Assembling genomes using SMRT sequencing

Bob VanBuren

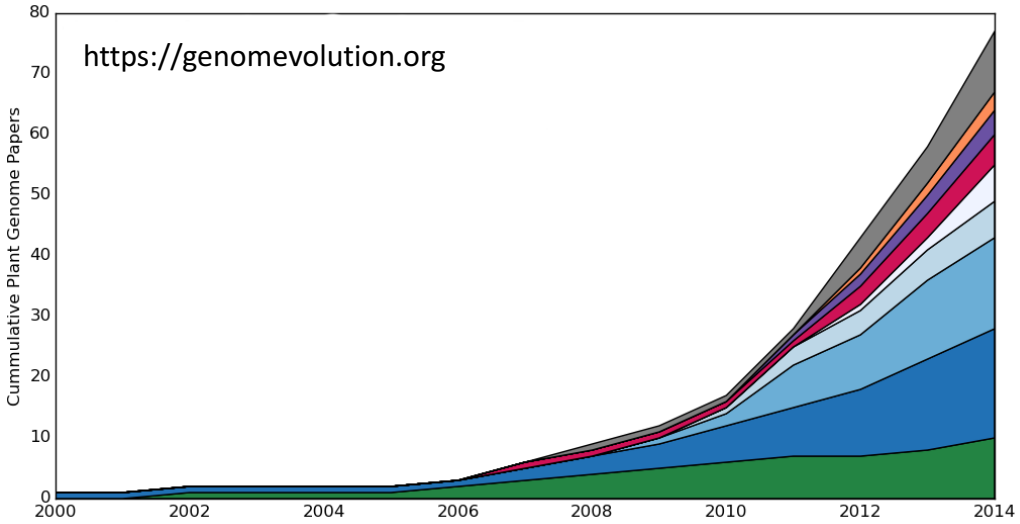
Sequenced plant genomes

Over 150 plant genomes have been released to date

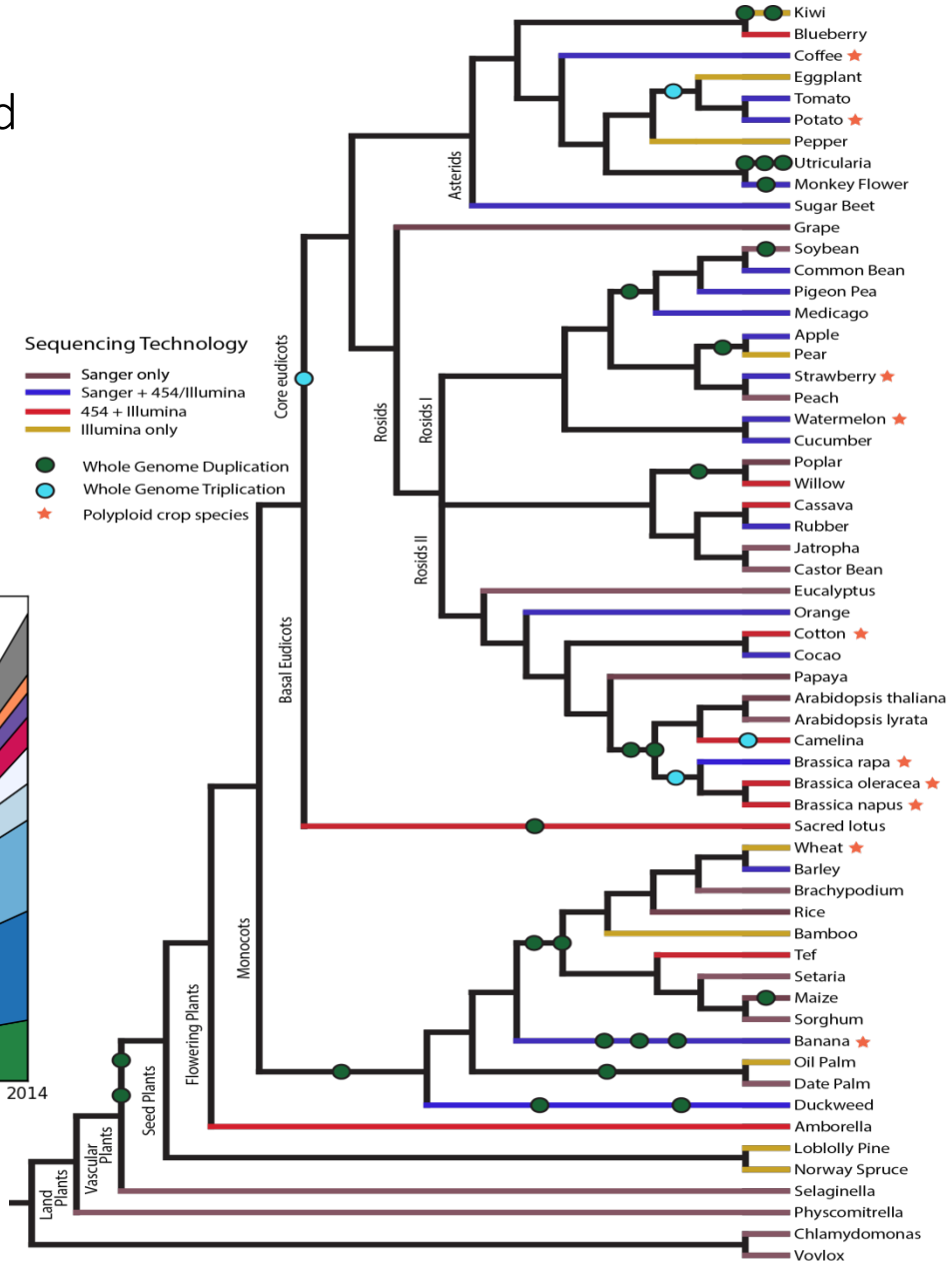
Several hundred more in various stages of completion

Early Sanger based genomes ‘gold standard’

Later NGS based genomes are lower quality



Third generation SMRT PacBio sequencing is facilitating a resurgence of ‘platinum standard’ reference genomes



Michael and VanBuren 2015 *COPB*

Limitations of Illumina sequencing

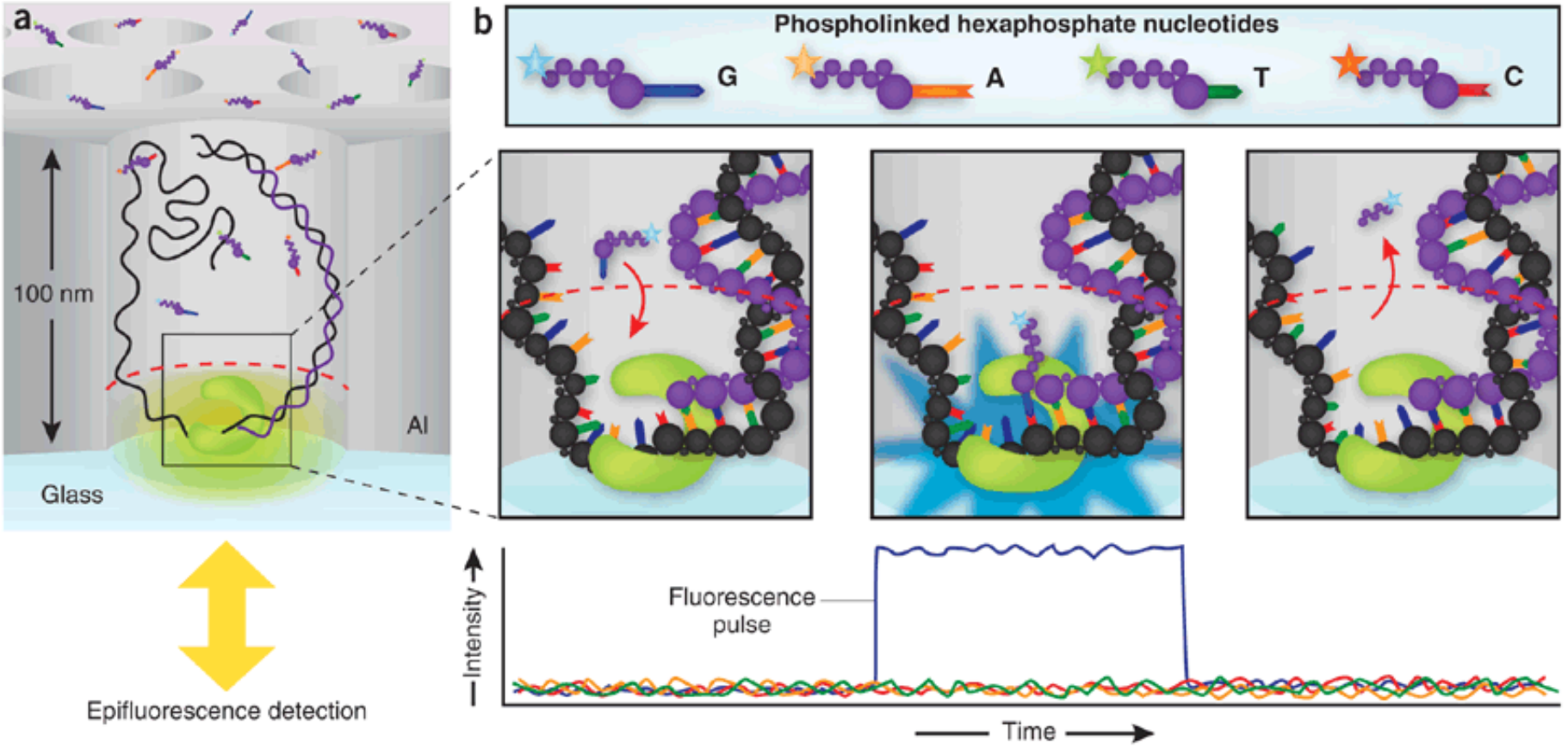
Assembly stats of the A subgenome in hexaploidy wheat

	1AS	1AL	2AS	2AL	3AS	3AL	4AS	4AL	5AS	5AL	6AS	6AL	7AS	7AL	∑
Assembly															
Chromosome size (Mbp)	275	523	391	508	360	468	317	539	295	532	336	369	407	407	5,727
Sequence (Mbp)	178.1	250	255.2	328.2	201.8	247.2	282.3	362	198.8	318.1	219.2	214.4	198	252.4	3,505.7
Coverage (*x*-fold)	0.65	0.48	0.65	0.65	0.56	0.53	0.89	0.67	0.67	0.60	0.65	0.58	0.49	0.62	0.62
L50 (bp)	2,242	2,639	2,398	2,688	1,404	1,346	2,782	3,053	3,509	2,078	2,669	2,154	1,470	2,271	
Repeat															
No. of contigs	34,793	26,746	34,722	45,893	33,943	43,823	32,079	64,364	19,719	47,572	28,041	34,030	44,175	35,586	542,486
L50	4,769	6,369	6,678	6,677	3,846	3,789	7,499	6,601	8,713	5,355	7,091	6,589	4,397	5,849	

Total wheat assembly contains ~1.6 million contigs.
Many imbedded gaps, likely missing genic regions.

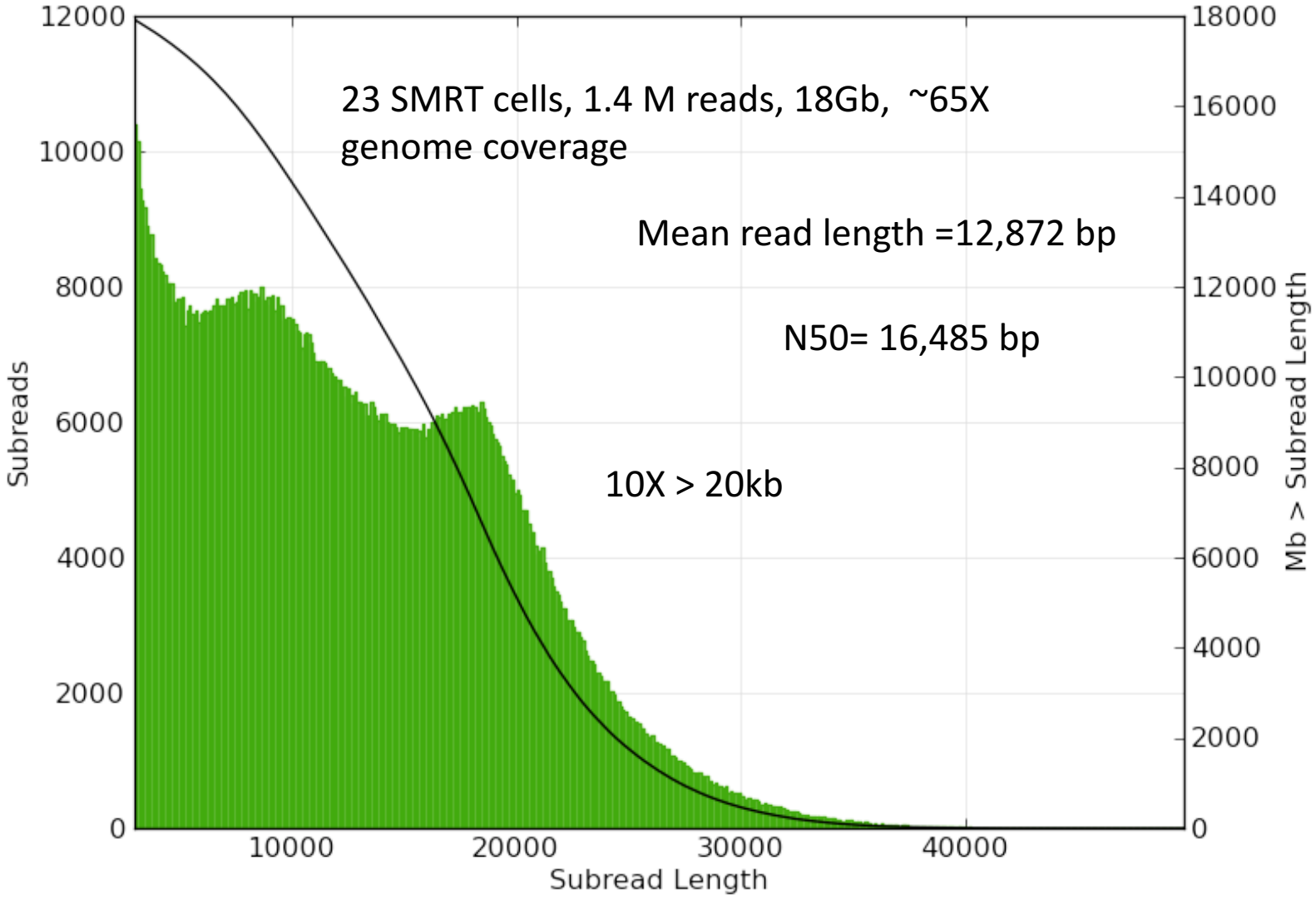
IWGSC, *Science* 2014

PacBio single molecule real time sequencing



Long reads (10kb-60kb), high throughput (1Gb/ flow cell), low cost (~$300/ flow cell)

P6C4 chemistry (PacBio)

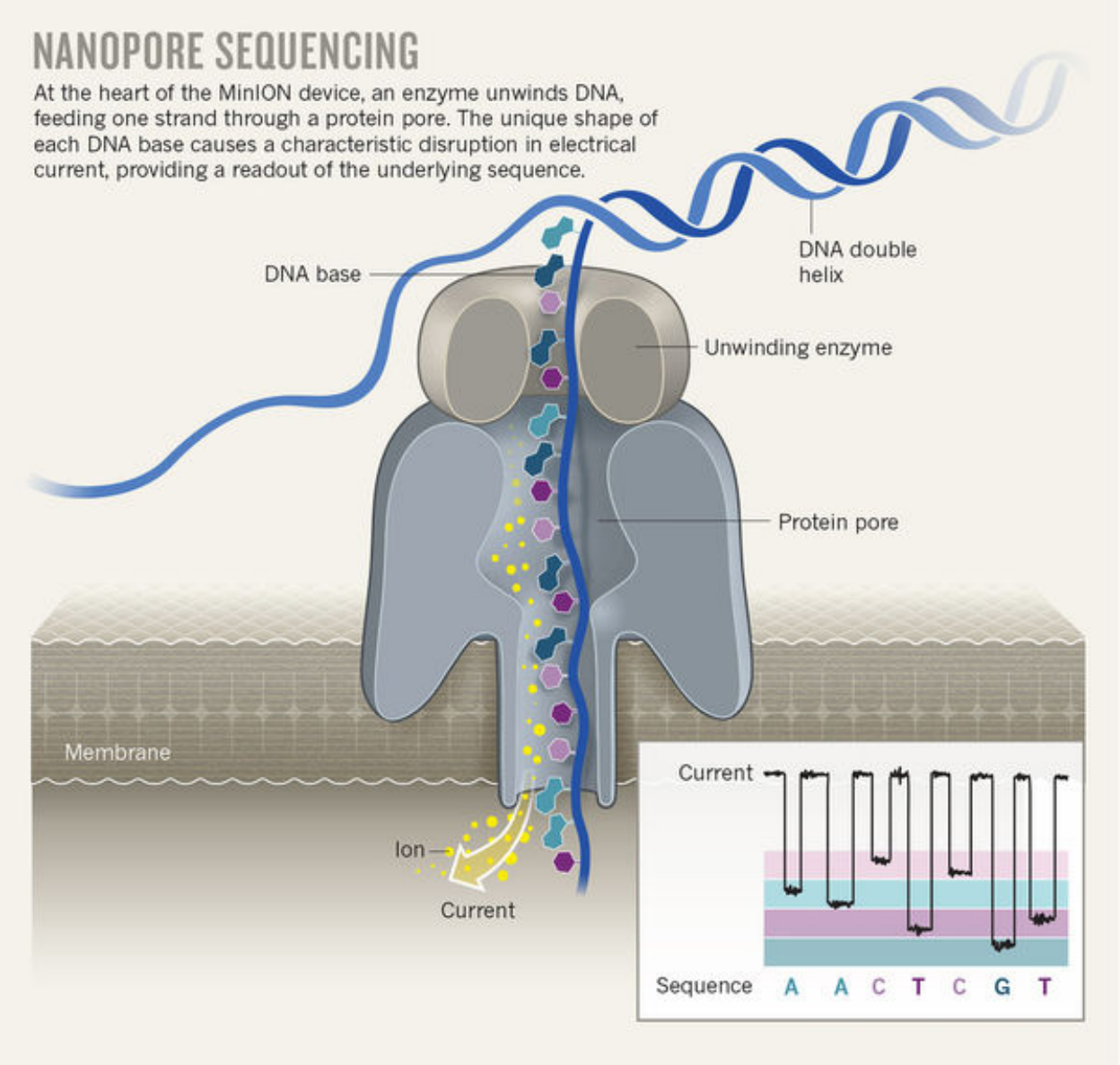


MinION single molecule real time sequencing

GridION X5

Long reads (10kb-600kb), high throughput
(10Gb/ flow cell/day),
low cost (~$1,000/ flow cell)

MinION single molecule real time sequencing



Running MinION

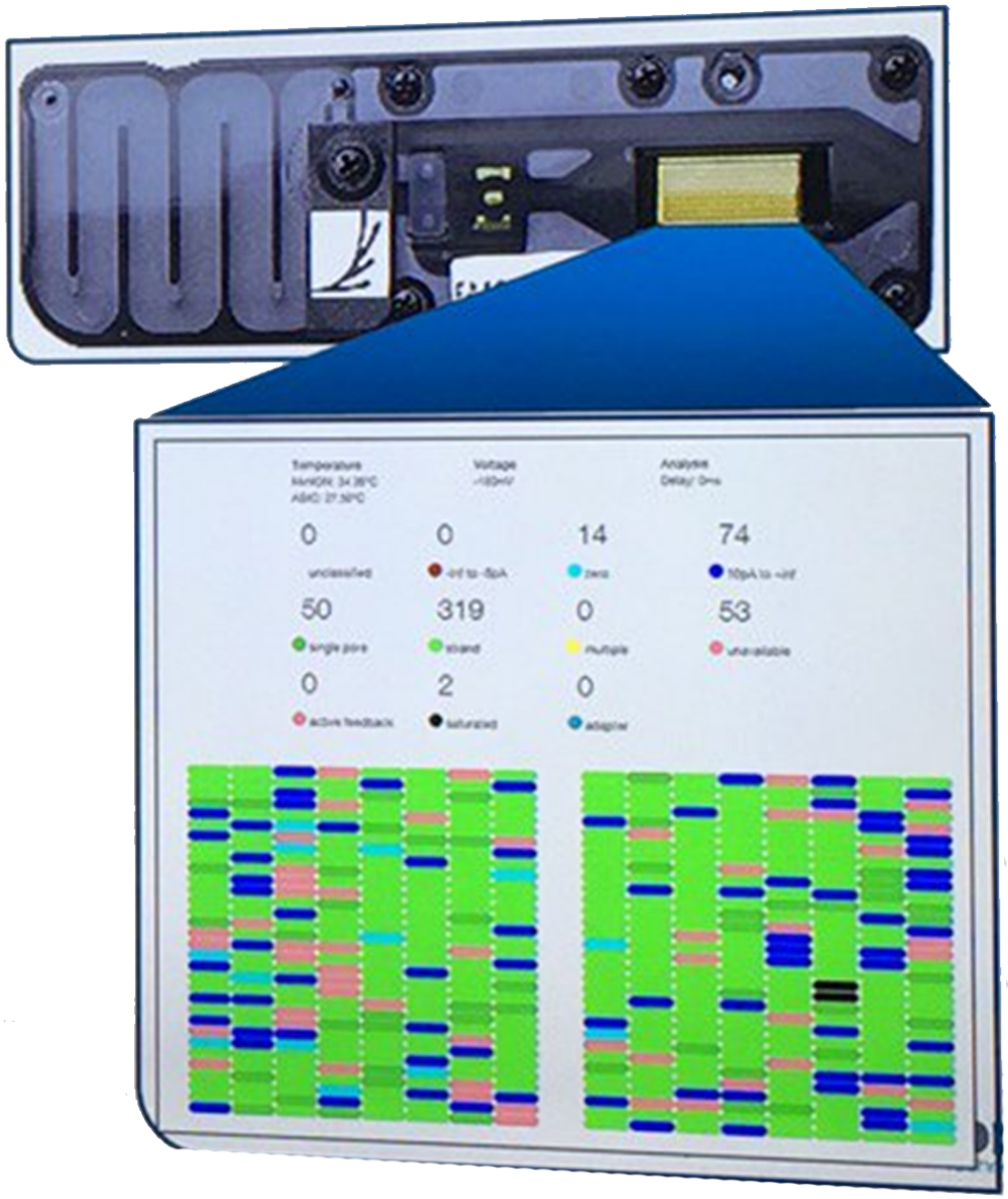
Isolate High Molecular Weight (HMW) gDNA (most important step)

Library prep (20 min or ~2 hours)

Load library on flow cell

Run 24-48 hours

Analyze

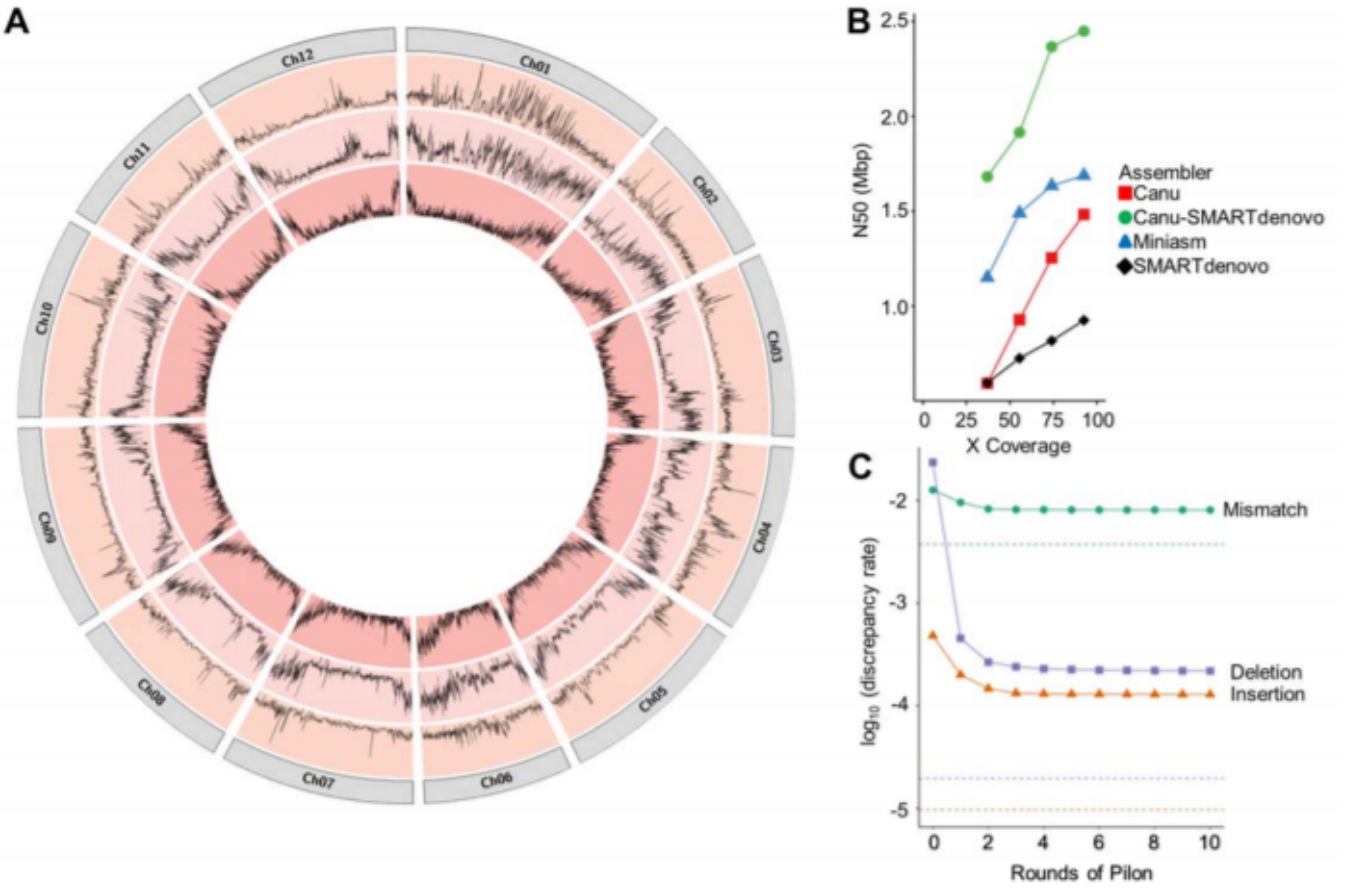


Nanopore based *Solanum pennellii* assembly

Contig N50 2.5 Mb

After 10 rounds of Illumina polishing, 0.02% errors

= 2 errors every 10kb



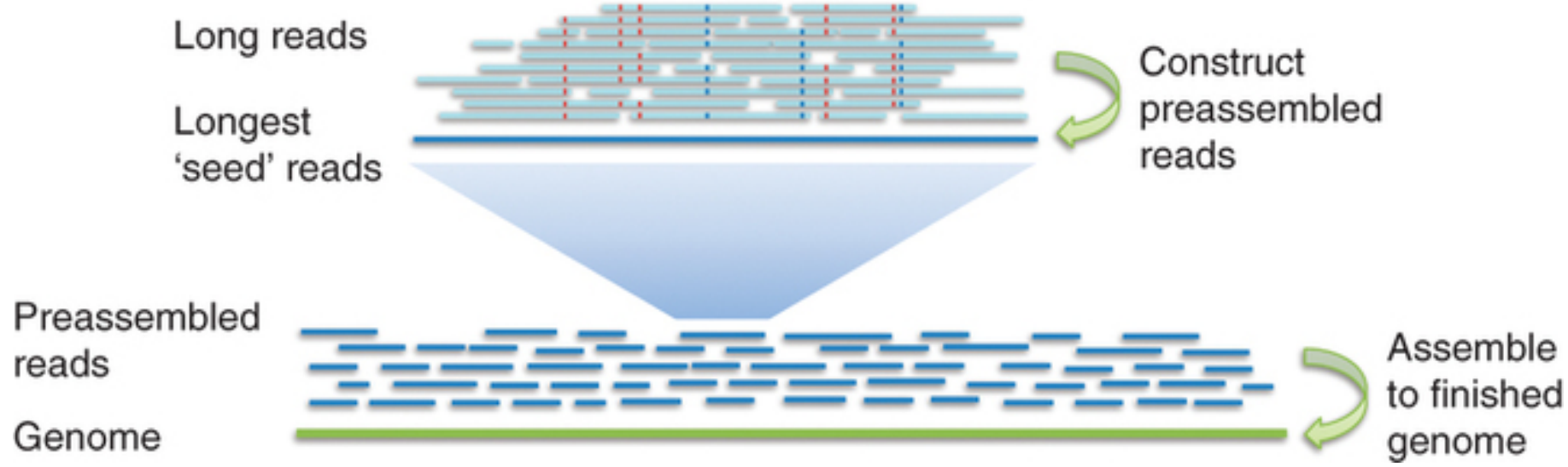
Schmidt et al. Plant Cell, 2017

SMRT Genome Assembly Workflow

SMRT sequencing reads have high error rate (8-20 %)

Errors are random for PacBio, semi random for Nanopore (homopolymer issues)

Long overlaps allow for high confidence alignment and error correction



Chin et al. Nature methods 2013.

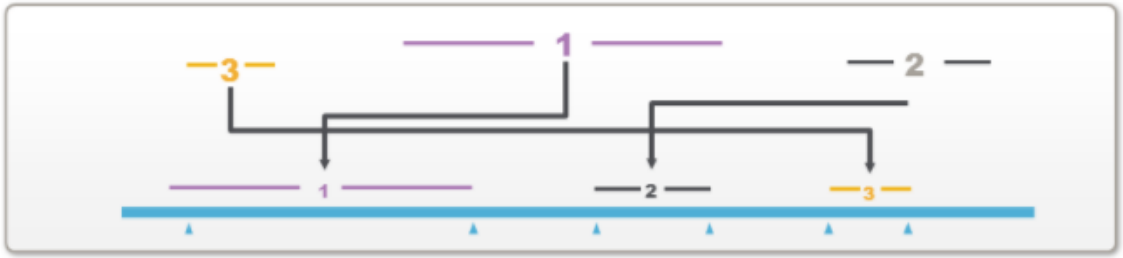
SMRT Genome Assembly Workflow

De novo Assembly

Complete genomes using only PacBio reads or combine technologies

Scaffold

Establish framework for genome and resolve ambiguities



Span Gaps

Polish genomic regions with up to 10x improvement

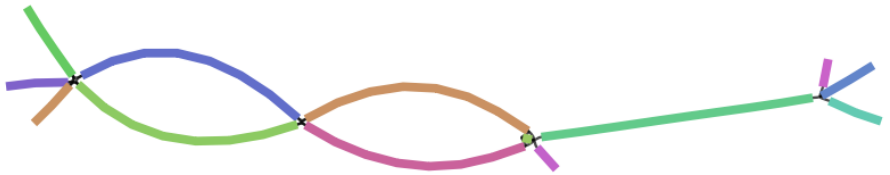
Assemble -> polish (Illumina) -> Scaffold -> gap fill -> polish (Illumina)

SMRT assembly programs

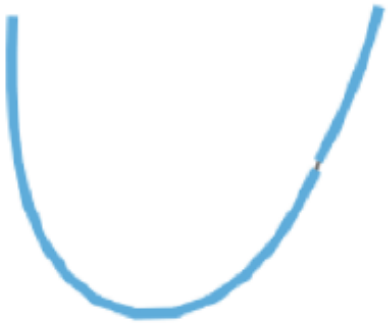
PacBio-only	
HGAP	A workflow to first preassemble reads, assemble the preassembled reads using Celera® Assembler, then polish using Quiver. • Supports up to 100 Mb from SMRT Portal, which is part of SMRT Analysis. • Larger genomes are possible from the command line using either smrtpipe.py or the Makefile-based smrtmake.
Falcon	An experimental diploid assembler, tested on multi Gb genomes. 2014 AGBT presentation by Jason Chin.
Canu	A fork of the Celera Assembler designed for high-noise single-molecule sequencing.
Celera® Assembler	Celera® Assembler 8.1 now offers a way to directly assemble subreads.
Sprai	A preassembly-based assembler that aims to generate longer contigs.

Evaluating quality (Graphical Fragment Assembly)

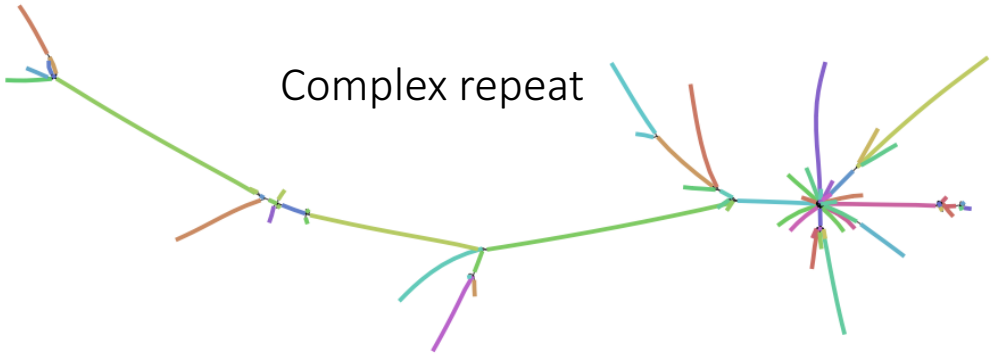
Each node represents a contig
Connections indicate ambiguities in the graph structure



Heterozygous bubble



Simple contig
(complete chromosome)



Complex repeat

Example PacBio projects

Gap filling

Finishing old reference genomes

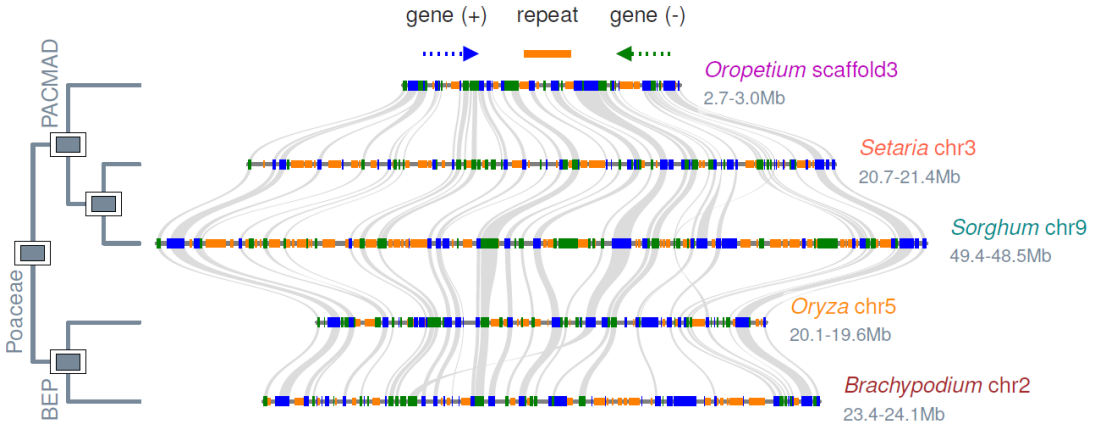
Heterozygosity

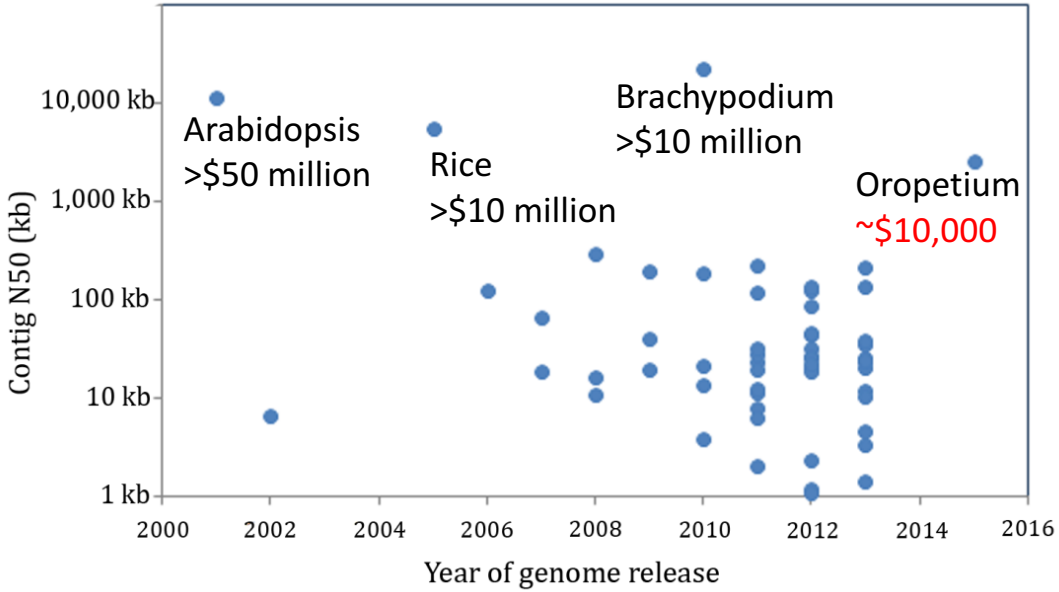
Polyploidy

First PacBio Plant genome: Oropetium

Smallest genome among the grasses (250Mb)

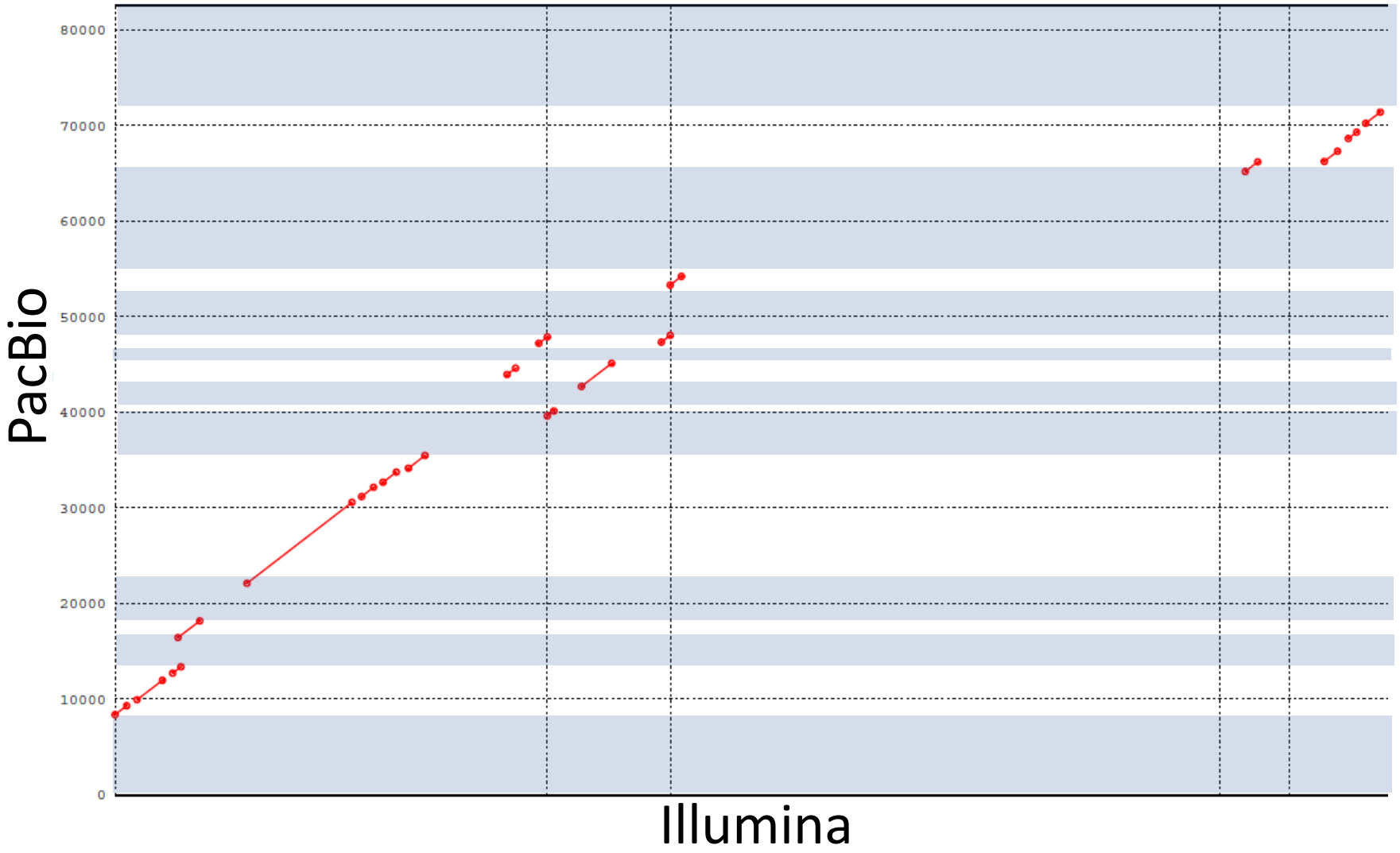
Output (Celera Assembler):	
Number of Polished Contigs	625
Max Contig Length	7,984,151 bp
N50 Contig Length	2,386,328 bp
Sum of Contig Lengths	243,174,629 bp





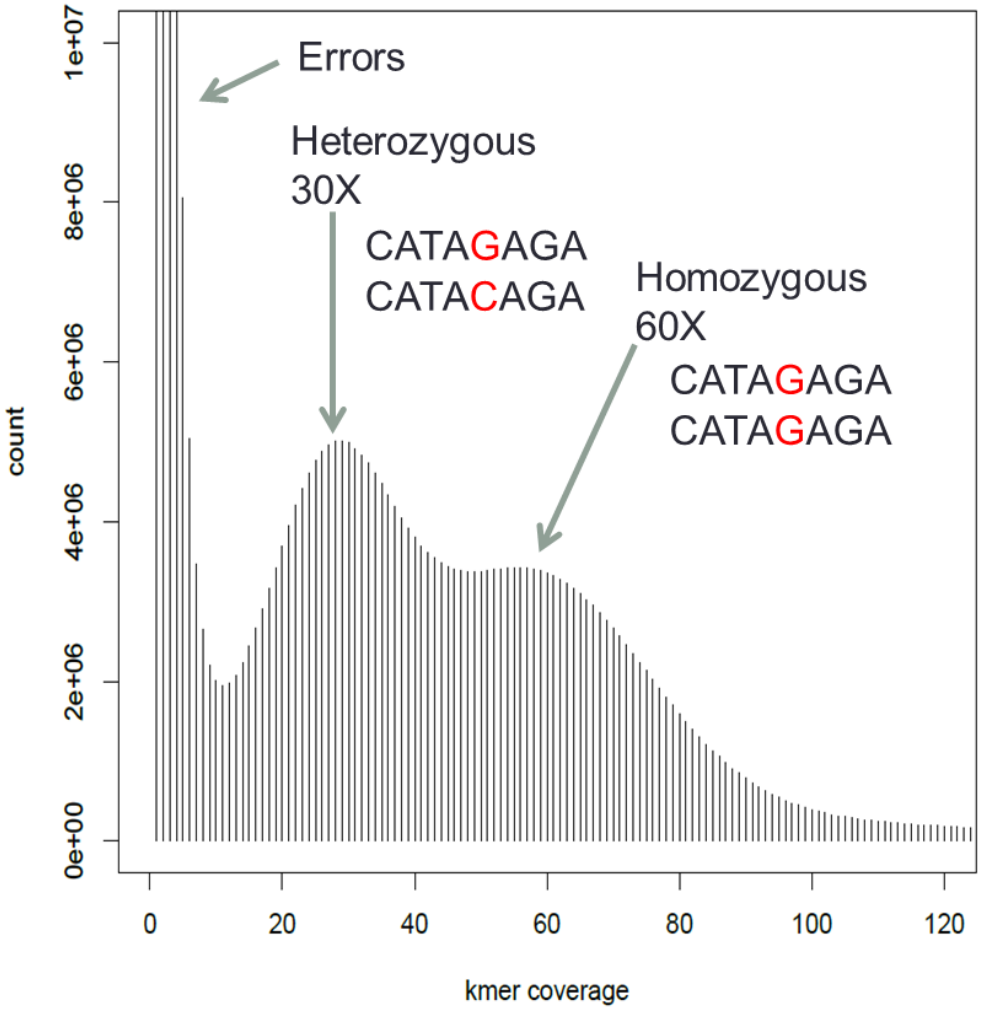
VanBuren et al. 2015 *Nature*

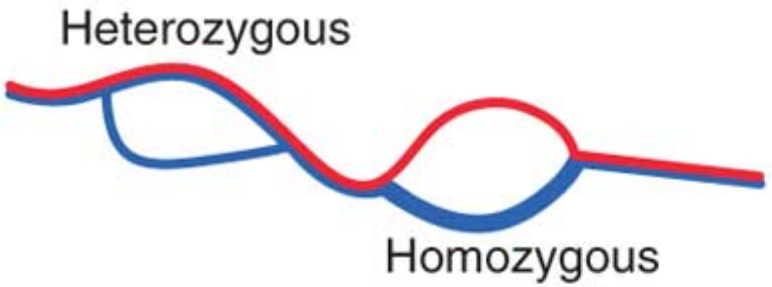
Pacbio vs Illumina assembly

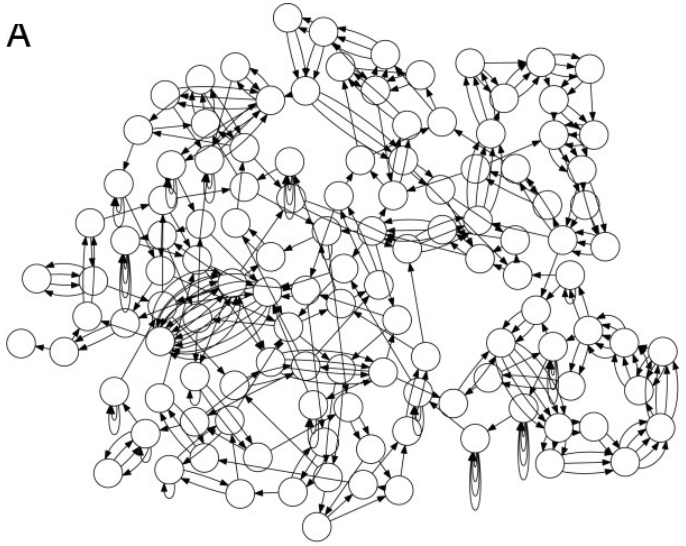


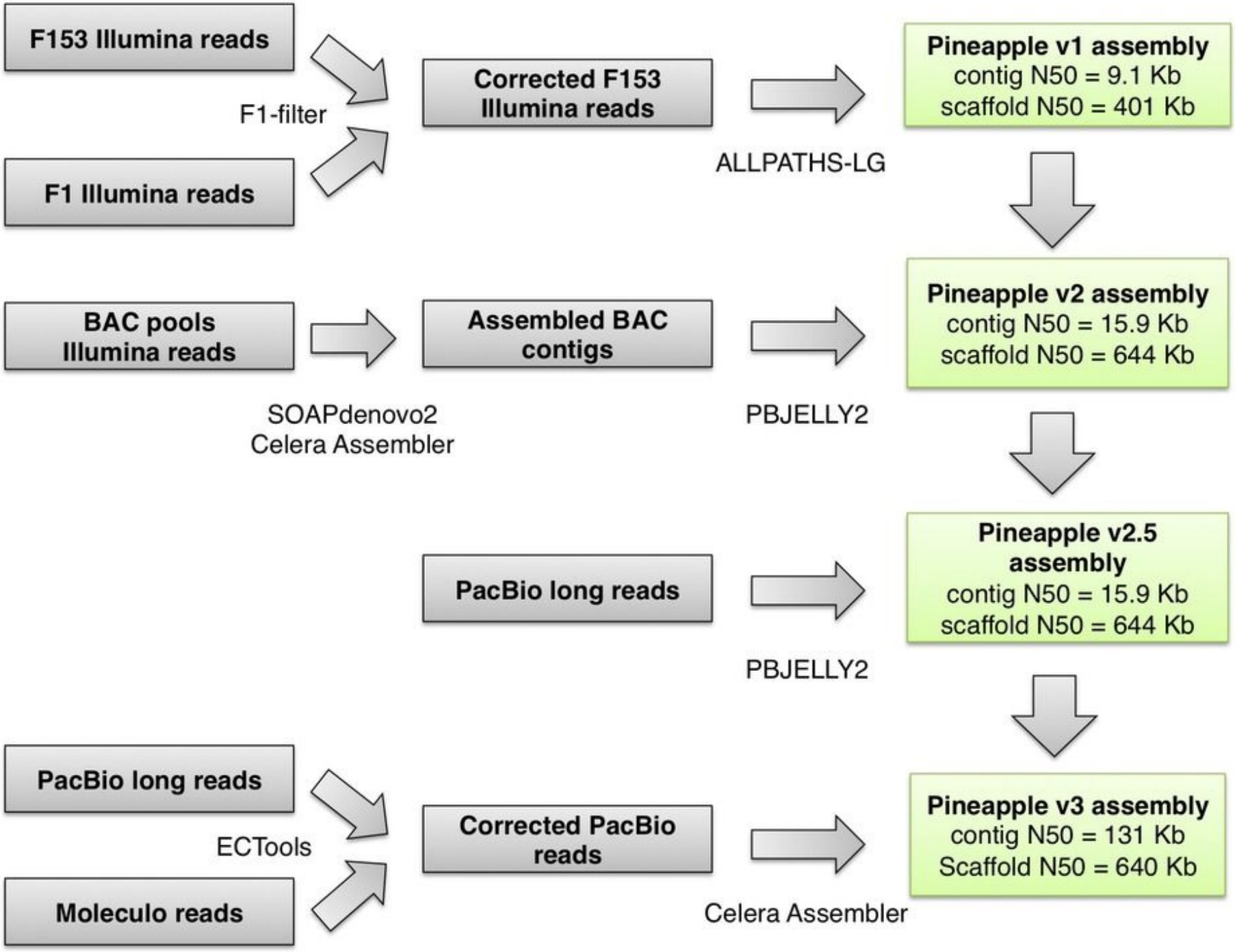
Gap filling in the pineapple genome (Low coverage PacBio)

The pineapple genome is highly heterozygotic (2.2%)



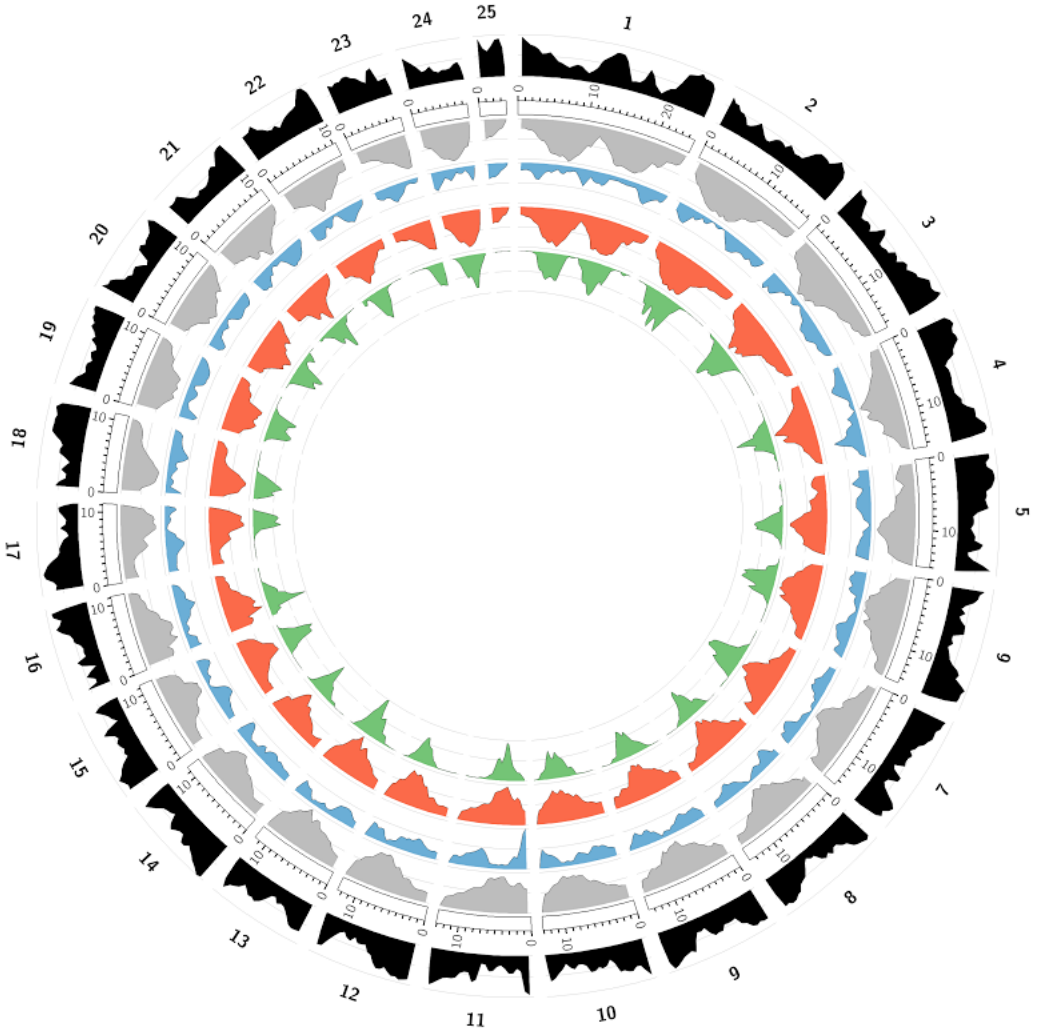




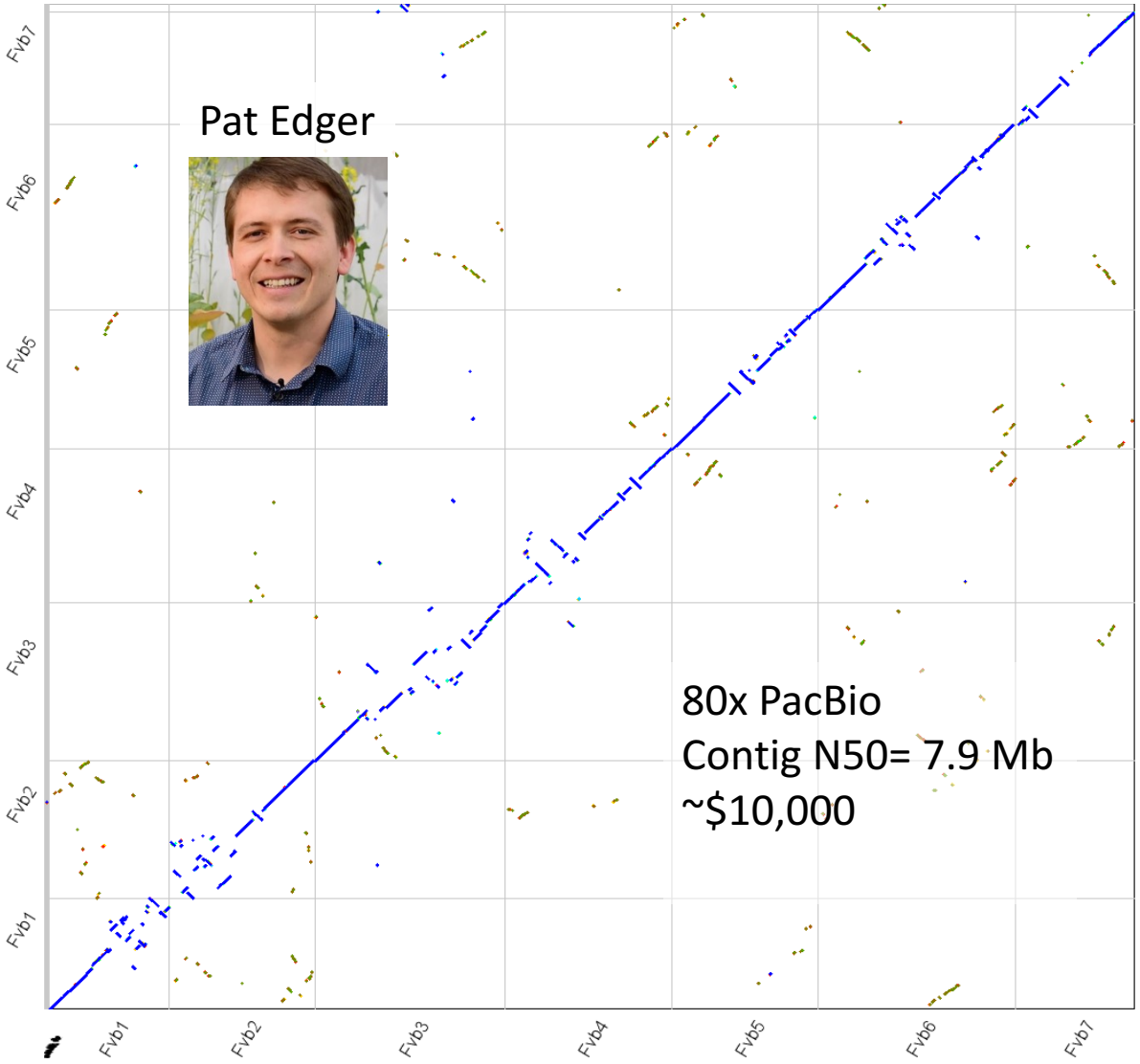

F153 Illumina reads
F1-filter
F1 Illumina reads
Corrected F153 Illumina reads
ALLPATHS-LG
Pineapple v1 assembly
contig N50 = 9.1 Kb
scaffold N50 = 401 Kb
BAC pools Illumina reads
SOAPdenovo2
Celera Assembler
Assembled BAC contigs
PBJELLY2
Pineapple v2 assembly
contig N50 = 15.9 Kb
scaffold N50 = 644 Kb
PacBio long reads
PBJELLY2
Pineapple v2.5 assembly
contig N50 = 15.9 Kb
scaffold N50 = 644 Kb
PacBio long reads
ECTools
Moleculo reads
Corrected PacBio reads
Celera Assembler
Pineapple v3 assembly
contig N50 = 131 Kb
Scaffold N50 = 640 Kb

Ultra-high density genetic map for anchoring genome

- Sequenced 91 F1 individuals to 10x coverage each
- Generated 296,896 high quality SNP markers
- Narrowed each recombination event to < 100 bp region.
- Anchored ~90% of the assembly to 25 chromosomes



Using PacBio to fix old reference genomes



Woodland strawberry (*Fragaria vesca*)

Sequenced in 2011

25 Mb new sequences

300 fold improvement in N50

1,500 new genes

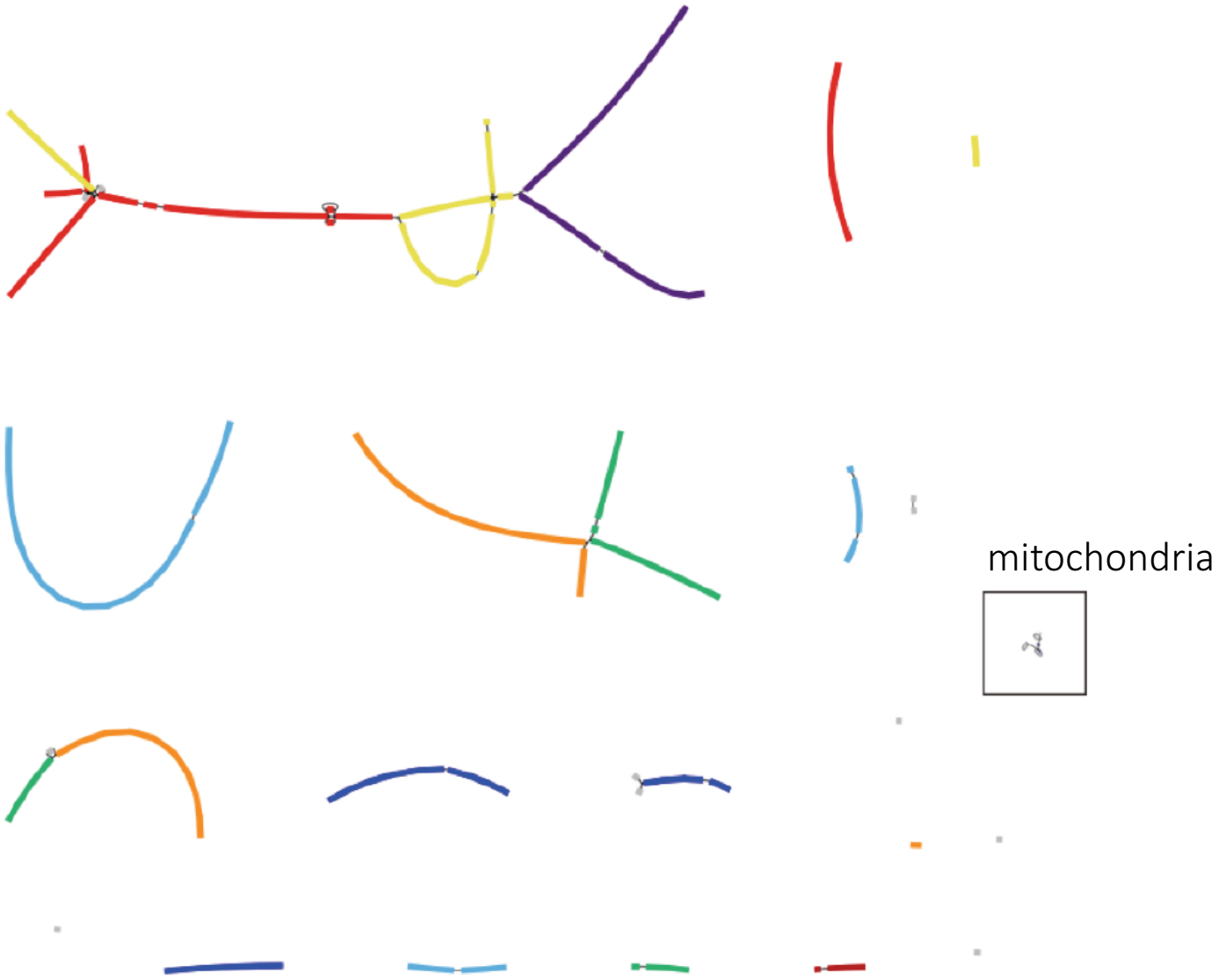
(1,100 Tandem duplicates)

1/4 genome scaffolded incorrectly

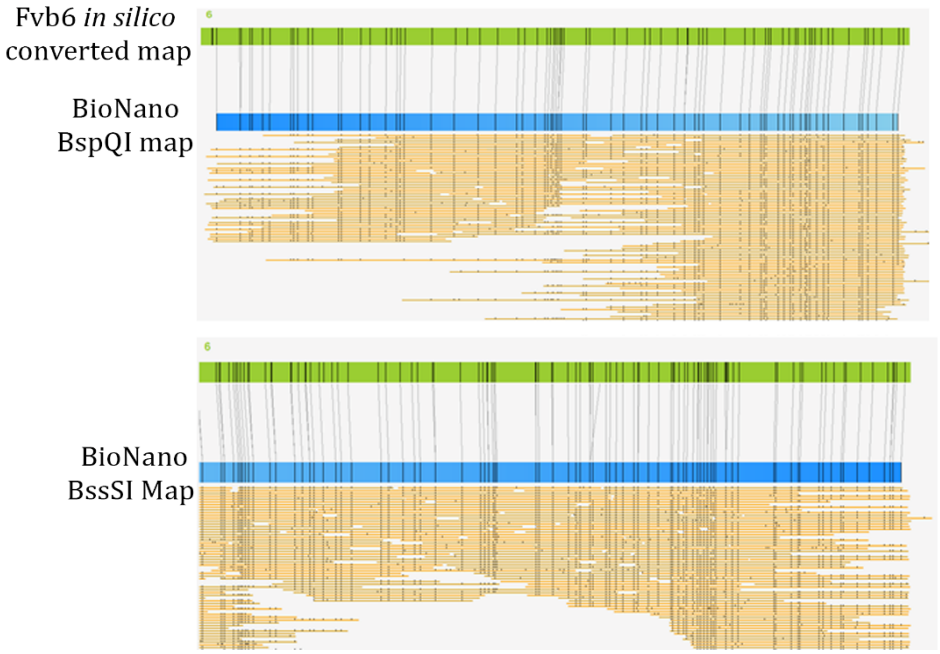
Edger, VanBuren, et al. *GigaScience*

F. vesca graph fragment assembly

Contigs are colored by chromosome



Scaffolding *F. vesca* using a Bionano genome map



Two enzyme map anchored most contigs into chromosomes (9 contigs for 7 chromosomes)

Terminal bionano maps correspond to telomere tracks

	Step 1 (Nt.BspQI)			Step 2 (Nb. BssSI)		
	Contig Count	N50 Mb	Total Length (Mb)	Contig Count	N50 Mb	Total Length (Mb)
Before merge: BioNano Genome Map	230	2.042	280.761	247	1.351	215.995
Before merge: NGS Genome Map	61	7.9	219.432	34	19.612	220.338
BNG contigs in hybrid Scaffold	149	2.739	219.964	245	1.368	214.527
NGS contigs in hybrid scaffold	47	7.261	218.555	13	19.612	219.462
Hybrid scaffold statistics	12	19.623	219.799	10	36.119	219.911
Hybrid scaffold plus not scaffolded BNG	93	19.047	280.595	12	36.119	221.38
Hybrid scaffold plus not scaffolded NGS	**33**	**19.623**	**220.675**	**31**	**36.119**	**220.788**

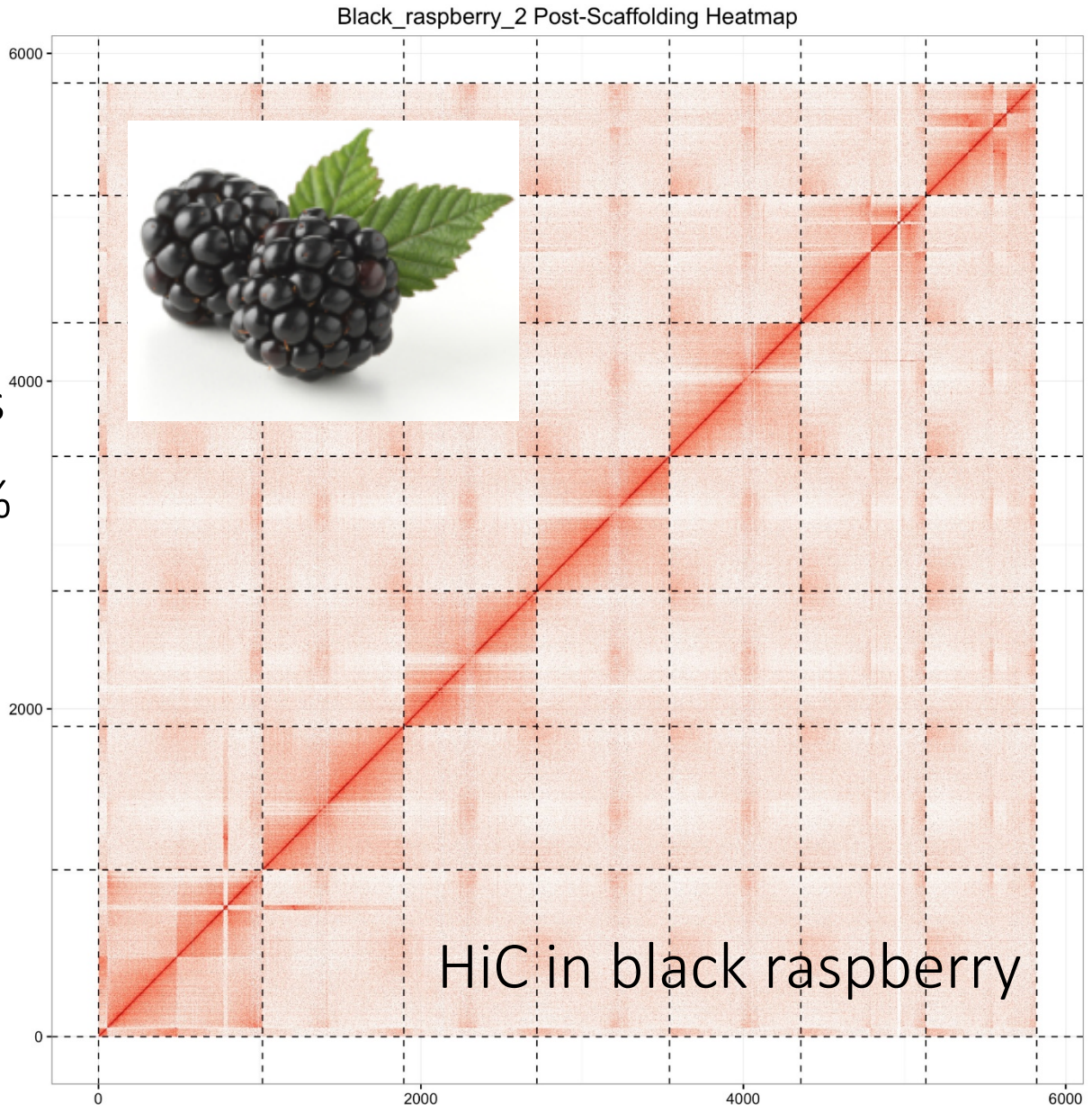
Using PacBio to fix old reference genomes

Contig N50 5.1 Mb and 235 contigs with a total assembly size of 287 Mb

50 Mb of new sequences

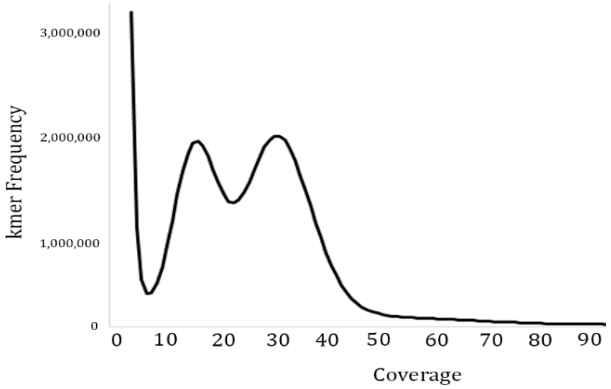
Hi-C map anchored 100% of contigs into 7 chromosomes.

Gap filling using PacBio produced several complete chromosomes.



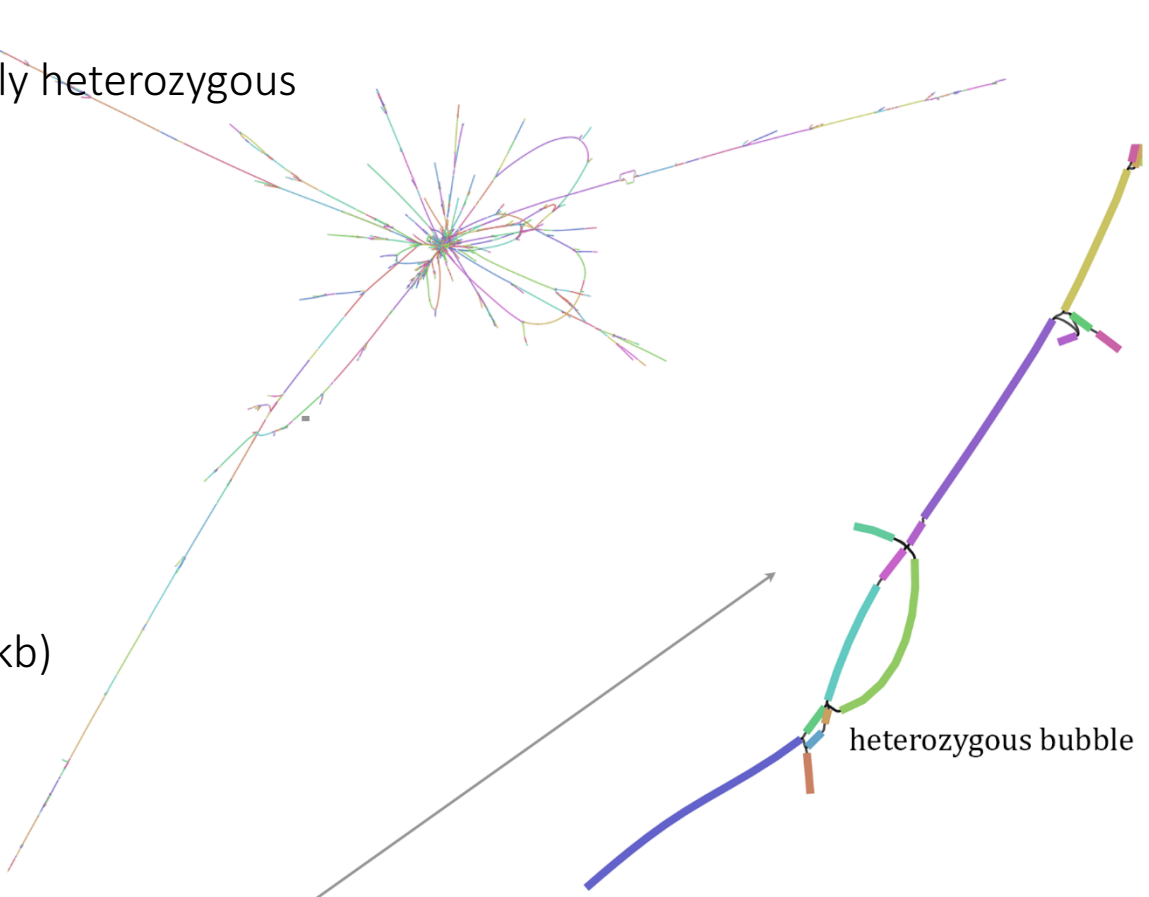
Selaginella lepidophylla

Small genome (90 Mb) but extremely heterozygous

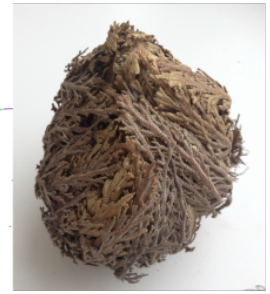


110x coverage PacBio data (N50 23kb)

Separate haplotypes assembled for much of the genome

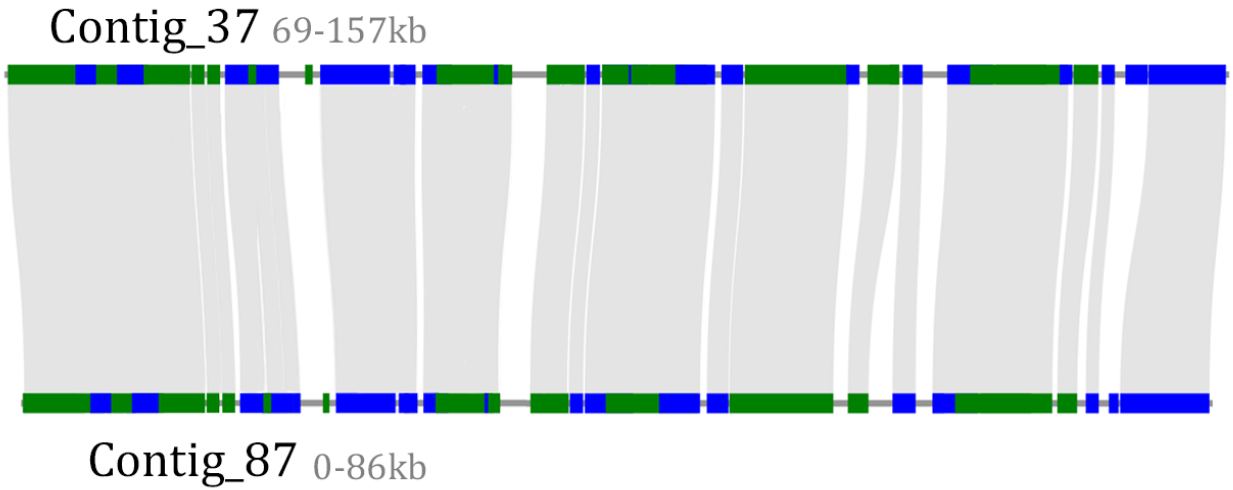


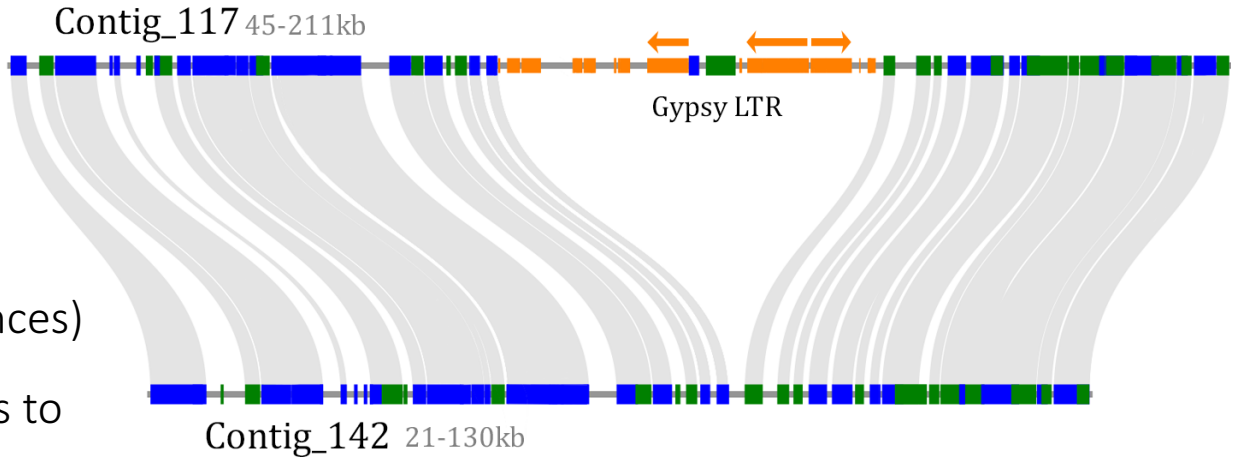
Assembly metric	#
Number of sequences:	1,149
Total size of sequences:	122,531,738
Longest sequence:	1430,794
Shortest sequence:	2,325
N50 sequence length:	163,247
N90 sequence length:	42,531



VanBuren et al submitted.

Extreme haplotype variation in *Selaginella*





Extensive structural variation

(8 Mb haplotype specific sequences)

Rapid bloat and purge of repeats to maintain small genome size

Haplotype variation is underestimated in most genomes

VanBuren et al submitted.

PacBio assembly of *Sedum album*

85X PacBio data

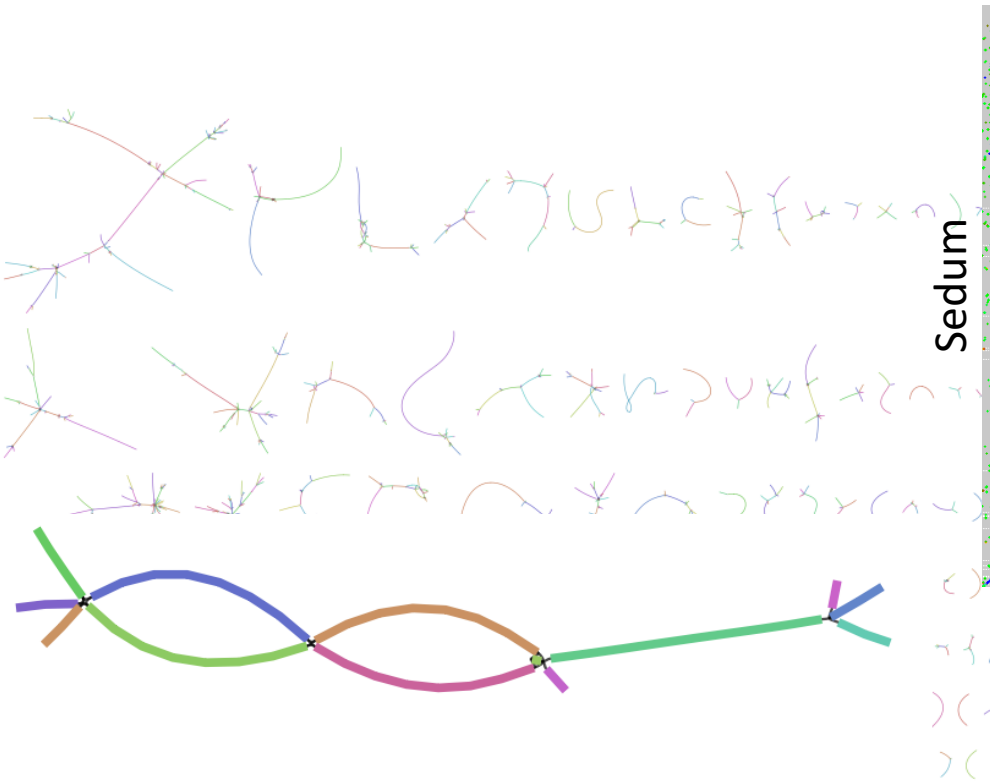
Contig N50 113,432 bp

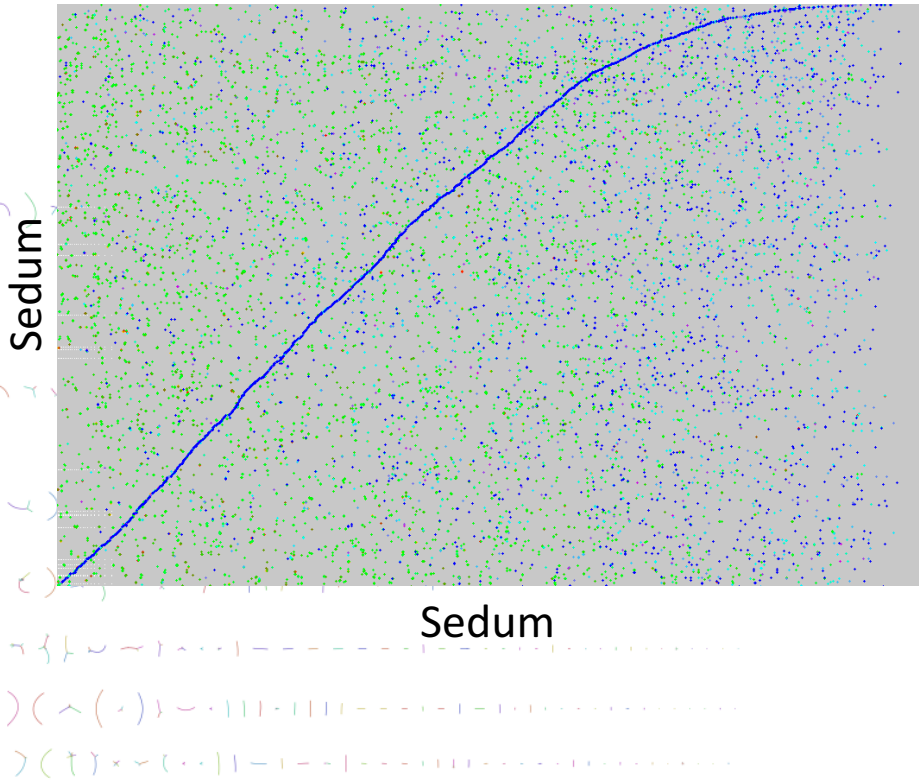
8,324 contigs

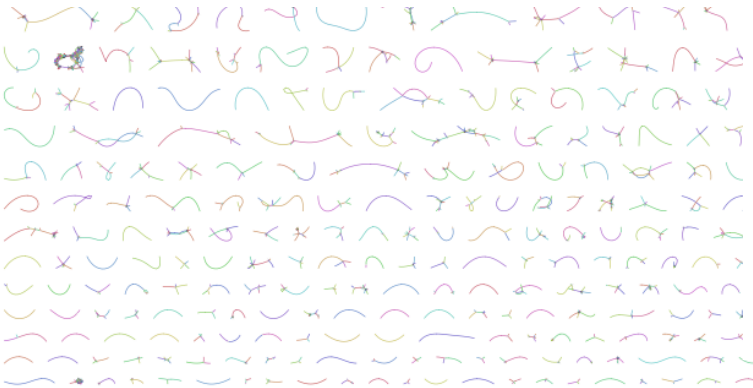
Total assembly size of 435 Mb

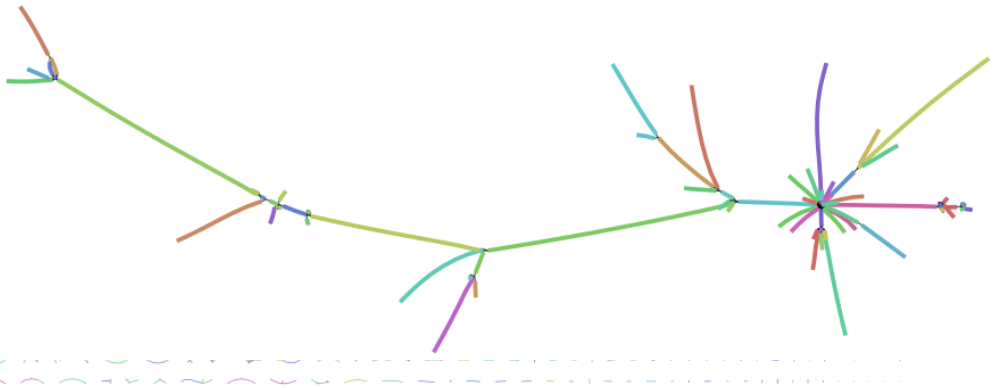
Flow cytometry estimate: 502 Mb

PacBio assembly of *Sedum album*

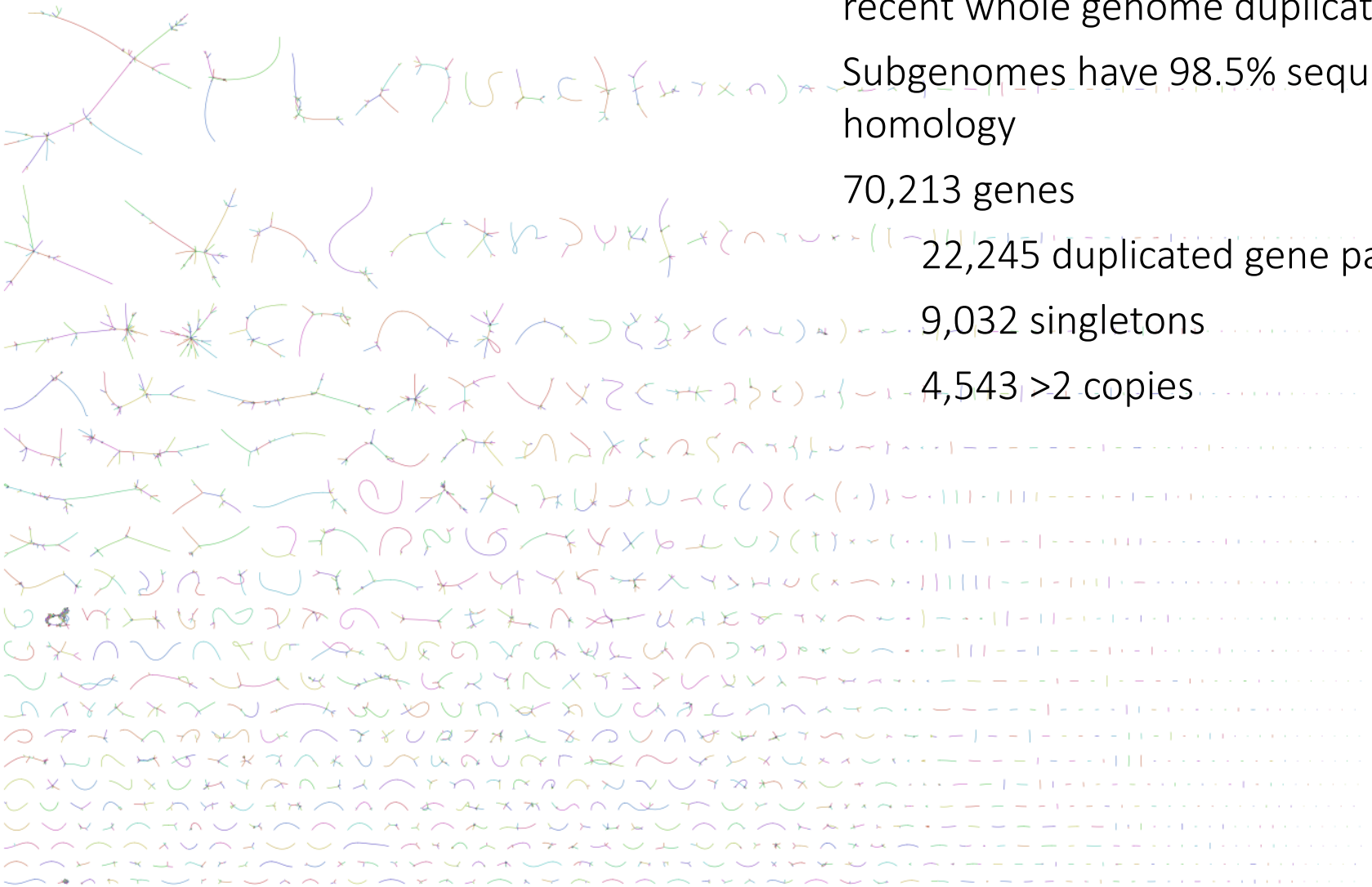








PacBio assembly of *Sedum album*



Sedum album is tetraploid with two recent whole genome duplications

Subgenomes have 98.5% sequence homology

70,213 genes

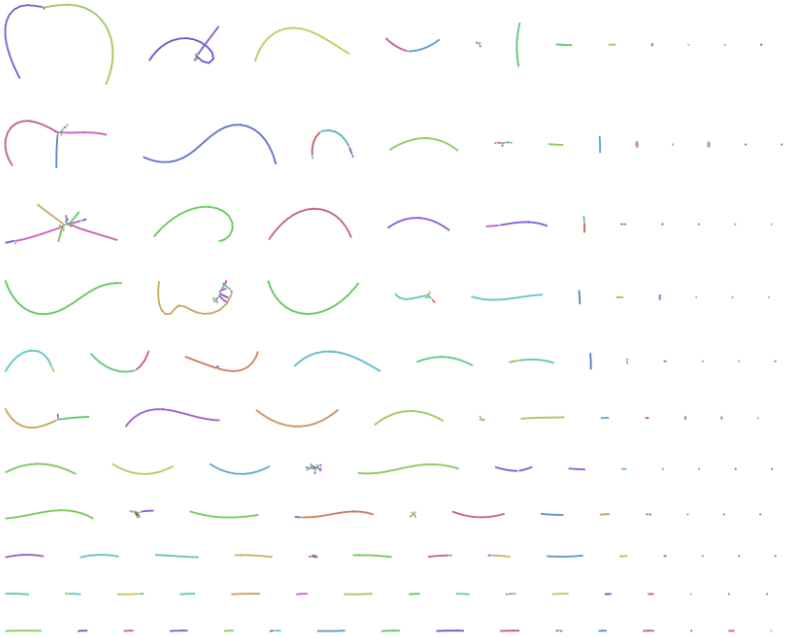
- 22,245 duplicated gene pairs
- 9,032 singletons
- 4,543 >2 copies

PacBio Sequel

Throughput is 5x more than RSII (cost is ~50% less)

Faster (shorter queue)

Similar quality and read lengths to RSII



Sequenced a 270 Mb genome:

Contig N50: 3.6 Mb, 327 contigs

Cost: ~$6,000 (Library + Sequencing)

Summary

- SMRT sequencing can be used to assemble ‘Platnum grade’ finished genomes economically
- PacBio (RSII or Sequel) is more consistent and less erroneous than Nanopore
- Nanopore is economical ($1,000 for starter kit) and can be used for sequencing RNA and detecting native methylation

Acknowledgements

MSU

Jennifer Wai
Pat Edger
Ning Jiang
Shujun Ou

Danforth Center

Todd Mockler
Malia Gehan
Doug Allen
Henry Priest

University of Illinois

Ray Ming

JCVI

Todd Michael