Generating the reference genome for American Chestnut

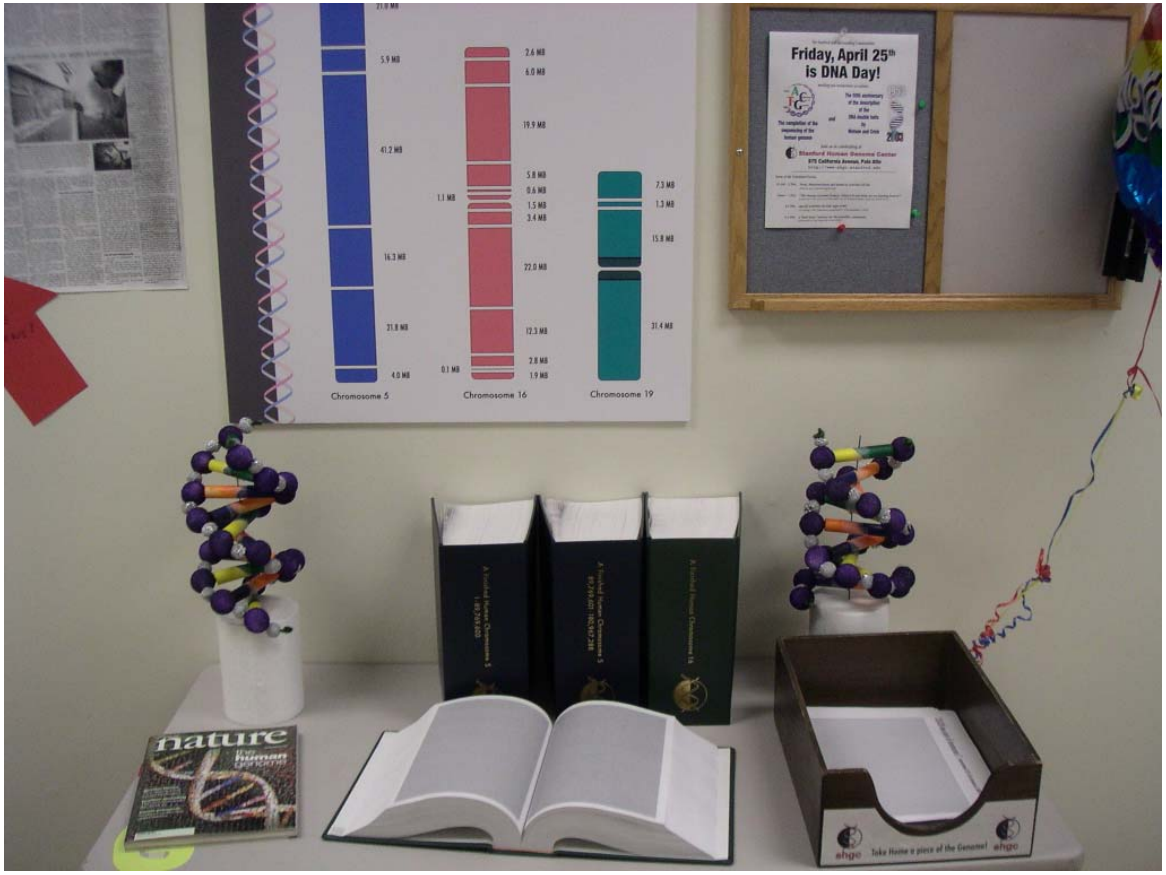
10-27-2018
TACF Meeting 2018
Jeremy Schmutz
HudsonAlpha Institute for Biotechnology

- Non-profit genomics research institute
- In Huntsville, Alabama
- 4 mission areas
  - Genomics Research
  - Clinical Genomics
  - Education & Outreach
  - Economic Development
- 213 staff & >558 in 35 Associate companies

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

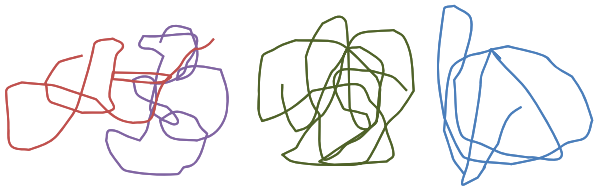
Leverage the synergy between discovery, education, and economic development in genomic sciences to improve the human condition around the globe.

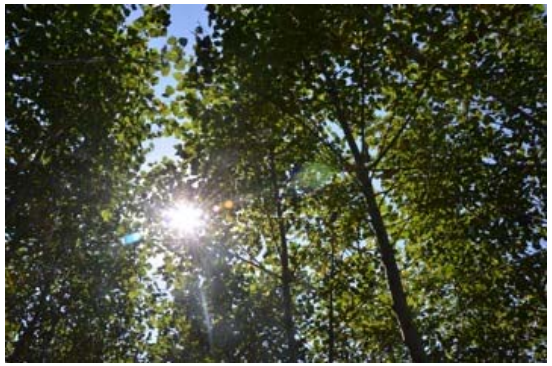
What is a reference genome?



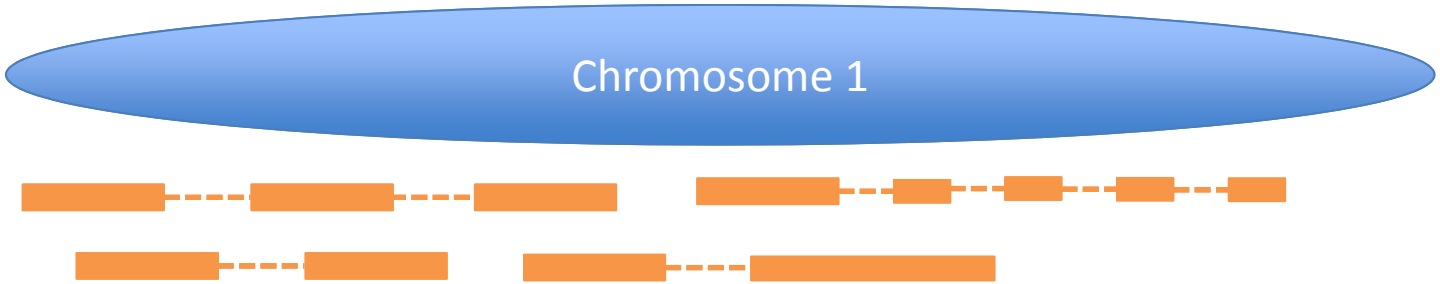
A book of genetic instructions for life- the first such copy is a *de novo* sequenced reference genome

Genomics

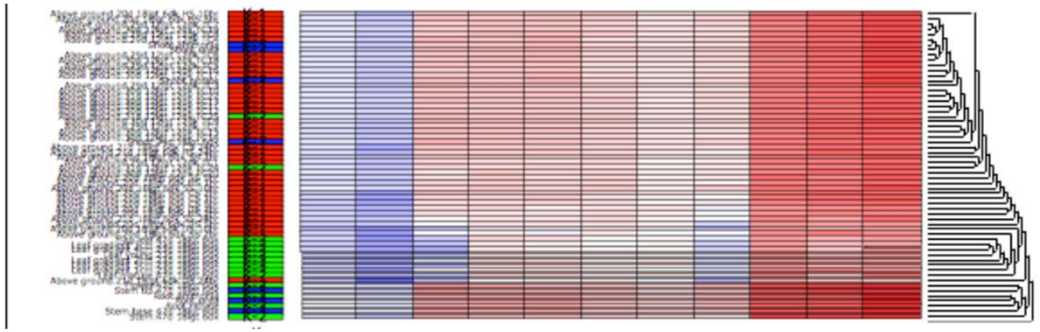




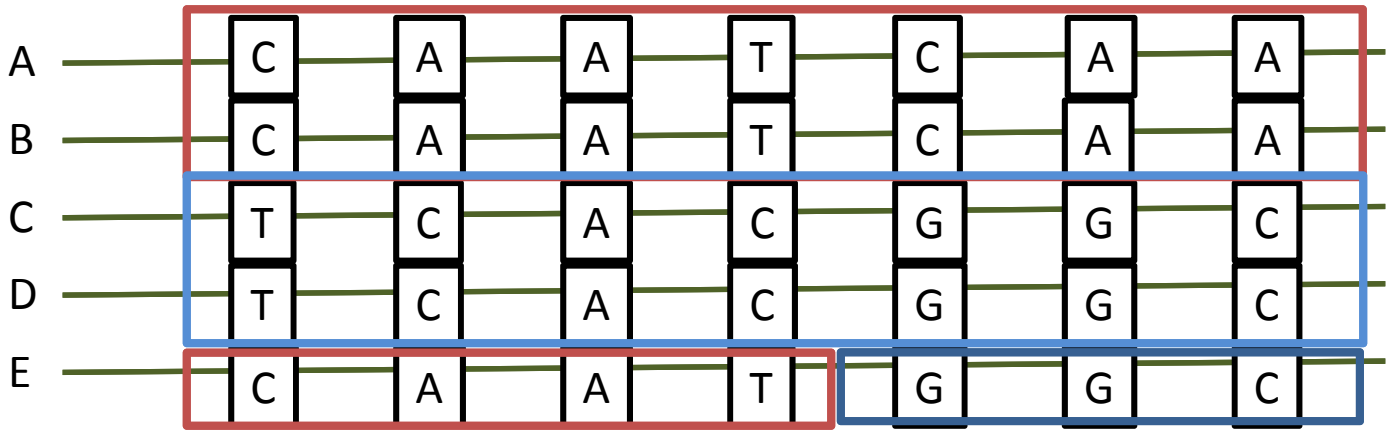
Genomics 2



RNA-seq



Reseq

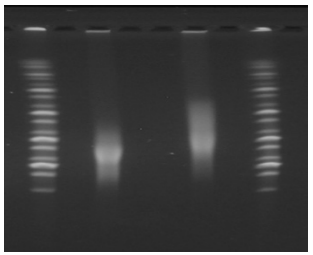


What can we do with an AC reference genome?

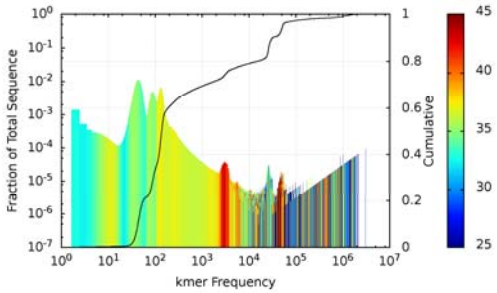
1. Understand the genetic material in the TACF breeding program
   - How do the original donor sources compare to the American chestnut
   - Compare latest offspring and identify sources of blight resistance
2. Understand the genetic diversity of remaining American chestnuts
3. Provide a platform for applying genomics to understand disease resistance and other links between DNA and tree traits

How to sequence a plant genome with PACBIO

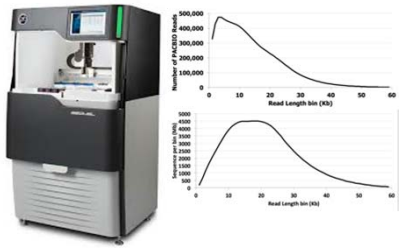
1. Get good DNA (50-150kb)



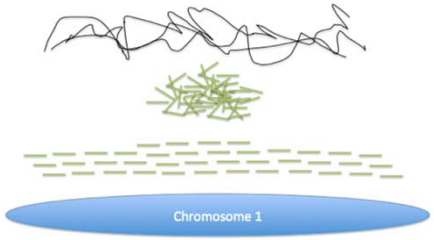
2. Assess the genome: Ploidy, het rate, repeat, organelle, contamination



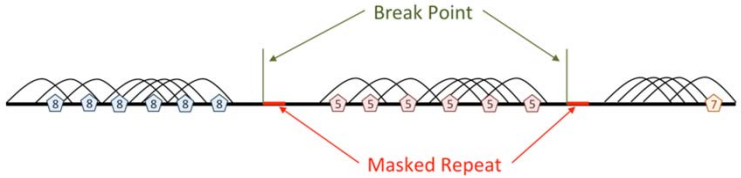
3. Generate long, high quality PACBIO libraries and collect data to 70-80x, more for outbreds



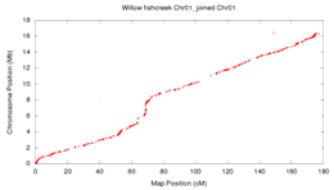
4. Assemble the genome (MECAT, Canu), quiver or arrow polish



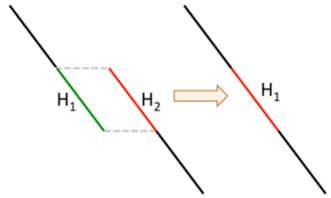
5. Break false joins



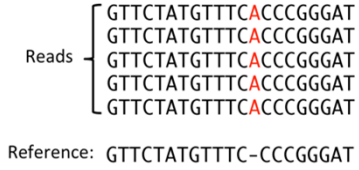
6. Integrate data and order into chromosomes



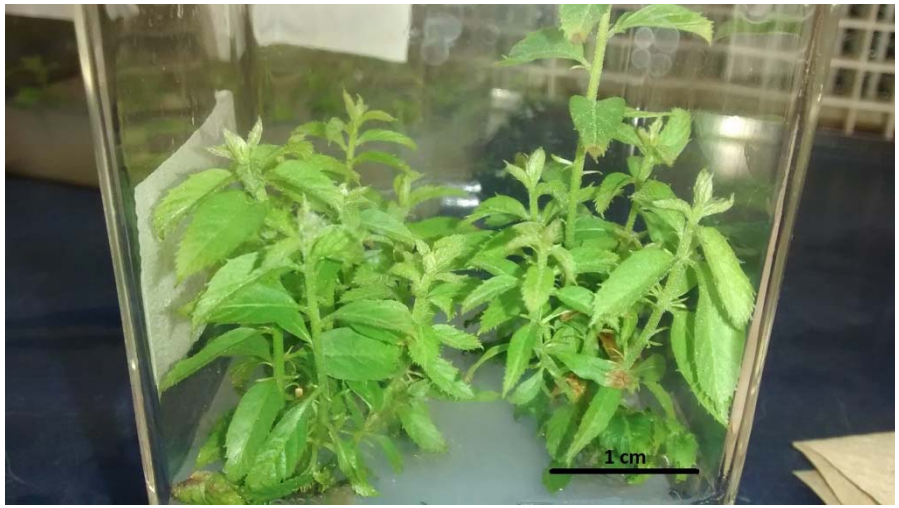
7. Address haplotype & assembly overlap



8. Final polish with Illumina to reduce homozygous errors



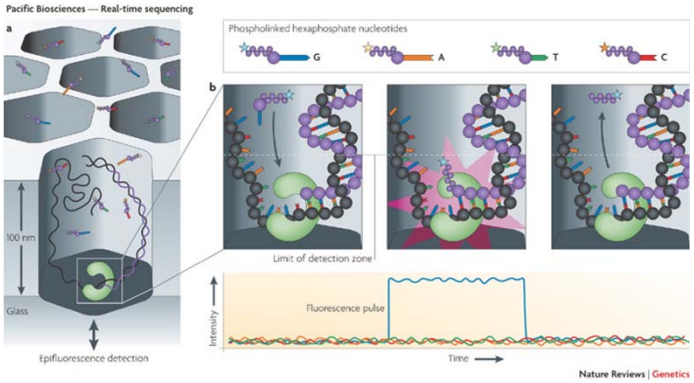
DNA Extraction



Andy Newhouse – SUNY-ESF

- Several attempts were made from test tissue and from Ellis1 tissue (from SUNY-ESF)
- Xiaoxia ‘Penny’ Xia at Clemson was successful – produced nice High Molecular Weight DNA

PACBIO based sequencing



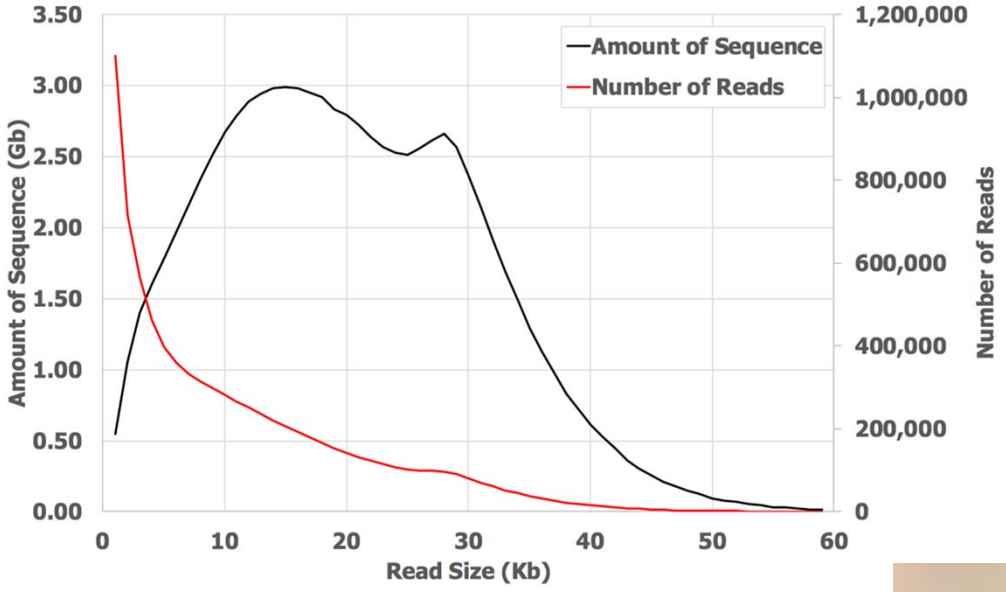
Pacific Biosciences single molecule sequencer

- Up to 50kb read length
- ~84% accuracy

>m160928_115127_42264_c101035962550000001823237911191676_s1_p0/60815

GTTGACCGCGCAGGCTTGGCGATTCCTTCGAGGTTAGAAGCTATATAGGCGGCCCATCGCCCATCGGCGTGCGCTCCGTCCACACCACCCCTC
ATCTCCGCCAGCGCCACTCGCATCGAGATCCGCGCGCCGTGCTCACCTCGCTCGCAAACTCCCGAGGTTAGGGTTAGAGCTTTCCGCTCAAGTC
CTTTTATGCTCCTATCCTCGCCGTACCGTGTGAGAGATGCGAGCGGCGACGATTTCCCCTTTTTTTATTTGAATTTTCGGGCCCTGTTATTGCGG
CGGCTAGATCTGCTTGTCGCTGGTGAGTAGGGGGTTCTCGTTCGGTTCCGATCACGCGAAAATCTGAGAACCCAACTAGTGGGGGGGCGTAG
ATCCGCCCGATTGCTGCGCATTGTCGCGCGTAAACGTCGAGATTTTTTTGTTACCCGTTTTTCCTCATGTGTTGTGGTATCGTTTTGGGAGGAGA
TTATTAGGGATAAGCTGTGATTACTTATGGGAAGGCGAGATGGTTGCCCATGTCGATTTGGCCGTGCAAAGTGCCGCAAATCTGAATGTTGGT
CGCCAACGGACTGCCCCTTTGATTTCTGATCTAGTCGCGCTCAGTCTGATGCTAACTCGGTTGTAACTGTGATTTCCTCGCTGCTTTTCGAAACA
TTGCGCCCGTGGGATTCGTTTCAAGCTGATGGCTTCACGATGTTGTTGAAAGATTGTTTAATACTGGAACTAGAAGTGTTAATAATATGGCTTA
TGTTGGTAGTGAATTAGTTTGGTTGGTGACGCAGTGTCCTTGAGGTAATAATTGTTGTCAATGGTTTATAGTCTGGTTAGCCTGTTGTAAACCTT
TTTGTTATGCTGTAATTGTATAAGTAGTATTACCTGCTTCGCTGGCATGGACATCGCATTCAAACTGATATAGTGATAATTAGGCCATATTTATG
GAGTTCAGTGTAATTATATGTGAACTTTTGGTGTACCCCAAAATAAAAATGTTTTACTGCAAATATATTTTTGTTGGAACTCAATTTACAATGTTC
TGGATATTTGCTCACAGATTTTGTGGGATGTTTGTGTGGTCACATCACATAGCTTTTTATTGTTTAGGTACTGTGAATAGAGCTAGAATTCTGTG
AATAGCGCCTCAGATATTGCATTGCCCTTTCCTGTTATTCCTGATTAGATGATACGTTTTCTACTGCTCTTGAAATTGCTCATATCAGTGTGTTTCT
TTGGGGTTATTTGCAGTAACGTCTATTCCCTTGTCAACATGTCTGCGGGAGGAGAATGTCTACATGGCCAAGCTGGCTGAGCAAGCAGAACGG
TACGAGGAAATGGGGTTGAGTACATGGAGAAGGTGGCTAGACCGTTGAATGTGGAAGAGCCACCGTTAGGAGCGTAACCTCCTATCTGTGTG
CTACAAGAATGTGATTGGGCCACGCCGTGCCCATGCGCAATTGTCTACTCCATTCGAGCAAGAAGGAGGAGGTCCCGCCAAGAATGAGGAAC
ATGTTGCTCAAGATTAAGGAGTACCGTGGCAAGATTGAGGCTGAGATTGAGCAAGCATTTGTGATGGTATCCTGAAGCTGCTTGACTTCGCAC
CTTGTCCTTCCTCCACTGCTGGCAGGAGTCCAAGGTGTTTTTACCTCAAGAGACGGTGATTATCACAGGTAGAAGTTTGAACACTATATTGTTG
CGTAGTTGACTCCATGATATAGTCTGTAATCATATATTAATTGGGTTTACTTTTTTATTTTTAATGTTTTGATTAAAGGTACCTTGCTGAGTTTAA
GACATGGTTAACTGTAGAGGCAAGGAATCTGCAGAGAGCACAATGGTAAGCTTATACAGGCAGCCGCAGGTTGGTTCATGTTATCACTGGCC
ATAACAATCATCTTTAATTTGATTATTGTTATTTCTTCCATAAGTAATGCAAATTTAATATCTCTTCTTAAAAATGTCTGTCCTATACCCTGTTAGG
ATATTGCATTGGCTGAACTGGCACCTAACTCATGCCATAAGGCTTGGACTTGCACTTAAACTTCTCAGTGTTCTTATTAATGAGATTCCTGAACC
TCCCGGACAAAGCTTTGCAACCTGCTAAGCAGGTAGGTAATCGCATTTGTCTTCTATTTGCTGTACCCGGTTTGAGTTGTAACTAGCTTCTGTGT
TGCTTGTTTGAATCTGGGCTGAAAGAAGGAAGGTGCAAATAGTATATGAATTGGTTGAAATACTCAAAGCACAGCGATTTTACCTATGGTTAA
CTTATGTGTCGGCTCTAGATTTCCAGGTGGAATTTGCGGATTCATTTGTTCAGTTCACTGCTATACATTTATCTCTTGTGATAATTGTATTGTTGG
TACAATGTGCCGCTCTTCGATGCTTTTTGGTGTGATGACATTTGGCACTTTTGCCACACAAACCAATTATGTTCCCCCGCAGCAGAAACCTTGCT
AGCAGCAGCAGCAGTAAAAATTTAATCATCACTCCTTCTTATTATGTATTATTGTTTGGTGTTATGTTGACCTACAAGAATTGCATGGTAATTTG
TTGCTGATTTGAGAAACTGATCATATCTAGATCACACGAAATGCATGATGTGTACAATTTGTGTAGCTGTTTTGGTTGGGGCGCAGTGGGTTTT
ATTTCCTGATCATATTCTTTGCAGGCATTTGATGAGGCTATTTCTGAGTTGACACAGCTTGGCGAGGAATCGCCTACAAAGAAGCACCTTGATT
AATGCCGGCTCCTGAGGGACACTTGACCCCTGTGGACCCTCGTACATCACGGTATCTCAGCAAGCTGAAACTTGGCTGTCCCTTAATGACATTA
CCCATACTCACGAGAATCTGATGCAGATATCATGAAATTTCCCCTGCAGGAGGAGGGCATTGAGGAGGCAAAGAAGCCACGAAAGGGTGATG
CAGGCGAGGGTCAGTAAGCCTGGTAAAGCCTTGAATGGATGGTTTCGGCATTTGTTCTGCCCATTTGTGTTGGACGCCATGTGCTGATGAGAA
GGTGTTGCTGCTCCATGGTTCAAGCTTCTGCCCTCCTTTCGTTCCTCCCTCTTCATGTTGTAGGCTGGTAGCGGTCAAATTAGCCCATTGGCTTTA
GTTGGCCCTGGGTCCATTGGACTGATGTACTGCTGTGAAGTCATCCGTTCTCGTTTTTGCTCTTGCAGTTGATGCGGTCTGTTTGGGTTGTGTTT
TTCGCATCTTTTTTATTTCCCAAACATGCAGTTTTAGATTTTGCTTAGGATGGGTTAGGCAGTAGGCTATGGCAAATATTCTGTTTCACTGCATG
ACGCCTTTGTCGCCCGCCCAAACAGCAGCATAATCGGCTCATGACCACTGGGATTCCGTGACGGATCTCCAGGCAGGTGTTTCCTTGCTGAAAA
GGGCTAAGTCGTCTAGAAACCGATTAGTTGCGCACTCGCTGTTTTGGCTACCGTGCGGGGCCCGTGCGAGTTTACCACACATAGTTGTGATCCT
TGCTGTGTCAATAGAATGATACAGGCTGGGTGCTATTTACCTTGTACTCGGATTCAAAAGAAAAGACTGTAGGGCAGAGGAAATGTAAAGACT
GGAAGGGCATAGGAAGTAGATTTGAAAATGAAATGTACATCATTCTTTCAGTAAAGACGGTAATAAGAGCAGCAAGGTATAATAAGTTTTAG
TCGCTAGCTAAATAAGCTATTGGGCGAGAAGAGCCGGCGTGGCCCCAGATCCATTCATGCTATCTCAAAATCAACTCGTTGAGAATAGCATAT
AGCTATTCTCAACAGCAGGTAAATTCATATCCCTACCATTTCAAACCCTGGAAACATAAGATTGTTTCTTAGCTGCTTAACAACCCCACAGGGCA
ATAGCAAGAAAATATTACACAACAGAATTTTCTCTTACGAATCCACAACTGGCCAATTGAAAAAGAAATAAAAAAAAAAAAAAAAAAAAAAAA
CTATTCCTAAGCTGTCTATCGAAAATCTCAGAAGATTGAACTCCACATTTGTTGCATAACACATTCAAAGAAACAAAGTGAGTAACCGTTTCTA
GTTCACAACAGAAACAGACAAGACTTGTCTTCTACATGTCTTCTAATACTACAGGTTTCCCTAGTGAATGAACTTGTTTTGGAAAAGTACAATA
GGAAAAATCTGACTCTAGAAGGCCTTCAGCCCCCAGAACGGCATGGTACATGCACTGGCATATAATTCCTCTAAAATTAATAACTCTATATAAA
GAGGAAGAAGAATAGAACACCATACCAGAGTATAATTGCCATCATCAGAAAAAACTAATAGTTGAGCCAGTTGAAACGCACCTCTAACCAGG
GATTCATTAATCATCAAATCAACACATCTCCTAAAGGTAACTTGAAGGTTTGTACCATCACAGATTGAAGAACTGATTAGTTTTCTCATTAACTA
GAATGTAGAATATCCCAGTACTGATGGACAACGTTGAAATTCCTAACGCAATTGTCTTCCCAAAACTTAACATTTTCCGCCCCATTACCAATTTT
CCTACCTACCACATCCTGTTGCAGCTTTAGCAGCCACATGAAGCCCCTTATAAAGAAATAGAAGCCCCAACTGTAAGCTACAAAAAATATTAGG
CCTACTTTGTATCATACTTGTGAATCAATAAAGCCTGGTTTCTAGAGCTTATGACTATCTAATATATCTCTAACCCAGTTCAAGCTTAACTATTGC
TTCAAATTAGGGATACCTAACCCTCCTAAATTCCTTTTAACATGCAAATTGTCCCATTCACTTATCAAAACGGGTCCGTTTGGATTCTTGGCAAGT
TTAGACCTCAGGTTGCTCATGGAAGTCATCATTGTAGTTTTGCTTCCAGATGTGGAAACTTAAAGCTCAGGATAAATGATAAGTAGCCTAGAAA
ATAACCTGCTTGTTCCTCACCTACTAACACAAATCCAAGCAGCTTCAAGCTTAACTATTGCTCCTCCAAAATTCCAACGATTTGAGGCCATCAAT
GCTGTTCAGGCTCAAACCACCTTCAACTATGGAATGTTCATCGATGAAAACAACATTTCCATAAGAGCTAAACGGTATGAAGCCGTATACACAA
TGGAACAGTGTGTACATGGTTATATGATACTGTACAACACCAACAAATGATGCATATATTTGTGCCTGTTATGCAAGTAAAAAATCAACACATA
TGAGTACAAGTAAAAAAGAAGAATCGTCAGTAATCAAACAATTAATCTGAGAAAGATACTGATAAACATAGAAAAATTGTACAGATAGAGCA
GACACTATCAAGAACAGACACGAGATAGGGAACATTGCTGGCGCCCAATGAGTAATATTGAGTGAGGGCATGTGCCGGATTTGTAAGTCTAA
TTCCATAACGTATTGCATAAATGACACAAGATGTAATTTCGTTGCAACTCACATGCATATTTCCCCTAGTGTAATATAATAGTAATAAACACATC
AACCATTACTACTAGATCTAGATAGTAAGATAAGATACAAAAAATATGTGGACAGTATAGGAAGGAACAATAAC

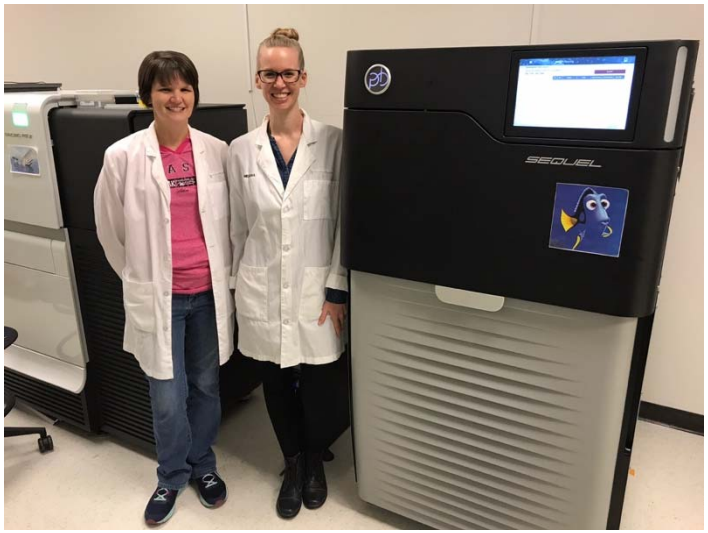
American Chestnut sequencing



126.7x PACBIO Coverage

Average Read 7,745 bp

Jane Grimwood @ HA



Lori Beth Boston, Melissa Williams @HA

MECAT and Canu

MECAT: fast mapping, error correction, de novo assembly for single molecule sequencing reads

- Superior performance compared to current tools with improved results
  - 10-70 X faster on reference genome, 5-20X on pairwise alignment, 5-20 X faster on error correction
- Collaboration between Sun Yat-sen University and Clemson
- Available at https://github.com/xiaochuanle/MECAT

Feng Luo

CLEMSON UNIVERSITY

中山大學 SUN YAT-SEN UNIVERSITY

Canu: Modified CELERA that uses MinHash to assess overlaps & localizes repeats to increase scaffolding accuracy

- auto-grid configuration and general ease of use
- Easy, flexible and accurate
- Available at https://github.com/marbl/canu

Brian Walenz

Adam Phillippy

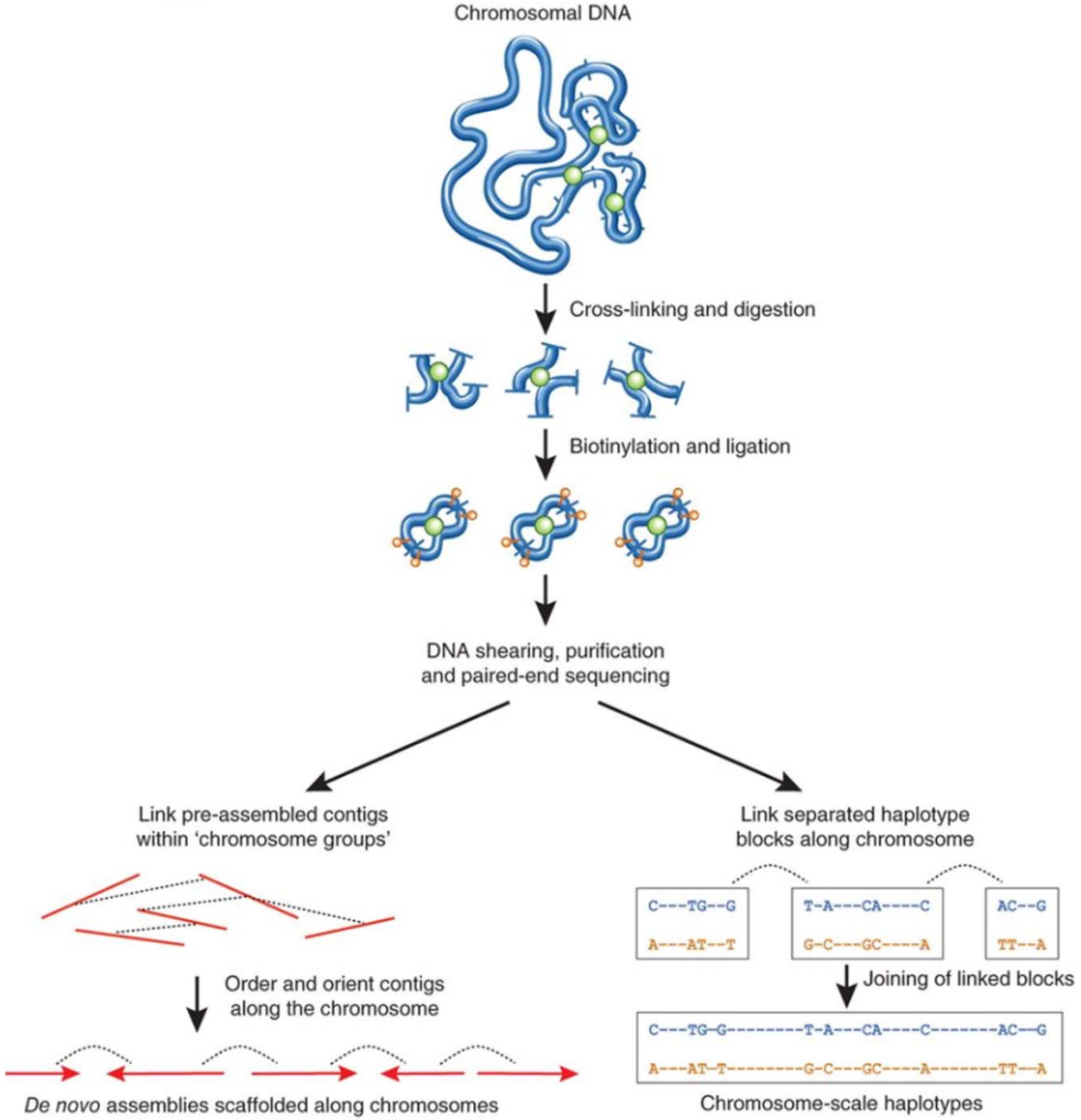
Sergey Koren

Xiao, Nature Methods (2017)

Koren, Genome Research (2017)

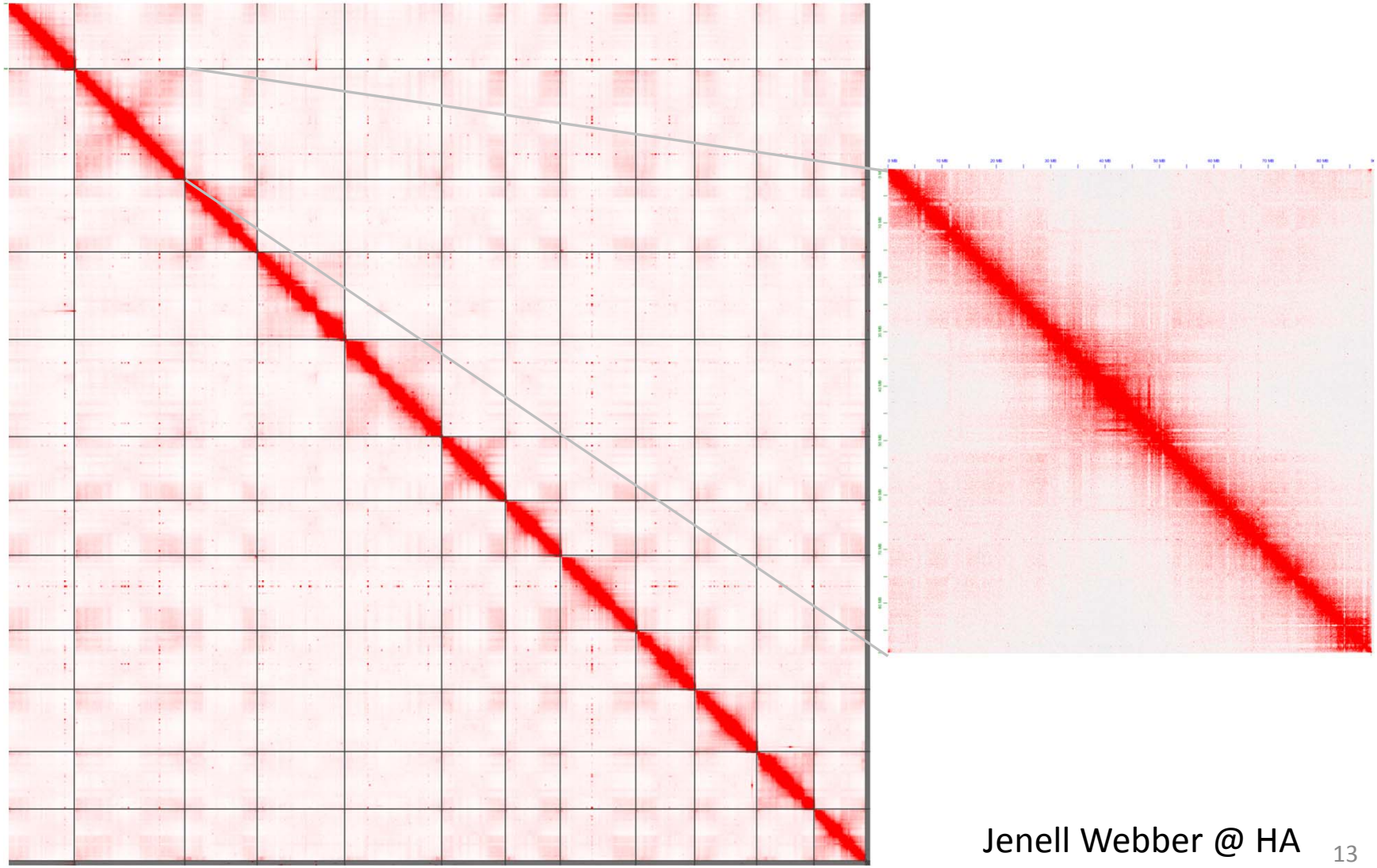
NIH National Human Genome Research Institute

Introducing HI-C for chromosome build



Nature Biotechnology **Volume 31**, pages 1099–1101 (2013)

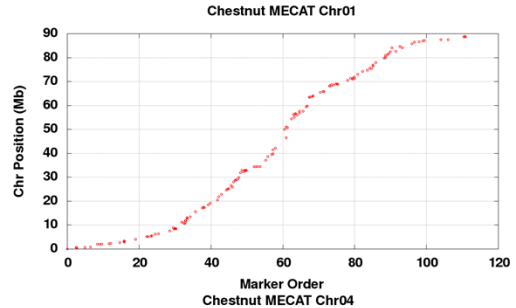
HI-C on AC ‘Ellis1’

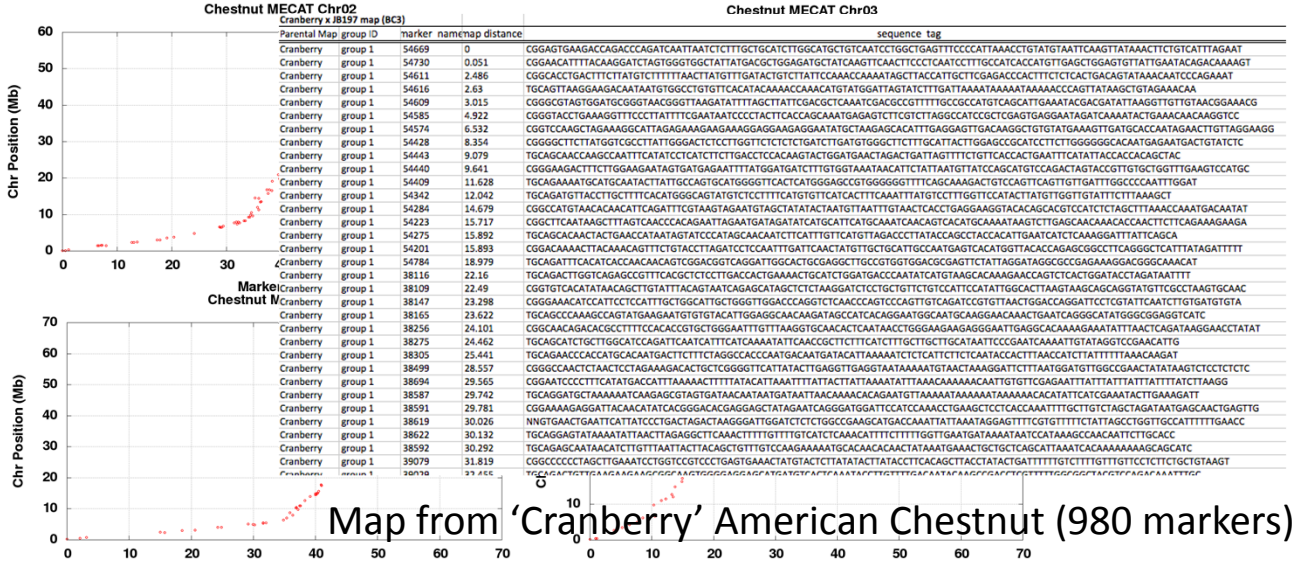


Juicer Juicebox

Jenell Webber @ HA

Mapping the genome





Cranberry x JB197 map (BC3)

Parental Map	group ID	marker_name	map distance	sequence_tag
Cranberry	group 1	54669	0	CGGAGTGAAGACCAGACCCAGATCAATTAATCTCTTTGCTGCATCTTGGCATGCTGTCAATCCTGGCTGAGTTTCCCCATTAAACCTGTATGTAATTCAAGTTATAAACTTCTGTCATTTAGAAT
Cranberry	group 1	54730	0.051	CGGAACATTTTACAAGGATCTAGTGGGTGGCTATTATGACGCTGGAGATGCTATCAAGTTCAACTTCCCTCAATCCTTTGCCATCACCATGTTGAGCTGGAGTGTTATTGAATACAGACAAAAGT
Cranberry	group 1	54611	2.486	CGGCACCTGACTTTCTTATGTCTTTTTAACTTATGTTTGATACTGTCTTATTCCAAACCAAAATAGCTTACCATTGCTTCGAGACCCACTTTCTCTCACTGACAGTATAAACAATCCCAGAAAT
Cranberry	group 1	54616	2.63	TGCAGTTAAGGAAGACAATAATGTGGCCTGTGTTCACATACAAAACCAAACATGTATGGATTAGTATCTTTGATTAAAATAAAAATAAAAACCCAGTTATAAGCTGTAGAAACAA
Cranberry	group 1	54609	3.015	CGGGCGTAGTGGATGCGGGTAACGGGTAAGATATTTTAGCTTATTCGACGCTCAAATCGACGCCGTTTTTGCCGCCATGTCAGCATTGAAATACGACGATATTAAGGTTGTTGTAACGGAAACG
Cranberry	group 1	54585	4.922	CGGGTACCTGAAAGGTTTCCCTTATTTTCGAATAATCCCCTACTTCACCAGCAAATGAGAGTCTTCGTCTTAGGCCATCCGCTCGAGTGAGGAATAGATCAAAATACTGAAACAACAAGGTCC
Cranberry	group 1	54574	6.532	CGGTCCAAGCTAGAAAGGCATTAGAGAAAGAAGAAAGGAGGAAGAGGAATATGCTAAGAGCACATTTGAGGAGTTGACAAGGCTGTGTATGAAAGTTGATGCACCAATAGAACTTGTTAGGAAGG
Cranberry	group 1	54428	8.354	CGGGGCTTCTTATGGTCGCCTTATTGGGACTCTCCTTGGTTCTCTCTCTGATCTTGATGTGGGCTTCTTTGCATTACTTGGAGCCGCATCCTTCTTGGGGGGCACAATGAGAATGACTGTATCTC
Cranberry	group 1	54443	9.079	TGCAGCAACCAAGCCAATTTCATATCCTCATCTTCTTGACCTCCACAAGTACTGGATGAACTAGACTGATTAGTTTTCTGTTCACCACTGAATTTCATATTACCACCACAGCTAC
Cranberry	group 1	54440	9.641	CGGGAAGACTTTCTTGGAAGAATAGTGATGAGAATTTTATGGATGATCTTTGTGGTAAATAACATTCTATTAATGTTATCCAGCATGTCCAGACTAGTACCGTTGTGCTGGTTTGAAGTCCATGC
Cranberry	group 1	54409	11.628	TGCAGAAAATGCATGCAATACTTATTGCCAGTGCATGGGGTTCACTCATGGGAGCCGTGGGGGGTTTTCAGCAAAGACTGTCCAGTTCAGTTGTTGATTTGGCCCCAATTTGGAT
Cranberry	group 1	54342	12.042	TGCAGATGTTACCTTGCTTTTCAACATGGGCAGTATGTCTCCTTTTCATGTGTTCATCACTTTCAAATTTATGTCCTTTGGTTCCATACTTATGTTGGTTGTATTTCTTTAAAGCT
Cranberry	group 1	54284	14.679	CGGCCATGTAACACAACATTCAGATTTCGTAAGTAGAATGTAGCTATATACTAATGTTAATTTGTAACTCACCTGAGGAAGGTACACAGCACGTCCATCTCTAGCTTTAAACCAAATGACAATAT
Cranberry	group 1	54223	15.717	CGGCTTCAATAAGCTTTAGTCAACCCACAGAATTAGAATGATAGATATCATGCATTCATGCAAATCAACAGTCACATGCAAAATAAGTCTTGAGCAACAAACACCAACTTCTTCAGAAAGAAGA
Cranberry	group 1	54275	15.892	TGCAGCACAACTACTGAACCATAATAGTATCCCATAGCAACAATCTTCATTTGTTCATGTTAGACCCTTATACCAGCCTACCACATTGAATCATCTCAAAGGATTTATTCAGCA
Cranberry	group 1	54201	15.893	CGGACAAAACTTACAAACAGTTCTGTACCTTAGATCTCCAATTTGATTCACTATAGTTGCTGCATTGCCAATGAGTCACATGGTTACACCAGAGCGGCCTTCAGGGCTCATTTATAGATTTTT
Cranberry	group 1	54784	18.979	TGCAGATTTCACATCACCAACAACAGTCCGACGGTCAGGATTGGCACTGCGAGGCTTGCCGTGGTGGACGCGAGTTCTATTAGGATAGGCGCCGAGAAAGGACGGGCAAACAT
Cranberry	group 1	38116	22.16	TGCAGACTTGGTCAGAGCCGTTTCACGCTCTCCTTGACCACTGAAAACTGCATCTGGATGACCCAATATCATGTAAGCACAAAGAACCAGTCTCACTGGATACCTAGATAATTTT
Cranberry	group 1	38109	22.49	CGGTGTCACATATAACAGCTTGTATTTACAGTAATCAGAGCATAGCTCTCTAAGGATCTCCTGCTGTTCTGTCCATTCCATATTGGCACTTAAGTAAGCAGCAGGTATGTTCGCCTAAGTGCAAC
Cranberry	group 1	38147	23.298	CGGGAAACATCCATTCCTCCATTTGCTGGCATTGCTGGGTTGGACCCAGGTCTCAACCCAGTCCCAGTTGTCAGATCCGTGTTAACTGGACCAGGATTCCTCGTATTCAATCTTGTGATGTGTA
Cranberry	group 1	38165	23.622	TGCAGCCCAAAGCCAGTATGAAGAATGTGTGTACATTGGAGGCAACAAGATAGCCATCACAGGAATGGCAATGCAAGGAACAAACTGAATCAGGGCATATGGGCGGAGGTCATC
Cranberry	group 1	38256	24.101	CGGCAACAGACACGCCTTTTCCACACCGTGCTGGGAATTTGTTTAAGGTGCAACACTCAATAACCTGGGAAGAAGAGGGAATTGAGGCACAAAAGAAATATTTAACTCAGATAAGGAACCTATAT
Cranberry	group 1	38275	24.462	TGCAGCATCTGCTTGGCATCCAGATTCAATCATTTCATCAAAATATTCAACCGCTTCTTTCATCTTTGCTTGCTTGCATAATTCCCGAATCAAAATTGTATAGGTCCGAACATTG
Cranberry	group 1	38305	25.441	TGCAGAACCCACCATGCACAATGACTTCTTTCTAGGCCACCCAATGACAATGATACATTAAAAATCTCTCATTCTTCTCAATACCACTTTAACCATCTTATTTTTTAAACAAGAT
Cranberry	group 1	38499	28.557	CGGGCCAACTCTAACTCCTAGAAAGACACTGCTCGGGGTTCATTATACTTGAGGTTGAGGTAATAAAAATGTAACTAAAGGATTCTTTAATGGATGTTGGCCGAACTATATAAGTCTCCTCTCTC
Cranberry	group 1	38694	29.565	CGGAATCCCCTTCATATGACCATTTAAAAACTTTTATACATTAAATTTTATTACTTATTAAAATATTTAAACAAAAAACAATTGTGTTCGAGAATTTATTTATTTATTTATTTTATCTTAAGG
Cranberry	group 1	38587	29.742	TGCAGGATGCTAAAAAATCAAGAGCGTAGTGATAACAATAATGATAATTAACAAAACACAGAATGTTAAAAATAAAAAATAAAAAACACATATTCATCGAAATACTTGAAAGATT
Cranberry	group 1	38591	29.781	CGGAAAAGAGGATTACAACATATCACGGGACACGAGGAGCTATAGAATCAGGGATGGATTCCATCCAAACCTGAAGCTCCTCACCAAATTTTGCTTGTCTAGCTAGATAATGAGCAACTGAGTTG
Cranberry	group 1	38619	30.026	NNGTGAACTGAATTCATTATCCCTGACTAGACTAAGGGATTGGATCTCTCTGGCCGAAGCATGACCAAATTATTAAATAGGAGTTTTCGTGTTTTTCTATTAGCCTGGTTGCCATTTTTTGAACC
Cranberry	group 1	38622	30.132	TGCAGGAGTATAAAATATTAACTTAGAGGCTTCAAACTTTTTGTTTTGTCATCTCAAACATTTTCTTTTTGGTTGAATGATAAAATAATCCATAAAGCCAACAATTCTTGCACC
Cranberry	group 1	38592	30.292	TGCAGAGCAATAACATCTTGTTTAATTACTTACAGCTGTTTGTCCAAGAAAAATGCACAACACAACTATAAATGAAACTGCTGCTCAGCATTAAATCACAAAAAAAAGCAGCATC
Cranberry	group 1	39079	31.819	CGGCCCCCCTAGCTTGAAATCCTGGTCCGTCCCTGAGTGAAACTATGTACTCTTATATACTTATACCTTCACAGCTTACCTATACTGATTTTTTGTCTTTTGTTTGTTCCTCTTCTGCTGTAAGT

Map from ‘Cranberry’ American Chestnut (980 markers)

Tatyana Zhebentyayeva
Clemson (now Penn State)

Integrated genome vs. map

What does the genome look like?

12 Chromosomes (99.4% bases)
702.2 Mb
12.8 Mb contig N50

Genome	Size (MB)	Contig N50	% in chrom
American Chestnut	702	12.8	99.4
Poplar V4	392	13.2	99.2
Wild Poplar	391	7.5	99.9
Willow	316	5.1	90.1
Tetraploid Switchgrass	1,125	5.5	97.2

Jerry Jenkins & Chris Plott @HA

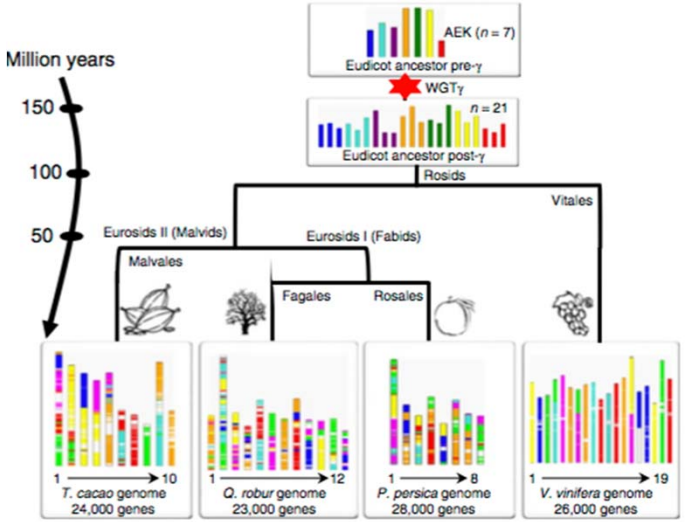
Only 72,952 more pages after this…

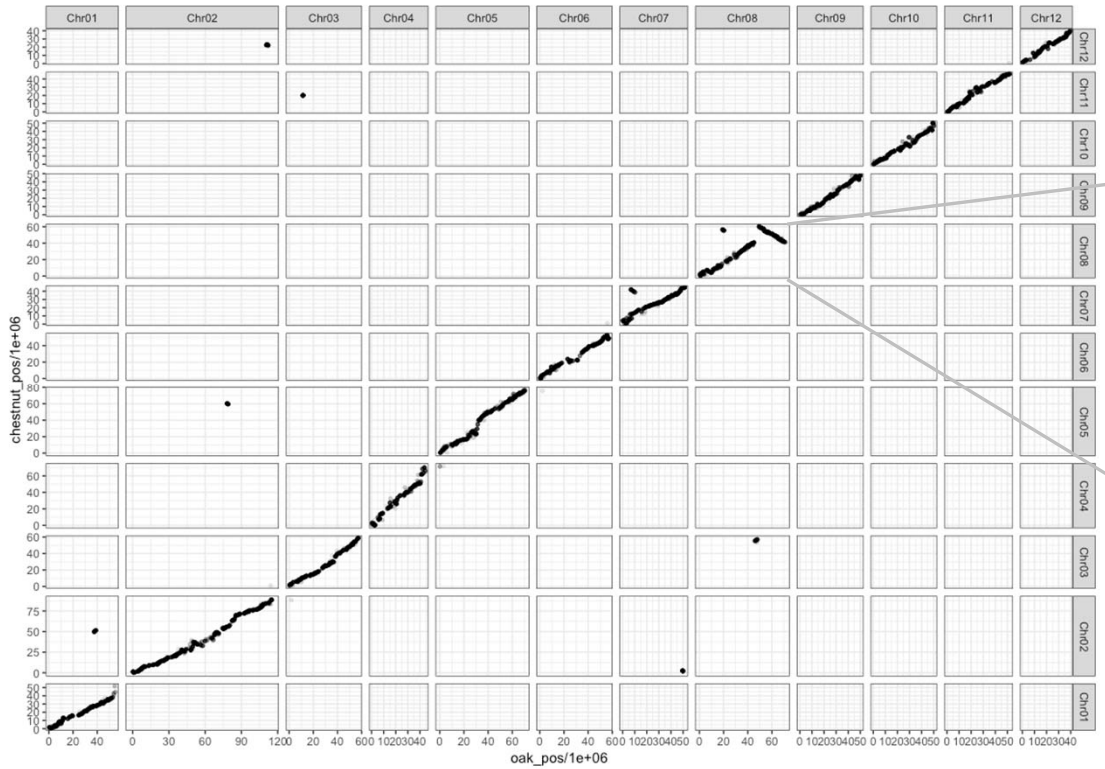
Comparisons to oak

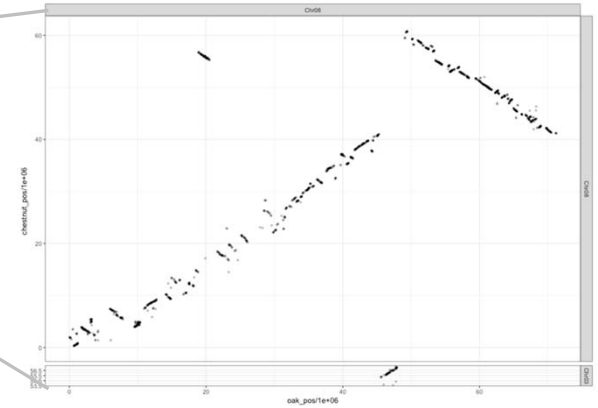
LETTERS
https://doi.org/10.1038/s41477-018-0172-3
nature plants
OPEN

Oak genome reveals facets of long lifespan

Christophe Plomion[1,24]*, Jean-Marc Aury[2,24], Joëlle Amselem[3,24], Thibault Leroy[1], Florent Murat[4], Sébastien Duplessis[5], Sébastien Faye[2], Nicolas Francillonne[3], Karine Labadie[2], Grégoire Le Provost[1], Isabelle Lesur[1,6], Jérôme Bartholomé[1], Patricia Faivre-Rampant[7], Annegret Kohler[5], Jean-Charles Leplé[8], Nathalie Chantret[9], Jun Chen[10], Anne Diévart[11,12], Tina Alaeitabar[3], Valérie Barbe[2], Caroline Belser[2], Hélène Bergès[13], Catherine Bodénès[1], Marie-Béatrice Bogeat-Triboulot[14], Marie-Lara Bouffaud[15], Benjamin Brachi[1], Emilie Chancerel[1], David Cohen[14], Arnaud Couloux[2], Corinne Da Silva[2], Carole Dossat[2], François Ehrenmann[1], Christine Gaspin[16], Jacqueline Grima-Pettenati[17], Erwan Guichoux[1], Arnaud Hecker[5], Sylvie Herrmann[18], Philippe Hugueney[19], Irène Hummel[14], Christophe Klopp[16], Céline Lalanne[1], Martin Lascoux[10], Eric Lasserre[20], Arnaud Lemainque[2], Marie-Laure Desprez-Loustau[1], Isabelle Luyten[3], Mohammed-Amin Madoui[2], Sophie Mangenot[2], Clémence Marchal[5], Florian Maumus[3], Jonathan Mercier[2], Célia Michotey[3], Olivier Panaud[20], Nathalie Picault[20], Nicolas Rouhier[5], Olivier Rué[16], Camille Rustenholz[19], Franck Salin[1], Marçal Soler[17,21], Mika Tarkka[15], Amandine Velt[19], Amy E. Zanne[22], Francis Martin[5], Patrick Wincker[23], Hadi Quesneville[3], Antoine Kremer[1] and Jérôme Salse[4]

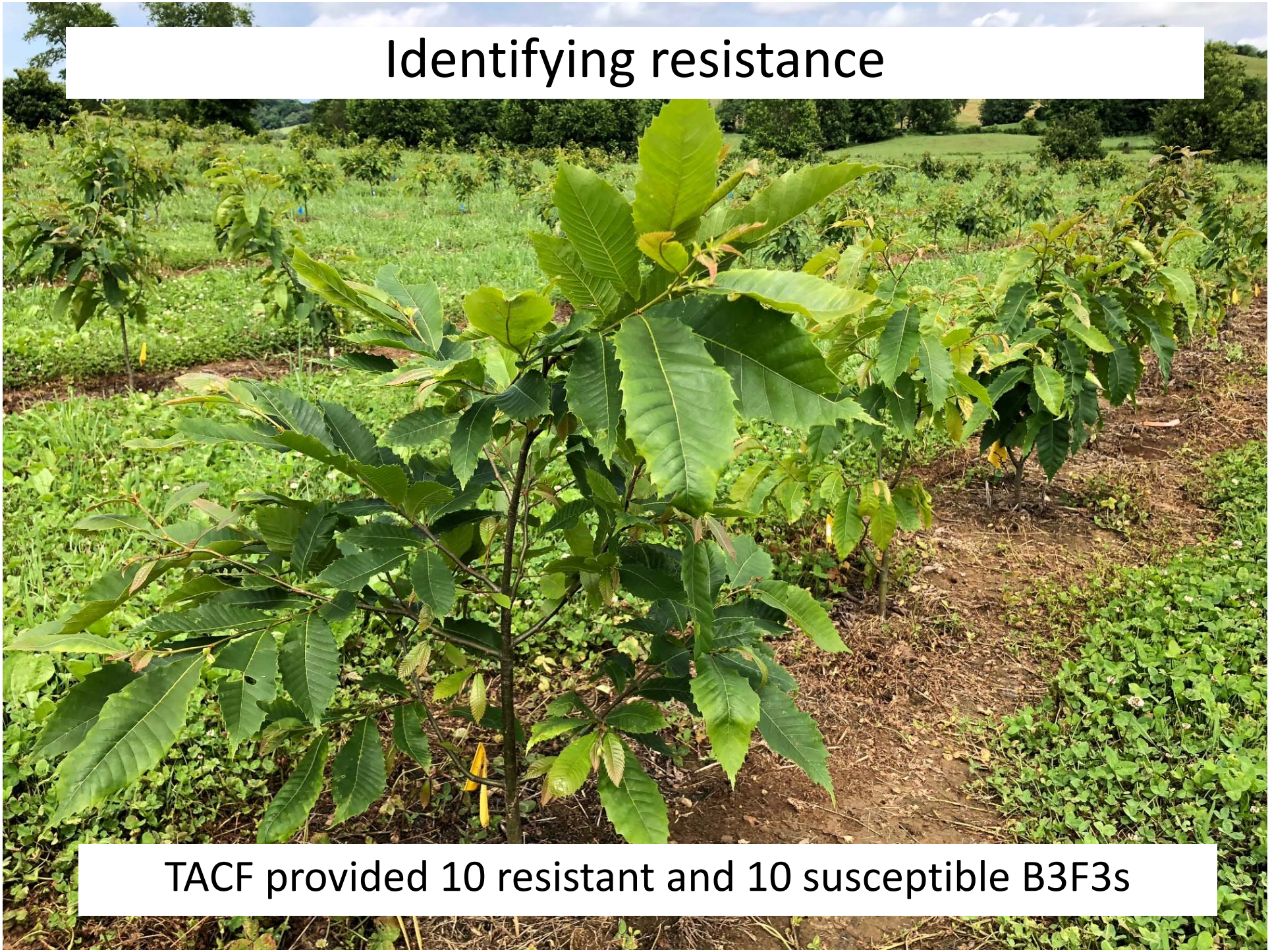






Pedunculate oak

Identifying resistance



TACF provided 10 resistant and 10 susceptible B3F3s

Searching for sources of resistance

Graves BC1

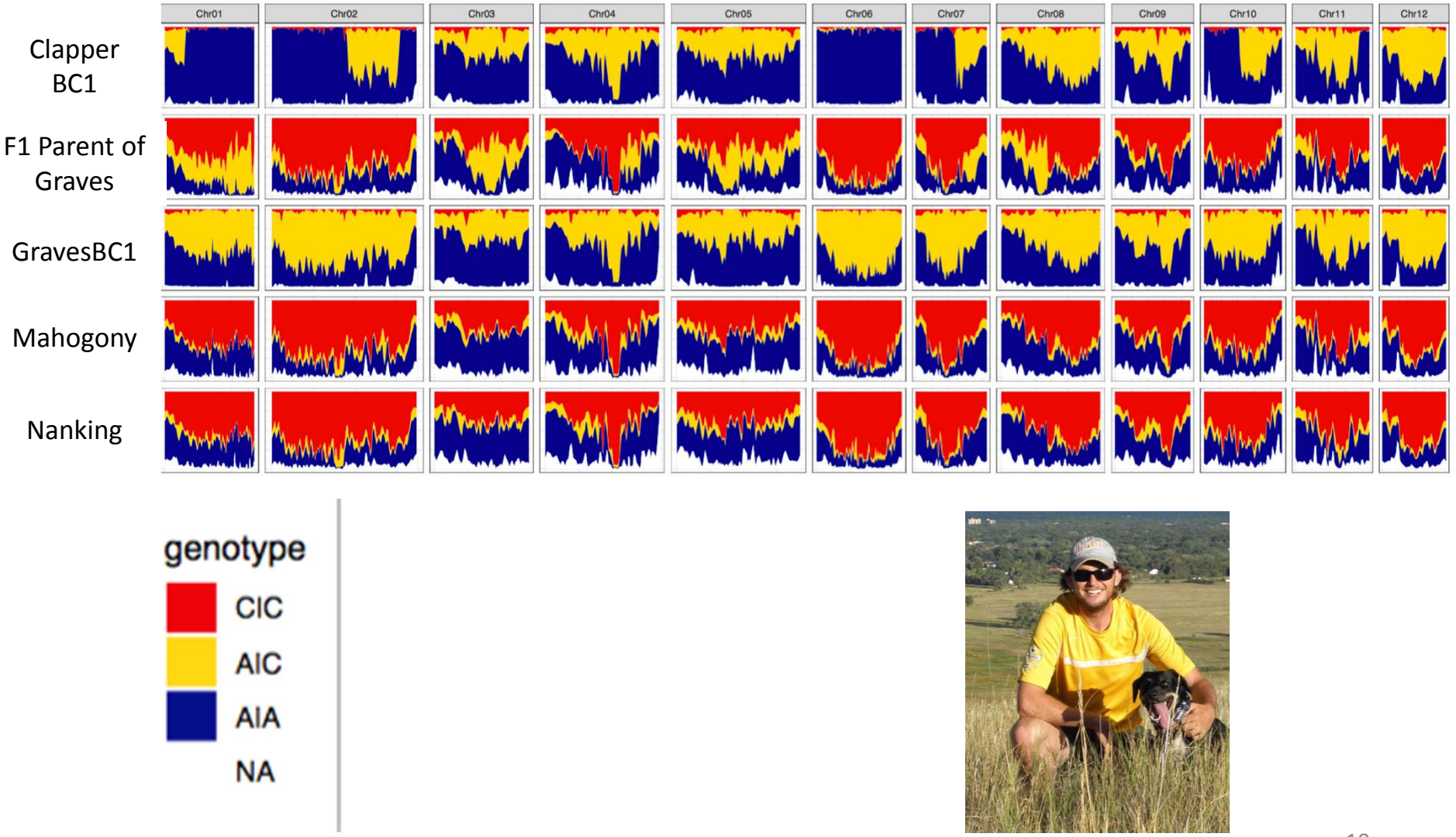
Clapper BC1

F1 Parent of Graves

"We were able to find the F1 parent of Graves (CAES S Lot R2T8) but the main trunk had died. There were new shoots springing up from the roots so we clipped some of the side branches to send assuming it is genetic material from the same tree, but I guess the results will determine if it is actually from the parent tree."

Jack Swatt

Genetics of the source trees



John Lovell

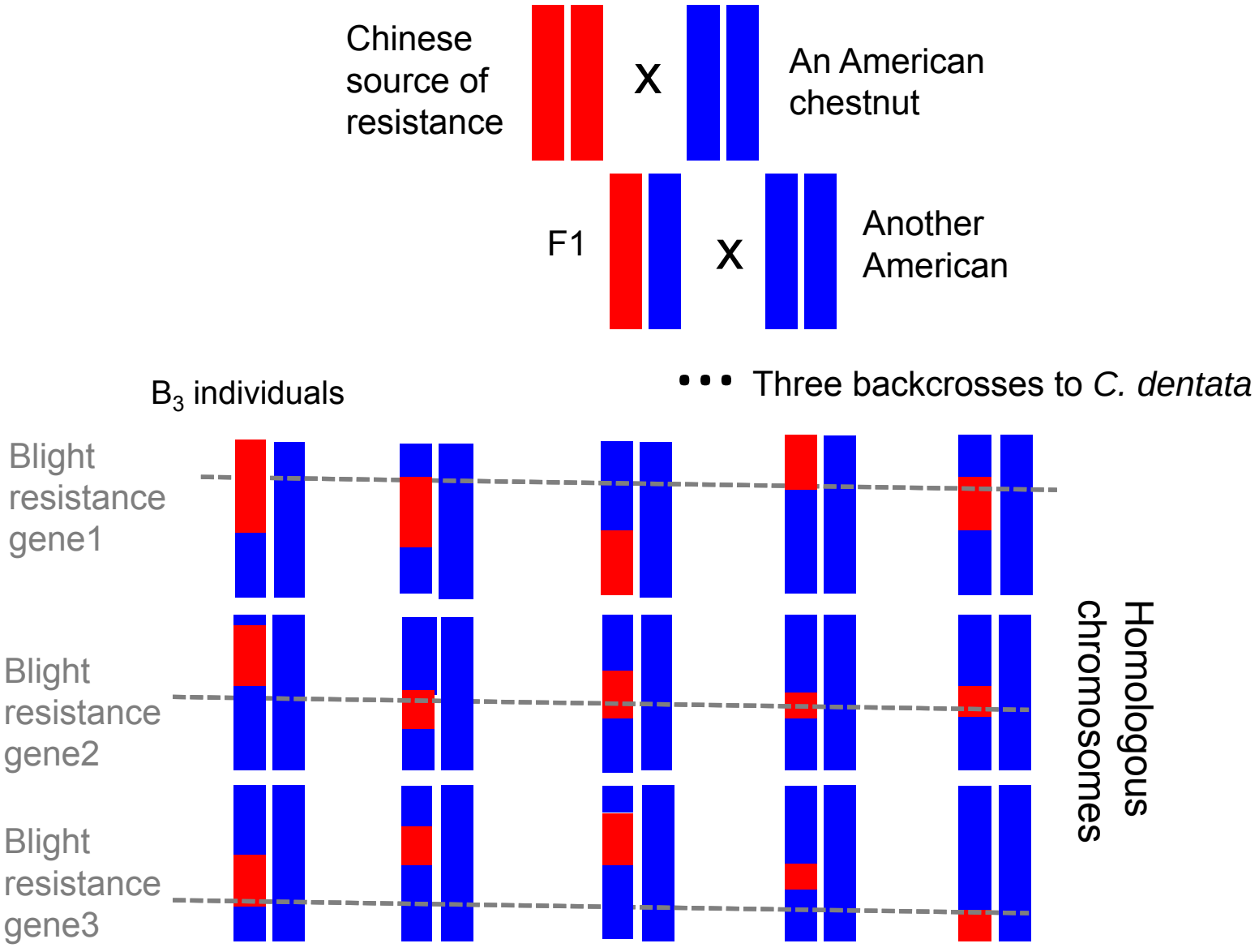
How was analysis done?

1. Mapped reads and called SNP variation (44.3M raw SNPS)
2. Processes SNPs via binomial test of counts
   - likely American genotype
   - likely American/Chinese heterozygote
   - likely Chinese genotype
3. Keep only markers that are very likely to have 100% genotyping accuracy (5,044,254 markers)
4. Reduce identical markers to informative set (47,508 markers)

Haplotype calling

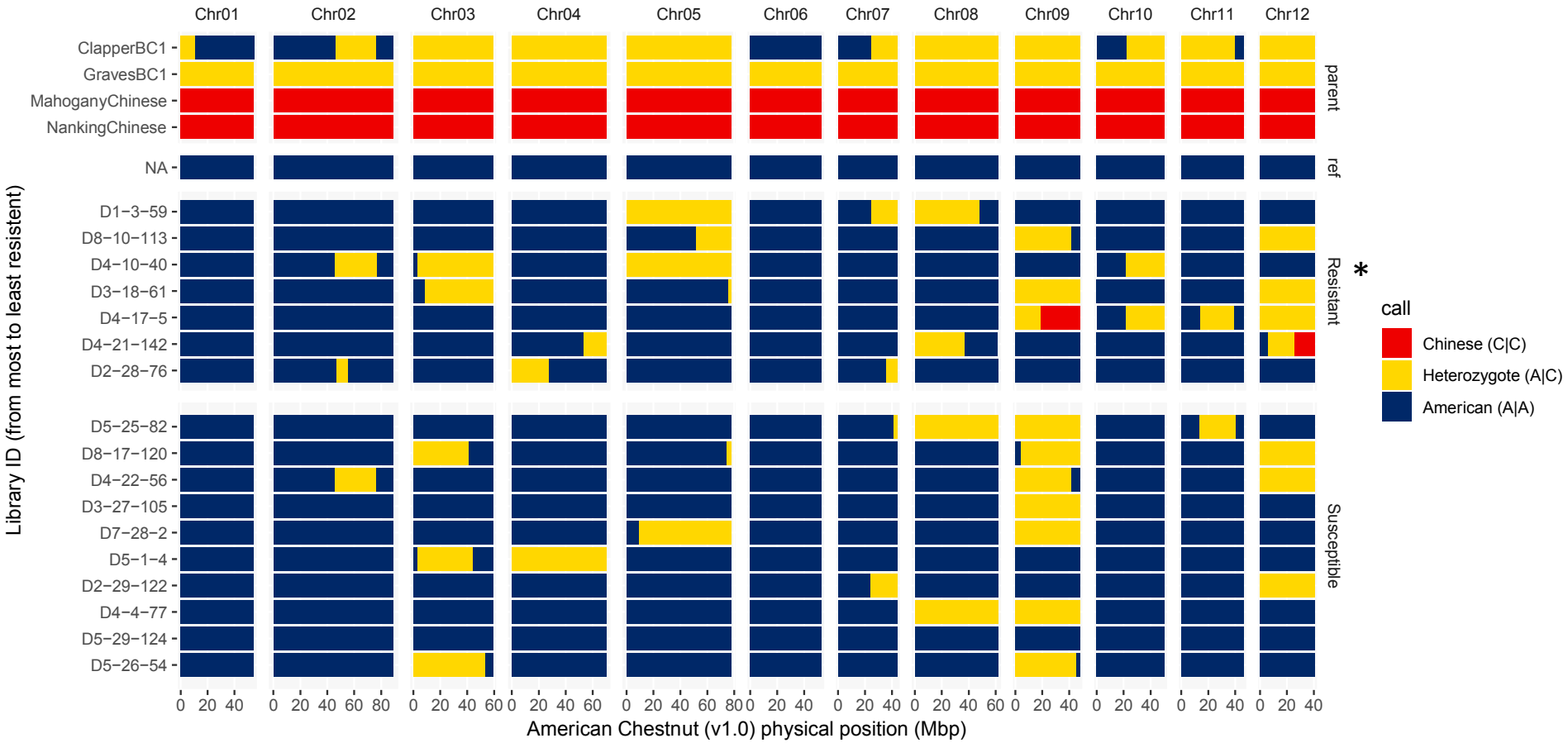
1. Split the genome into non-overlapping 1000-marker bins
2. Train simple neural net classifier and predict the genotypes
3. Build haplotype blocks
4. Apply random forest machine learning probabilistic genotype for assignment
5. Calculating mean probabilities of assignment across overlapping marker bins
6. Compile haplotype block break points where logistic regression of probability of assignment crosses 0.5

Progeny tested B3F3s



Jared Westbrook

Haplotype analysis



* Minas 3 resistant F1 hybrids

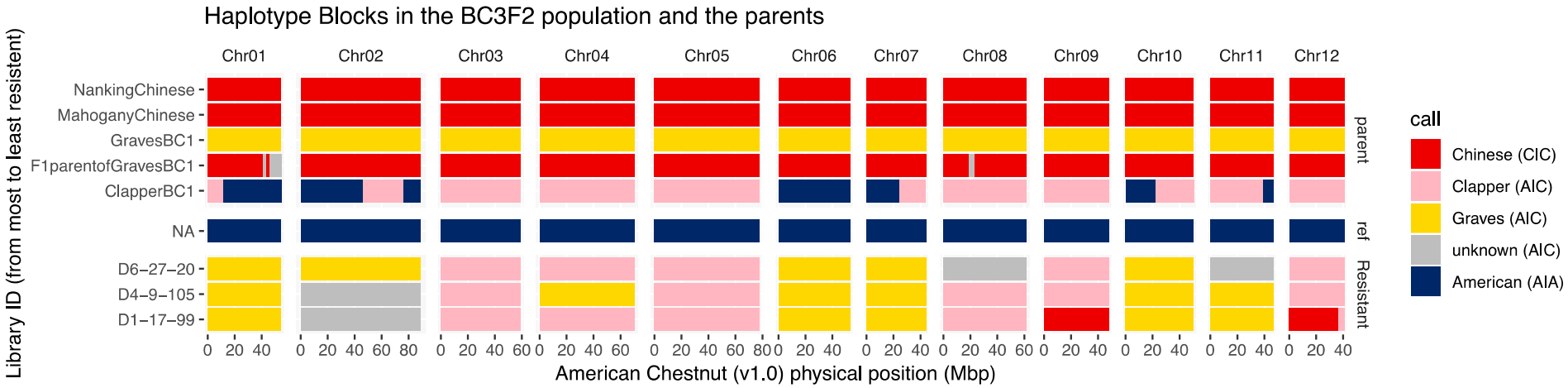
Retained Chinese in BC3F2s

	% Chinese	
D1-3-59	**10.5**	**R**
D8-10-113	**7.8**	**R**
D4-10-40	**14.0**	**R**
D3-18-61	**10.3**	**R**
D4-17-5	**12.4**	**R**
D4-21-142	**7.6**	**R**
D2-28-76	**3.4**	**R**
D5-25-82	**10.0**	**S**
D8-17-120	**9.3**	**S**
D4-22-56	**8.1**	**S**
D3-27-105	**3.5**	**S**
D7-28-2	**8.5**	**S**
D5-1-4	**8.1**	**S**
D2-29-122	**4.4**	**S**
D4-4-77	**7.9**	**S**
D5-29-124	**0.0**	**S**
D5-26-54	**7.1**	**S**

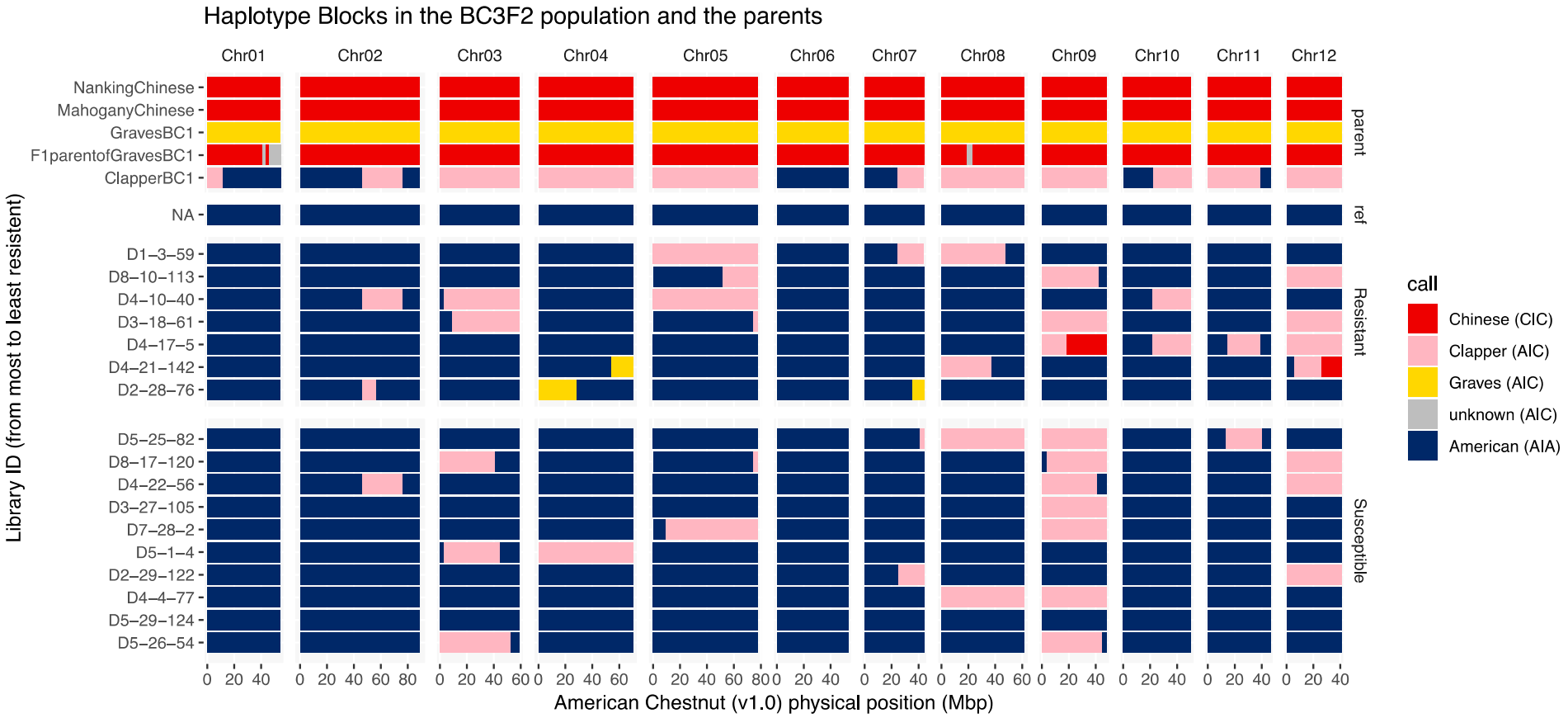
Resistant = 9.4%

Susceptible = 7.4%

Which parents contributed the F1 pollution in the BC3F2s?



Which parents contributed to the BC3F2s?



Main takeaways

- Resistant trees contain between 3 to 14% Chinese
- There is small amount of graves in the clapper orchard and 3/10 resistant are Chinese hybrids (not surprising as these will be highly resistant)
- Sorry Jack, the F1 parent of graves is a mostly Chinese tree and The Graves BC1 is likely an F1

- A sample of 7 is not enough to identify likely regions of resistance
  - High fidelity tracking of Chinese haplotypes
  - Hints, but not clear regions associated with resistance in B3F2s

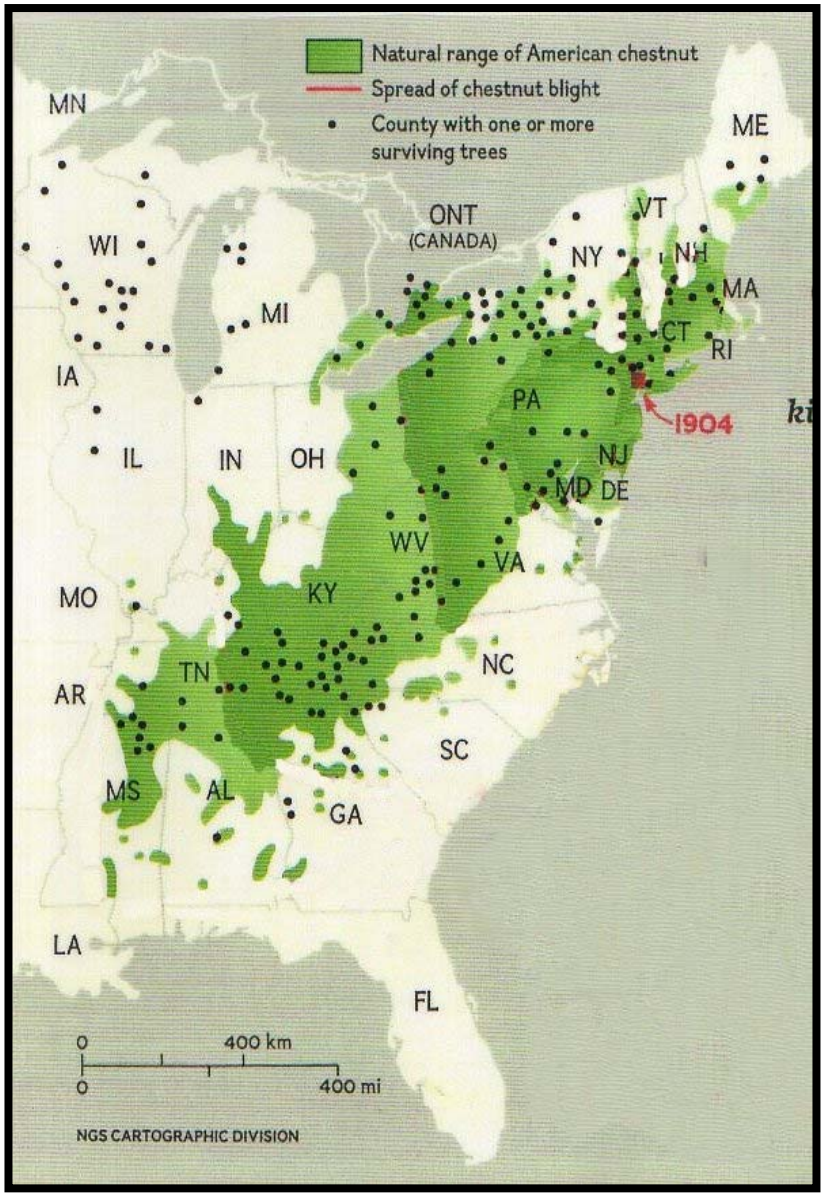
Where do we go from here for B3F2s?

1. Jared can now analyze ~2,000 light sampled B3F2s to identify roughly the Chinese contributions
2. We can reseq additional resistant B3F2s to get more complete data from MeadowBrook
3. We can survey a portion of chapter B3F2s to evaluate Chinese retention at high coverage ($300 per tree) and expand this to more at lower pass sequencing (~$100 per tree)
4. If we can validate the parent Chinese trees contributing resistance, reference genomes would be useful to identify unique Chinese material

- Some of these proposals are currently funding limited

Where do we go from here understanding molecular mechanism of infection?

- We now have a genome but need high quality annotation!
- We can now do real transcriptomics of infection to identify genes involved in plant response
- We can deep survey at $300 per wild tree American Chestnut diversity for accurate demographics and adaptation analysis



- Please contact me at jschmutz@hudsonapha.org for access to genome sequence under Ft. Lauderdale
- Once the annotation is finalized, it will be publicly available (hopefully at Phytozome, comparative plant genomics portal)

CROPS
CONFERENCE

June 3 - 6, 2018

HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY • HUNTSVILLE, AL USA

Improving agriculture through genomics

www.CROPSconference.org

PRESENTED BY

HudsonAlpha Institute for Biotechnology

Institute of Plant Breeding, Genetics and Genomics
College of Agricultural & Environmental Sciences
University of Georgia

The University of Georgia

Keynote Speakers:

Loren Rieseberg, Washington State University, WA, USA
Patrick Schnable, Iowa State University, IA, USA

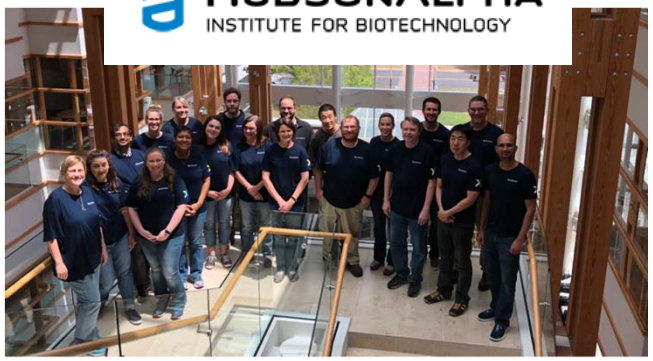
Conference Speakers:

Jean-Philippe Vielle-Calzada, LangeBio Cinvestav, Guanajuato, México
Andrea Eveland, Danforth Plant Center, MO USA
Maria Salas-Fernandez, Iowa State University, IA USA
Tom Juenger, University of Texas, Austin, TX USA
Benjamin Kilian, Global Crop Diversity Trust, Bonn Germany
Amy Lezzoni, Michigan State University, East Lansing, MI USA
Alice MacQueen, University of Texas, Austin, TX USA
Allison Miller, Saint Louis University, St. Louis, MO USA
Sergey Nuzhdin, University of Southern California, CA USA
Philipp Simon, ARS USDA, Madison, WI USA
Arti Singh, Iowa State University, IA USA
Mark Sorrells, Cornell University, NY USA
Nils Stein, Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany
Mark Tester, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
Gerald Tuskan, Oak Ridge National Laboratory, TN USA
Esther van der Knaap, University of Georgia, GA USA

Contributors

- Tissue and materials
  - Andy Newhouse & Bill Powell (SUNY-ESF)
  - Jack Swatt (CT chapter president)
  - Chris Saski & Xiaoxia Xia (Clemson)
  - Dave Kudrna (Arizona)
  - Qian Zhang & Jason Holliday (VT)
  - Warren Chatwin & Jeanne Romero-Severson (Notre Dame)
- Mapping
  - Tetyana Zhebentyayev (Clemson)
- Sequencing
  - Jane Grimwood, Lori Beth Boston, Melissa Williams, Ada Stewart, Jenell Webber (HudsonAlpha)
- Assembly and Analysis
  - Jerry Jenkins, Chris Plott, John Lovell (HudsonAlpha)
- TACF
  - Jared Westbrook
  - Lisa Thomson

HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY



U.S. DEPARTMENT OF ENERGY | Office of Science

National Science Foundation
Directorate for Biological Sciences (BIO)

THE AMERICAN CHESTNUT FOUNDATION®

Colcom Foundation