bionano GENOMICS

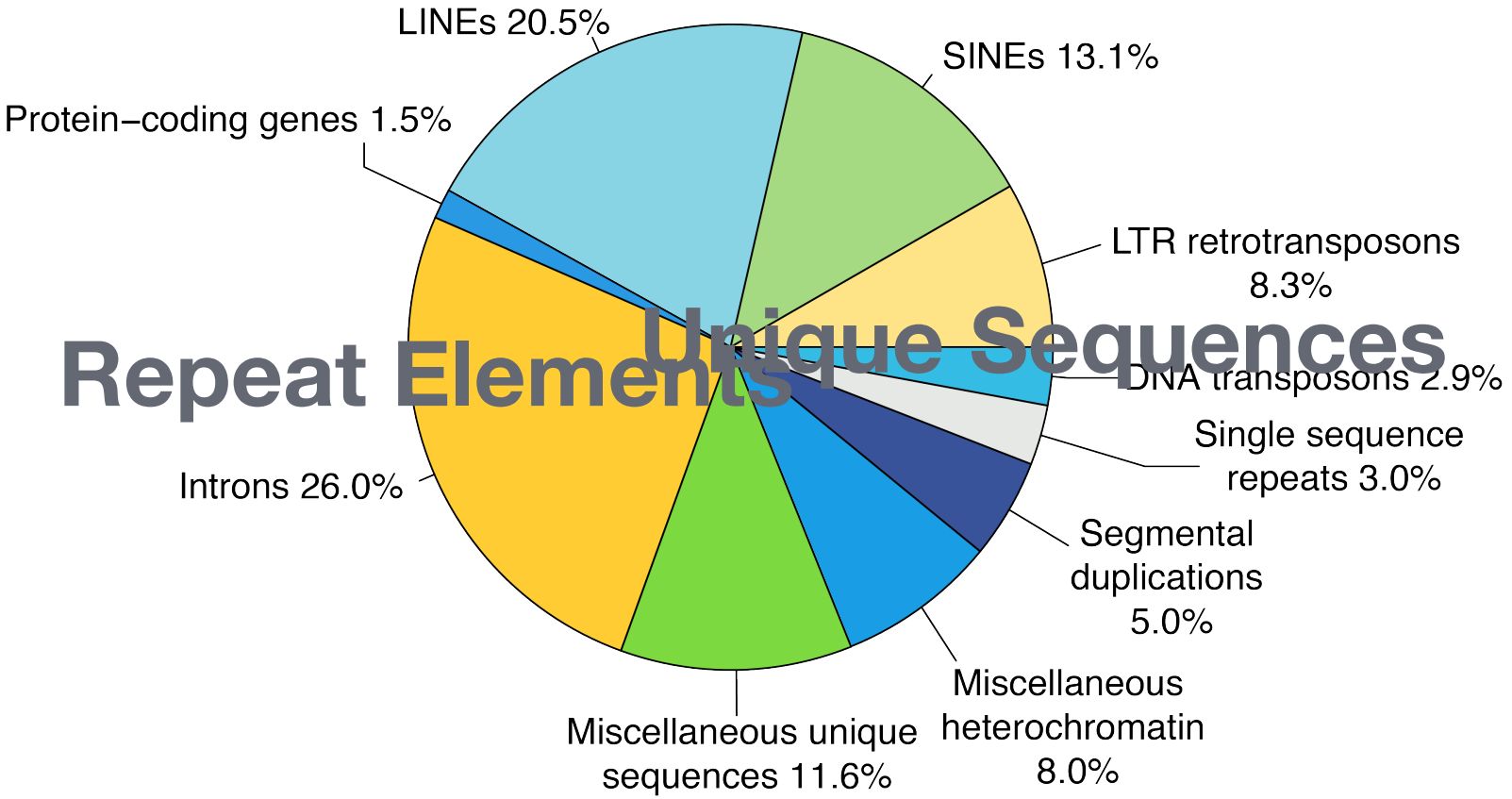
Improve Genome Accuracy and Contiguity using Bionano Next-Generation Mapping

Sven Bocklandt, Ph.D.

i5k Webinar Series

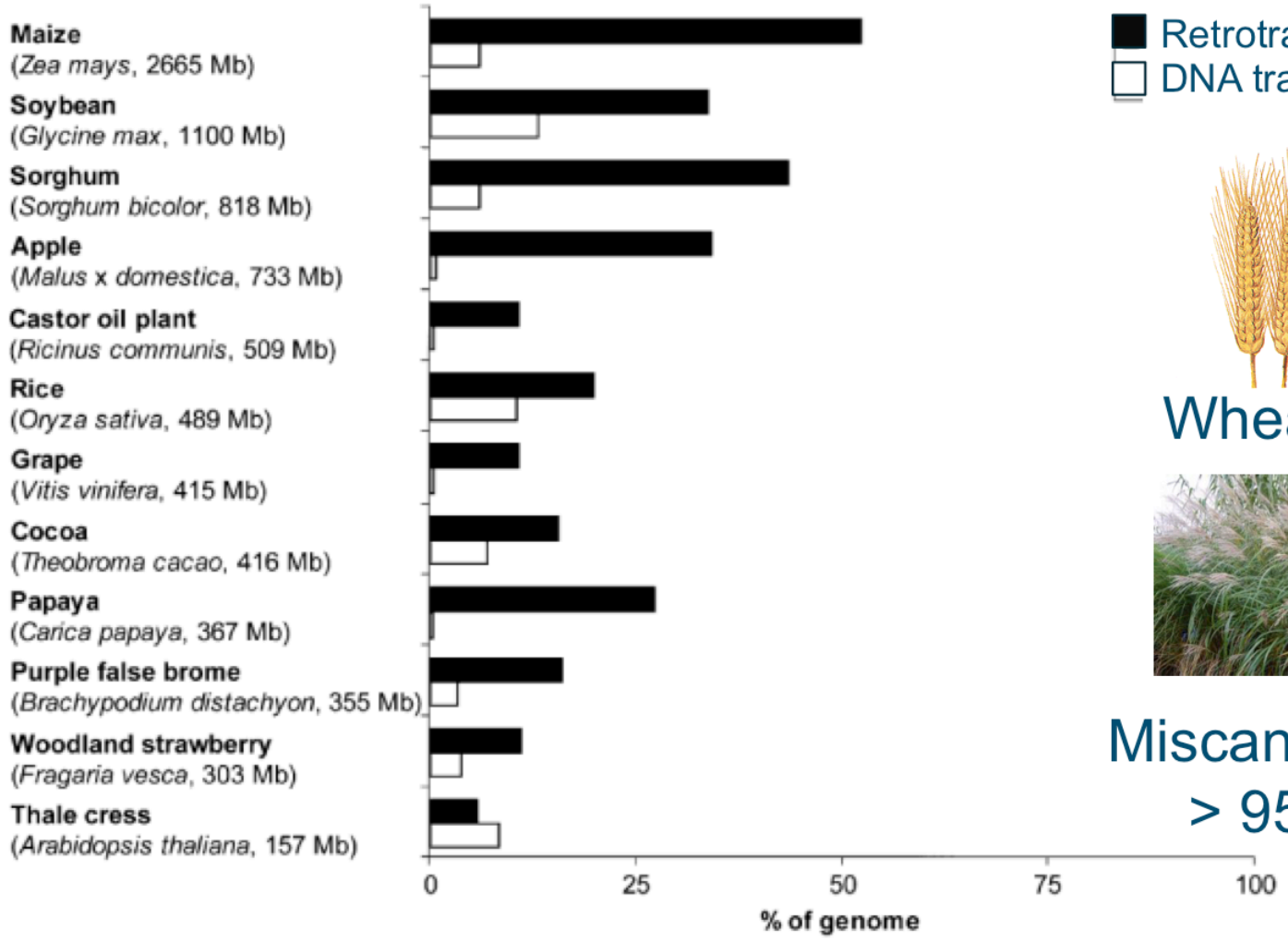
March 1, 2017

Animal genomes are highly repetitive



bionano GENOMICS

Plant genomes are highly repetitive



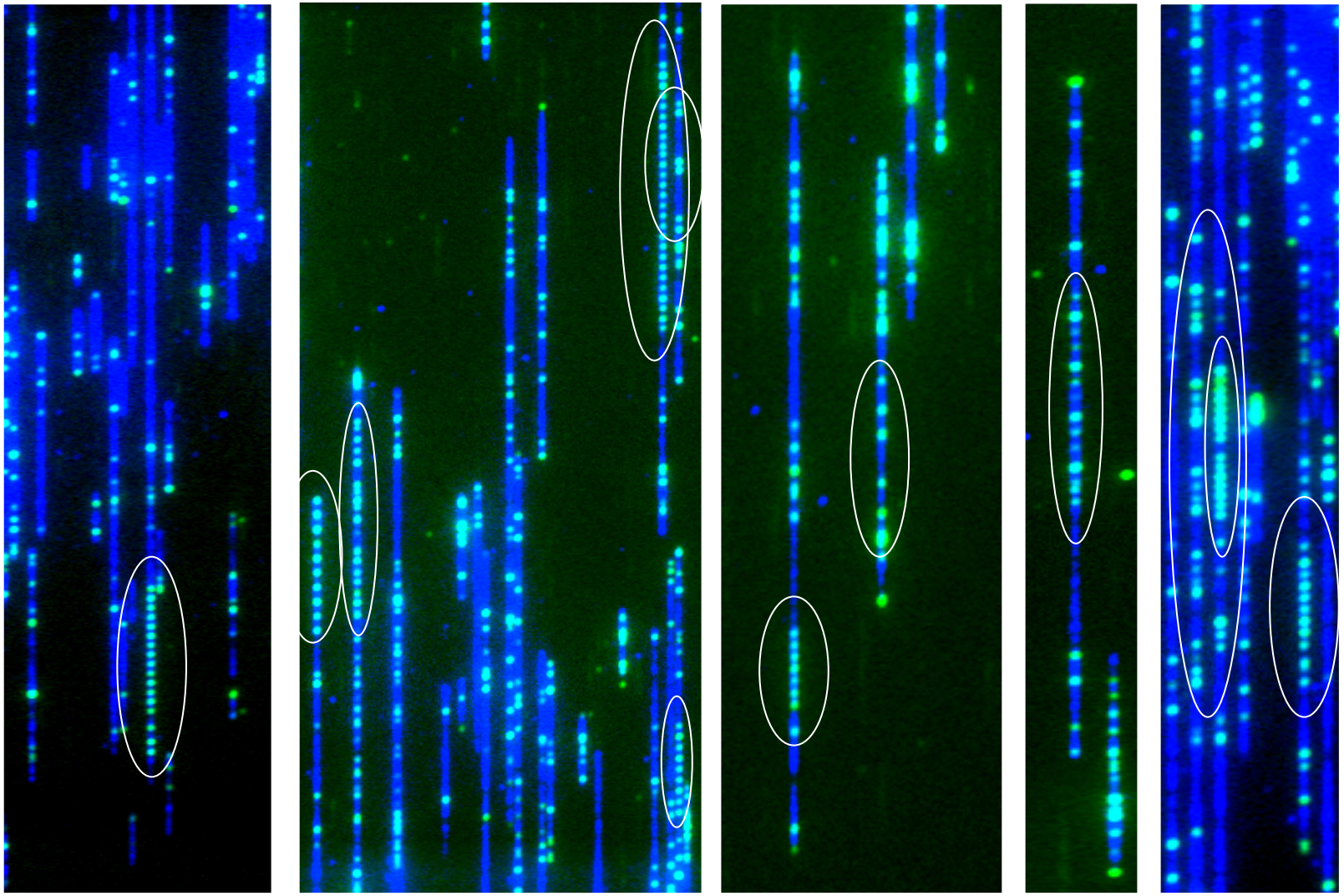
Wheat > 80%

Miscanthus (biofuel)
> 95%

Chr. Res (2011) 19:939-53

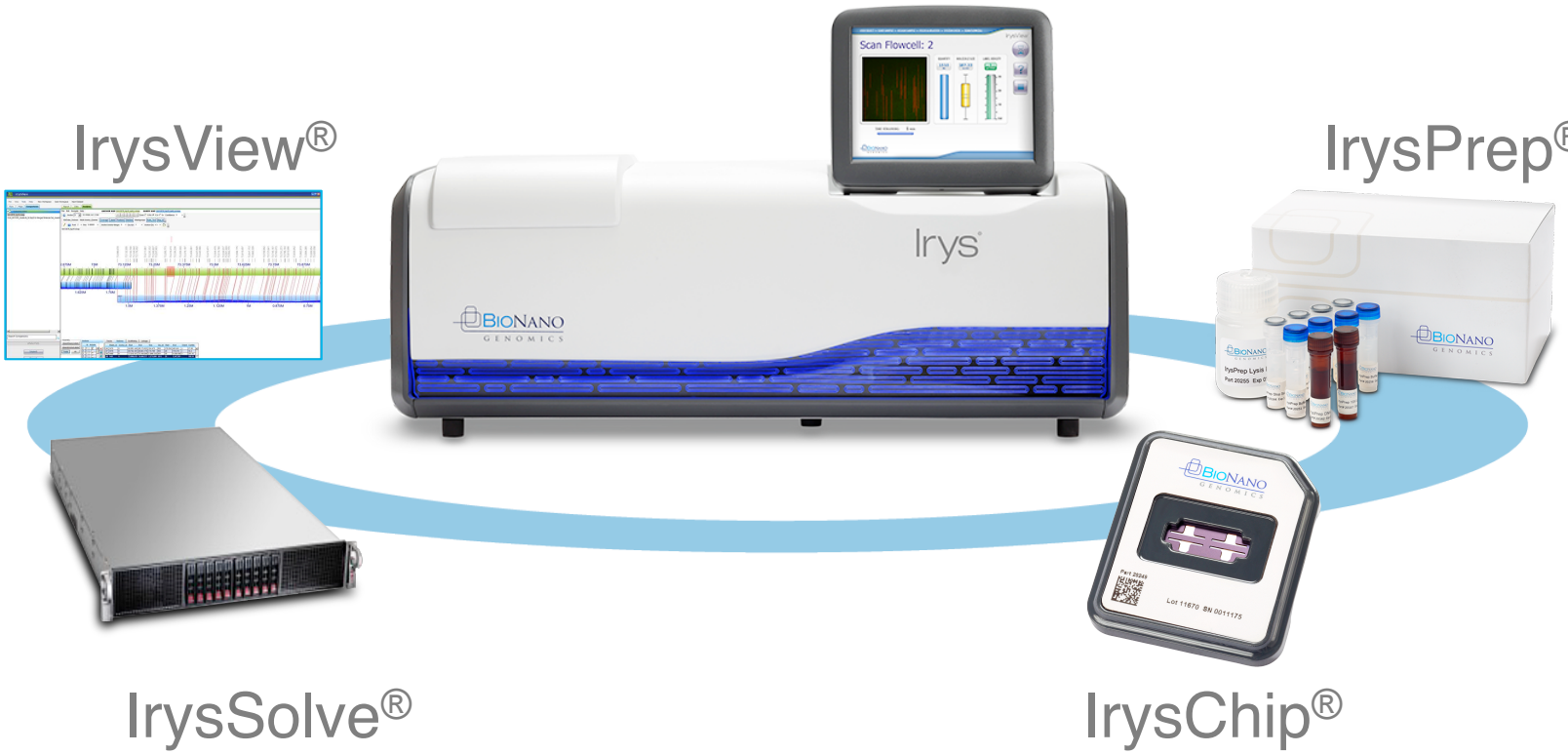
bionano GENOMICS

The Dark Matter of the Genome



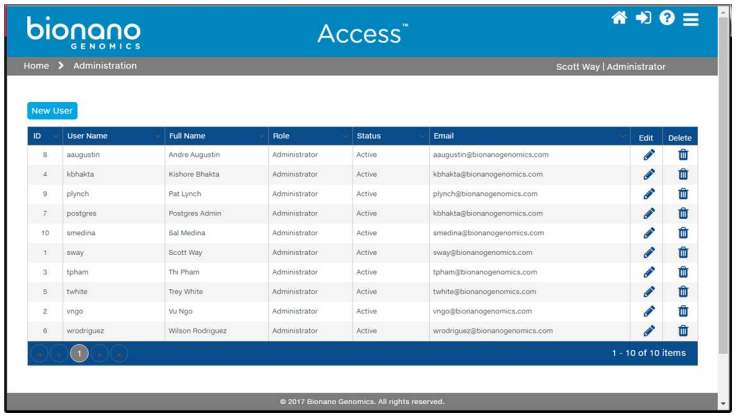
bionano GENOMICS

The Irys® System



New instrument, new analysis and visualization tools

Saphyr



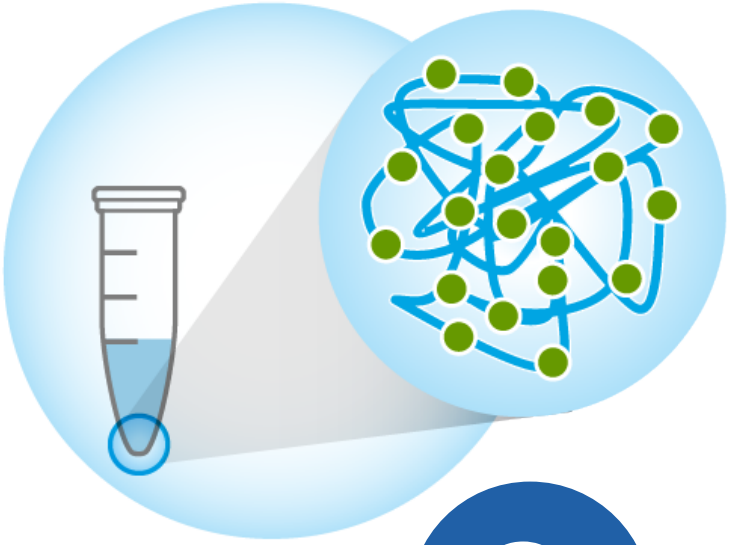
Bionano Solve 3.0

Bionano Access

Workflow

Extraction of long DNA molecules

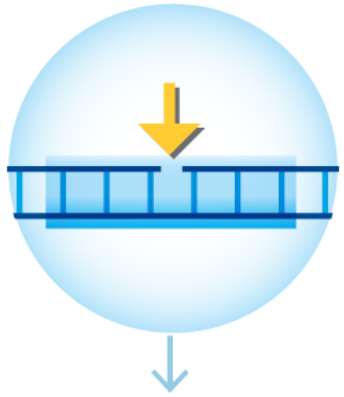
1



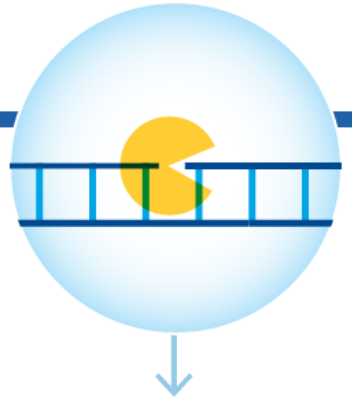
2

Label DNA at specific sequence motifs

A nicking endonuclease creates single-strand nicks at recognition sites



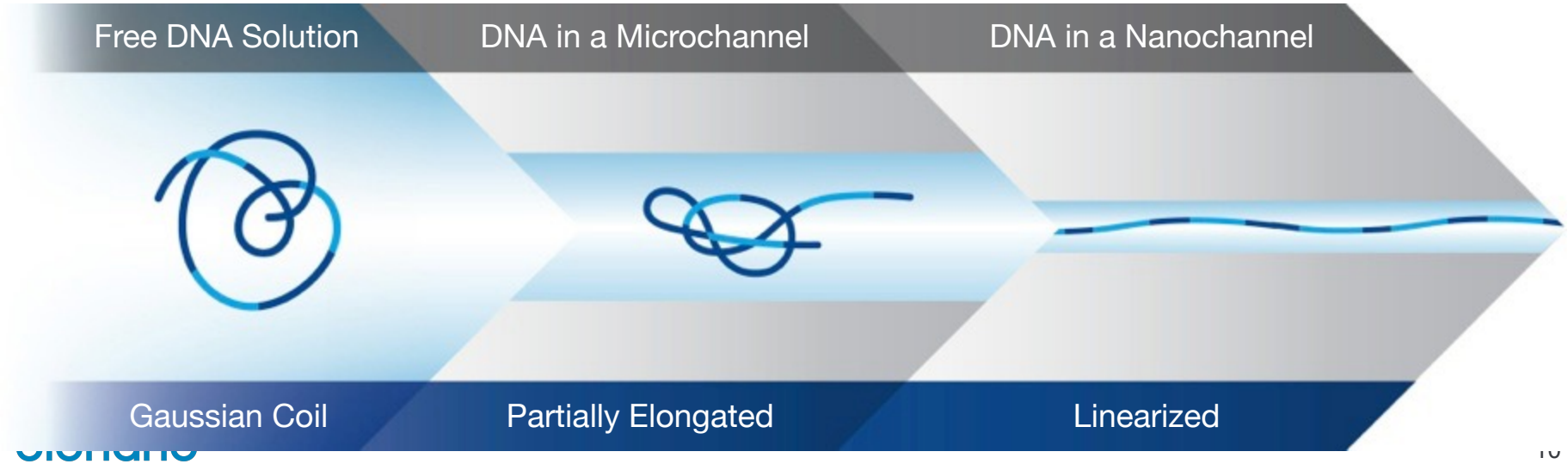
Polymerase initiates strand displacement and polymerization



Fluorescent nucleotides are incorporated to label recognition sites

Saphyr Chip linearizes DNA in NanoChannel array

3

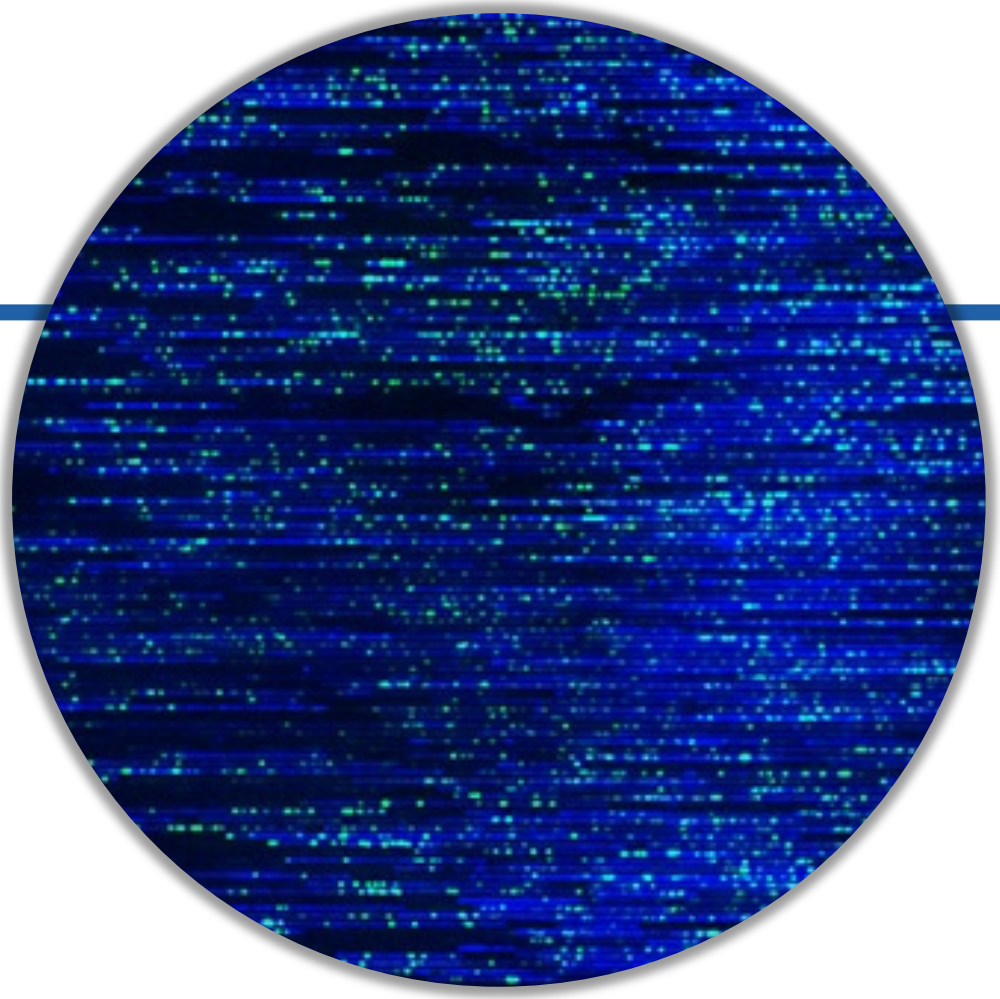


4

Single molecules are cycled through NanoChannel arrays and imaged

Molecules and labels detected in images

5



De novo consensus genome maps are assembled

6

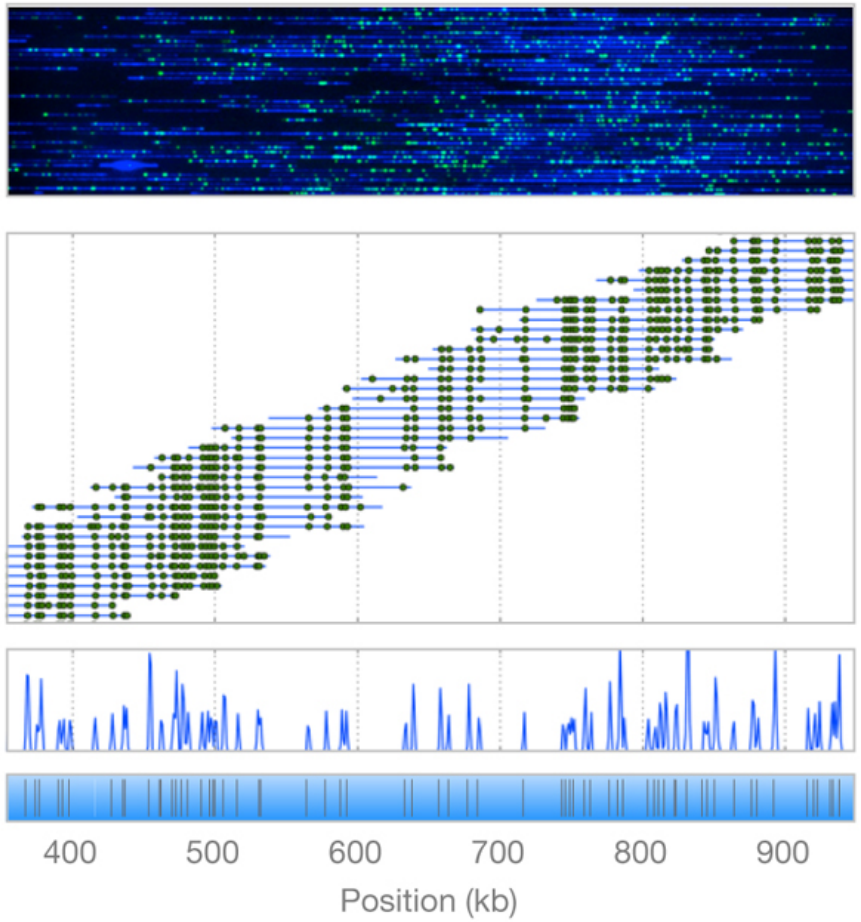
Raw Image Data

Conversion to Molecules

Population of Molecules

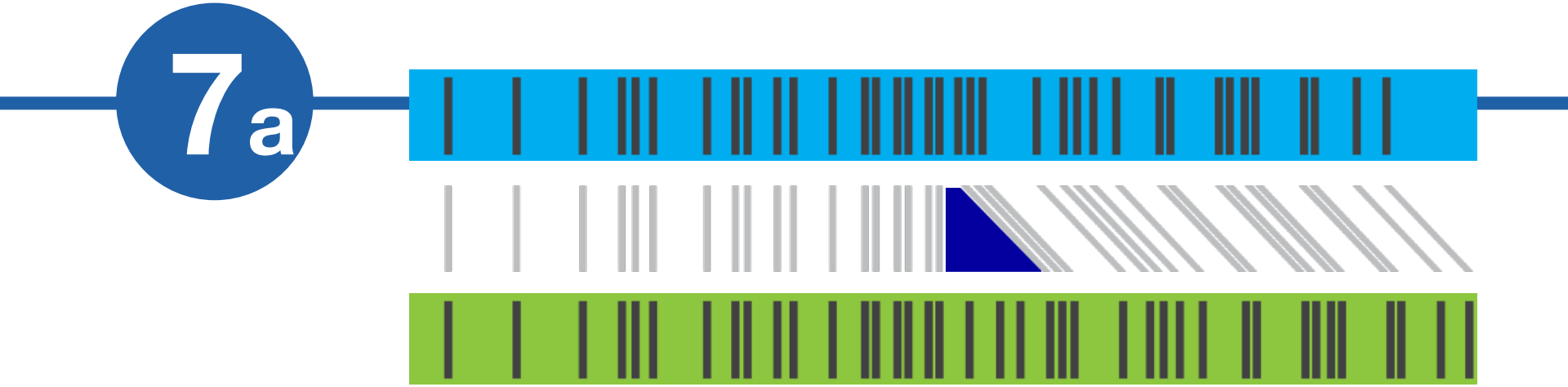
Aggregation & Assembly

Consensus Genome Map



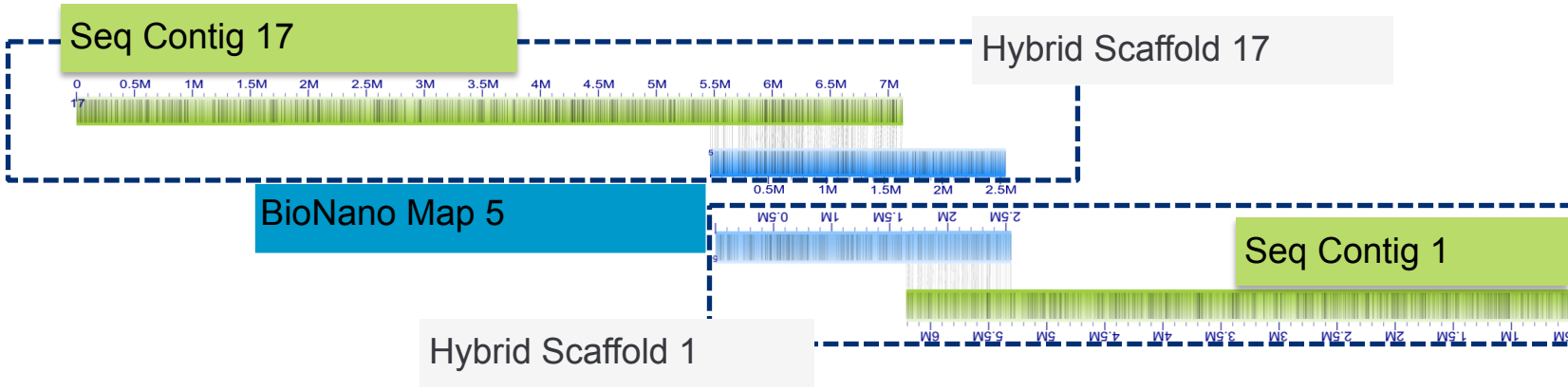
bionano GENOMICS

Structural variants called by comparing maps to reference or each other

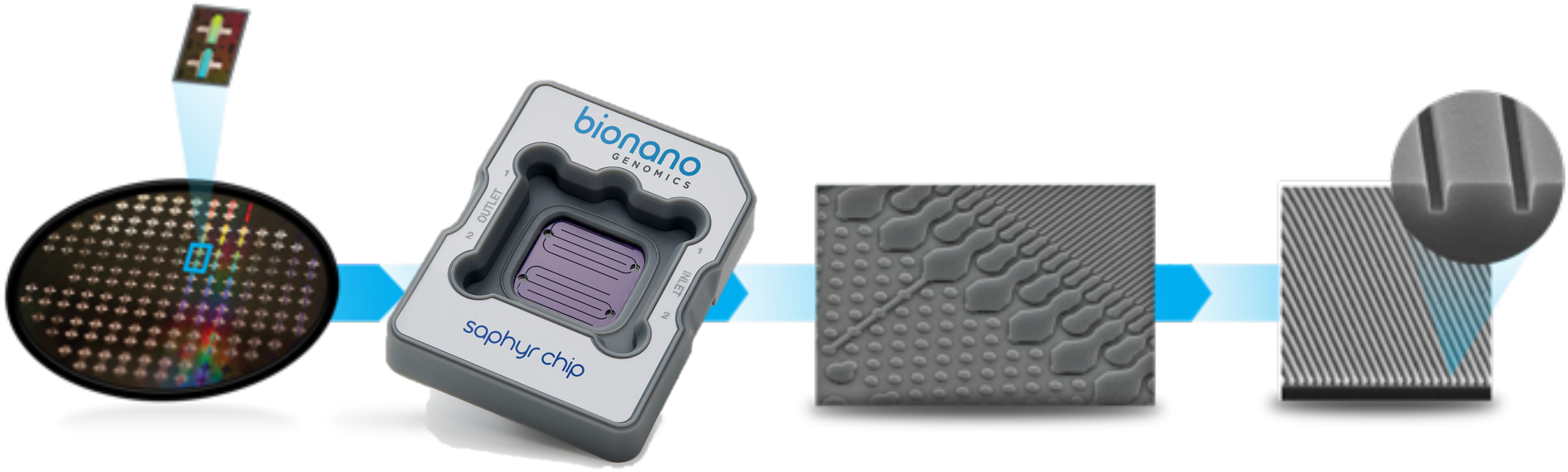


Genome maps are used to scaffold sequence contigs

7b



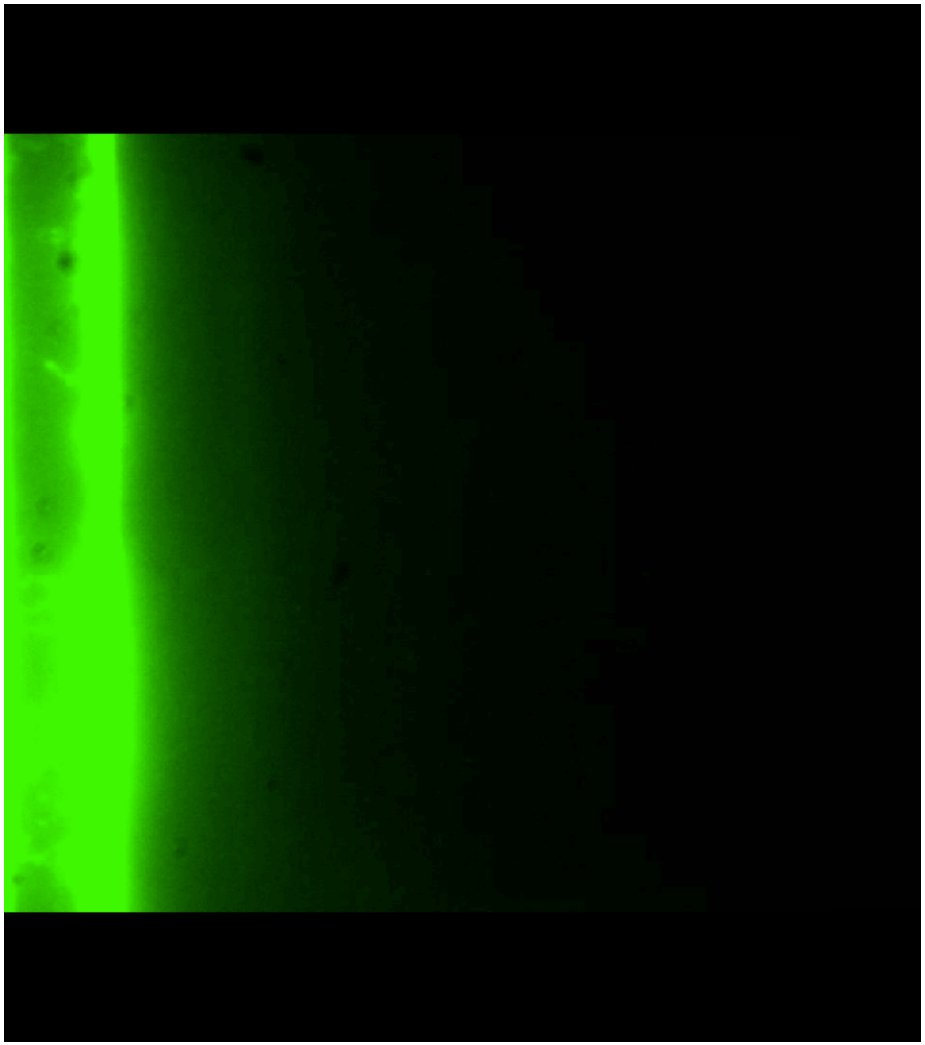
NanoChannel Arrays on Silicon



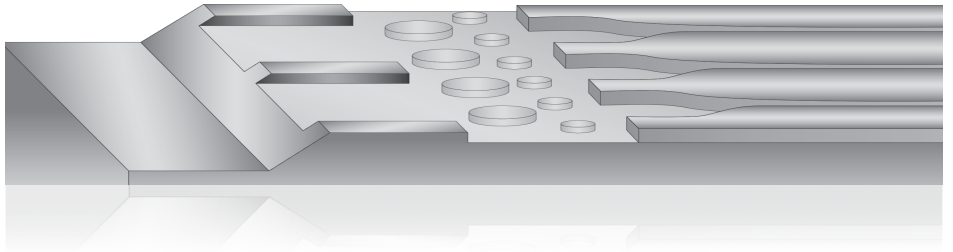
The Saphyr Chip

- 120,000 parallel nanochannels linearize long DNA in solution
- Leverages mature semiconductor manufacturing

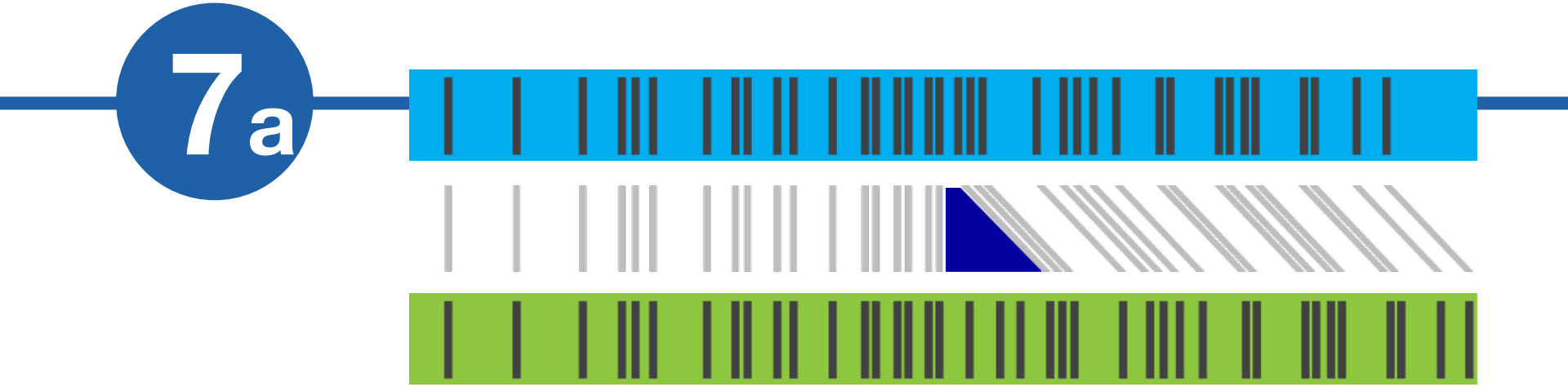
DNA Linearization in NanoChannel Arrays



Saphyr Chip

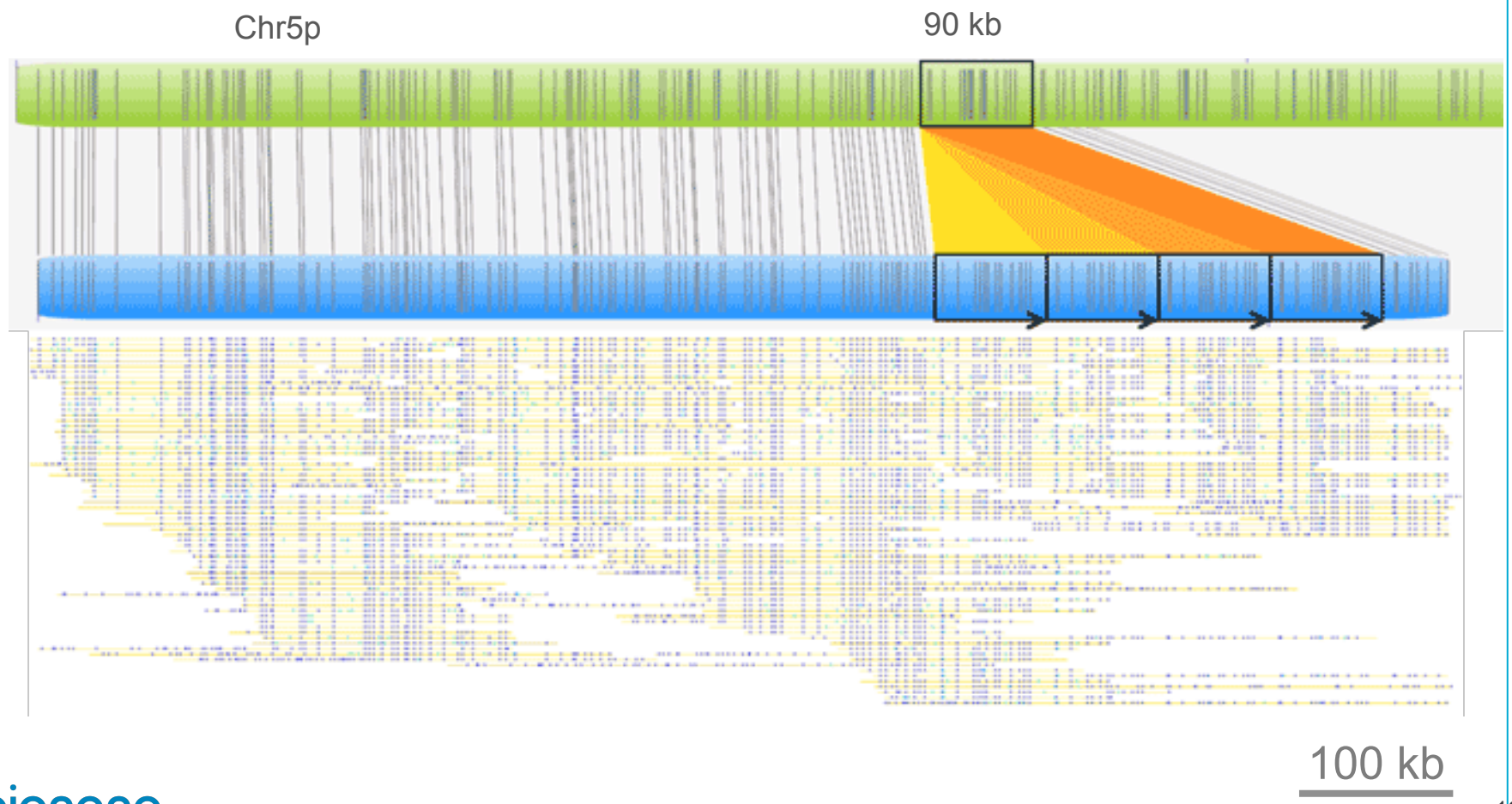


Structural variants called by comparing maps to reference or each other



Biologically Relevant Tandem Amplification Readily Detected

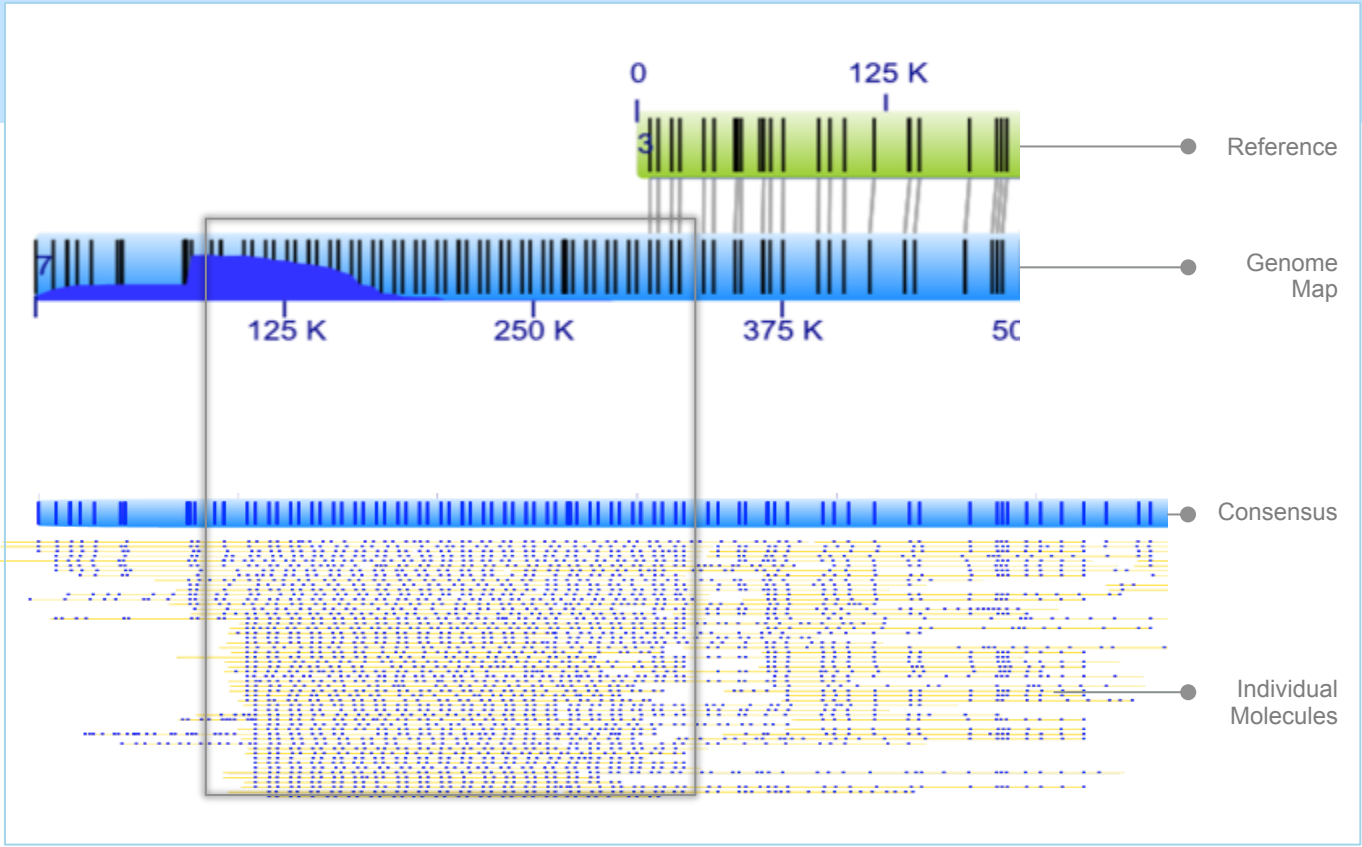
- Positional information and precise accounting of CNVs



Penetrating yeast sub-telomeric repeats

- Sub-telomeric region covered by our contigs extending beyond the end of chr3 in reference (*S. pombe*)

Schizosaccharomyces pombe
Genome Size:
13.8 Mb



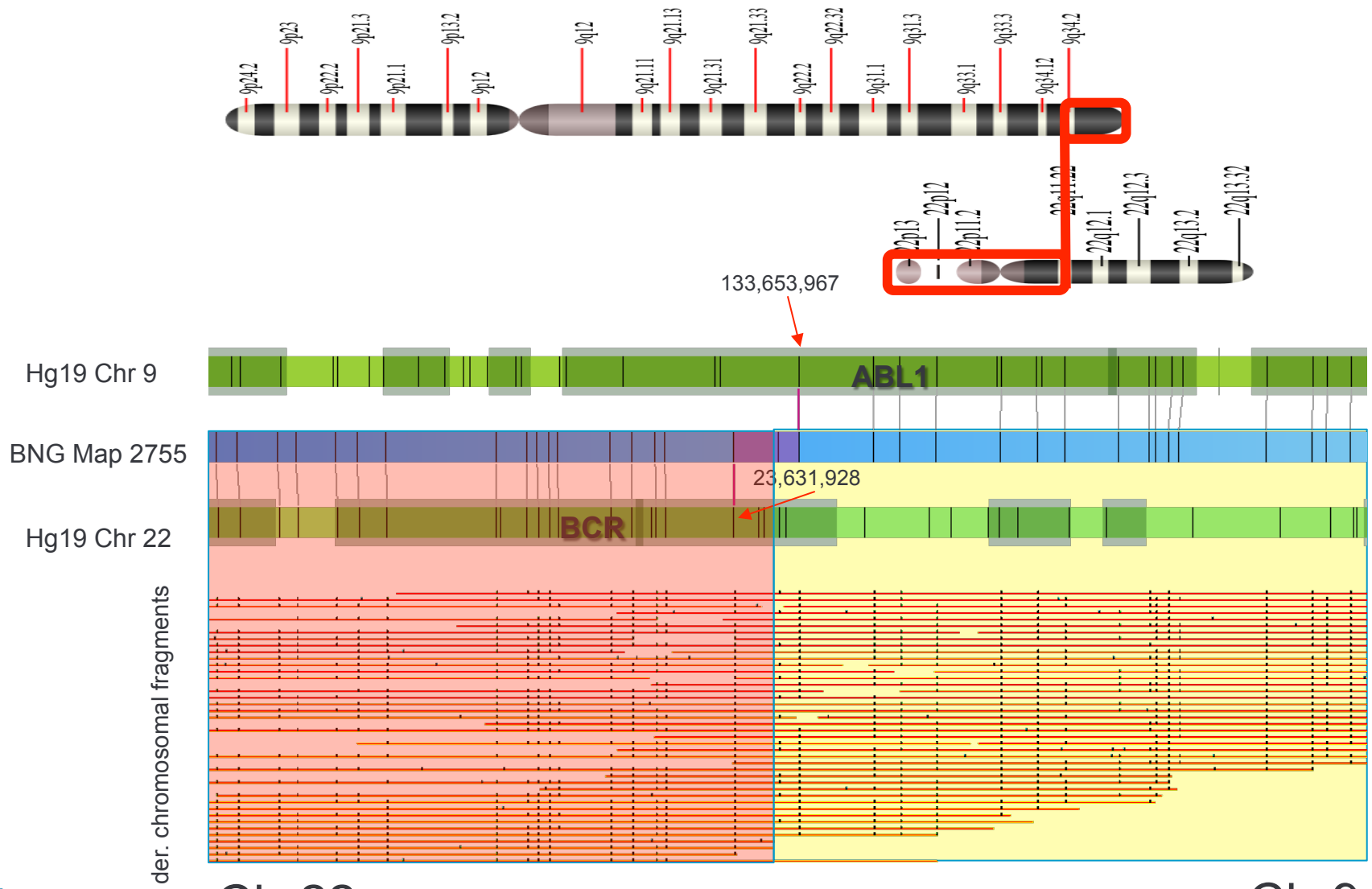
90%

99% sensitivity for homozygous insertions/ deletions

87% sensitivity for heterozygous insertions/ deletions

Sensitivity to detect large insertions and deletions

Translocation Detection



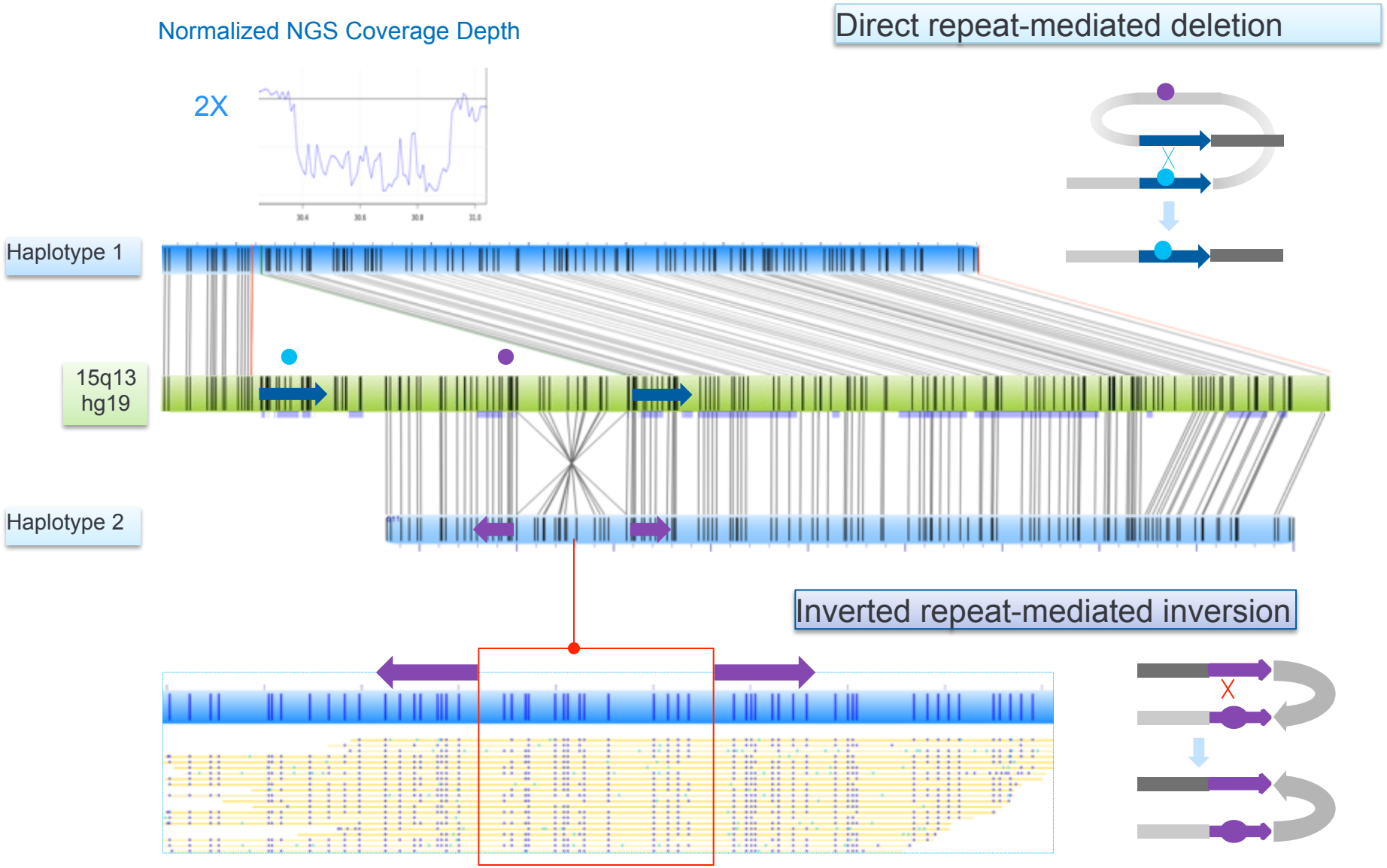
bionano GENOMICS

Balanced or Unbalanced

98%

Sensitivity to detect translocations

Visualization of Complex Rearrangements



Long molecules extending outside the Inverted Repeat confirm the inversion.

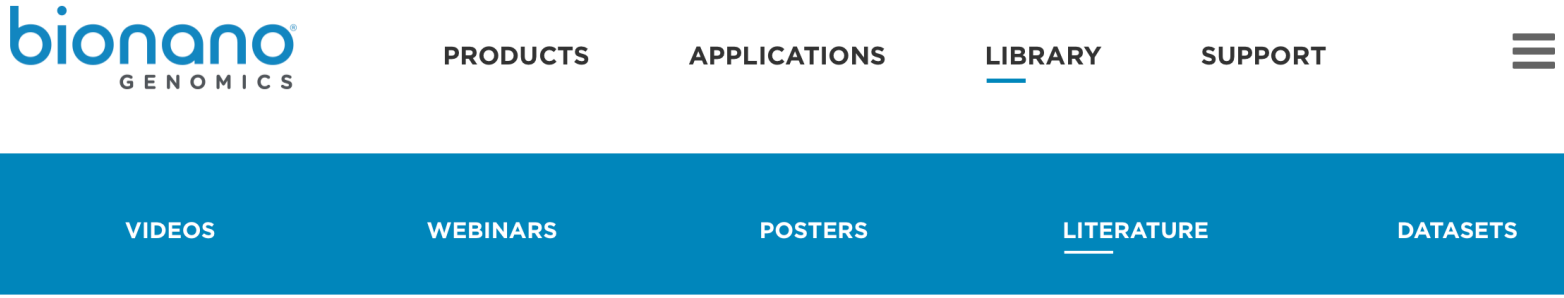
Inversions

CNVs

Repeat Array Sizing

Complex Rearrangements

Bionano White Papers



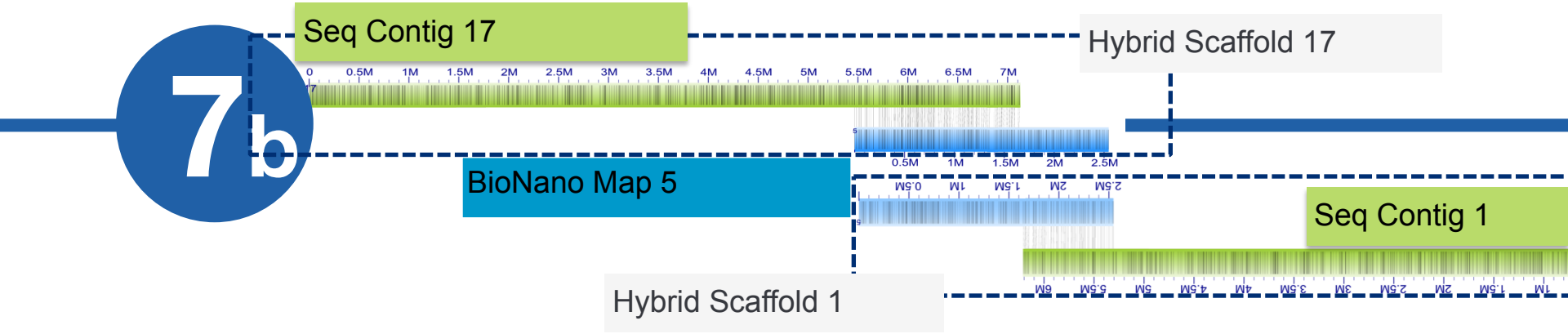
LITERATURE

Assembling High Quality Human Genomes: Going Beyond the '$1,000 genome'

This Case Study demonstrates the power of combining 2 single molecule technologies to produce Gold-quality genomes. Those allow the discovery of substantial amount of structural variation unique to individuals and populations otherwise not accessed by other short-read technologies.

First Comprehensive View of Maize Genome Reveals Regulatory and Structural Mechanisms

Genome maps are used to scaffold sequence contigs



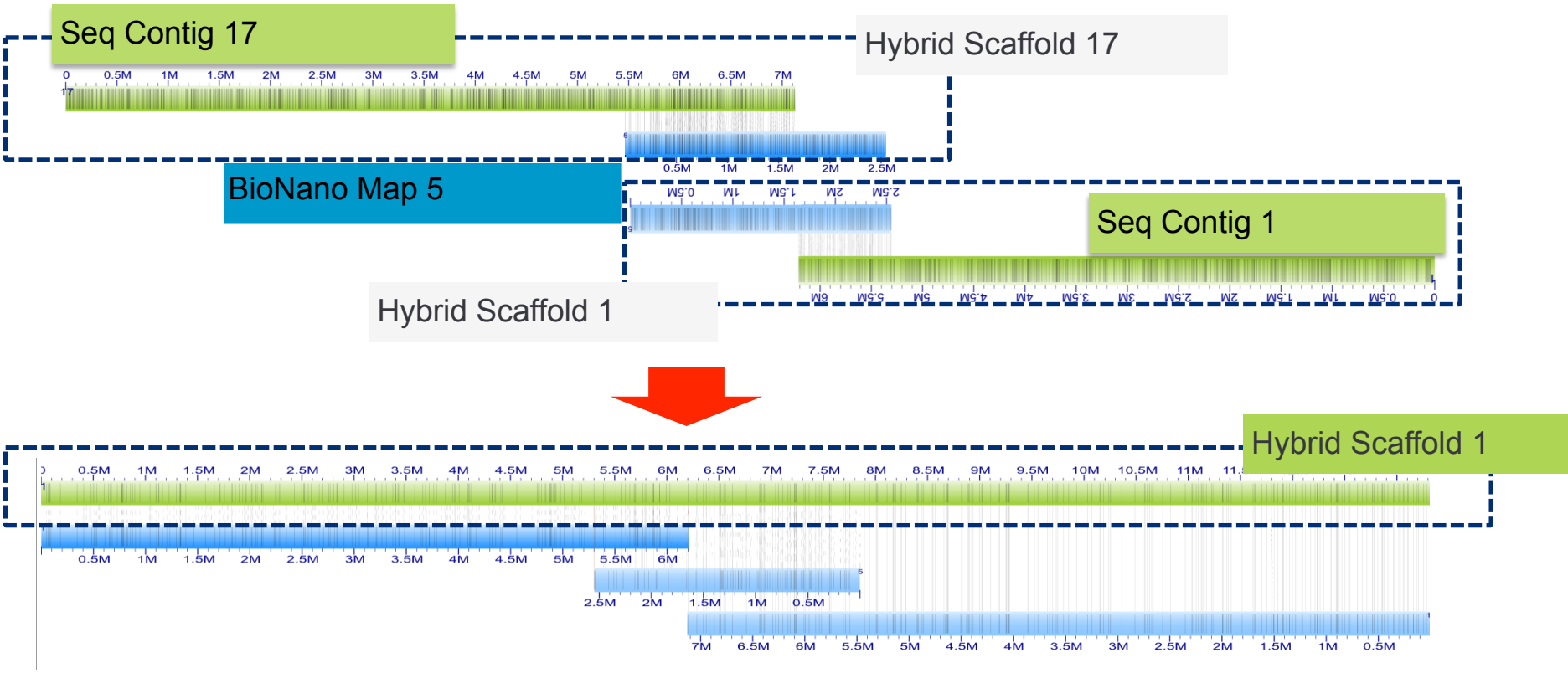
up to **100x**

Improvement in contiguity

Correct assembly errors

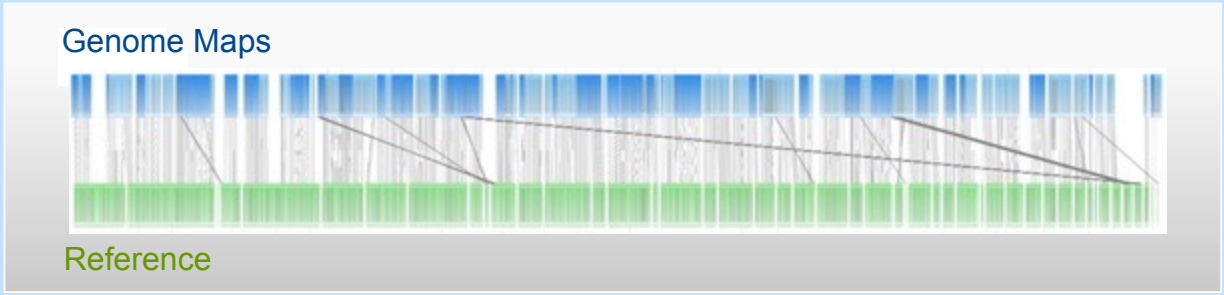
Reduce contig number

Hybrid Scaffold



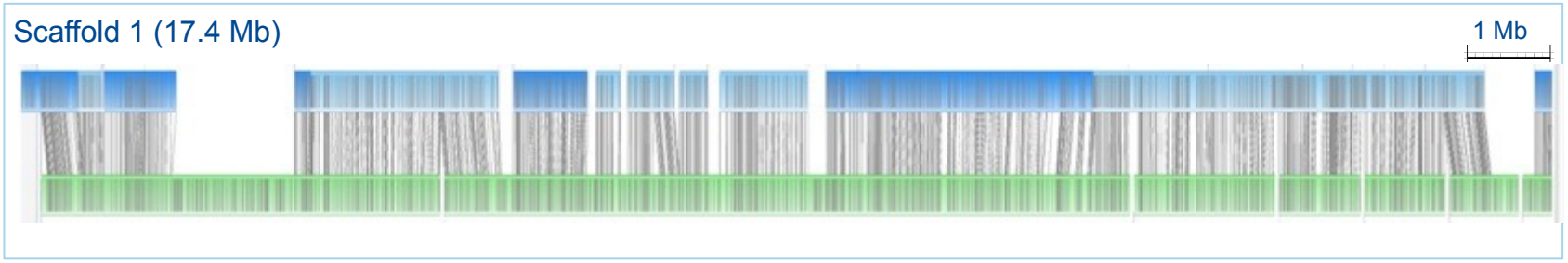
Scaffolding Spider Mite Assembly (Sanger)

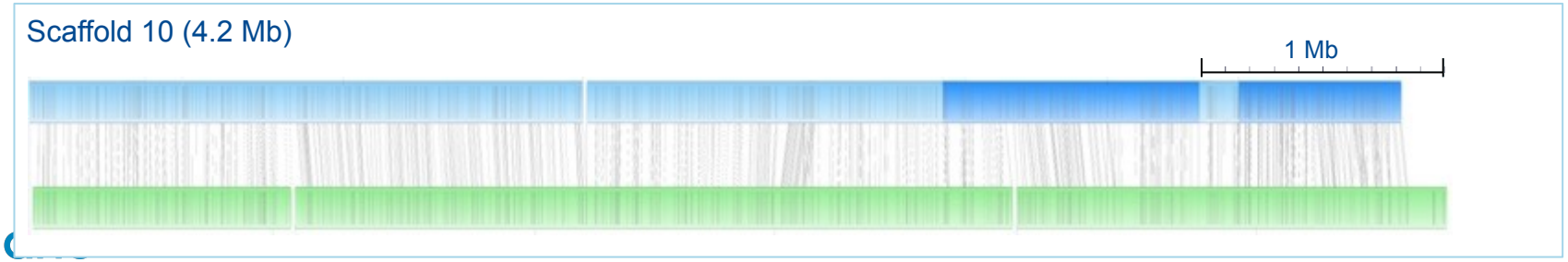
Whole Genome *De Novo* Assembly



Tetranychus urticae
Genome Size:
~90 Mb

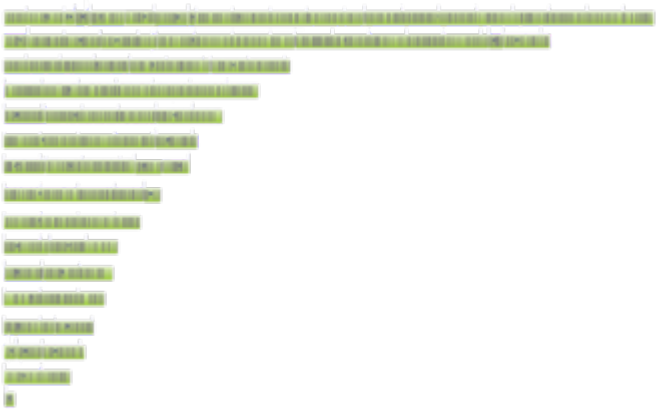
Scaffold Examples





GENOMICS

Spider Mite Genome Map Assembly

Sequence Assembly	Bionano Genome Map Assembly
44 Scaffolds > 100 kb	*15 Scaffolds > 100 kb*
	

Tetranychus urticae
Genome Size:
~90 Mb

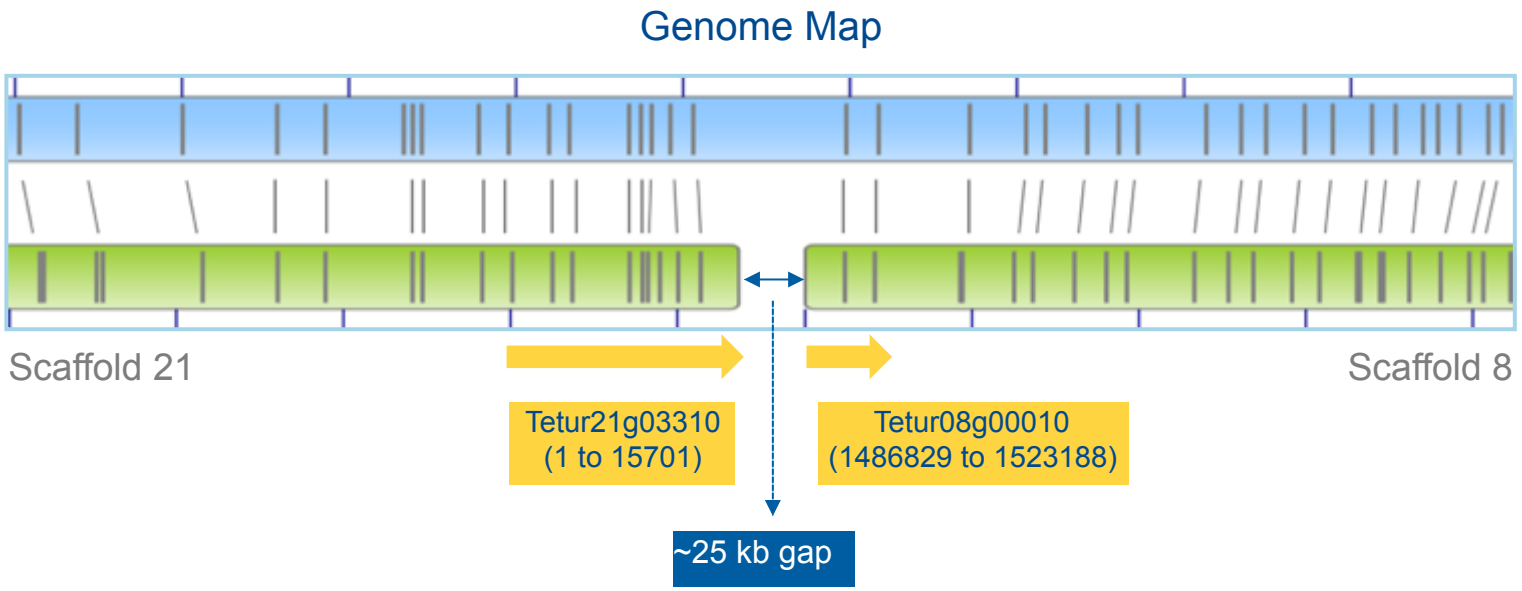
	Size	Scaffold N50	Largest Scaffold	Number of Scaffolds (>100 kb)
Sequence Assembly	*90.8 Mb (95% of genome)*	*3 Mb*	*7.8 Mb*	*44*
Genome Map Assembly	*90.8 Mb*	*6.8 Mb*	*17.4 Mb*	*15*

bionano GENOMICS

Assembly of Difficult Repeat-Rich Silk Genes

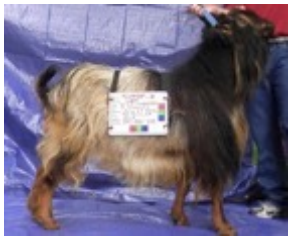
- Silk genes contain long repeating amino acid sequences encoded by repeating nucleotide sequences, making them difficult to assemble and often breaking contigs.

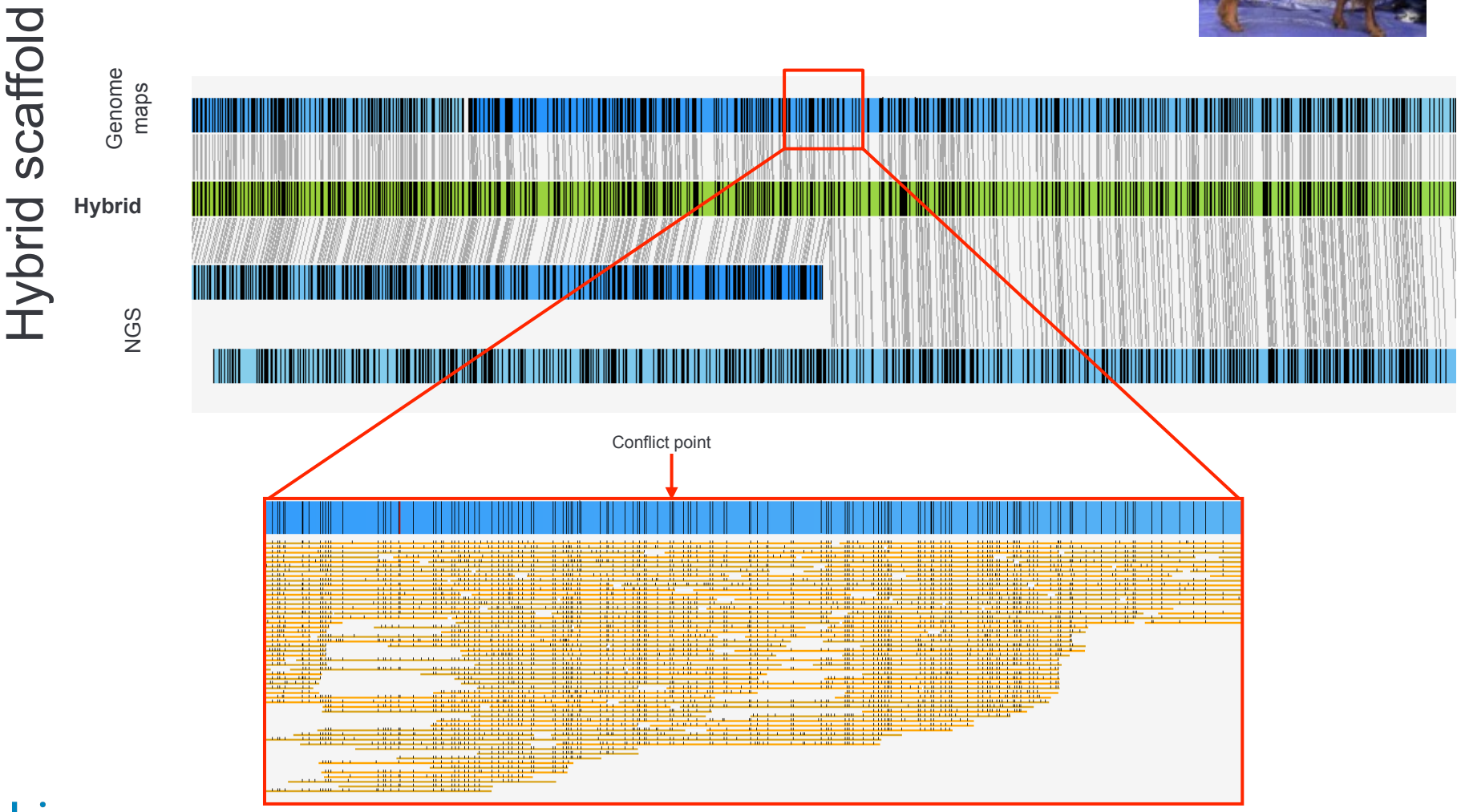
Tetranychus urticae
Genome Size:
~90 Mb



bionano GENOMICS

Sequence Assembly Error Correction (PacBio)
Evaluation of Conflicting Alignments



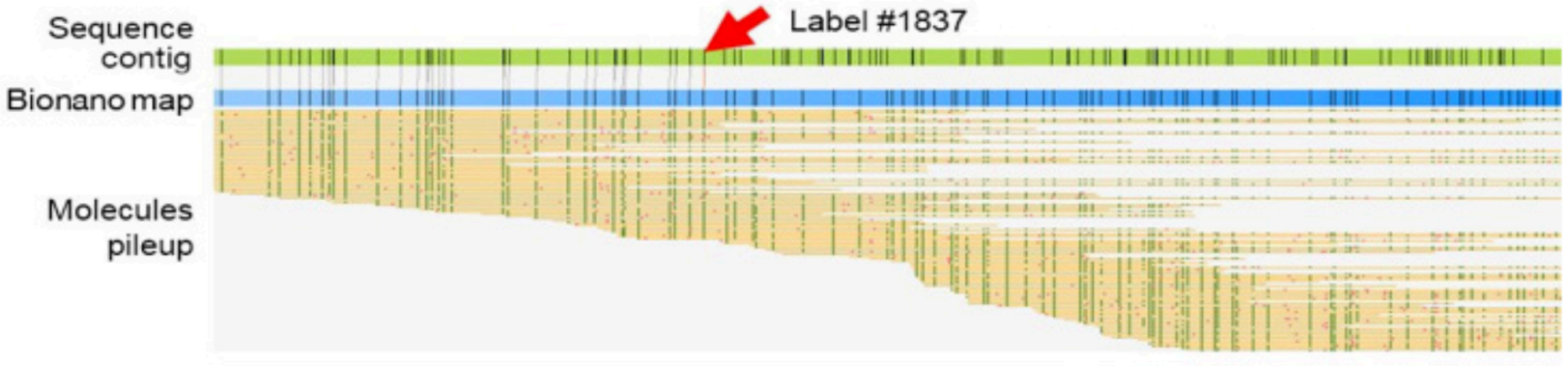


Long intact molecules provide strong support for the accuracy of the Bionano assembly.

*Bickhart and Rosen, USDA

Assembly Conflicts and Resolution

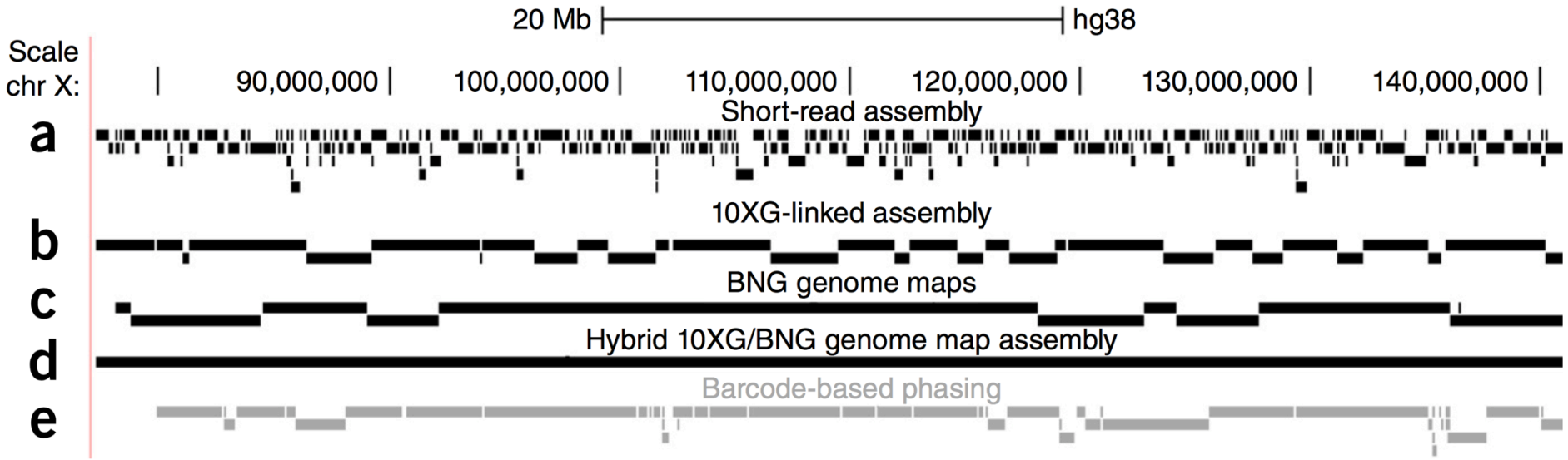
The Bionano hybrid scaffold pipeline detects and resolves chimeric joins



Species	# of Cuts on Sequence Confirmed/Total	# of Cuts on Bionano Confirmed/Total
Human NA12878	4 / 6 (67%)	1 / 1 (100%)
Goat	66 / 79 (84%)	11 / 16 (69%)
Maize	24 / 26 (92%)	12 / 13 (92%)

A hybrid approach for *de novo* human genome sequence assembly and phasing

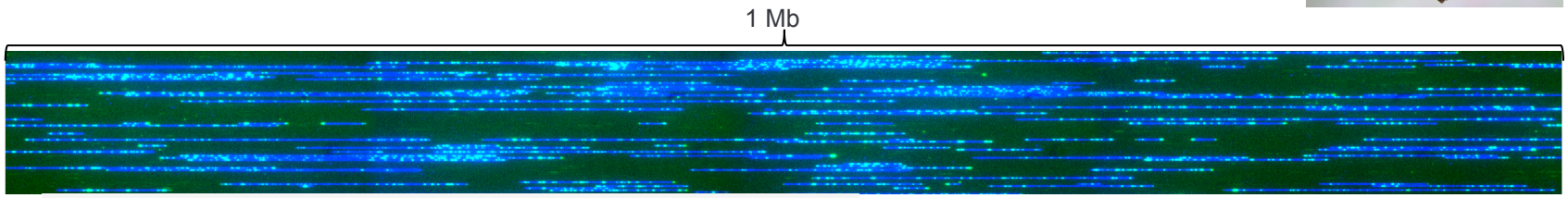
Yulia Mostovoy[1], Michal Levy-Sakin[1], Jessica Lam[1], Ernest T Lam[2], Alex R Hastie[2], Patrick Marks[3], Joyce Lee[2], Catherine Chu[1], Chin Lin[1], Željko Džakula[2], Han Cao[2], Stephen A Schlebusch[4], Kristina Giorda[3], Michael Schnall-Levin[3], Jeffrey D Wall[5] & Pui-Yan Kwok[1,5,6]

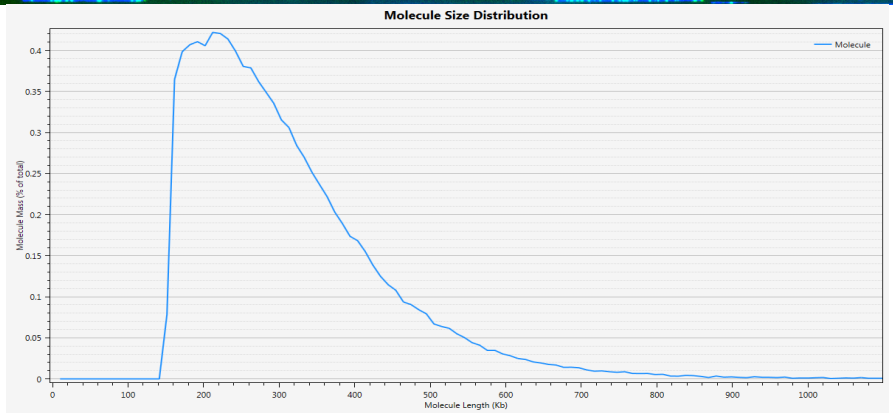


Nature Methods, May 2016; doi:10.1038/nmeth.3865

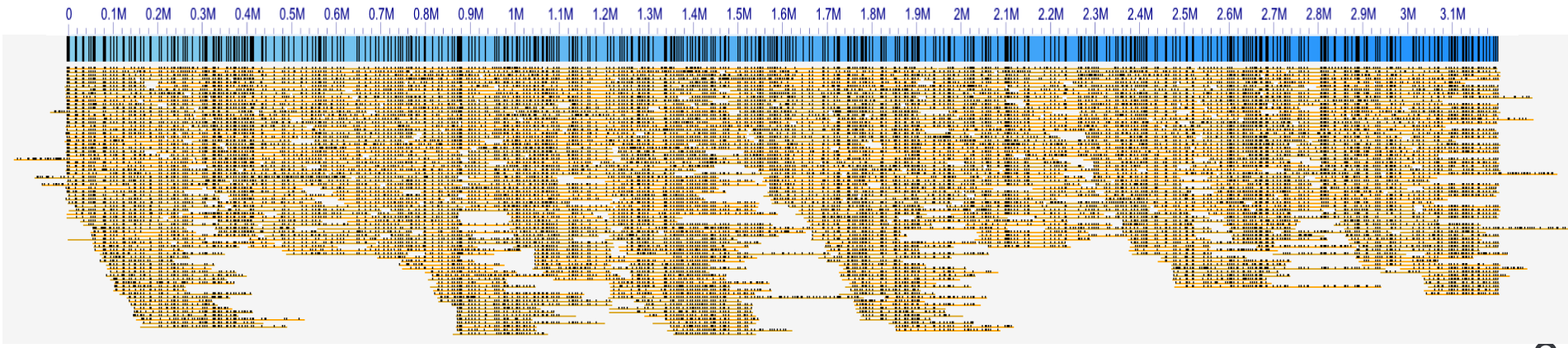
bionano GENOMICS

De novo Assembly - Hummingbird (PacBio)



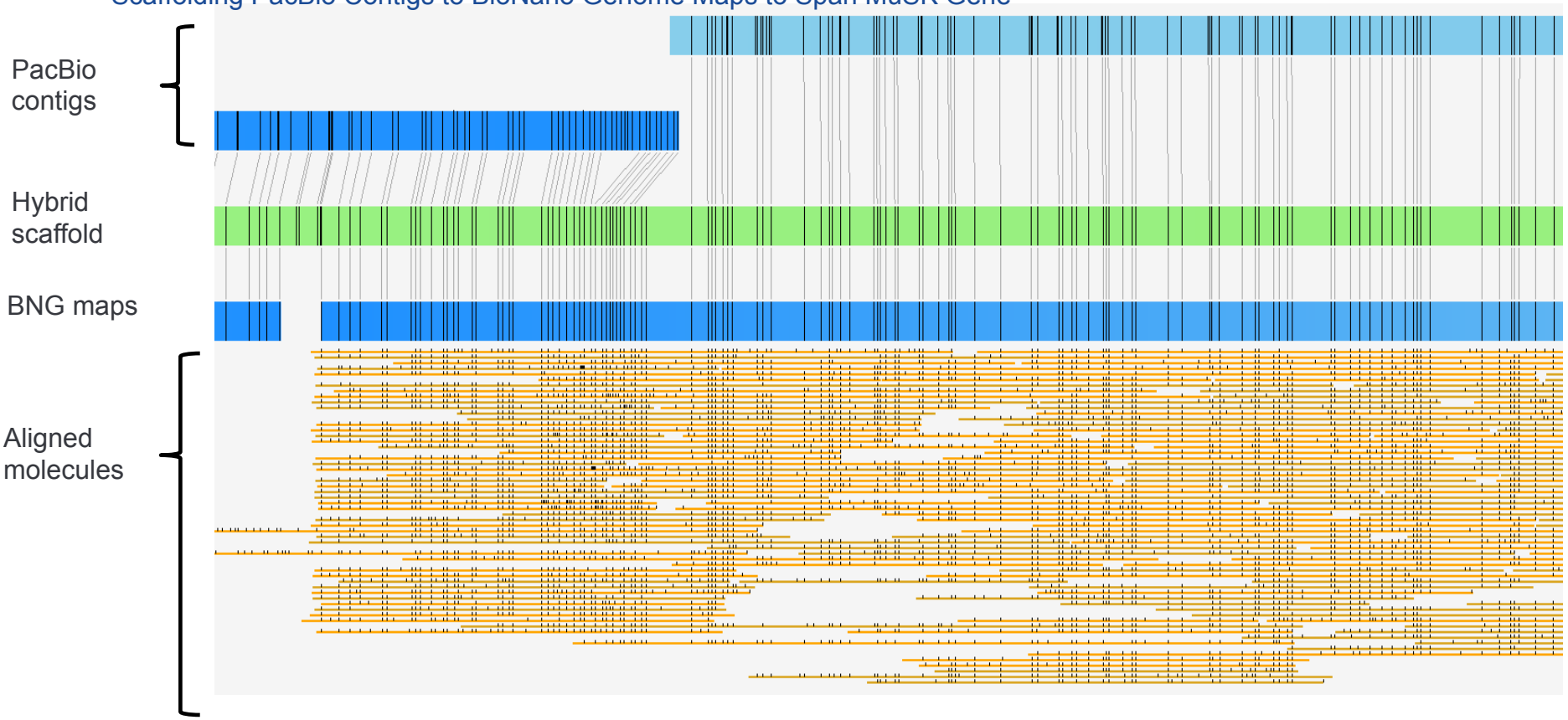


	BspQI
Data input (molecules >150kb)	95 Gb
Single molecule N50	245 kb
Genome map N50	0.7 Mb
Number of genome maps	1814
Total length	1.04 Gb



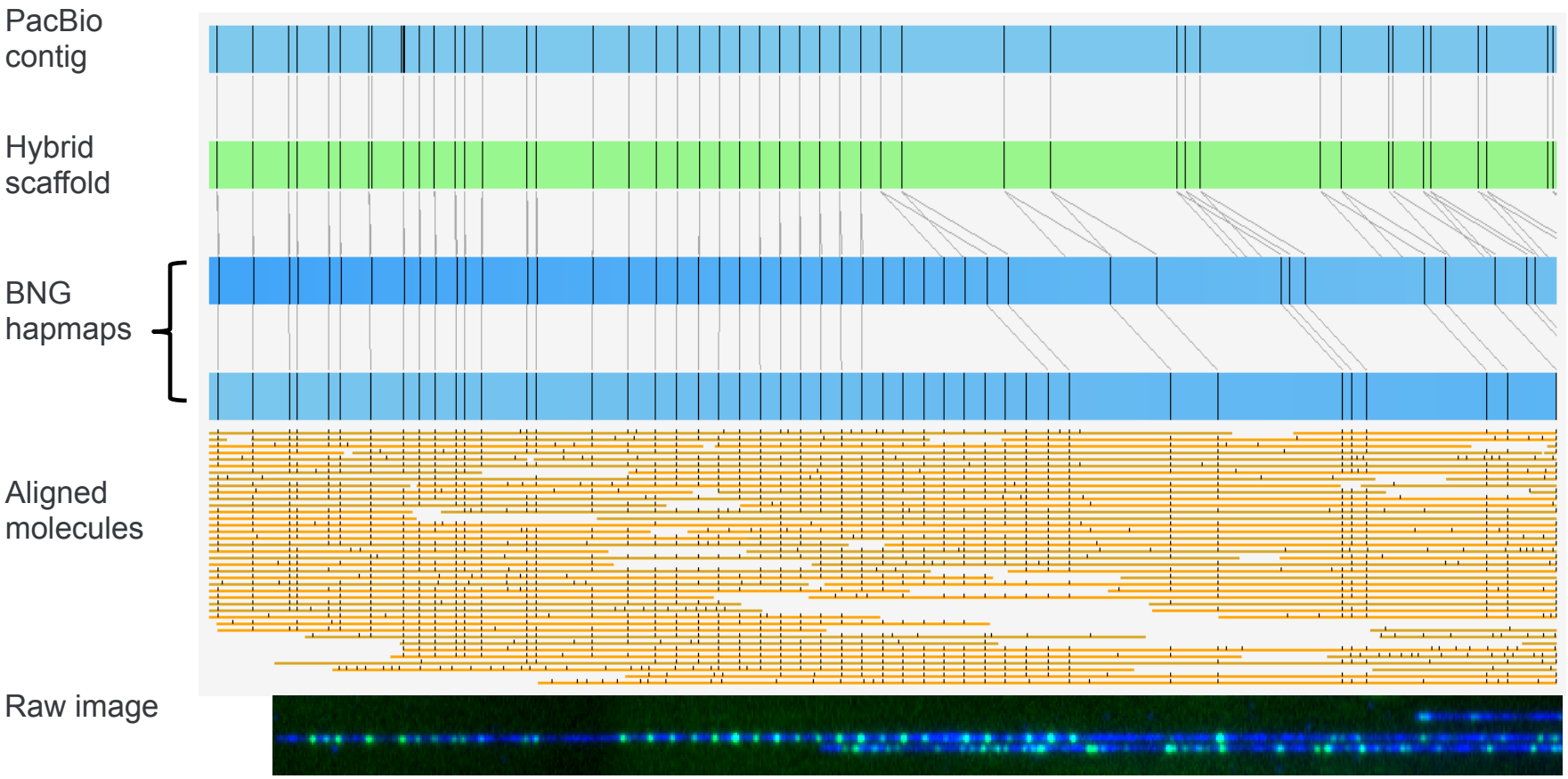
Gene Assembly by Scaffolding

- Scaffolding PacBio Contigs to BioNano Genome Maps to Span MuSK Gene



Repeat Haplotypes Spanning and Sizing

- *De novo* Assembly of Multiple Alleles and Their Correct Copy Numbers in MARK1 Gene



bionano GENOMICS

The Value of Combining Data Types

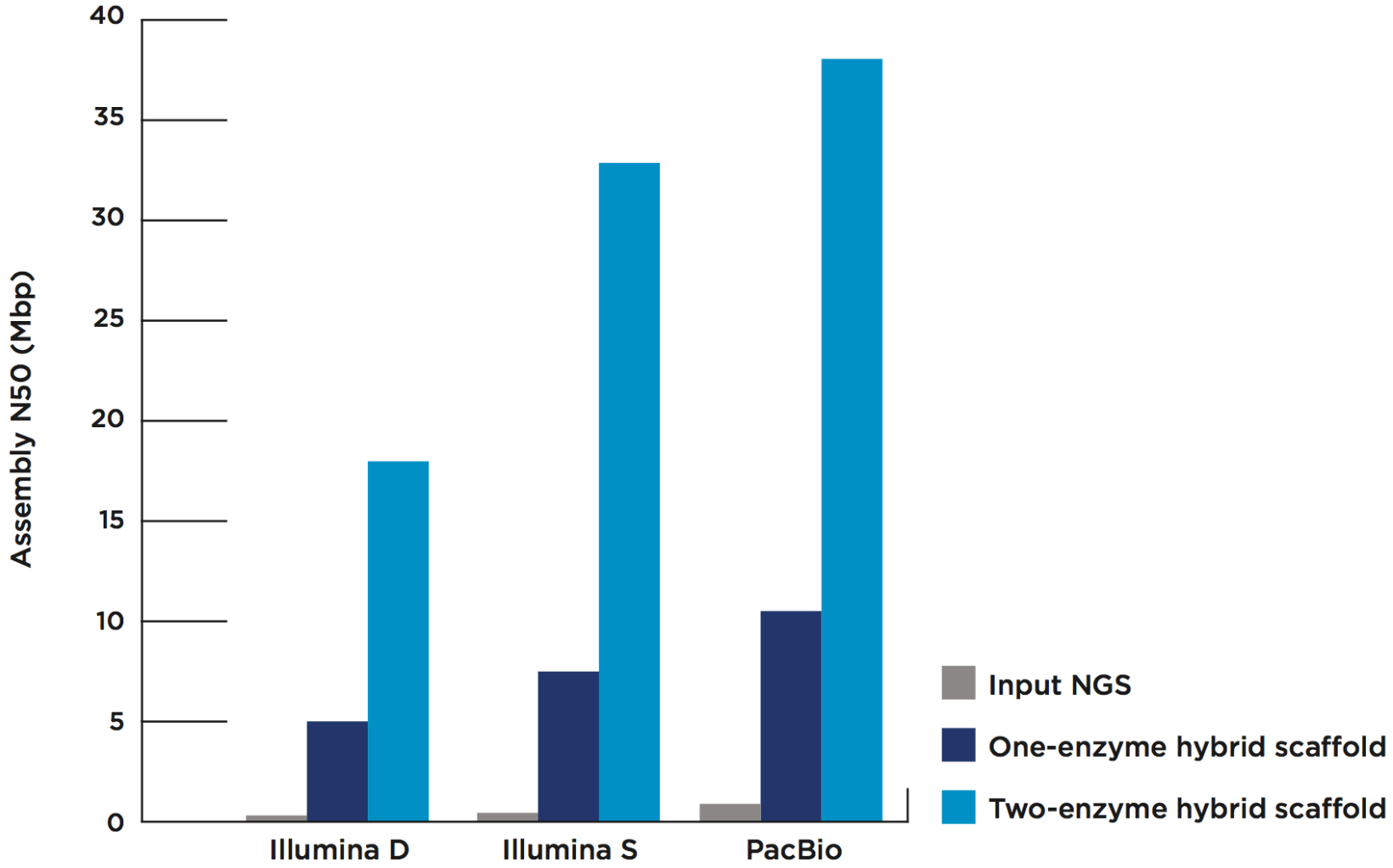
PacBio and BioNano Combined

PB Seq*	BNG*	# Contigs	N50 (MB)	Max Length (Mb)	Total Length (Mb)
90x	*0*	*7764*	*9.5*	*33.8*	*3045*
50x	0	9417	7.0	39.7	2956
30x	0	12341	3.5	24.1	3051
0	100x	2613	1.5	8.6	2833
0	50x	2864	1.3	8.5	2746
90x	100x	278	22.9	75.8	2822
50x	100x	386	18.5	77.3	2797
30x	100x	473	12.3	47.5	2805
90x	50x	317	21.3	74.0	2796
50x	50x	408	17.4	77.0	2744
30x	**50x**	**532**	**11.5**	**44.0**	**2761**

bionano GENOMICS

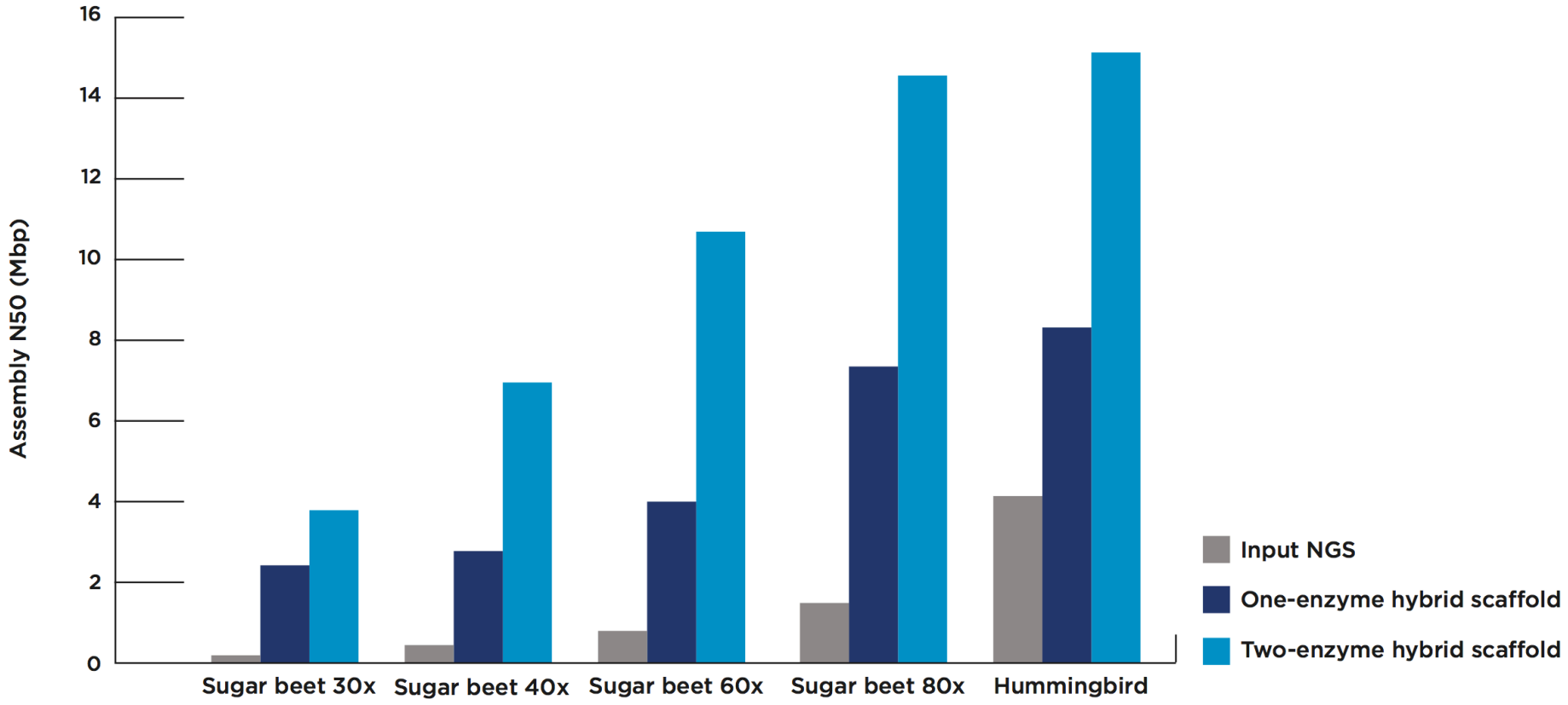
Higher Levels of Contiguity Using Two-Enzyme Hybrid Scaffolding

Assembly contiguity can be further increased by performing hybrid scaffolding with maps using two separate nicking enzymes.



bionano GENOMICS

Higher Levels of Contiguity Using Two-Enzyme Hybrid Scaffolding



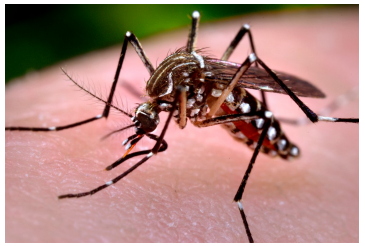
bionano GENOMICS

Scaffolding Human 10xGenomics Assembly

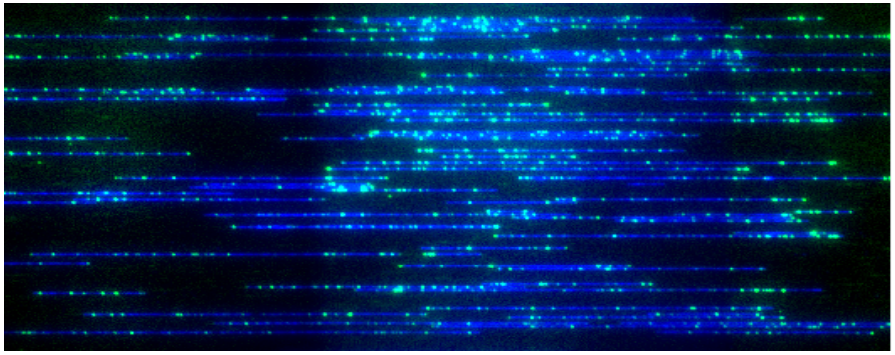
Assembly	Total Map Length (Gb)	Number of Scaffolds	Scaffold N50 (Mb)	Longest Scaffold (Mb)
illumina	2.79	14,047	0.59	5.57
10x	2.81	5,697	7.03	37.9
BNG	2.93	1,079	4.59	26.6
Hybrid (combined illumina, 10x, BNG)	2.86	170	33.5	99.96

Source: Nature Methods, May 2016; doi:10.1038/nmeth.3865

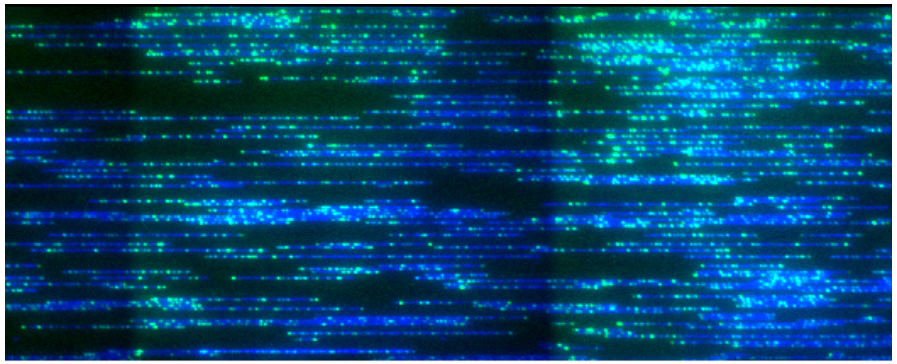
De Novo Assembly of *Aedes aegypti* – Vector for Dengue Fever, Zika and Yellow Fever Viruses

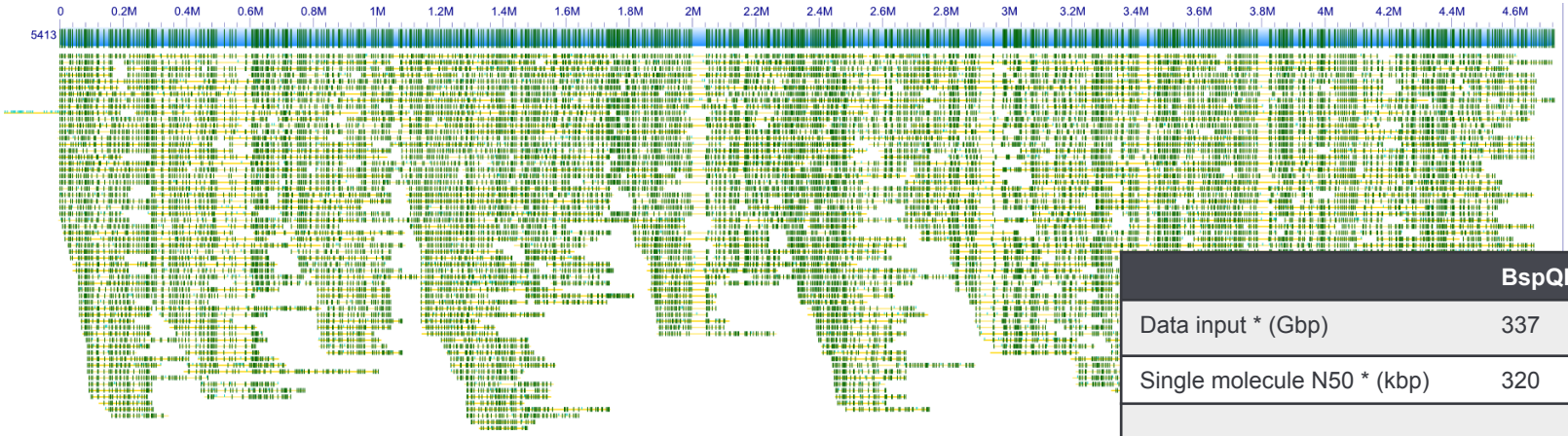


Nt.BspQI



Nb.BssSI





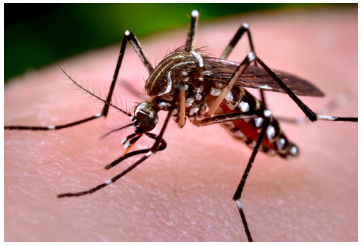
	BspQI	BssSI
Data input * (Gbp)	337	216
Single molecule N50 * (kbp)	320	245
Genome map N50 (Mbp)	1.47	0.75
Number of genome maps	1606	3152
Total length (Mbp)	1779	1967

*molecules >150kb

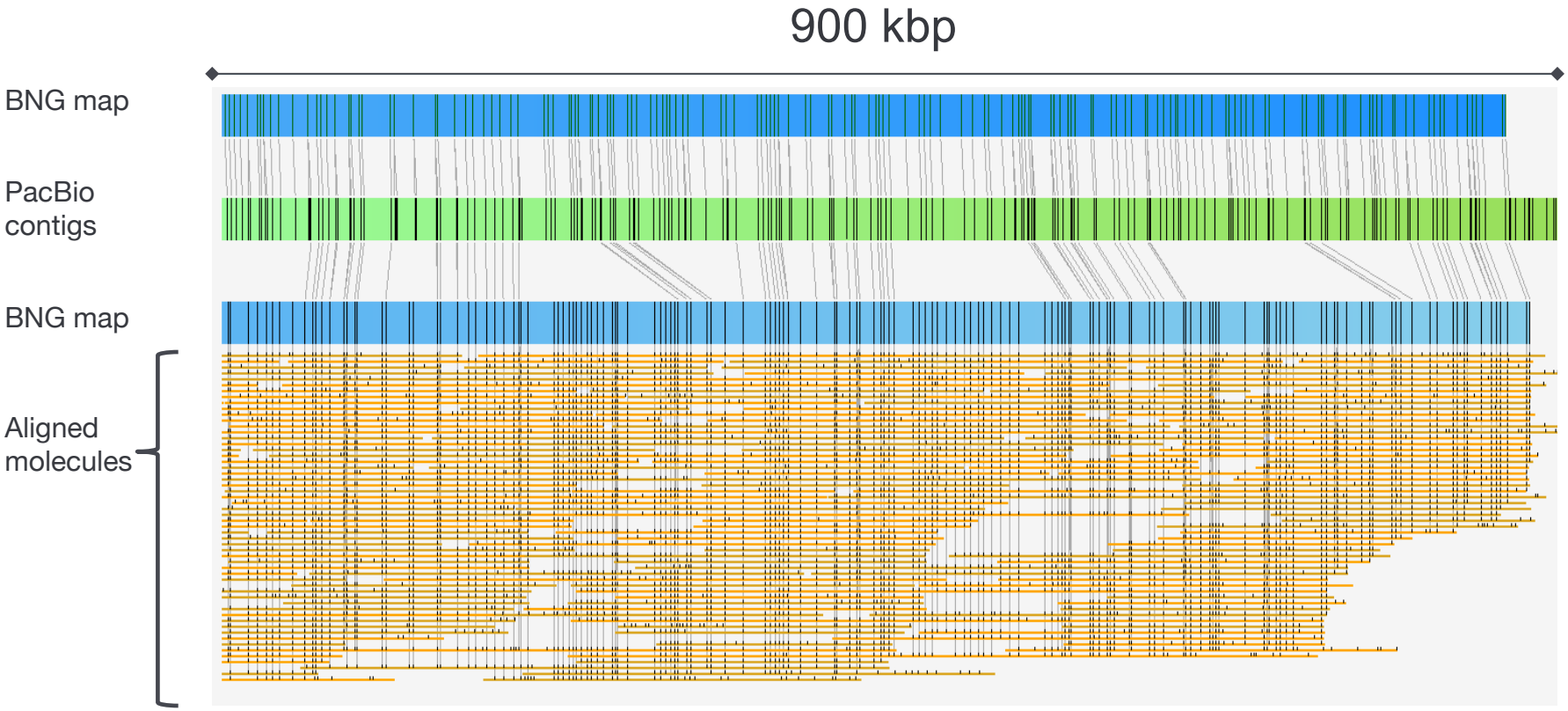
bionano GENOMICS

THE ROCKEFELLER UNIVERSITY
Science for the benefit of humanity

Comparison of Bionano maps to PacBio



Very high amount of heterozygosity in these individuals resulted in sequence assemblies with high error and double expected (haploid) genome size



bionano GENOMICS

THE ROCKEFELLER UNIVERSITY
Science for the benefit of humanity

Primary assembly statistics

	Assembler	Input data	Total size	Primary size	Contig number	version	Contig NG50*
Vectorbase	ARACHNE	~8x shotgun	1.31 Gb	1.31 Gb	36205	AaegL3	83.4 kb
PacBio	Falcon-UNZIP	PacBio	2.04 Gb	1.69 Gb	3967	v1.0	1.88 Mb
				1.45 Gb	3462	v1.1	1.67 Mb
NIH	canu	PacBio + Yale	2.7 Gb	-	19087	v1.0	2.46 Mb
				1.30 Gb	1585	v1.1	2.37 Mb
		PacBio only	2.2 Gb	-	11839	V2.0	2.56 Mb

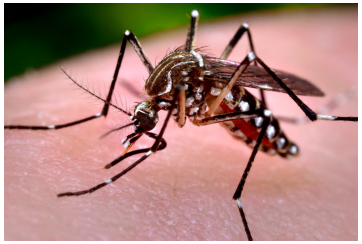
Scaffolded assembly statistics

	Method	Input assembly	Original contig number	New contig number	Original contig NG50*	New scaffold NG50*	'Cuts'
BioNano	'Irys' optical mapping	Falcon v1.0 (all contigs)	7790	508	1.90 Mb	10.7 Mb	647
		canu v2.0 (all contigs)	11839	9813	2.56 Mb	15.3 Mb	506
		canu v1.0 (all contigs)	19087	16889	0.63 Mb	12.6 Mb	
Dovetail	'Chicago' proximity ligation	Falcon v1.1	3462	3079	1.43 Mb†	1.40 Mb†	595
		Canu v2.1	1504	1288	2.48 Mb†	2.34 Mb†	410
		Canu v1.1	1585	1211	2.36 Mb†	2.26 Mb†	264

* NG50 as calculated by QUAST assuming a genome size of 1.30 Gb
† N50 from Dovetail genomics report

Leslie Vosshall
Rockefeller

Testing *Aedes aegypti* Sequences by Bionano Hybrid Scaffold



Several sequence assemblies were produced and scaffolded to determine the highest quality

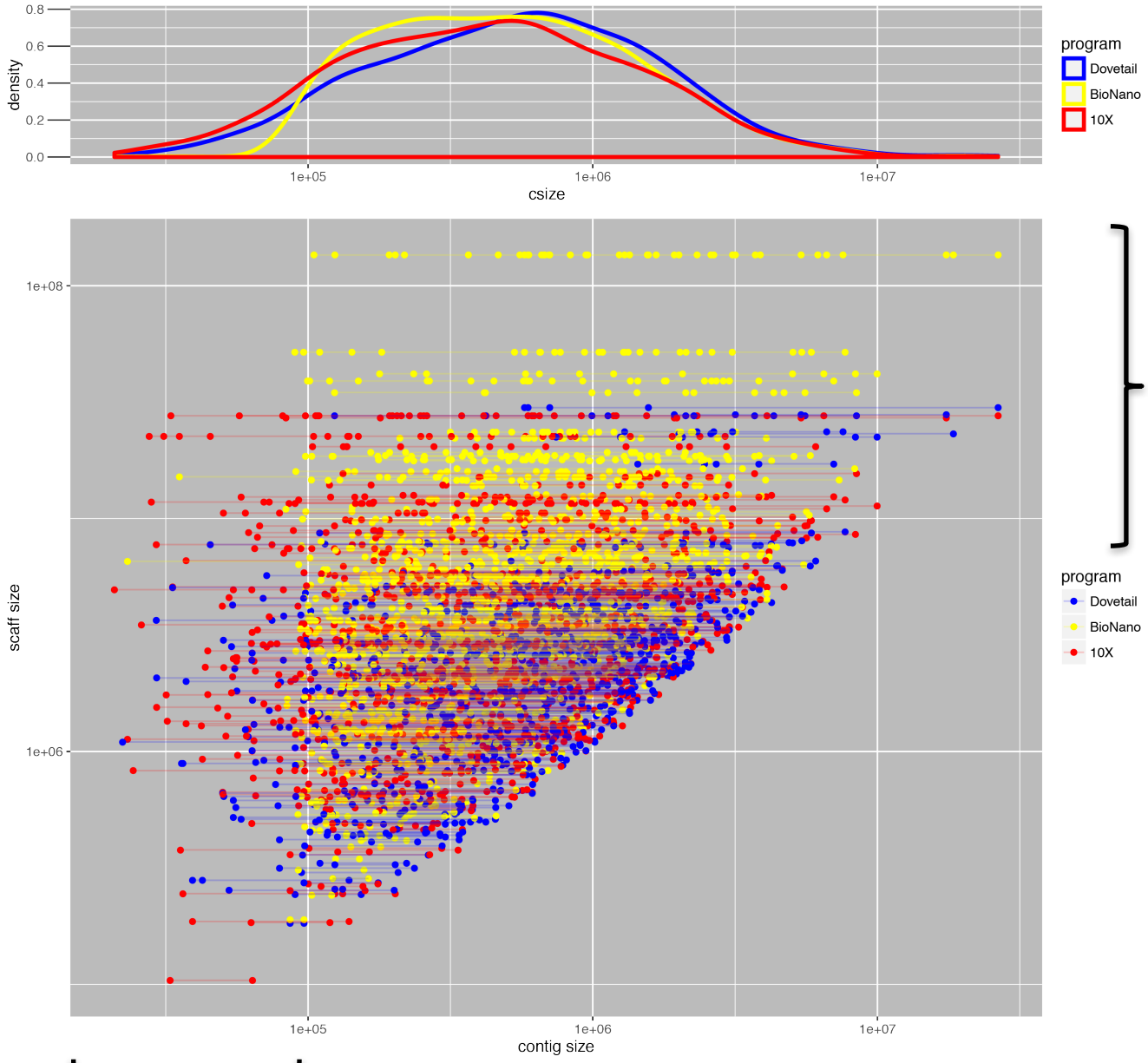
After selecting Canu 2.0, the two enzyme hybrid scaffold was run

	Falcon (p only) BspQI	Falcon (p+a) BspQI	Canu 1.0 BspQI	Canu 2.0 BspQI
Input Seq length (Gbp)	1.70	2.05	2.75	**2.25**
Amount Seq used (Gbp)	1.39 (82 %)	1.55 (76%)	1.71 (62%)	**1.57 (70%)**
Hybrid N50 (Mbp)	4.63	4.67	4.42	**5.69**
NGS cuts	499	495	470	**366**
BNG cuts	3	3	6	**13**

Hybrid Info	Bionano Nb.BssSI	Bionano Nt.BspQI	Original NGS	Hybrid	2nd Hybrid
N contigs	3044	1606	11839	1324	565
Contig N50 (Mbp)	0.86	1.47	0.85	2.79	7.26
Total contig length (Mbp)	2095.63	1779.04	2253.74	2100.87	1925.92

bionano GENOMICS

THE ROCKEFELLER UNIVERSITY
Science for the benefit of humanity



BioNano scaffs orders of magnitude larger

Large scaffolds not just composed of few large contigs

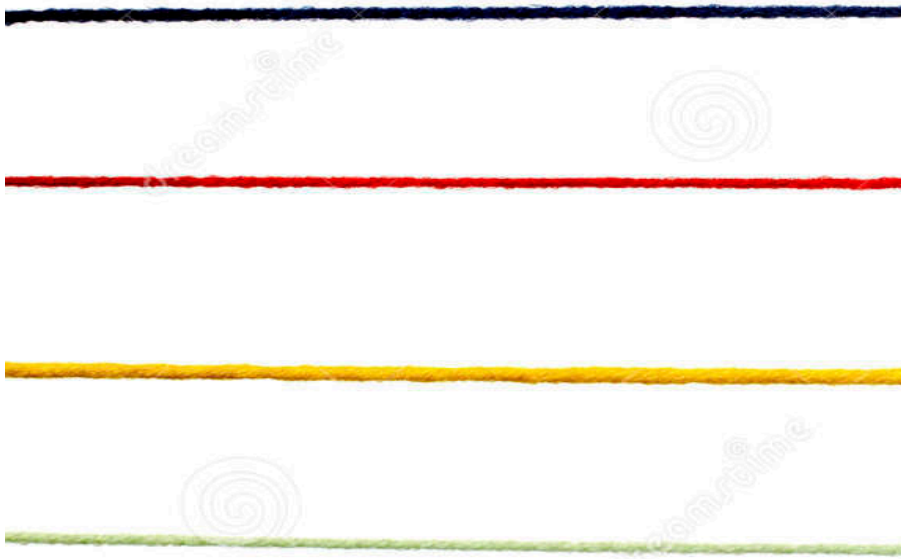
Only Dovetail / 10X scaffolding contigs < 100kb
lower limit of resolution for BioNano?

Dan Neafsey
BROAD INSTITUTE

What about Hi-C?

Hi-C

Bionano



All recent human reference-quality genome publications use Bionano data

	AK1	HX1	NA12878	NA12878	NA24385	GRCh38
Sequencing	PacBio	PacBio	Illumina + 10x Genomics	PacBio	PacBio	Sanger
Scaffolding	Bionano	Bionano	Bionano	Bionano	Bionano two-enzyme	multiple
Input N50 (Mbp)	17.92	8.325	7.03	1.56	4.7	56.41
Hybrid Scaffold N50 (Mbp) scaffold	44.85	21.979	33.5	26.83	80.46	67.79
Fold Improvement after Bionano hybrid scaffold	2.5x	2.6x	4.8x	17.2x	17.1x	

White Paper Series

Hybrid Scaffolding Improves Genome Assembly Accuracy and Contiguity

Next-Generation Mapping Reveals True Long Range Structure of the Genome and Reduces Sequencing Costs

Generating high-quality finished genomes remains challenging. Accurate identification of structural variation with minimal gaps is difficult or impossible using short-read sequencing technologies alone.

The genomes of most higher organisms are highly repetitive. Two thirds of human and most mammalian genomes consist of repeats, and many plant genomes have even higher repeat content. Short reads usually fail to span long repeat arrays or disambiguate different copies of interspersed repeats that are not spanned. These failures can limit contig length and introduce chimeric joins and other assembly errors.

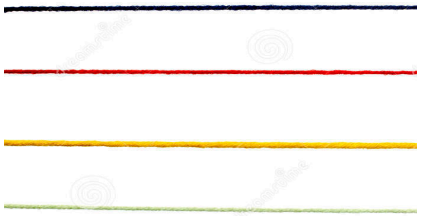
The widespread use of next-generation sequencing

Only extremely long molecules, ranging in size from hundreds of thousands to millions of base pairs, provide accurate structure of the genome. Bionano Genomics' Next-Generation Mapping (NGM) visualizes long DNA molecules in their native state. Long range genomic structural information is preserved and directly observed instead of algorithmically inferred as in sequencing approaches. These long labeled molecules are *de novo* assembled into physical maps spanning the whole genome. The resulting order and orientation of sequence elements in the map can be used for anchoring NGS contigs and detecting structural variation.

Conclusions

- **Bionano genome maps show the true structure of the genome**
- **Large structural variants and rearrangements are detected with much higher sensitivity than using sequence based methods**
- **Combining NGS and Bionano NGM data produces assemblies of the highest quality**
- **Bionano hybrid scaffolding is agnostic to the sequence technology used**
- **Bionano is the ONLY non-sequencing based scaffolding technology capable of correcting sequencing-type errors**
- **Including Bionano mapping data into *de novo* genome assemblies has become a *de facto* standard**

Thank you



Contact:

Sven Bocklandt
sbocklandt@bionanogenomics.com

bionano GENOMICS