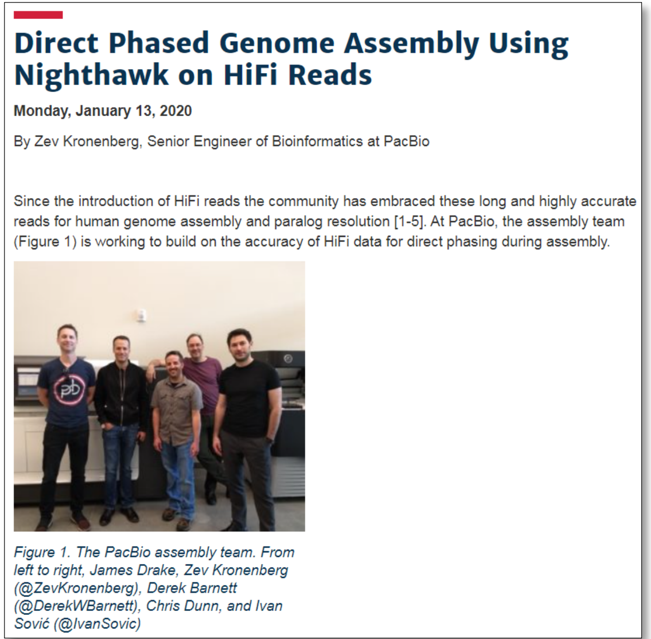
- Visualization of a de Bruijn graph for a pile of reads of a Drosophila HiFi dataset
  - k = 23
  - Red dots – heads of heterozygous bubbles



IPA WORKFLOW



- **Nighthawk**
  - More info on Nighthawk in our blog post:
    - https://www.pacb.com/blog/direct-phased-genome-assembly-using-nighthawk-on-hifi-reads/



Direct Phased Genome Assembly Using Nighthawk on HiFi Reads

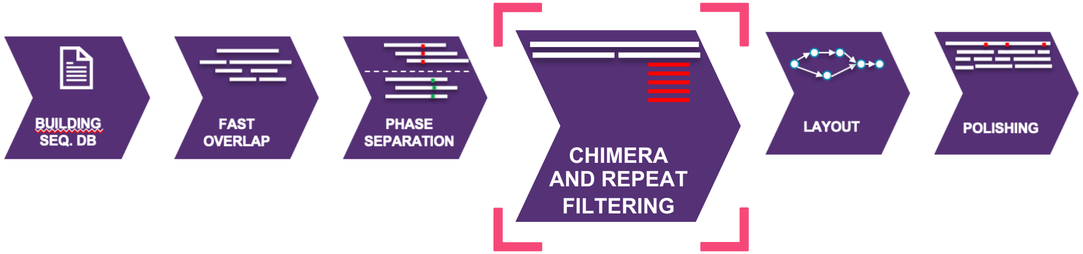
Monday, January 13, 2020

By Zev Kronenberg, Senior Engineer of Bioinformatics at PacBio

Since the introduction of HiFi reads the community has embraced these long and highly accurate reads for human genome assembly and paralog resolution [1-5]. At PacBio, the assembly team (Figure 1) is working to build on the accuracy of HiFi data for direct phasing during assembly.

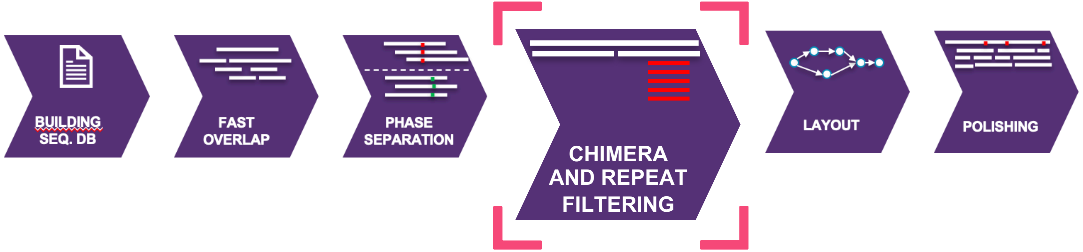
Figure 1. The PacBio assembly team. From left to right, James Drake, Zev Kronenberg (@ZevKronenberg), Derek Barnett (@DerekWBarnett), Chris Dunn, and Ivan Sović (@IvanSovic)

IPA WORKFLOW



- Chimera and repeat filtering

IPA WORKFLOW



BUILDING SEQ. DB

FAST OVERLAP

PHASE SEPARATION

CHIMERA AND REPEAT FILTERING

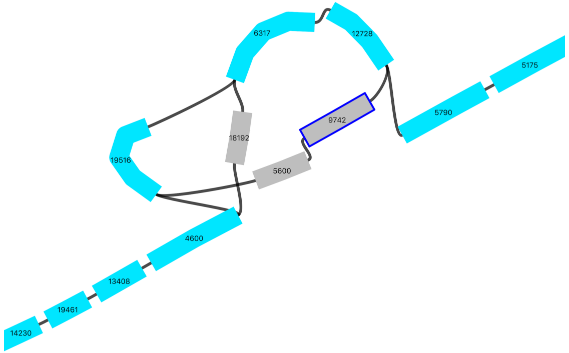
LAYOUT

POLISHING

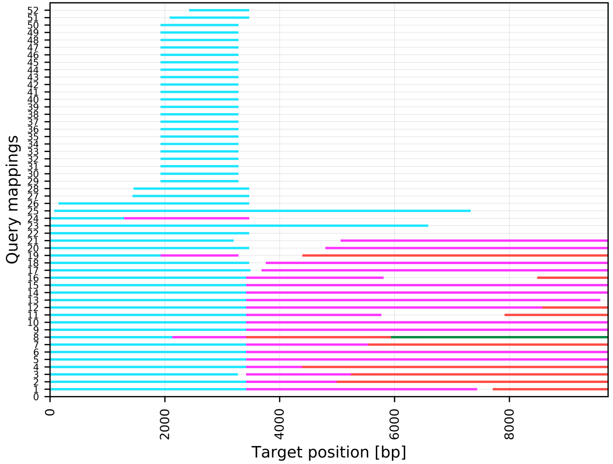
Chimera and repeat filtering

- Small fraction of HiFi reads are molecular chimeras
- Filtering improves contiguity and reduces misassemblies

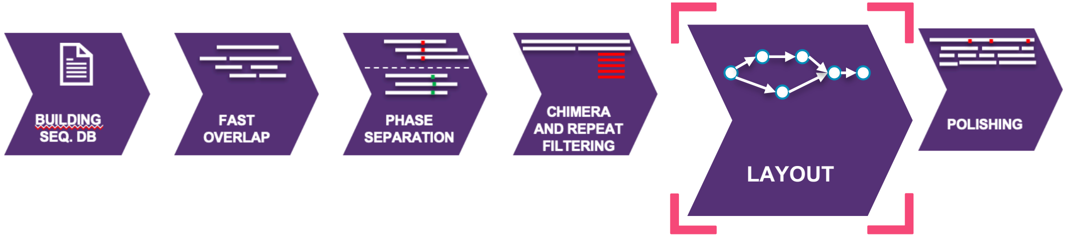
Visualization of chimeric joins in an assembly graph



Overlap pile demonstrating chimera detection

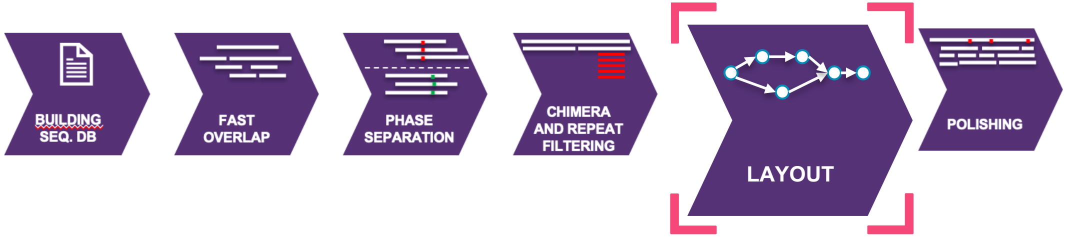


IPA WORKFLOW



- Layout

IPA WORKFLOW

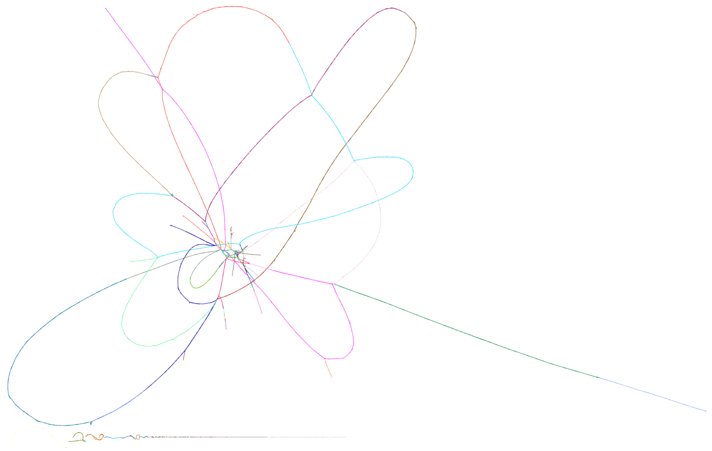


Layout

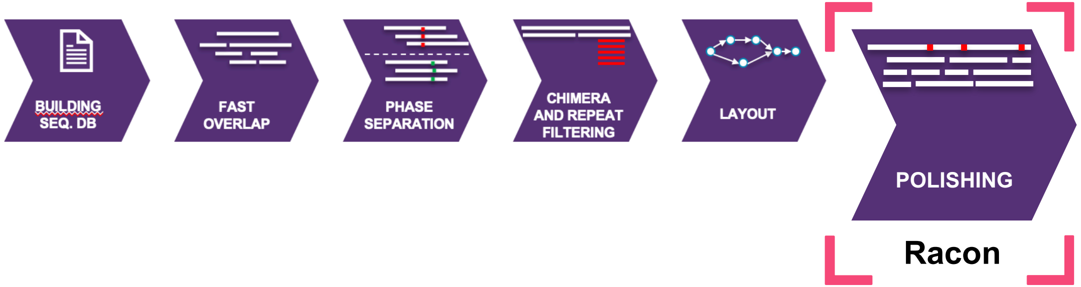
- String graph based
- Polyploid aware
- Primary and associate contig sets

- Phase-aware read tracking
  - Reads assigned to contigs based on phased overlaps
  - Important for polishing

String graph for a Drosophila HiFi dataset

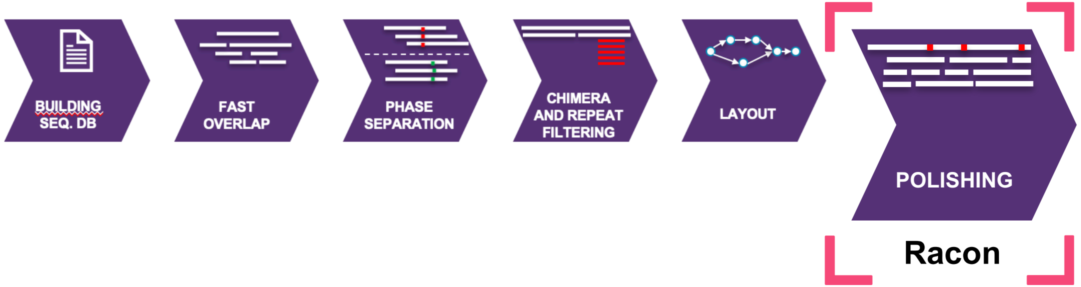


IPA WORKFLOW



- Polishing

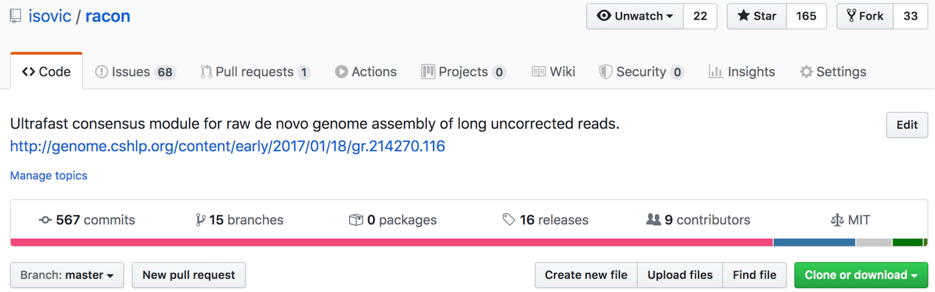
IPA WORKFLOW



– Polishing

- Consumes read-to-contig assignment
- Phase-aware
- Assignment-based mapping using **Pbmm2**

- **Racon**
- Possibility of GPU acceleration


isovic / racon

Ultrafast consensus module for raw de novo genome assembly of long uncorrected reads.
http://genome.cshlp.org/content/early/2017/01/18/gr.214270.116

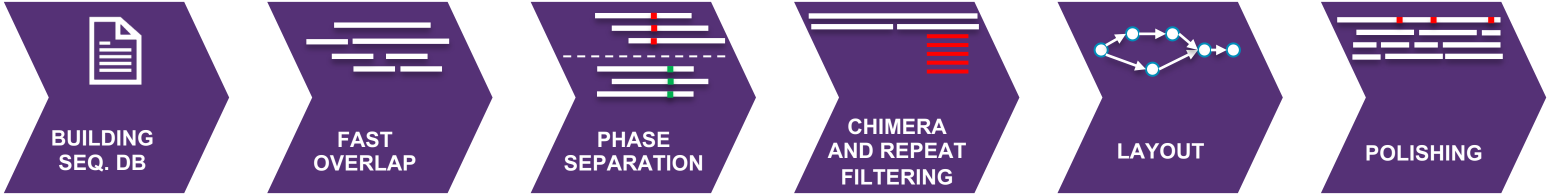
Manage topics

567 commits | 15 branches | 0 packages | 16 releases | 9 contributors | MIT

https://github.com/isovic/racon

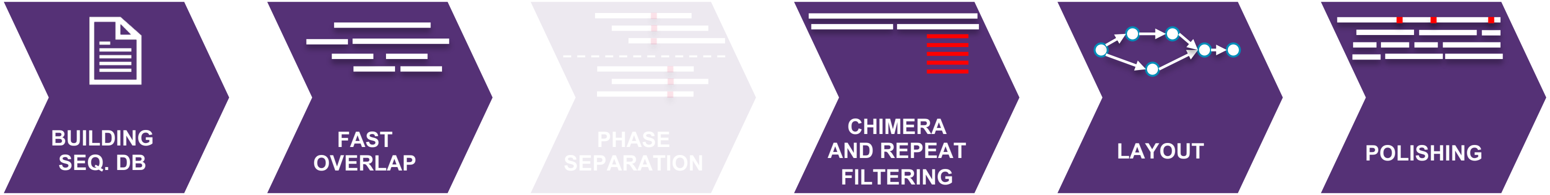
IPA WORKFLOW

- Full workflow with phasing and polishing



IPA WORKFLOW

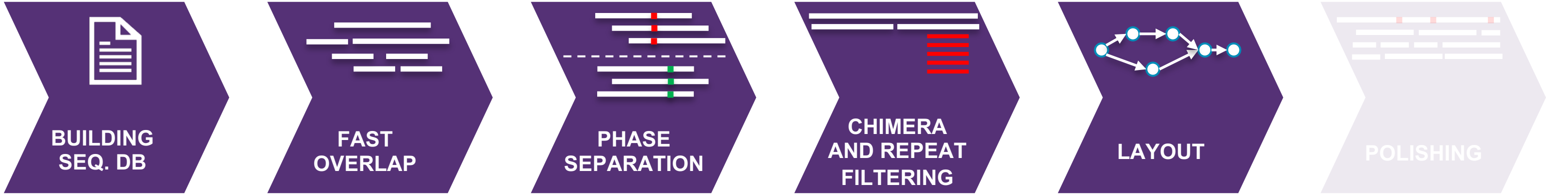
- **Haploid workflow – phasing can optionally be switched on/off**



IPA WORKFLOW

- Polishing can optionally be switched on/off
  - Fast draft assembly

Phased workflow



BUILDING SEQ. DB

FAST OVERLAP

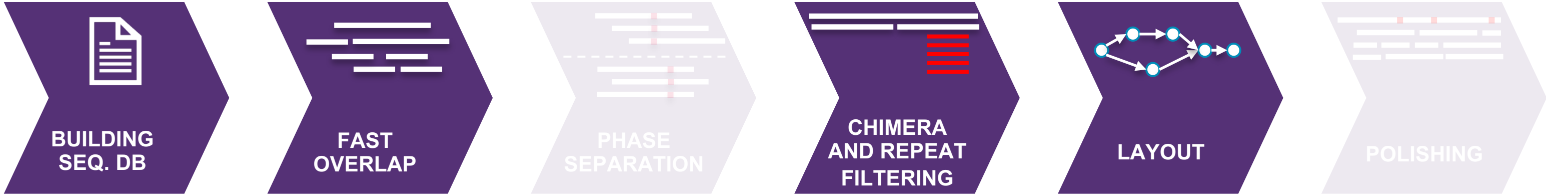
PHASE SEPARATION

CHIMERA AND REPEAT FILTERING

LAYOUT

POLISHING

Haploid workflow



BUILDING SEQ. DB

FAST OVERLAP

PHASE SEPARATION

CHIMERA AND REPEAT FILTERING

LAYOUT

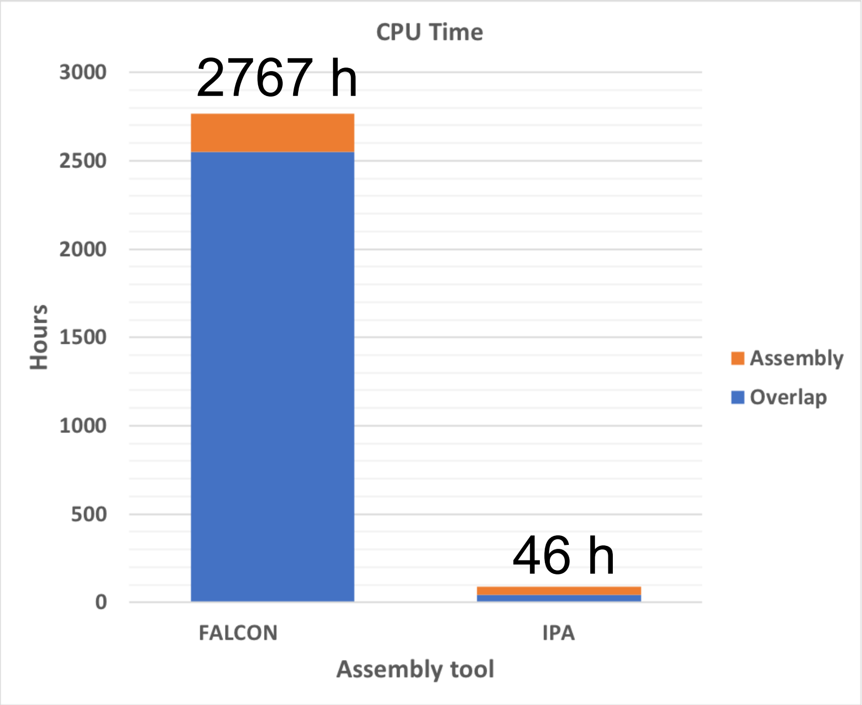
POLISHING

RESULTS

RESULTS: HUMAN ASSEMBLY IS VERY FAST

HPRC HG002 34x Dataset – Haploid workflow without polishing

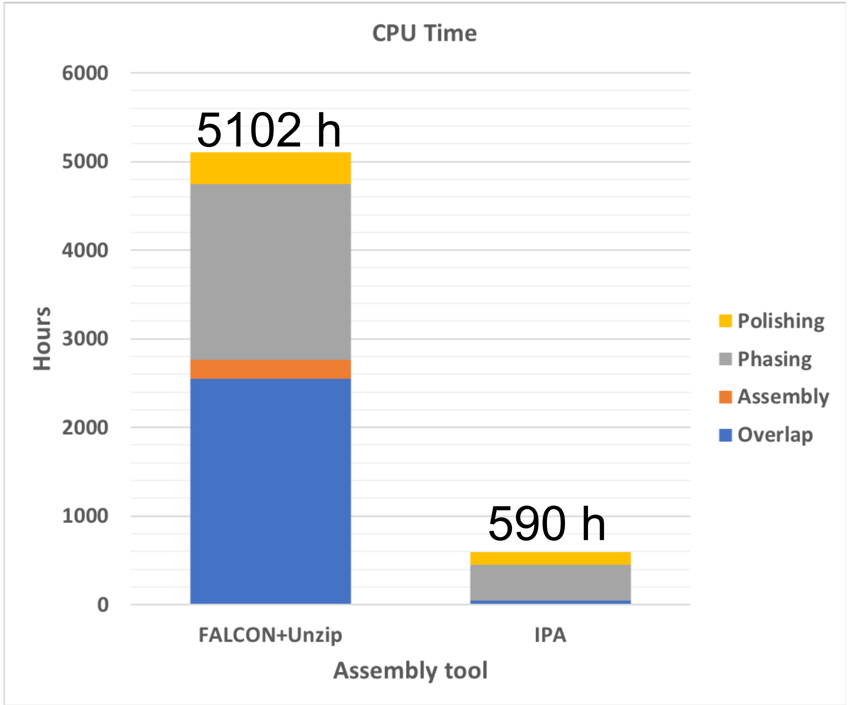
	FALCON	IPA
N50 [Mbp]	31.37	38.81
Max length [Mbp]	110.15	110.72
Total length [Gbp]	2.96	3.06
CPU time [h]	2767	46
		60x Faster!



RESULTS: LONG PHASE BLOCKS IN HUMAN, HIGH BASE QV

HPRC HG002 34x Dataset – Phased workflow with polishing

	FALCON-Unzip		IPA (Phased)	
	primary	haplotigs	primary	haplotigs
N50 [Mbp]	31.40	0.191	33.75	0.352
Max length [Mbp]	110.12	1.62	110.94	2.30
Total length [Gbp]	2.95	1.99	3.02	1.85
CPU time [h]	5102		590	
			8.64x Faster!	



RESULTS: LONG PHASE BLOCKS IN HUMAN, HIGH BASE QV

HPRC HG002 34x Dataset – Phased workflow with polishing

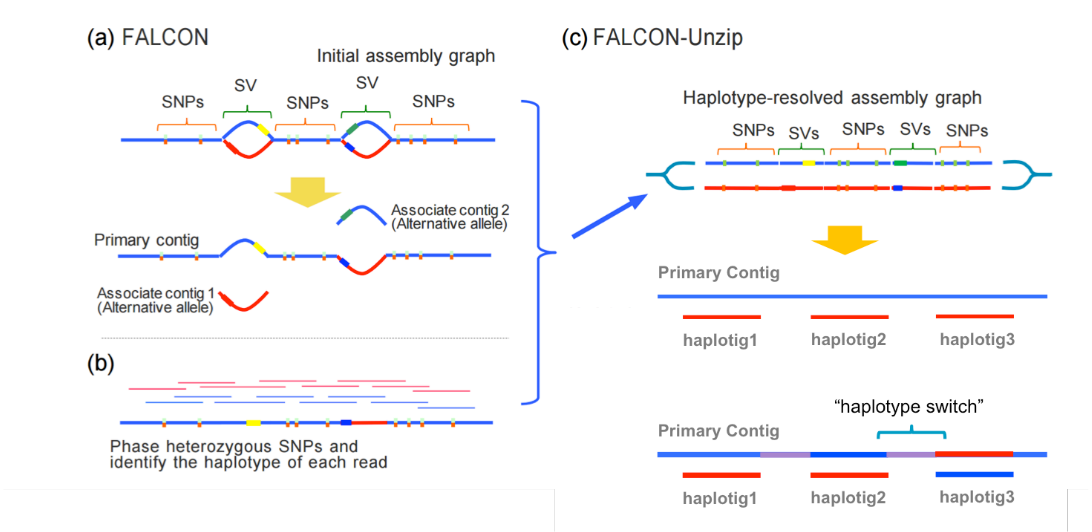
	FALCON-Unzip		IPA (Phased)	
	primary	**haplotigs**	**primary**	**haplotigs**
N50 [Mbp]	31.40	0.191	33.75	0.352
Max length [Mbp]	110.12	1.62	110.94	2.30
Total length [Gbp]	2.95	1.99	3.02	1.85
CPU time [h]	5102		590	
			8.64x Faster!	

- Primary contig pile is slightly larger than expected haploid genome size
- Fully phased regions of the graph can appear as separate graph components

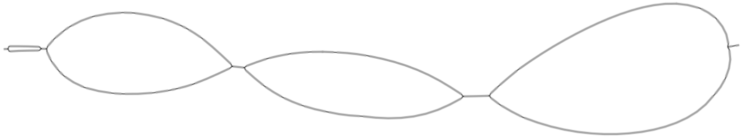
RESULTS: LONG PHASE BLOCKS IN HUMAN, HIGH BASE QV

Purging “duplicate” haplotigs from the primary contig set

- Common expectation when phasing contigs



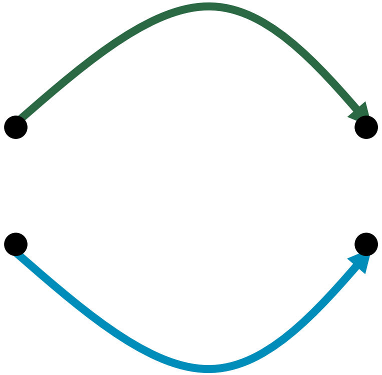
Example of a phased assembly graph



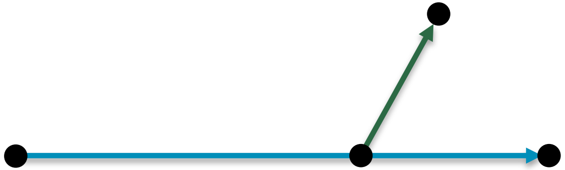
RESULTS: LONG PHASE BLOCKS IN HUMAN, HIGH BASE QV

Purging “duplicate” haplotigs from the primary contig set

Fully phased graph component



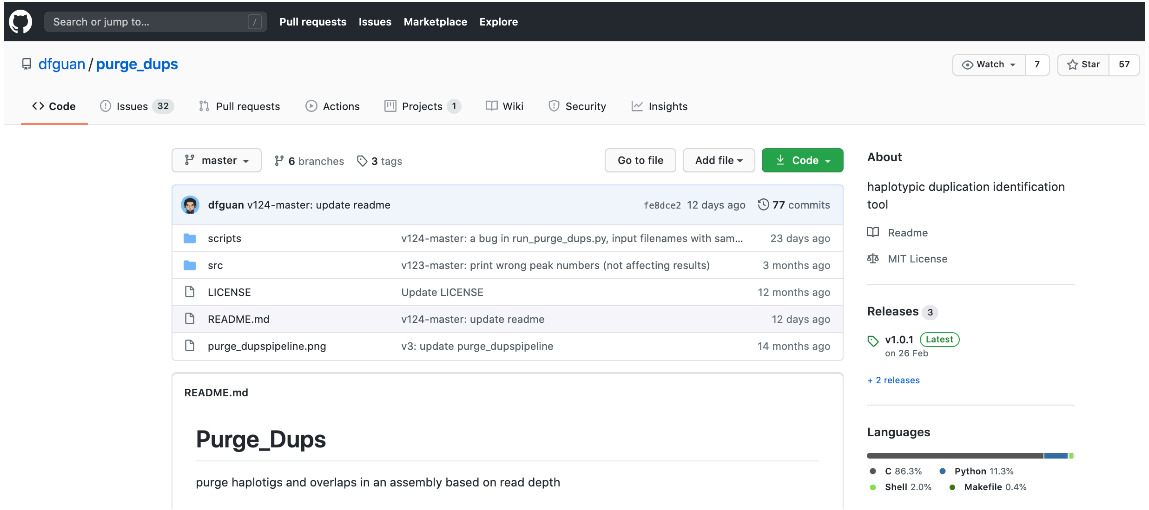
Artifacts in the graph - spurs



RESULTS: LONG PHASE BLOCKS IN HUMAN, HIGH BASE QV

Purging “duplicate” haplotigs from the primary contig set

- Happens to all current assembly tools
- Remedy – publicly available tool “purge_dups”



dfguan / purge_dups

6 branches 3 tags

dfguan v124-master: update readme fe8dce2 12 days ago 77 commits

scripts	v124-master: a bug in run_purge_dups.py, input filenames with sam...	23 days ago
src	v123-master: print wrong peak numbers (not affecting results)	3 months ago
LICENSE	Update LICENSE	12 months ago
README.md	v124-master: update readme	12 days ago
purge_dupspipeline.png	v3: update purge_dupspipeline	14 months ago

README.md

Purge_Dups

purge haplotigs and overlaps in an assembly based on read depth

About

haplotypic duplication identification tool

Readme

MIT License

Releases 3

v1.0.1 Latest
on 26 Feb

+ 2 releases

Languages

● C 86.3% ● Python 11.3%
● Shell 2.0% ● Makefile 0.4%

RESULTS: GREAT HAPLOTIG SEPARATION WITH PURGE DUPS

HPRC HG002 34x Dataset – Phased workflow with polishing

	FALCON-Unzip + Purge dups		IPA (Phased) + Purge dups	
	primary	**haplotigs**	**primary**	**haplotigs**
N50 [Mbp]	33.25	0.195	34.48	0.353
Max length [Mbp]	110.12	1.62	110.94	4.12
Total length [Gbp]	2.87	1.98	2.88	1.94
Base QV	50.6	49.9	50.6	50.2
Phase accuracy	0.706	0.997	0.720	0.980
BUSCO of primary	C:95.1% S:94.2%,D:0.9%		C:95.2% S:94.2%,D:1.0%	

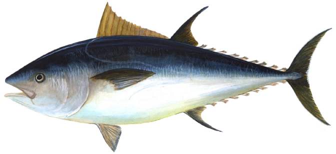
RESULTS: GREAT HAPLOTIG SEPARATION WITH PURGE DUPS

HPRC HG002 34x Dataset – Phased workflow with polishing

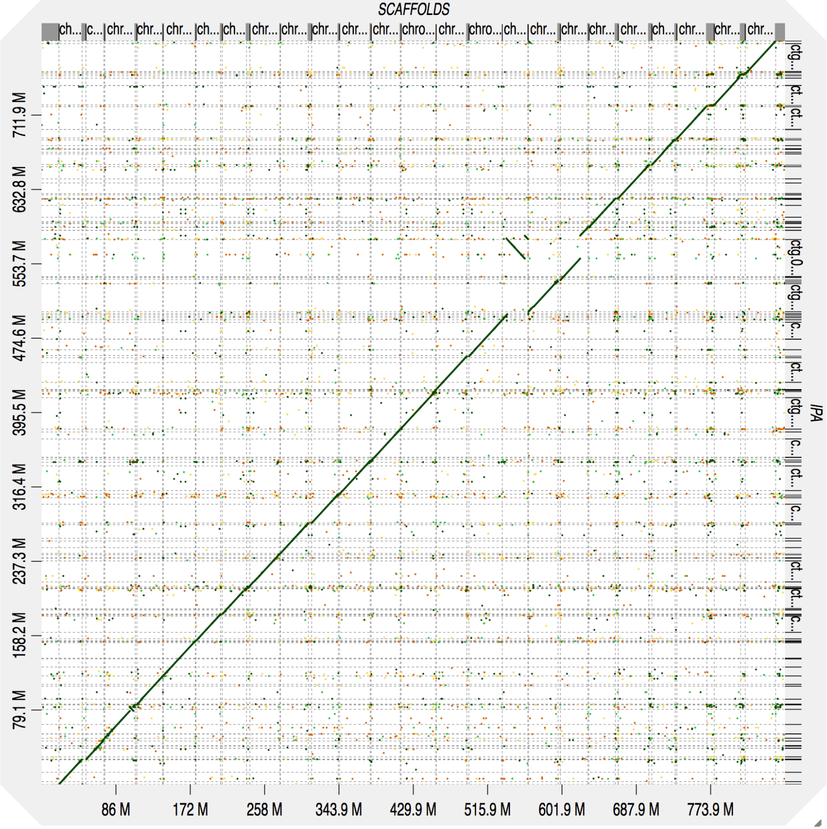
	FALCON-Unzip + Purge dups		IPA (Phased) + Purge dups	
	primary	haplotigs	primary	haplotigs
N50 [Mbp]	33.25	0.195	34.48	0.353
Max length [Mbp]	110.12	1.62	110.94	4.12
Total length [Gbp]	2.87	1.98	2.88	1.94
Base QV	50.6	49.9	50.6	50.2
Phase accuracy	0.706	0.997	0.720	0.980
BUSCO of primary	C:95.1% S:94.2%,D:0.9%		C:95.2% S:94.2%,D:1.0%	

RESULTS: HIGHLY ACCURATE CONTIG ASSEMBLY

Atlantic Bluefin Tuna – Phased workflow



	IPA (Phased)		IPA + purge_dups	
	primary	haplotigs	primary	haplotigs
N50 [Mbp]	9.34	3.70	13.80	3.83
Max length [Mbp]	39.38	13.95	39.38	19.49
Total length [Gbp]	1.26	0.280	0.791	0.744
BUSCO of primary	C:97.4% S:44.6%,D:52.8%		C:97.7% S:95.1%,D:2.6%	

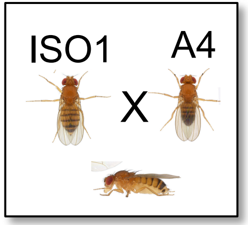


* Credit for the ABFT data, reference assembly and scaffolding: Paul Peluso[1], Greg Concepcion[1], Jay Ghurye[2], Nathan Truelove[3] and Barbara Block[4]

[1] Pacific Biosciences, [2] Dovetail Genomics, [3] Monterey Bay Aquarium Research Institute, [4] Hopkins Marine Station

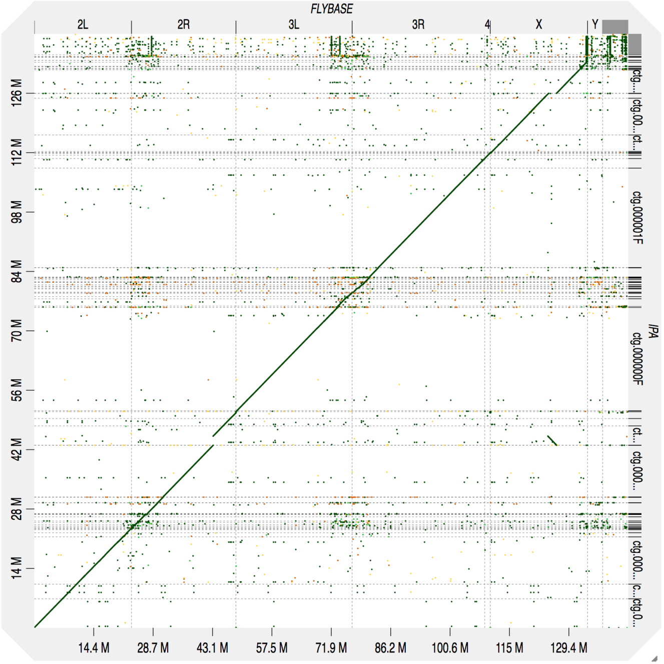
* Cabanettes F, Klopp C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958 https://doi.org/10.7717/peerj.4958

RESULTS: HIGH PHASE ACCURACY



Drosophila melanogaster F1 – Phased and polished

	Hifiasm + purge_dups		IPA + purge_dups	
	primary	haplotigs	primary	haplotigs
N50 [Mbp]	22.55	1.28	13.49	2.42
Max length [Mbp]	28.13	6.81	23.47	12.48
Total length [Mbp]	160.19	149.87	134.19	115.26
Base QV	48.1	47.4	47.97	46.87
Phase accuracy	0.788	0.998	0.826	0.999
BUSCO of primary	C:98.5% S:98.0%,D:0.5%		C:98.7% S:98.2%,D:0.5%	

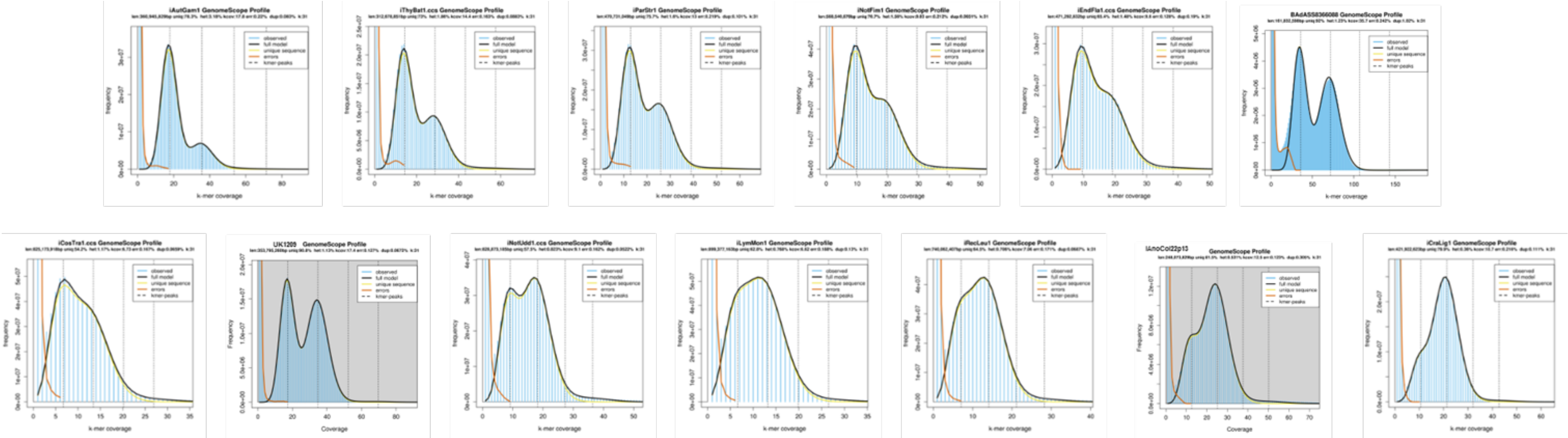


* Cabanettes F, Klopp C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958 https://doi.org/10.7717/peerj.4958

RESULTS: BUG GENOMES

Testing on real-world samples - butterflies, moths & mosquitoes

Darwin Tree of Life, Sanger


iAutGam1 GenomeScope Profile
iThyBat1.ccs GenomeScope Profile
iParStr1 GenomeScope Profile
iNotFim1 GenomeScope Profile
iEndFla1.ccs GenomeScope Profile
BAdASS8366088 GenomeScope Profile
iCosTra1.ccs GenomeScope Profile
UK1209 GenomeScope Profile
iNotUdd1.ccs GenomeScope Profile
iLymMon1 GenomeScope Profile
iRecLeu1 GenomeScope Profile
iAnoCol22p13 GenomeScope Profile
iCraLig1 GenomeScope Profile
observed
full model
unique sequence
errors
kmer-peaks
frequency
k-mer coverage
Coverage

https://www.darwintreeoflife.org
https://github.com/darwintreeoflife/darwintreeoflife.data

RESULTS: BUG GENOMES

ALL SAMPLES: IPA *vs.* FALCON

	Primary										Haplotigs									
	# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV		# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV	
Species	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**	**Falcon**	**IPA**
ilAutoGam1	85	79	10.97	12.01	381	368	99.0%	99.2%	47.1	47.1	378	1613	6.98	9.09	330	379	93.8%	97.9%	46.3	44.9
ilCosmTra1	2173	1924	0.79	0.88	872	862	97.9%	97.5%	44.4	45.3	3932	3505	0.27	0.32	717	742	85.6%	89.8%	44.1	43.3
ilCranLig1	220	267	7.32	3.50	438	436	99.0%	98.9%	46.1	46.8	4218	1241	0.07	0.42	246	266	57.3%	67.7%	39.1	45.6
ilEndoFla1	623	489	1.46	1.92	492	489	98.6%	99.2%	46.5	45.9	2110	1587	0.29	0.56	375	418	82.7%	89.0%	45.9	46.3
ilLymaMon1	251	301	10.31	5.77	917	912	99.2%	99.2%	46.5	46.8	5525	3162	0.17	0.47	610	633	64.6%	66.4%	41.6	43.8
ilNoctFim1	382	307	3.12	3.75	576	572	98.2%	98.9%	47.6	46.3	1493	1152	0.86	1.62	514	545	83.9%	94.9%	45.1	48.1
ilNotoUdd1	1128	809	1.52	2.07	829	814	98.9%	98.3%	45.5	46.8	4115	2791	0.26	0.69	629	780	75.3%	94.0%	42.8	45.6
ilParaStr1	218	129	4.42	6.92	480	481	99.4%	99.4%	48.0	47.4	1166	1309	0.64	1.78	413	431	84.3%	84.7%	47.1	46.9
ilRecuLeu1	1268	1247	1.09	0.98	748	746	98.5%	98.2%	44.4	45.1	4890	4029	0.12	0.23	416	534	54.1%	70.4%	42.0	43.4
ilThyaBat1	186	88	3.29	7.15	316	316	98.5%	98.9%	46.2	47.0	650	1073	0.89	2.77	296	327	91.6%	96.0%	46.9	45.3
ilVaneAta1	80	48	10.13	12.12	368	368	99.1%	99.3%	47.9	48.3	2185	1444	0.14	4.80	205	369	60.1%	97.7%	45.3	46.9
idAnopAqu88	260	79	18.22	15.05	188	181	99.2%	97.8%	49.5	50.6	1056	1833	0.11	4.02	90	211	43.2%	94.2%	48.9	43.8
idAnopCol22p13	272	189	5.48	5.01	260	260	99.4%	99.2%	49.4	49.5	1112	2474	0.12	0.23	77	148	27.1%	52.5%	43.9	40.3

RESULTS: BUG GENOMES

ALL SAMPLES: IPA *vs.* FALCON

	Primary										Haplotigs									
	# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV		# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV	
Species	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA
ilAutoGam1	85	79	10.97	12.01	381	368	99.0%	99.2%	47.1	47.1	378	1613	6.98	9.09	330	379	93.8%	97.9%	46.3	44.9
ilCosmTra1	2173	1924	0.79	0.88	872	862	97.9%	97.5%	44.4	45.3	3932	3505	0.27	0.32	717	742	85.6%	89.8%	44.1	43.3
ilCranLig1	220	267	7.32	3.50	438	436	99.0%	98.9%	46.1	46.8	4218	1241	0.07	0.42	246	266	57.3%	67.7%	39.1	45.6
ilEndoFla1	623	489	1.46	1.92	492	489	98.6%	99.2%	46.5	45.9	2110	1587	0.29	0.56	375	418	82.7%	89.0%	45.9	46.3
ilLymaMon1	251	301	10.31	5.77	917	912	99.2%	99.2%	46.5	46.8	5525	3162	0.17	0.47	610	633	64.6%	66.4%	41.6	43.8
ilNoctFim1	382	307	3.12	3.75	576	572	98.2%	98.9%	47.6	46.3	1493	1152	0.86	1.62	514	545	83.9%	94.9%	45.1	48.1
ilNotoUdd1	1128	809	1.52	2.07	829	814	98.9%	98.3%	45.5	46.8	4115	2791	0.26	0.69	629	780	75.3%	94.0%	42.8	45.6
ilParaStr1	218	129	4.42	6.92	480	481	99.4%	99.4%	48.0	47.4	1166	1309	0.64	1.78	413	431	84.3%	84.7%	47.1	46.9
ilRecuLeu1	1268	1247	1.09	0.98	748	746	98.5%	98.2%	44.4	45.1	4890	4029	0.12	0.23	416	534	54.1%	70.4%	42.0	43.4
ilThyaBat1	186	88	3.29	7.15	316	316	98.5%	98.9%	46.2	47.0	650	1073	0.89	2.77	296	327	91.6%	96.0%	46.9	45.3
ilVaneAta1	80	48	10.13	12.12	368	368	99.1%	99.3%	47.9	48.3	2185	1444	0.14	4.80	205	369	60.1%	97.7%	45.3	46.9
idAnopAqu88	260	79	18.22	15.05	188	181	99.2%	97.8%	49.5	50.6	1056	1833	0.11	4.02	90	211	43.2%	94.2%	48.9	43.8
idAnopCol22p13	272	189	5.48	5.01	260	260	99.4%	99.2%	49.4	49.5	1112	2474	0.12	0.23	77	148	27.1%	52.5%	43.9	40.3

- □ within 10%
- 🟩 >10% better
- 🟩 >25% better
- 🟥 worse

RESULTS: BUG GENOMES

ALL SAMPLES: IPA *vs.* FALCON

	Primary										Haplotigs									
	# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV		# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV	
Species	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA
ilAutoGam1	85	79	10.97	12.01	381	368	99.0%	99.2%	47.1	47.1	378	1613	6.98	9.09	330	379	93.8%	97.9%	46.3	44.9
ilCosmTra1	2173	1924	0.79	0.88	872	862	97.9%	97.5%	44.4	45.3	3932	3505	0.27	0.32	717	742	85.6%	89.8%	44.1	43.3
ilCranLig1	220	267	7.32	3.50	438	436	99.0%	98.9%	46.1	46.8	4218	1241	0.07	0.42	246	266	57.3%	67.7%	39.1	45.6
ilEndoFla1	623	489	1.46	1.92	492	489	98.6%	99.2%	46.5	45.9	2110	1587	0.29	0.56	375	418	82.7%	89.0%	45.9	46.3
ilLymaMon1	251	301	10.31	5.77	917	912	99.2%	99.2%	46.5	46.8	5525	3162	0.17	0.47	610	633	64.6%	66.4%	41.6	43.8
ilNoctFim1	382	307	3.12	3.75	576	572	98.2%	98.9%	47.6	46.3	1493	1152	0.86	1.62	514	545	83.9%	94.9%	45.1	48.1
ilNotoUdd1	1128	809	1.52	2.07	829	814	98.9%	98.3%	45.5	46.8	4115	2791	0.26	0.69	629	780	75.3%	94.0%	42.8	45.6
ilParaStr1	218	129	4.42	6.92	480	481	99.4%	99.4%	48.0	47.4	1166	1309	0.64	1.78	413	431	84.3%	84.7%	47.1	46.9
ilRecuLeu1	1268	1247	1.09	0.98	748	746	98.5%	98.2%	44.4	45.1	4890	4029	0.12	0.23	416	534	54.1%	70.4%	42.0	43.4
ilThyaBat1	186	88	3.29	7.15	316	316	98.5%	98.9%	46.2	47.0	650	1073	0.89	2.77	296	327	91.6%	96.0%	46.9	45.3
ilVaneAta1	80	48	10.13	12.12	368	368	99.1%	99.3%	47.9	48.3	2185	1444	0.14	4.80	205	369	60.1%	97.7%	45.3	46.9
idAnopAqu88	260	79	18.22	15.05	188	181	99.2%	97.8%	49.5	50.6	1056	1833	0.11	4.02	90	211	43.2%	94.2%	48.9	43.8
idAnopCol22p13	272	189	5.48	5.01	260	260	99.4%	99.2%	49.4	49.5	1112	2474	0.12	0.23	77	148	27.1%	52.5%	43.9	40.3

□	within 10%
🟩	>10% better
🟩	>25% better
🟥	worse

Very similar in size,
completeness & accuracy

RESULTS: BUG GENOMES

ALL SAMPLES: IPA *vs.* FALCON

	Primary										Haplotigs									
	# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV		# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV	
Species	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA	Falcon	IPA
ilAutoGam1	85	79	10.97	12.01	381	368	99.0%	99.2%	47.1	47.1	378	1613	6.98	9.09	330	379	93.8%	97.9%	46.3	44.9
ilCosmTra1	2173	1924	0.79	0.88	872	862	97.9%	97.5%	44.4	45.3	3932	3505	0.27	0.32	717	742	85.6%	89.8%	44.1	43.3
ilCranLig1	220	267	7.32	3.50	438	436	99.0%	98.9%	46.1	46.8	4218	1241	0.07	0.42	246	266	57.3%	67.7%	39.1	45.6
ilEndoFla1	623	489	1.46	1.92	492	489	98.6%	99.2%	46.5	45.9	2110	1587	0.29	0.56	375	418	82.7%	89.0%	45.9	46.3
ilLymaMon1	251	301	10.31	5.77	917	912	99.2%	99.2%	46.5	46.8	5525	3162	0.17	0.47	610	633	64.6%	66.4%	41.6	43.8
ilNoctFim1	382	307	3.12	3.75	576	572	98.2%	98.9%	47.6	46.3	1493	1152	0.86	1.62	514	545	83.9%	94.9%	45.1	48.1
ilNotoUdd1	1128	809	1.52	2.07	829	814	98.9%	98.3%	45.5	46.8	4115	2791	0.26	0.69	629	780	75.3%	94.0%	42.8	45.6
ilParaStr1	218	129	4.42	6.92	480	481	99.4%	99.4%	48.0	47.4	1166	1309	0.64	1.78	413	431	84.3%	84.7%	47.1	46.9
ilRecuLeu1	1268	1247	1.09	0.98	748	746	98.5%	98.2%	44.4	45.1	4890	4029	0.12	0.23	416	534	54.1%	70.4%	42.0	43.4
ilThyaBat1	186	88	3.29	7.15	316	316	98.5%	98.9%	46.2	47.0	650	1073	0.89	2.77	296	327	91.6%	96.0%	46.9	45.3
ilVaneAta1	80	48	10.13	12.12	368	368	99.1%	99.3%	47.9	48.3	2185	1444	0.14	4.80	205	369	60.1%	97.7%	45.3	46.9
idAnopAqu88	260	79	18.22	15.05	188	181	99.2%	97.8%	49.5	50.6	1056	1833	0.11	4.02	90	211	43.2%	94.2%	48.9	43.8
idAnopCol22p13	272	189	5.48	5.01	260	260	99.4%	99.2%	49.4	49.5	1112	2474	0.12	0.23	77	148	27.1%	52.5%	43.9	40.3

□ within 10%

🟩 >10% better

🟩 >25% better

🟥 worse

Very similar in size, completeness & accuracy

🟦 Improved haplotype separation

🟦 >90% haplotype resolved

RESULTS: BUG GENOMES

ALL SAMPLES: IPA *vs.* HICANU[1]

	Primary										Haplotigs									
	# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV		# contigs		Contig N50 (Mb)		Size (Mb)		BUSCO C		QV	
Species	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA	HiCanu	IPA
ilAutoGam1	128	79	12.17	12.01	375	368	99.1%	99.2%	47.1	47.1	643	1613	8.14	9.09	356	379	97.0%	97.9%	46.5	44.9
ilCosmTra1		1924		0.88		862		97.5%		45.3		3505		0.32		742		89.8%		43.3
ilCranLig1	211	267	4.94	3.50	435	436	99.1%	98.9%	48.7	46.8	2268	1241	0.36	0.42	425	266	95.2%	67.7%	46.6	45.6
ilEndoFla1	662	489	1.56	1.92	488	489	99.3%	99.2%	46.8	45.9	2155	1587	0.39	0.56	442	418	95.0%	89.0%	46.1	46.3
ilLymaMon1	157	301	13.59	5.77	912	912	99.2%	99.2%	47.3	46.8	4208	3162	0.57	0.47	950	633	96.3%	66.4%	45.7	43.8
ilNoctFim1	783	307	1.88	3.75	577	572	98.8%	98.9%	48.2	46.3	1258	1152	0.69	1.62	530	545	96.8%	94.9%	47.6	48.1
ilNotoUdd1	1147	809	1.56	2.07	826	814	99.1%	98.3%	46.1	46.8	3193	2791	0.49	0.69	802	780	95.4%	94.0%	46.2	45.6
ilParaStr1	105	129	11.77	6.92	482	481	99.1%	99.4%	48.3	47.4	789	1309	1.94	1.78	465	431	95.5%	84.7%	47.5	46.9
ilReculeu1		1247		0.98		746		98.2%		45.1		4029		0.23		534		70.4%		43.4
ilThyaBat1	214	88	3.31	7.15	319	316	98.4%	98.9%	44.3	47.0	900	1073	0.81	2.77	318	327	97.4%	96.0%	43.7	45.3
ilVaneAta1	242	48	12.18	12.12	372	368	99.1%	99.3%	48.6	48.3	769	1444	4.31	4.80	357	369	97.7%	97.7%	47.8	46.9
idAnopAqu88		79		15.05		181		97.8%		50.6		1833		4.02		211		94.2%		43.8
idAnopCol22p13		189		5.01		260		99.2%		49.5		2474		0.23		148		52.5%		40.3

□ within 10%

🟩 >10% better

🟩 >25% better

🟥 worse

Very similar in size,
completeness & accuracy

Darwin TREE of LIFE

🟦 Improved haplotype separation

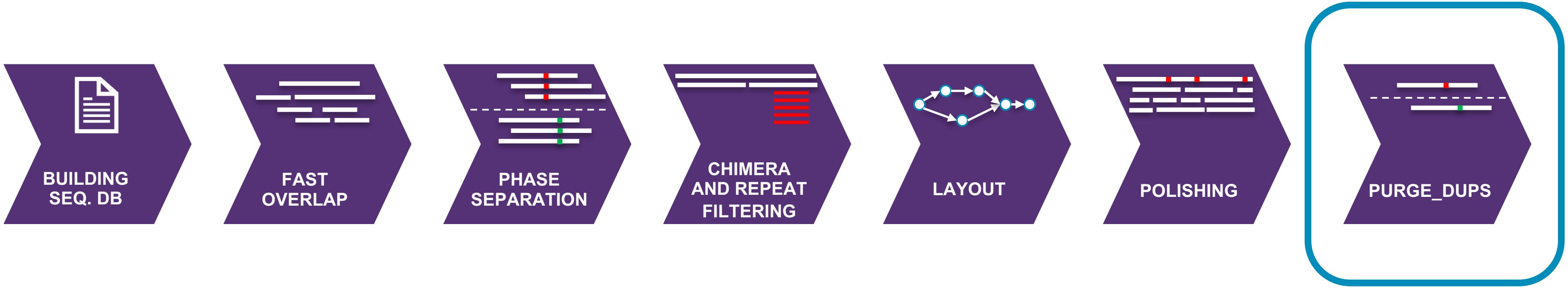
🟦 >90% haplotype resolved

[1] https://github.com/darwintreeoflife/darwintreeoflife.data/

UPCOMING FEATURES

UPCOMING FEATURES

- Integration of “purge_dups” into the workflow

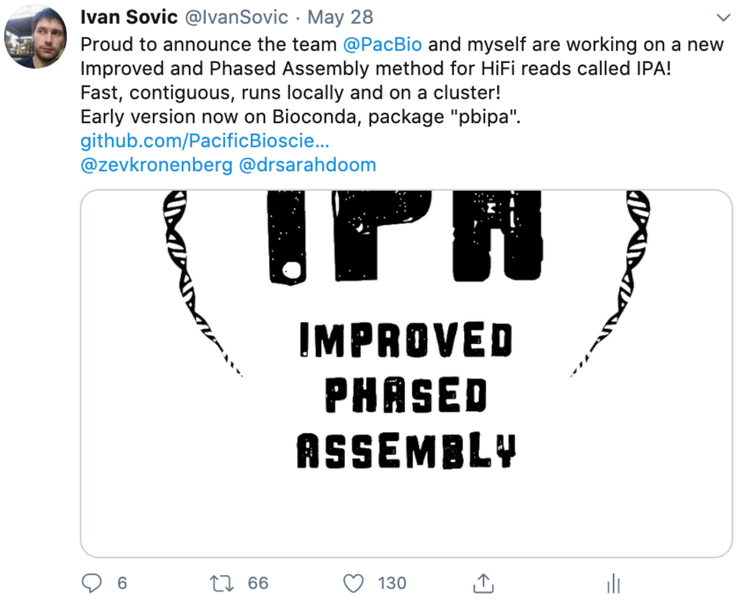


- Phasing improvements
- Read tracking improvements for better polishing

AVAILABILITY, INSTALLATION AND USAGE

AVAILABILITY

- IPA available on Bioconda!
- More details and documentation available here:
  - https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assembler
  - https://github.com/PacificBiosciences/pbipa
  - https://github.com/PacificBiosciences/pbbioconda



Ivan Sovic @IvanSovic · May 28

Proud to announce the team @PacBio and myself are working on a new Improved and Phased Assembly method for HiFi reads called IPA!
Fast, contiguous, runs locally and on a cluster!
Early version now on Bioconda, package "pbipa".
github.com/PacificBioscie...
@zevkronenberg @drsarahdoom

INSTALLATION AND USAGE

- Installation

```
conda create -n ipa -c bioconda -c conda-forge -c defaults
conda activate ipa
conda install pbipa
```

- Run assembly on a local machine:

```
ipa local --nthreads 24 --njobs 1 -i <reads.fasta>
```

- Run assembly on an SGE cluster:

```
ipa dist --nthreads 24 --njobs 40 -i <reads.fasta> \
      --cluster-args 'qsub -S /bin/bash -N ipa.{rule} -cwd -q default -pe smp {params.num_threads} -e
qsub_log/ -o qsub_log/ -V'
```

SUMMARY

- IPA delivers highly accurate and contiguous assemblies, with high speed and accurately phased haplotig regions
  - Generates true haplotigs constructed through a phasing process

- Polishes the phased genome to achieve **>Q50** accuracy!

- Further evaluations and developments ongoing

- Potential for IPA & HiCanu to learn from each other

- Ease of use!

- **Work in progress:**
  - Integrate ”purge_dups” directly into the workflow
  - Improve contiguity of the phased assembly
  - Optimization of all stages

THANK YOU!

IPA
IMPROVED
PHASED
ASSEMBLY

PACBIO®

Darwin TREE of LIFE

IPA TEAM

Ivan Sović
Zev Kronenberg
Christopher Dunn
Sarah Kingan
Derek Barnett
James Drake
Jonas Korlach

PACBIO

Armin Töpfer
Paul Peluso
Greg Concepcion

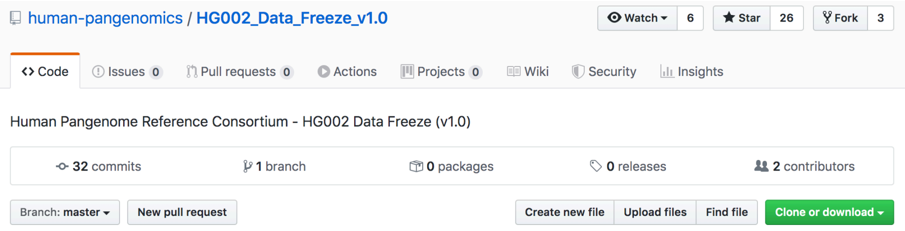
COLLABORATORS

Jay Ghurye
Nathan Truelove
Barbara Block
Mara Lawniczak
Darwin Tree of Life Project

PUBLIC HIFI DATA

HG002 Human Pan-Genome Reference Consortium

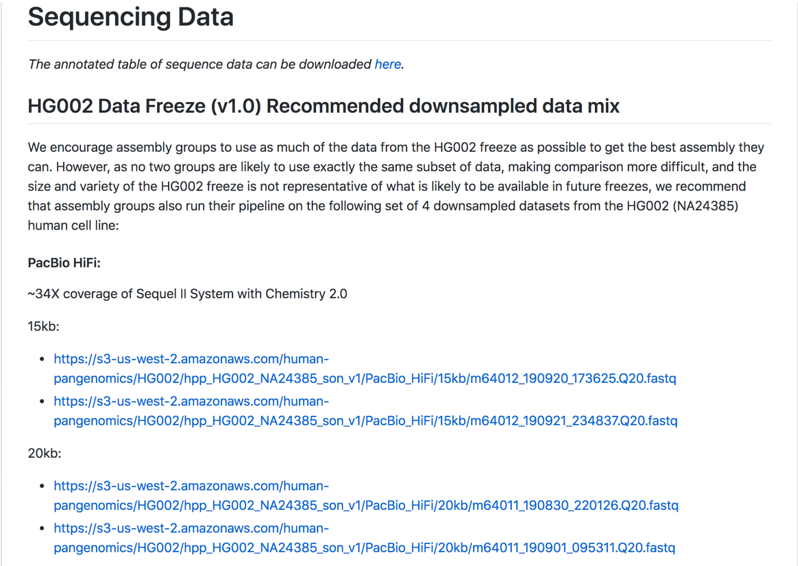
- 4 cells: 2 cells 20kb and 2 cells 15kbp
- ~34x coverage
- https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0



human-pangenomics / **HG002_Data_Freeze_v1.0**

Human Pangenome Reference Consortium - HG002 Data Freeze (v1.0)

32 commits · **1** branch · **0** packages · **0** releases · **2** contributors



Sequencing Data

The annotated table of sequence data can be downloaded here.

HG002 Data Freeze (v1.0) Recommended downsampled data mix

We encourage assembly groups to use as much of the data from the HG002 freeze as possible to get the best assembly they can. However, as no two groups are likely to use exactly the same subset of data, making comparison more difficult, and the size and variety of the HG002 freeze is not representative of what is likely to be available in future freezes, we recommend that assembly groups also run their pipeline on the following set of 4 downsampled datasets from the HG002 (NA24385) human cell line:

PacBio HiFi:

~34X coverage of Sequel II System with Chemistry 2.0

15kb:

- https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/15kb/m64012_190920_173625.Q20.fastq
- https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/15kb/m64012_190921_234837.Q20.fastq

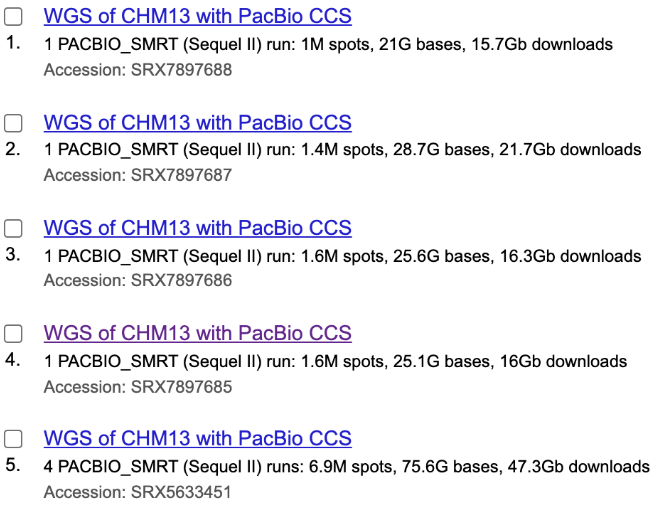
20kb:

- https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/20kb/m64011_190830_220126.Q20.fastq
- https://s3-us-west-2.amazonaws.com/human-pangenomics/HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/20kb/m64011_190901_095311.Q20.fastq

PUBLIC HIFI DATA

CHM13 data from the HiCanu preprint

- 5 HiFi datasets
- https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA530776



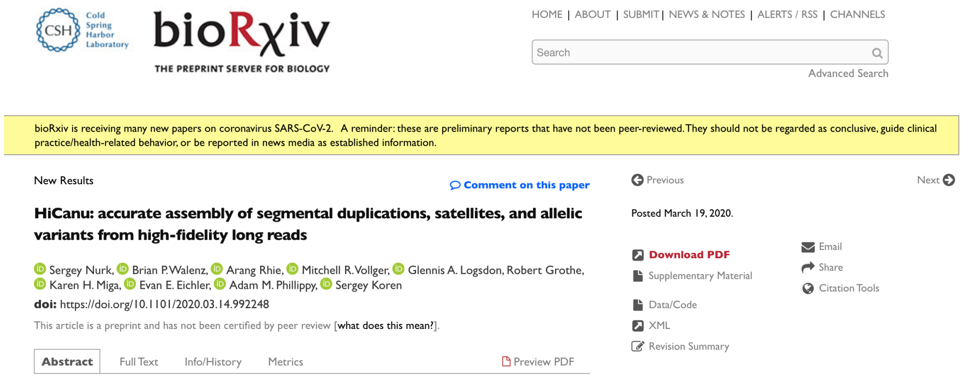
□ WGS of CHM13 with PacBio CCS
1. 1 PACBIO_SMRT (Sequel II) run: 1M spots, 21G bases, 15.7Gb downloads
Accession: SRX7897688

□ WGS of CHM13 with PacBio CCS
2. 1 PACBIO_SMRT (Sequel II) run: 1.4M spots, 28.7G bases, 21.7Gb downloads
Accession: SRX7897687

□ WGS of CHM13 with PacBio CCS
3. 1 PACBIO_SMRT (Sequel II) run: 1.6M spots, 25.6G bases, 16.3Gb downloads
Accession: SRX7897686

□ WGS of CHM13 with PacBio CCS
4. 1 PACBIO_SMRT (Sequel II) run: 1.6M spots, 25.1G bases, 16Gb downloads
Accession: SRX7897685

□ WGS of CHM13 with PacBio CCS
5. 4 PACBIO_SMRT (Sequel II) runs: 6.9M spots, 75.6G bases, 47.3Gb downloads
Accession: SRX5633451



CSH Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS | CHANNELS
Search
Advanced Search

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

Comment on this paper
Previous
Next

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads

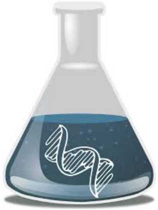
Posted March 19, 2020.

Sergey Nurk, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, Sergey Koren

doi: https://doi.org/10.1101/2020.03.14.992248

This article is a preprint and has not been certified by peer review [what does this mean?].

PUBLIC HIFI DATA

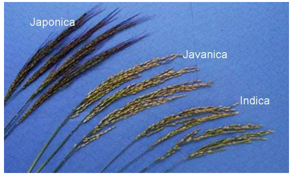


HG002

15 kb + 20 kb library

6 SMRT Cell 8M

Data: PRJNA586863

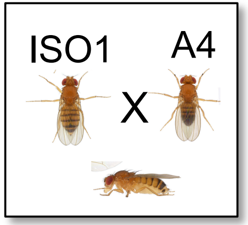


***Oryza sativa indica* MH63**

17 kb + 24 kb library

2 SMRT Cell 8M

Data: PRJNA573706



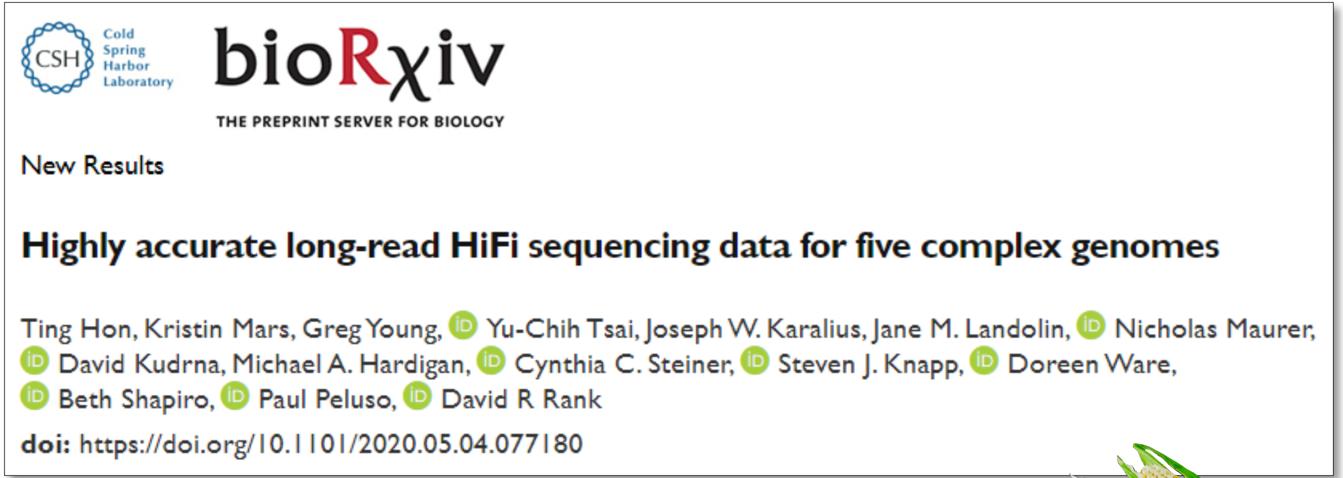
Drosophila melanogaster F1

19 kb + 24 kb library

2 SMRT Cell 8M

Data: PRJNA573706

PUBLIC HIFI DATA



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Highly accurate long-read HiFi sequencing data for five complex genomes

Ting Hon, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, David Kudrna, Michael A. Hardigan, Cynthia C. Steiner, Steven J. Knapp, Doreen Ware, Beth Shapiro, Paul Peluso, David R Rank

doi: https://doi.org/10.1101/2020.05.04.077180

https://www.biorxiv.org/content/10.1101/2020.05.04.077180v1

PACBIO®

www.pacb.com