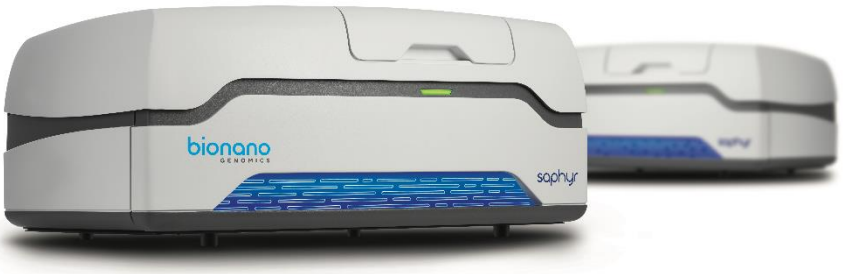
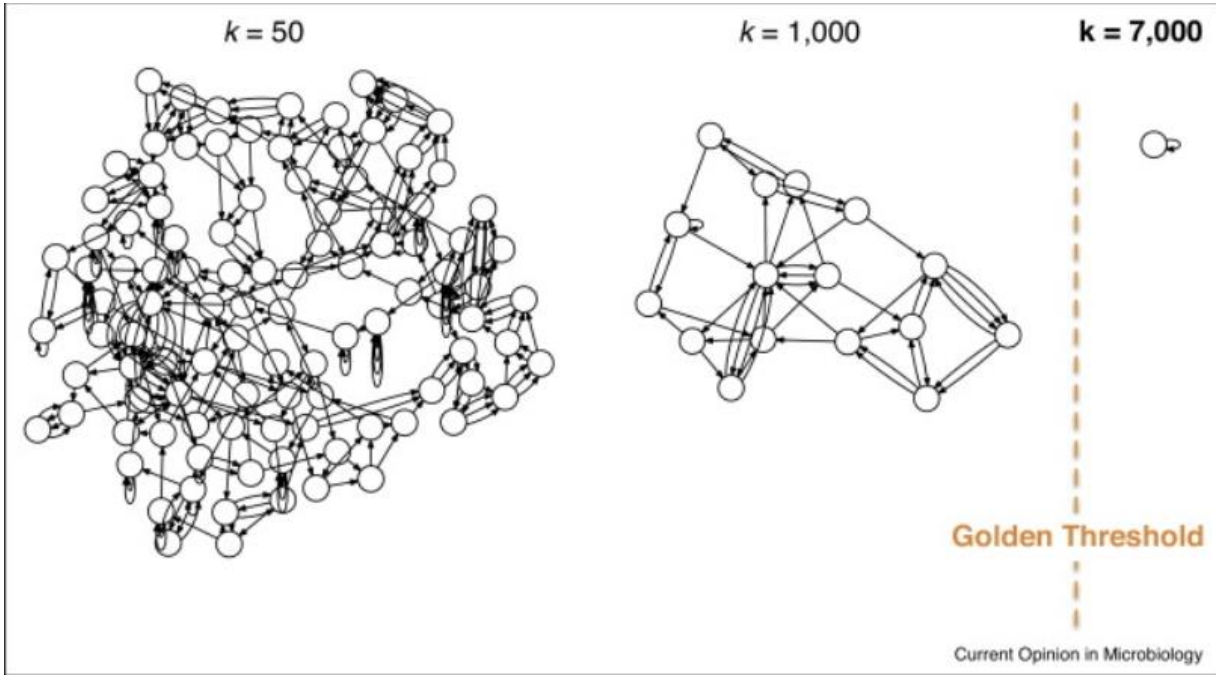
Bionano optical mapping for accurate genome assembly, comparative genomics, and haplotype segregation



Alex Hastie, PhD Sr Director, Scientific Strategy and Customer Solutions | Bionano Genomics
ahastie@bionanogenomics.com

Genome assembly



Requirements for accurate comprehensive genome assembly

- Accuracy sufficient for differentiation of unique and repeated sequences including homologous chromosomes
- Read lengths sufficient to span repeats
- Ideal: highly accurate sequence reads spanning from end to end of chromosomes (or at least spanning the longest repeat element in the genome, ~3-5Mbp)

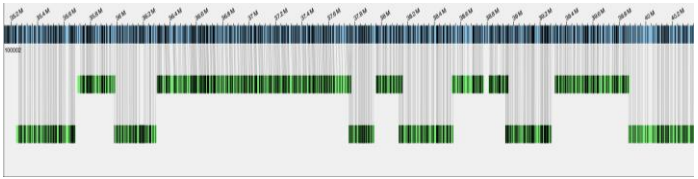
- Current core technologies:
  - **ILMN – contigs in the kbp range**
  - **PacBio/ONT/10X – contigs/scaffolds in the Mbp-10Mbp range**
  - **Bionano – maps in the 20-200 Mbp range**
  - **Hi-C scaffolds can span centromeres**

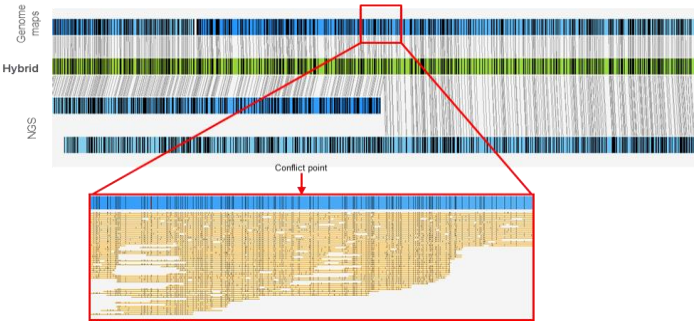
Genomics Technologies are Compatible

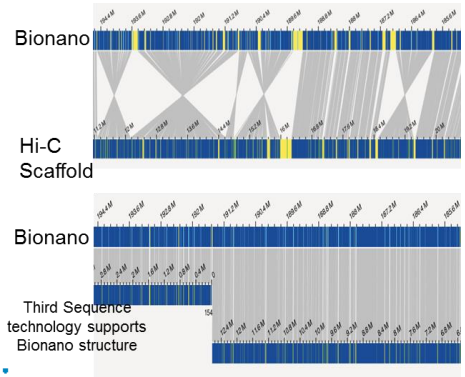
Bionano scaffolds and corrects sequence contigs and scaffolds

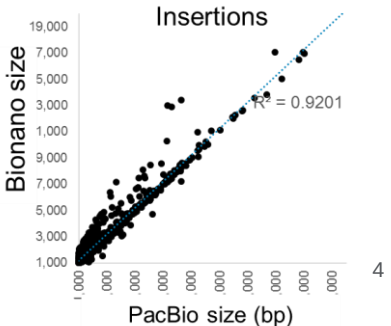
Hi-C spans centromeres

Bionano validates and improves assemblies









bionano GENOMICS

Bionano Access software assembles *de novo* genome maps

6

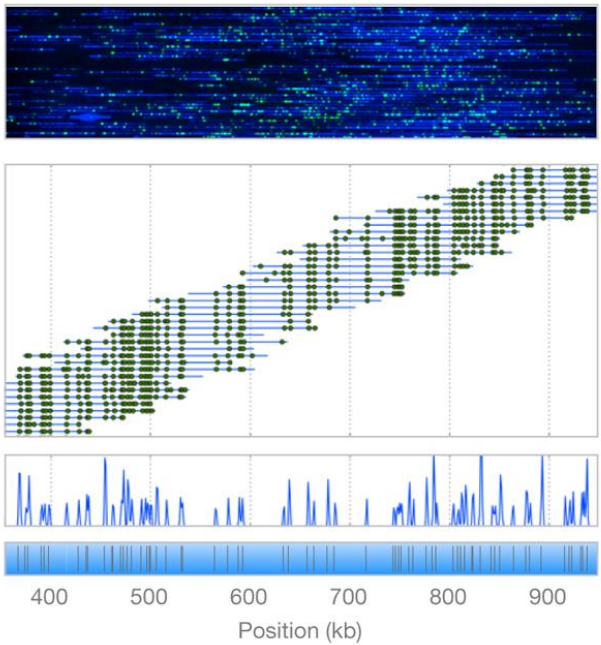
Raw Image Data

Conversion to Molecules

Population of Molecules

Aggregation & Assembly

Consensus Genome Map



Correct assembly errors
Reduce fragment number

up to

1000x

improvement in contiguity

OPE
doi:10.1038/nature229

Improved maize reference genome with single-molecule technologies

Yinping Jiao[1], Paul Peluso[2], Jinghua Shi[3], Tiffany Liang[3], Michelle C. Stitzer[4], Bo Wang[1], Michael S. Campbell[1], Joshua C. Stein[1], Xuehong Wei[1], Chen-Shan Chin[2], Katherine Guill[5], Michael Regulski[1], Sunita Kumari[1], Andrew Olson[1], Jonathan Gent[6], Kevin L. Schneider[7], Thomas K. Wolfgruber[7], Michael R. May[8], Nathan M. Springer[9], Eric Antoniou[1], W. Richard McCombie[1], Gernot G. Presting[7], Michael McMullen[5], Jeffrey Ross-Ibarra[10], R. Kelly Dawe[6], Alex Hastie[3], David R. Rank[2] & Doreen Ware[1,11]

Complete and accurate reference genomes and annotations provide fundamental tools for characterization of genetic and functional variation[1]. These resources facilitate the determination of biological processes and support translation of research findings into improved and sustainable agricultural technologies. Many reference genomes for crop plants have been generated over the past decade, but these genomes are often fragmented and missing complex repeat regions[2]. Here we report the assembly and annotation of a reference genome of maize, a genetic and agricultural model species, using single-molecule real-time sequencing and high-resolution optical mapping. Relative to the previous reference genome[3], our assembly features a 52-fold increase in contig length and notable improvements in the assembly of intergenic spaces and centromeres. Characterization of the repetitive portion of the genome revealed

research, which will enable increases in yield to feed the growing wor
population. The current assembly of the maize genome, based
Sanger sequencing, was first published in 2009 (ref. 3). Although th
initial reference enabled rapid progress in maize genomics[1], the orig
nal assembly is composed of more than 100,000 small contigs, ma
of which are arbitrarily ordered and oriented, markedly complicati
detailed analysis of individual loci[6] and impeding investigation of inte
genic regions crucial to our understanding of phenotypic variation
and genome evolution[9,10].

Here we report a vastly improved *de novo* assembly and annotati
of the maize reference genome (Fig. 1). On the basis of 65× singl
molecule real-time sequencing (SMRT) (Extended Data Fig. 1),
assembled the genome of the maize inbred line B73 into 2,958 contig
in which half of the total assembly is made up of contigs larger th

Unique Species Mapped Using Bionano

Actinomyces	black flying fox	Cuccumber	Honey bee	Monarch Butterfly	Raspberry	Tobacco
aerobic	Brewer's Yeast	Deer	Hooded Crow	Monk Seal	Red Algea	Tomato
Amaranthus	Burying Beetle	Diatom	Horse	Mosquitoe	Rice	Tomato (Arcanum)
anaerobic bacteria	Cabbage	Drosophila	Horseshoe bat	Mouse (blood -1 FC, Blk6)	Rice (Japanica)	Trypanosoma
Anna's hummingbird	Cabernet Sauvignon	Duckweed	Human Blood	Mouse Lemur	S.cryophilus†	Valley Oak
Arabidopsis	Cercospora beticola	Durum wheat (3x)	Human Cell-line	Mus Musculus	S.pombe	Water sample (Microbiome)
BAC clones	Carion Crow	Dust mite	Human Leukocytes	Neisseria Meningitis	Sea Goose Berry	Wasp (Cotesia plutellae)
Bacteria	Chicken	E.coli	Jade (Succulent)	Organpipe cactus (S. thurberi)	Seasquirt	Wheat
Banana	Chickpea	Eggplant	Jellyfish	Ornithorhynchus Anatinus (Platypus)	Soy Bean	Wheat Nuclei
Barley	Chimpanzee	Enterobacter	Kalanchoe (succulent)	Papaya	Spider Mite	Wheat Rust (Fungus)
Bat (Leaf-nosed)	Chrysanthemum	eucalyptus	Kashmir flour beetle	Paprika	Staphylococcus Aureus	Wild Rice
Bat (Vampire)	Clover	Fire ant	Kingfish	Pea	Stickleback Fish	Wild tobacco
Bat (Cave myotis)	Coffee	Fish	Komodo dragon	Peanut	Stawberry (Diploid)	Woodchuck
Basil	Corn	Fusobacter	Leishmania	Pedicoccous	Strawberry (Pentaploid)	Xanthomonas
Blackcap	Cotton D genome	Gardnerella vaginalis	Lettuce	Pichia	Streptomyces	Yarrowia
Brasilian Grass	Cotton strain	Geotrichum	Lizard	Pit Viper	Sugar Beet	
Broccoli	Cotton (G. sturtianum)	Goat grass	Maize	Plasmodium	Sugar beet Fungus	
Bed Bug†	Cotton (G. herbaceum cv. Wagad)	Gonium	Manduca (horn worm) 500 Mb	Platypus	(Pythium)	
Beef Tapeworm (Taenia saginata)	Cotton (G. hirsutum cv. Maxxa)	Grapevine (Nebiolo)	Marine Viruses	Planaria	Sugar Cane	
black flour beetle	Crow	Haemonchus contortus	marmoset	Rabbit	T. cruzi	

De novo Assembly of Diverse Genomes

- A representative subset of genomes mapped on the Saphyr system

Sample	Molecule N50 (kbp)	Bionano Map N50 (Mbp)
NA12878	293	55.9
Human Fresh Blood	307	56.9
Ferret	262	66.1
Mouse	280	101
Pig	335	65.2
Blackbird	243	21.6
Hummingbird	310	38.7
Kakapo	247	69.3
Fish	245	22.3
Brassica	270	12.4
Durum Wheat	364	13
Farro	300	32.7
Maize B73	260	100
Soybean	246	23
Strawberry	241	13.3
Sunflower	317	178.3

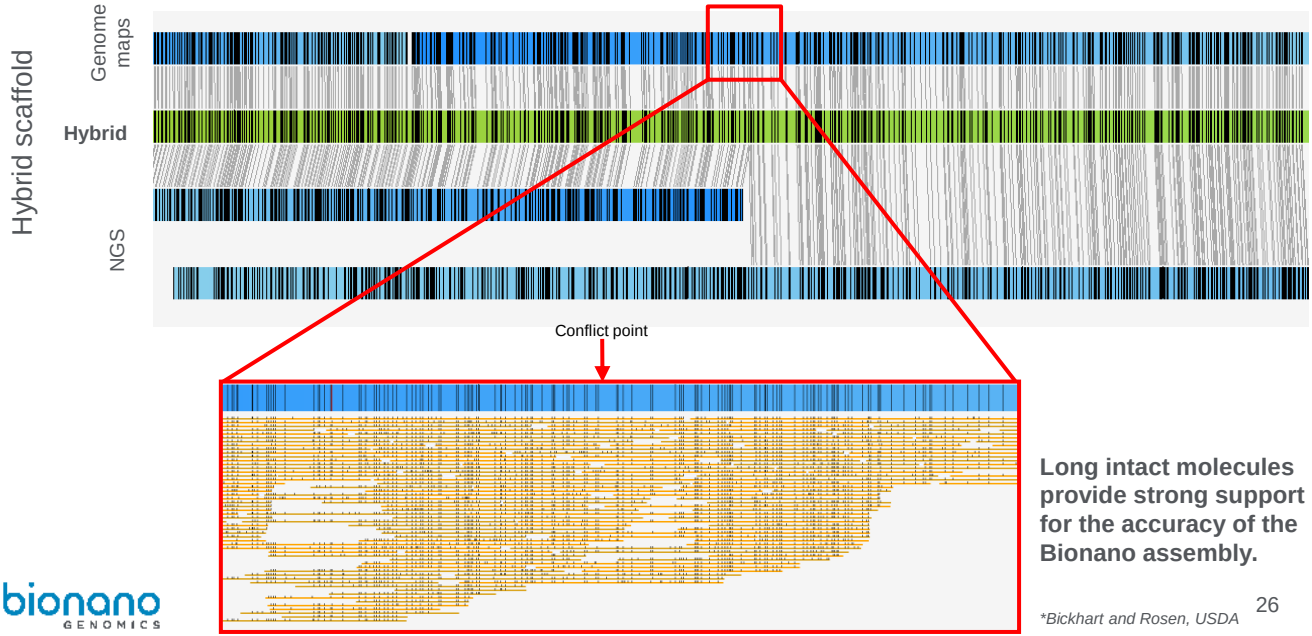
Highest Quality Long Scaffolds

Species	NGS N50 (Mbp)	Scaffold N50 (Mbp)	% NGS Anchored
Maize B73	1.19	100	99.5%
Sorghum	3.05	34	95.9%
Kakapo	4.34	71	95.9%
Blackbird	1.47	42	95.0%

- Enabling projects like the maize pangenome – Corteva; Kelly Dawe
  - High quality
  - High contiguity
  - Low cost
  - Fast turn-around

Highest Quality Hybrid Scaffolding

- Sequence assembly error flagging and correction
  - Chimeric NGS contigs
  - NGS contig overlaps
- Accurate N-gap sizing

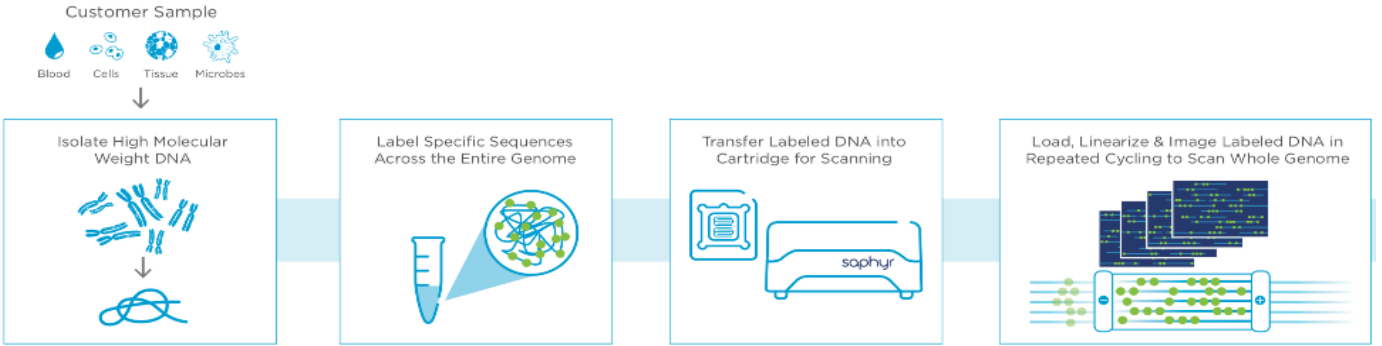


Long intact molecules provide strong support for the accuracy of the Bionano assembly.

**Bickhart and Rosen, USDA*

Recent Bionano Advancements

Tissue preservation and small input amounts (10mg)



Automatable in-solution UHMW DNA isolation for up to 12 samples/day

- blood and cells
- Animal tissue
- Plants not yet supported

Significant increase in data throughput and yields

- Scale to handle the wheat pangenome

Sample Prep Updates

SP DNA Isolation from Frozen Nucleated Blood in Ethanol

Sample	Sample ID	Blood Volume (in 50% ETOH)	DNA [ ] (ng/ul)	Yield (ug)	N50 >20kb	N50 >150kb	Labels/ 100kb	Map Rate	PLV	NLV
Tern	1	50 µl	110	12.1	176	250	16.0	76.5%	7.6%	9.5%
Crow	2	50 µl	202	22.2	225	309	15.1	77.8%	3.2%	9.3%

- ✓**Technote coming**

SP DNA Isolation from Frozen Animal Blood (Non-nucleated)

- **RBCs with smaller mcvs are not as easy to lyse by freeze/thaw**
- **Employ 1x RBC lysis with RBC lysis reagent (Qiagen)**
- ✓**Technote coming**

SP DNA Isolation from Fresh Frozen Animal Tissue

- **5-20 mg of fresh frozen rat tissue**
  - ✓ Kidney, Lung, Liver, Prostate, Colon, Bladder, Thyroid
  - – Skeletal muscle, Breast
  - ? Ovaries, Uterus, Testes

SP DNA Isolation from RT Preserved Animal Tissue

- **All Protect (Qiagen)**

2 genomes/ day

6 genomes/day

More to come

bionano GENOMICS

*Mammalian sized genomes

Increased throughput

Saphyr gen 1 -> 1300 Gbp x 2 flow cells per ~48-72 hours

Saphyr gen 2 -> 320 Gbp x 6 flow cells per 24 hours

or -> 1300 Gbp x 3 flow cells per 48 hours

Applications


Mexico City favors a scientist for mayor
New books to beat the summer heat
Carbon-based molecules on Mars
Science
$15
8 JUNE 2018
AAAS
GREAT APE
GENOMES
Deciphering humanity from our closest relatives

Primate genomes assembly and variation

- PacBio sequencing and contig assembly
- Bionano scaffolding and error correction (dual NLRS)
- Hi-C, BAC, FISH scaffolding and validation

TABLES

Table 1. Assembly statistics for the great ape genomes.

Ape assembly	CHM13_HSAv1[d] (human)	YRI_HSAv1 (human)	Clint_PTRv1 (chimpanzee)	GSMRT3.2 (gorilla)	Susie_PABv1 (orangutan)
Estimated depth[a]	72	116	124	86.3	94.9
Subread length N50 (kbp)[b]	16.2	13.4	17.4	18.6	16.6
Contig (number)	1,923	3,645	4,912	15,997	5,771
Assembly size (Gbp)	2.88	2.88	2.99	3.08	3.04
Contig length >3 Mbp (Gbp)	2.65	2.27	2.48	2.42	2.51
Contig N50 (Mbp)	29.26	6.60	12.76	10.02	11.27
Scaffold N50 (Mbp)	83.02	ND	59.55	ND	101.33
Longest contig (Mbp)[c]	81	27	80	36	53
BAC concordance	97.11%	97.73%	99.13%	96.85%	96.75%
Bionano breaks [# contigs]	122 [49]	ND	152 [68]	ND	49 [32]
Sequence accuracy (QV)	36	31	33-38	30-38	28-33
Iso-Seq transcripts	710,974	ND	565,691	881,801	528,145
Contigs in AGP	ND	ND	651	794	598
Contigs aligned to GRCh38[e] [Gbp]	407 [2.8]	1,167 [2.8]	656 [2.8]	907 [2.8]	524 [2.8]

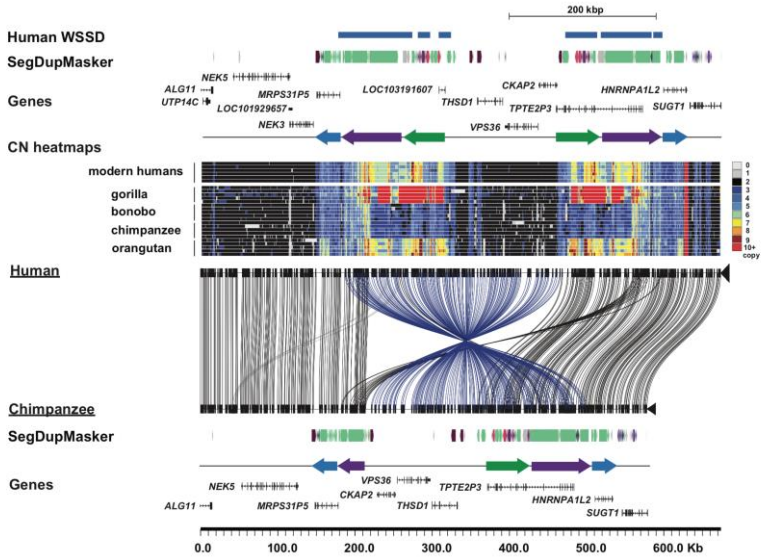
Detection of large inversions required Bionano

- Bionano was required to detect 29 large inversions including pericentromeric species specific inversions.
- “chromosomes constructed using Bionano and FISH captured all nine pericentric inversions correctly”

- A ~265 kbp inversion on chromosome 13q14.3 detected b optical mapping in chimpanzee (annotated blue lines). The inverted region is flanked by larg ~180 kbp inverted SD blocks tha vary with respect to copy numbe among great apes

Table A11: Pericentric inversion breakpoints captured by Bionano scaffolds.

chr	start	end	size	Scaffold spanning breakpoint (L)	(R)
chr2 fusion	113,000,000	113,000,000		na	na
chr4	44,813,133	84,898,851	40,085,718	yes	yes
chr5	23,020,320	93,926,801	70,906,481	yes	no
chr9	68,643,184	86,184,102	17,540,918	no	no
chr12	20,782,790	67,910,583	47,127,793	no	yes
chr15	25,108,810	29,751,155	4,642,345	yes	no
chr17	15,523,701	49,485,678	33,961,977	no	no
chr18	2,146,810	12,914,783	10,767,973	yes	yes



bionano GENOMICS

Bionano DLS and ONT assembly

Article | Open Access | Published: 19 November 2018

A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping

Stéphane Deschamps ✉, Yun Zhang, Victor Llaca, Liang Ye, Abhijit Sanyal, Matthew King, Gregory May & Haining Lin ✉

Nature Communications **9**, Article number: 4844 (2018) | Download Citation ⬇

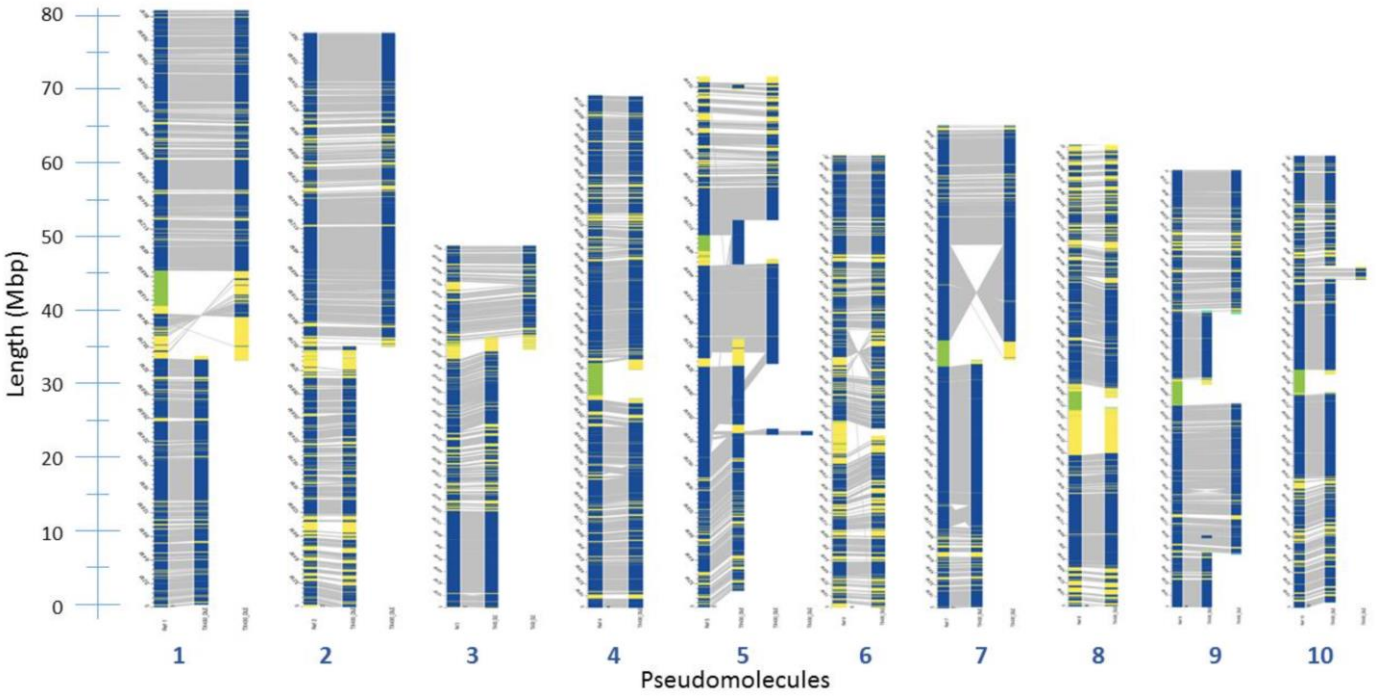
Bionano DLS and ONT assembly

A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping

	Original DLS Genome Map	Original ONT contigs	ONT Contigs in Hybrid Scaffold	ONT Contigs Not in Hybrid Scaffold	Hybrid Scaffolds
Number of Contigs	79	723	500	363	30
Total Length (Mbps)	719.339	671.867	644.44	25.117	661.06
Minimum Contig Length (Mbp)	0.225	0.009	0.058	0.00006	0.086
Maximum Contig Length (Mbp)	47.659	16.337	16.337	1.549	52.621
N_{50} Contig Length (Mbp)	33.773	3.053	2.991	0.13	33.35

140 contig errors were corrected

Structural changes are immediately recognized



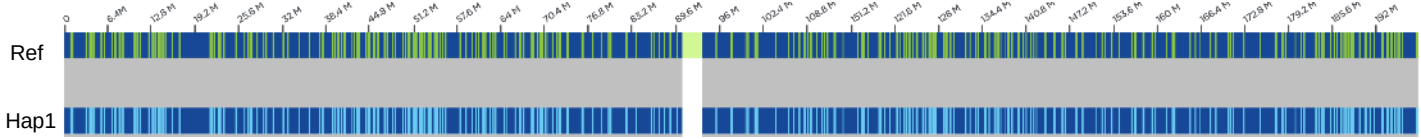
Resolving heterozygosity

De novo haplotype map resolution is possible – how about scaffolding

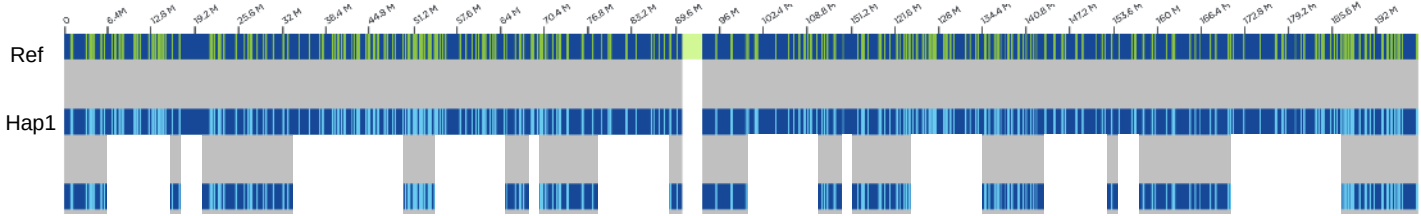
- Sequence assembly should have both alleles represented
- Map and sequence assembly should be put into phase, can be partially handled by Bionano hybrid scaffolder
- Bionano hybrid scaffold
  - **HS can perform diploid hybrid scaffold only with high heterozygosity, so far**
  - **Diploid aware scaffolder may be needed to optimize scaffolding**
- A map guided sequence assembler would be ideal

Resolving heterozygosity with only one individual

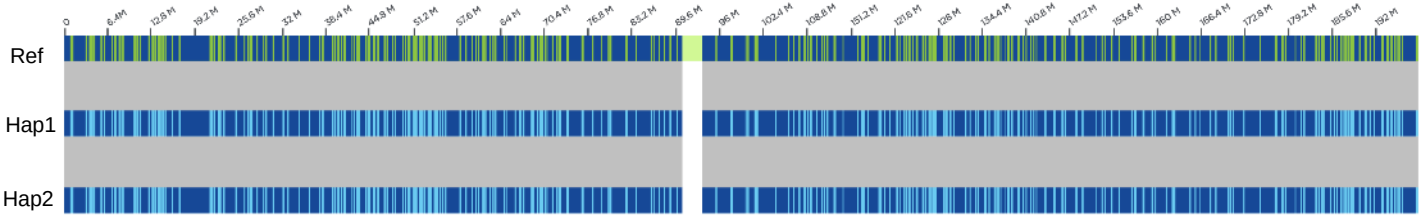
Basic assembly (recommended for scaffolding) -> 1N



Extend and split – resolving large (>30kbp) het. sites -> 1-2N



Extend and split + haplotype refinement - resolves >500bp het. sites (for human SV studies) -> 2N



bionano GENOMICS

Phasing is based on SVs at molecule ranges

Trio binning of cattle genomes

- Trio binning is an approach championed by Adam Phillippy and others
  - **Long read sequencing of a heterozygous individual**
  - **Short read sequencing of parents**
  - **Select maternal and paternal reads by alignment of parental ILMN sequences**

- Bionano aimed to produce trio binned genome maps in order to improve phasing, more accurately scaffold contigs, and to identify maternal and paternal variants between individuals

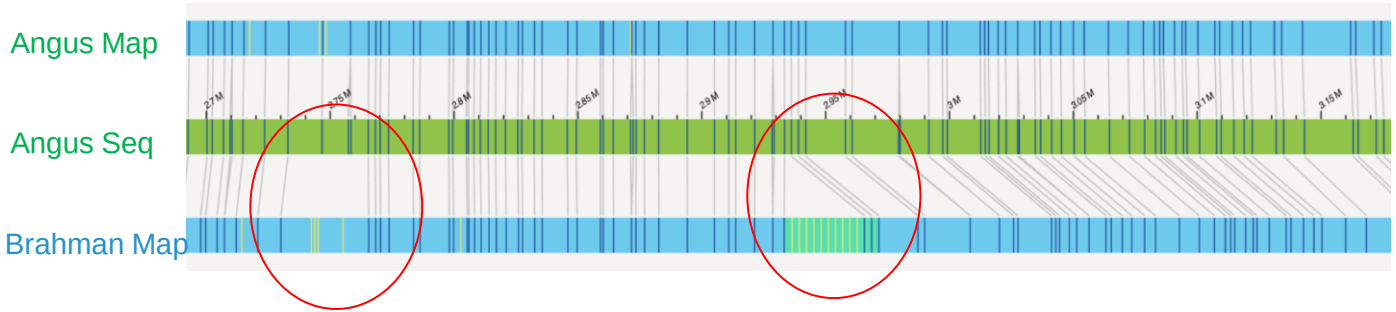
- University of Adelaide, NHGRI, USDA, PacBio, Bionano Genomics, Phase Genomics, EMBL, University of Maryland
  - **Wai Yee Low (Lloyd)**
  - **John Williams**

Trio binning

- Molecules were aligned to the Angus and Brahman sequences.
- Based on the alignment score of each molecule to each sequence, they were assigned to the pool (Angus/Brahman) that they have better alignment score with.
- To maintain equivalent coverage throughout each genome, molecules that aligned to both sequences with similar scores (difference <2) were split by half and assigned to each pool (Angus/Brahman).
- Molecules that did not align to either sequence were also split by half and assigned to each pool (Angus/Brahman).

Assembly of Angus and Brahman cattle

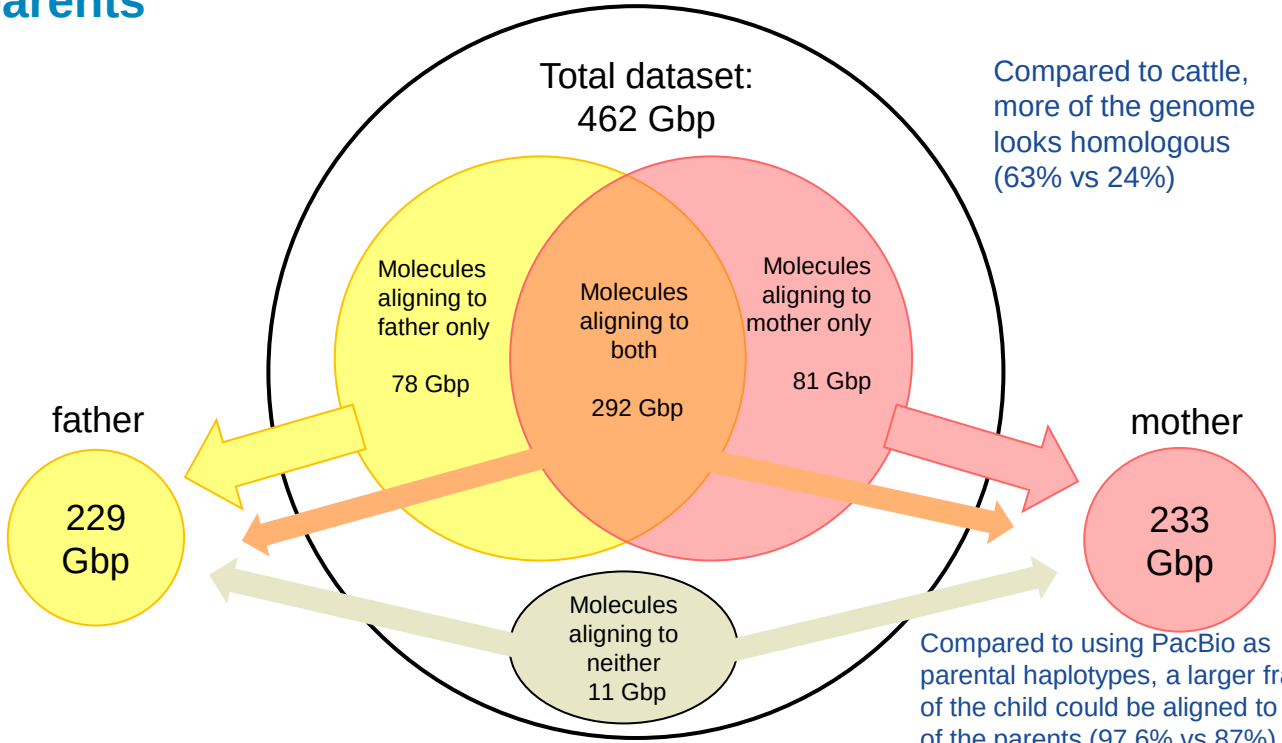
	Expect	Trio binning Angus	Trio binning Brahman	ES	ES + Hap refinement
Total length	~3 Gbp	2.79 Gbp	2.87 Gbp	3.37 Gbp	5.86 Gbp
Map N50		33.97 Mbp*	28.62 Mbp*	71 Mbp	75 Mbp
Ins-Del vs Angus		388	6183	3325	6577



Caution: Haplotype refinement is only designed for human SV studies, we have seen over splitting with some non-human genomes. Optimization of parameters may be needed.

bionano GENOMICS

Human trio binning for reference genomes – Bionano for parents



Compared to cattle, more of the genome looks homologous (63% vs 24%)

Compared to using PacBio as parental haplotypes, a larger fraction of the child could be aligned to one of the parents (97.6% vs 87%)

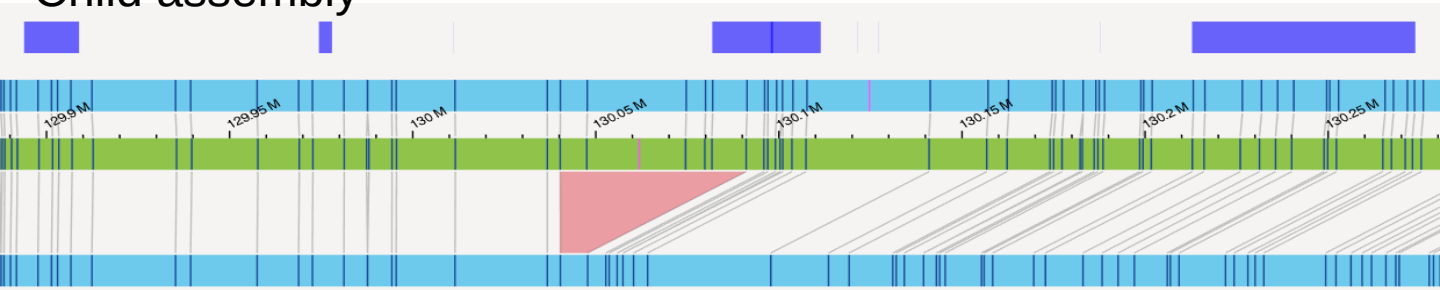
- Molecules were aligned to the parental genome maps
- Based on the alignment score of each molecule to each genome maps, they were assigned to the pool (father/mother) that they have better alignment score with.
- To maintain equivalent coverage throughout each genome, molecules that aligned to both genome maps with similar scores (difference <2) were split by half and assigned to each pool (father/mother).
- Molecules that did not align to either sequence were also split by half and assigned to each pool (father/mother).

bionano GENOMICS

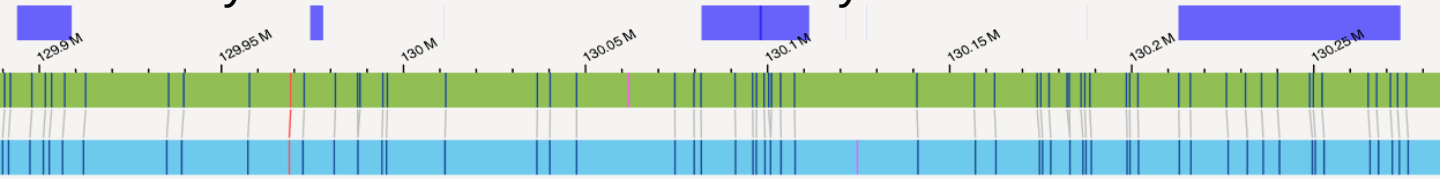
Erich Jarvis, Olivier Fedrigo, Arang Rhie, Adam Philippy, Joyce Lee

Parental Allele Assemblies after trio binning

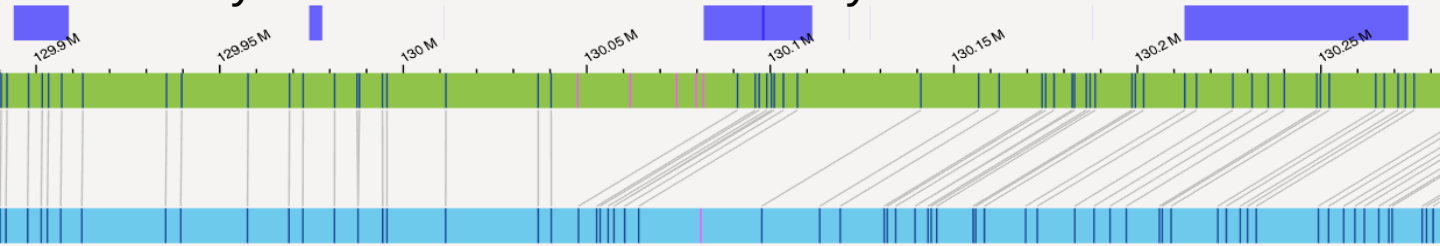
Child assembly



Paternally selected molecule assembly



Maternally selected molecule assembly



Hybrid scaffoldings of binned assemblies

Father allele selected:

Statistic	BNG	Sequence	Sequence used in hybrid scaffold	Hybrid scaffold
Number of maps	318	2538	1004	146
N50 (Mb)	51.13	12.53	12.69	79.78
Total length (Mb)	2951.80	2736.57	2686.29	2760.57

113 cuts were performed on 55 sequence contigs. 10 cuts were performed on 10 Bionano maps.

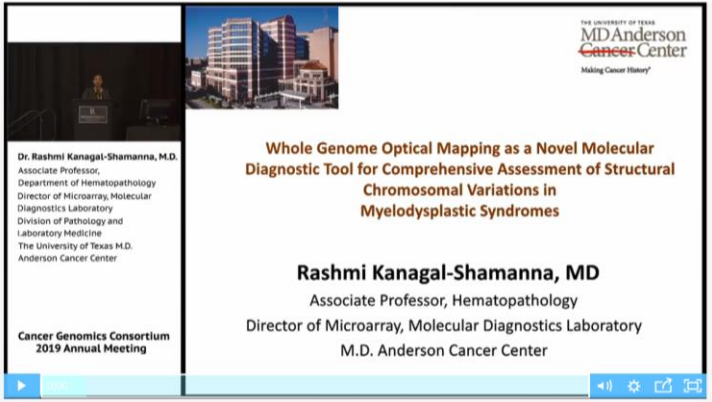
Mother allele selected:

Statistic	BNG	Sequence	Sequence used in hybrid scaffold	Hybrid scaffold
Number of maps	279	1953	792	112
N50 (Mb)	43.39	12.12	12.06	69.80
Total length (Mb)	2946.92	2864.24	2783.77 (97.2%)	2824.67

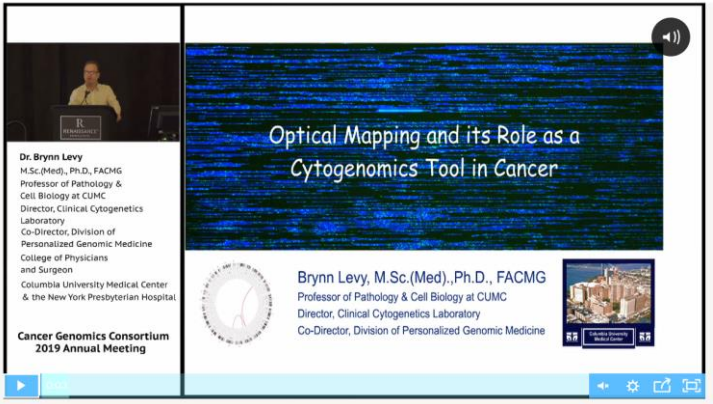
109 cuts were performed on 59 sequence contigs. 7 cuts were performed on 7 Bionano maps.

Structural variation

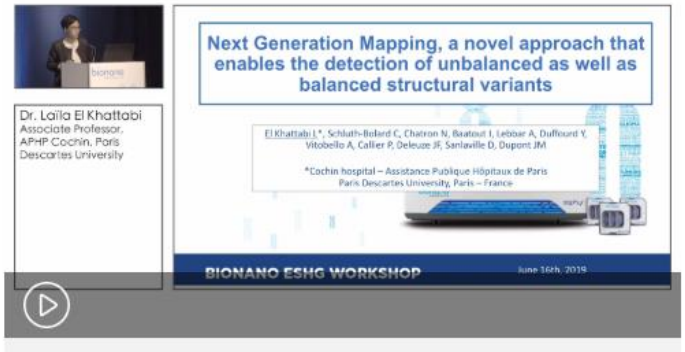
Cancer and genetic disease

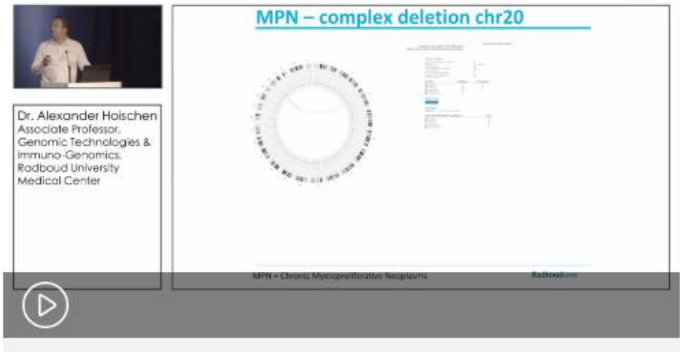


Dr Rashmi_converted_small

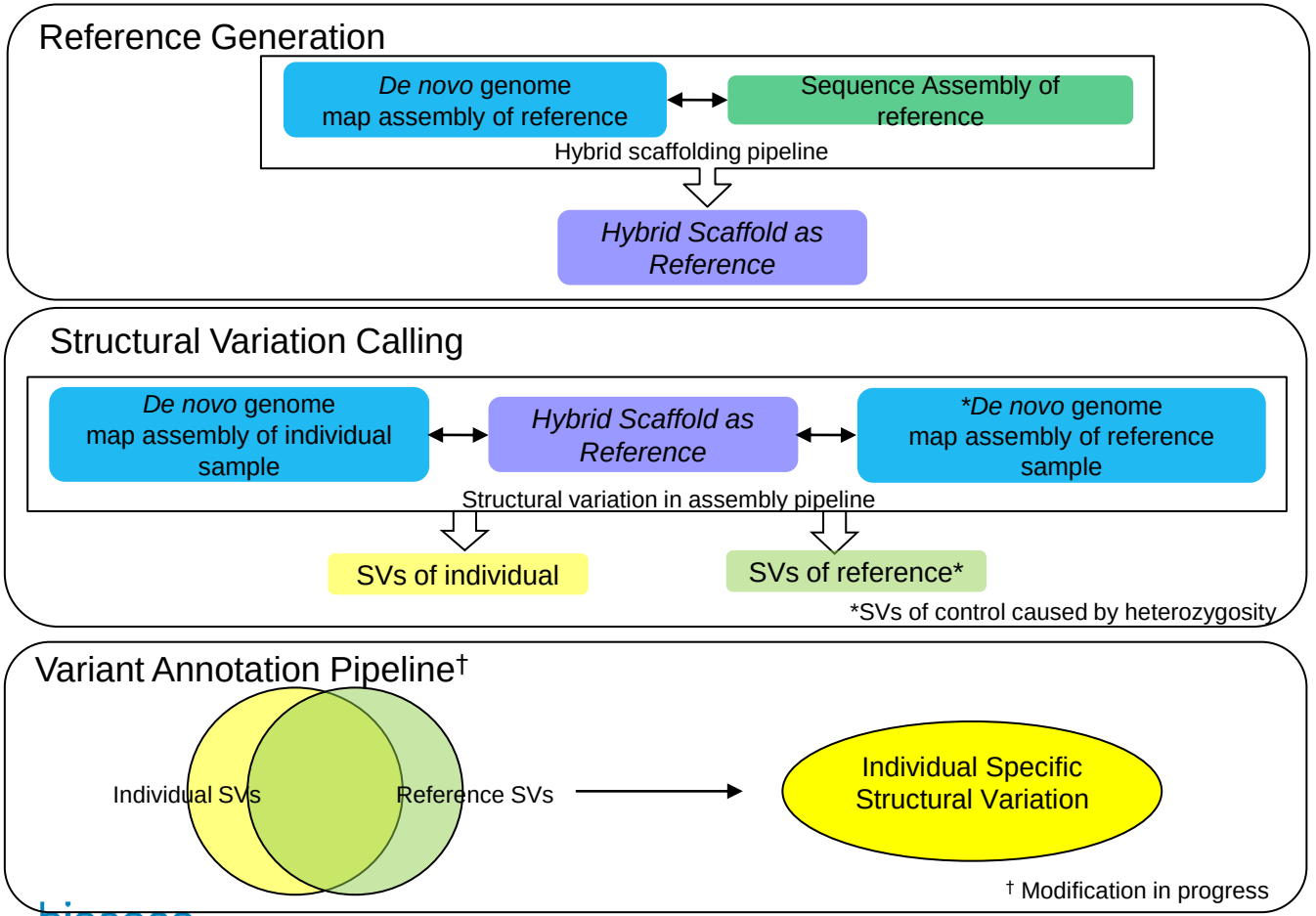


Dr. Brynn Levy

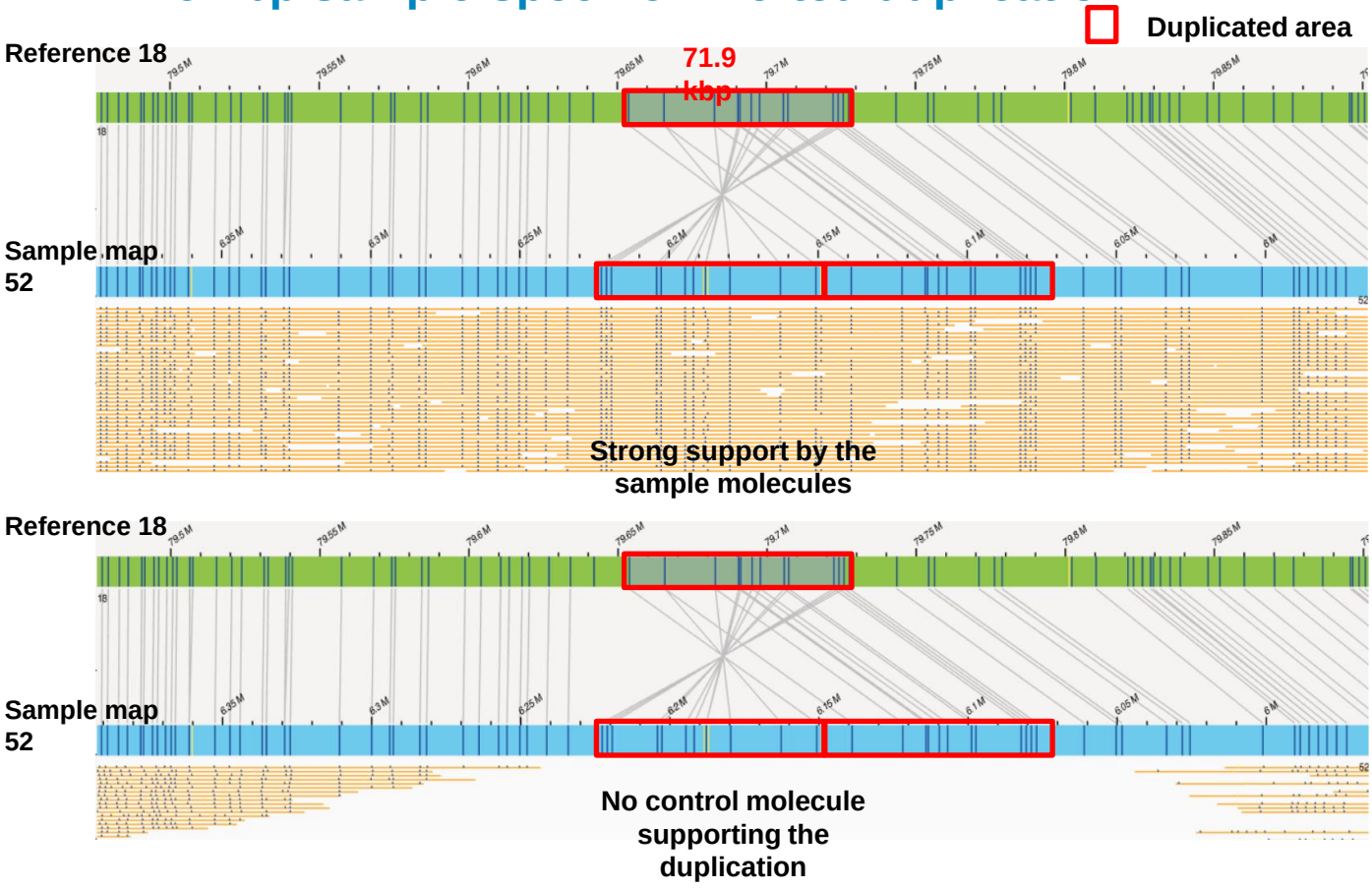




Analysis Workflow: Reference Generation and Sample Specific Structural Variation



A 71.9 kbp sample-specific inverted duplication



Large genomes

Axolotl – 32 Gbp genome

nature

Table 1 Comparison of assembly contiguity statistics in axolotl, spruce and pine genomes

From: The axolotl genome and the evolution of key tissue formation regulators

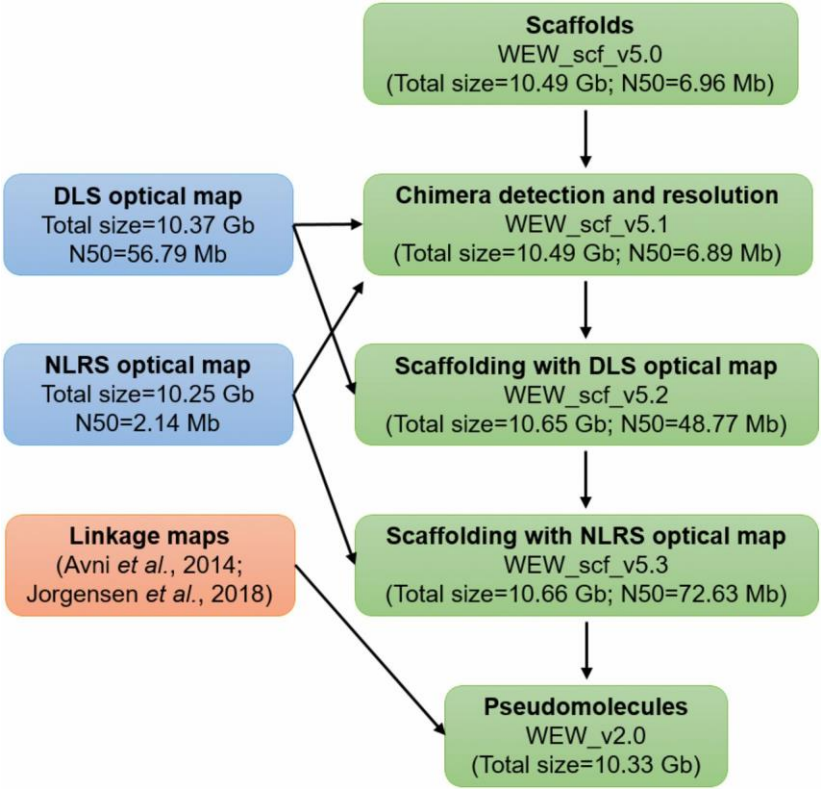
	Axolotl (*A. mexicanum*)	White spruce (*Picea glauca*)	Loblolly pine (*Pinus taeda*)
Assembly size (Gb)	32.4 (28.4 in contigs)	24.6	20.6
Genome size (Gb)	32	20	22
Chromosomes	14	12	12
Sequencing technology	PacBio; Optical map	Illumina; cDNAs	Illumina; PacBio; Fosmid DiTag
Coverage	32×	65×	68× Illumina; 12× PacBio
Assembler	MARVEL	ABySS	MaSuRCA
Contig N50 (bp)	216,277	6,644	25,361
Number of contigs	217,461	5,252,090	2,445,689
Scaffold N50 (bp)	3,052,786	54,661	107,036
Number of scaffolds	125,724	3,033,322	1,496,869

bionano GENOMICS

Done using older NLRS technology

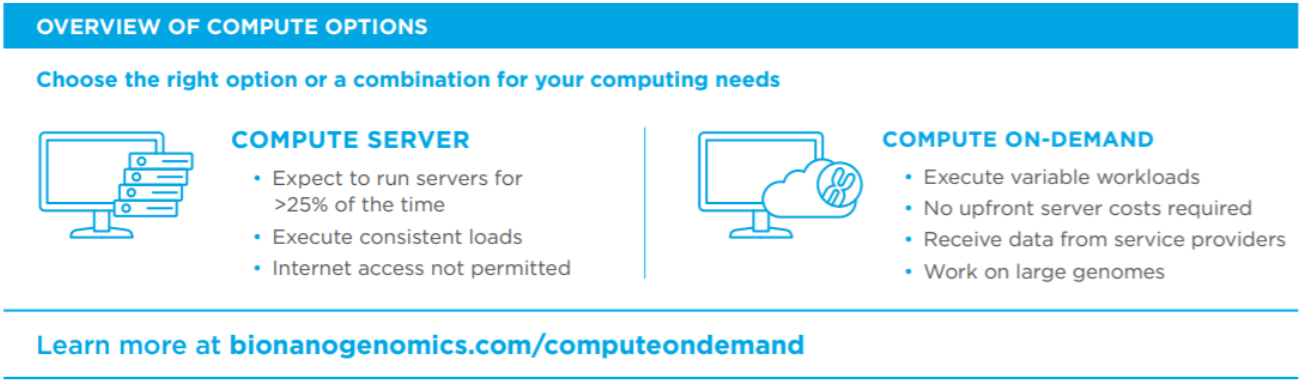
Nowoshilow, S., Schloissnig, S., Fei, J. et al. The axolotl genome and the evolution of key tissue formation regulators. Nature 554, 50–55 (2018). https://doi.org/10.1038/nature25458

Emmer Wheat



Improved Genome Sequence of Wild Emmer Wheat Zavitan with the Aid of Optical Maps
Tingting Zhu, Le Wang, Juan C. Rodriguez, Karin R. Deal, Raz Avni, Assaf Distelfeld, Patrick E. McGuire, Jan Dvorak and View ORCID ProfileMing-Cheng Luo
G3: GENES, GENOMES, GENETICS *March 1, 2019 vol. 9 no. 3 619-624; https://doi.org/10.1534/g3.118.200902*

Compute solutions



Current max data input is
- 5 Gbp for RVP
- 2.2 Gbp for de novo

Largest genome assembled to date on COD are
- Oat - 11 Gbp (2.3 Tbp input data)
- Bread wheat - 15 Gbp (2.2 Tbp input data)
- Larger genomes – contact Bionano

https://bionanogenomics.com/wp-content/uploads/2019/09/Bionano-Genomics-Compute-On-Demand-One-Sheet.pdf

Conclusions

Bionano genome maps show the true structure of the genome

DLS labeling chemistry yields up to chromosome length maps

Combining NGS and Bionano data produces assemblies of the highest quality

Bionano hybrid scaffolding is agnostic to the sequence technology used. No matter your sequencing strategy, we make your assembly better

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

Bionano optical mapping for accurate genome assembly, comparative genomics, and haplotype segregation

Alex Hastie, PhD Sr Director, Scientific Strategy and Customer Solutions | Bionano Genomics
ahastie@bionanogenomics.com