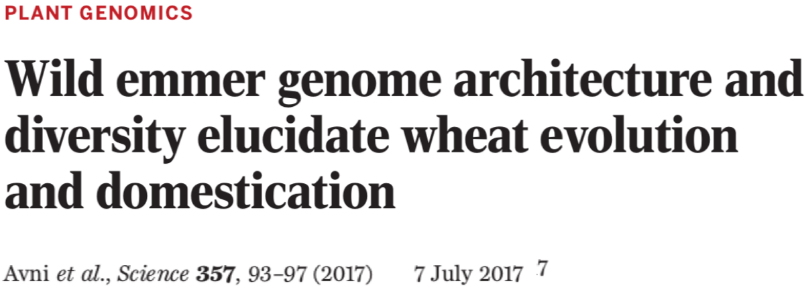
No Magic Involved: Chromosome-Scale Sequence Assembly of Wheat Genomes with Open-Source Tools

Martin Mascher

IPK Gatersleben

January 12th, 2019

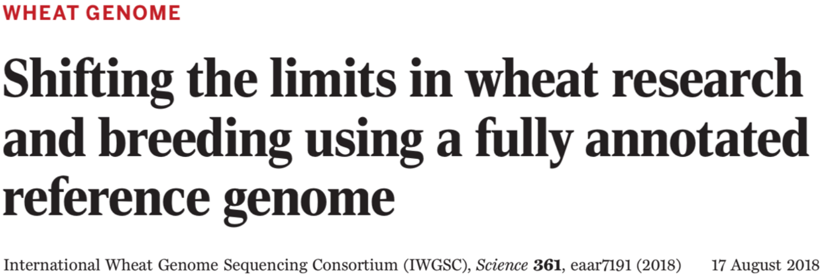
Chromosome-scale assembly in wheat



PLANT GENOMICS

Wild emmer genome architecture and diversity elucidate wheat evolution and domestication

Avni *et al.*, *Science* **357**, 93–97 (2017)   7 July 2017 [7]

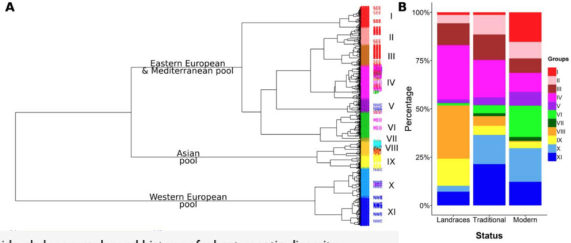


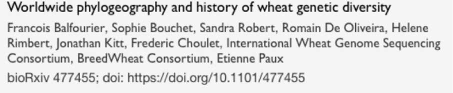
WHEAT GENOME

Shifting the limits in wheat research and breeding using a fully annotated reference genome

International Wheat Genome Sequencing Consortium (IWGSC), *Science* **361**, eaar7191 (2018)   17 August 2018

Why more wheat genomes?

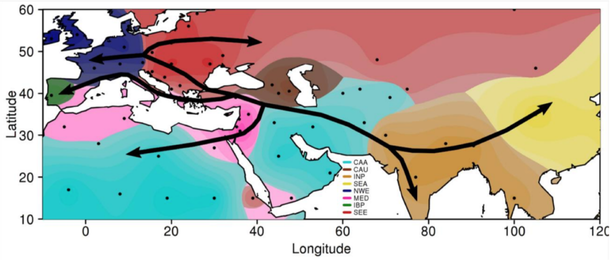




Worldwide phylogeography and history of wheat genetic diversity

Francois Balfourier, Sophie Bouchet, Sandra Robert, Romain De Oliveira, Helene Rimbert, Jonathan Kitt, Frederic Choulet, International Wheat Genome Sequencing Consortium, BreedWheat Consortium, Etienne Paux

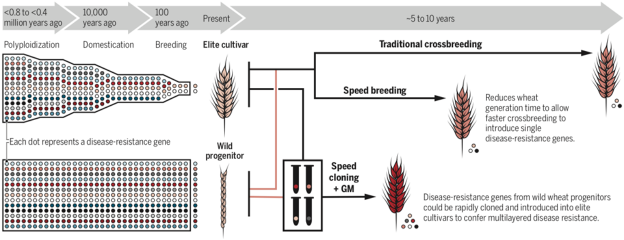
bioRxiv 477455; doi: https://doi.org/10.1101/477455



Why more wheat genomes?

Introducing disease resistance to wheat

Genetic diversity for disease resistance in wheat has been lost through bottlenecks imposed by polyploidization, domestication, and breeding. Resistance genes from wild relatives can be incorporated into elite cultivars by crossbreeding, which is accelerated by speed breeding, and speed cloning with genetic modification (GM).



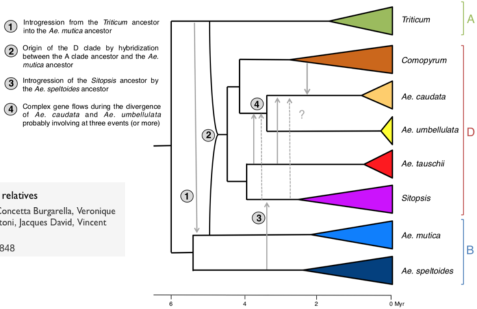
Reduces wheat generation time to allow faster crossbreeding to introduce single disease-resistance genes.

Disease-resistance genes from wild wheat progenitors could be rapidly cloned and introduced into elite cultivars to confer multilayered disease resistance.

By **Brande B. H. Wulff**[1] *and* **Kanwarpal S. Dhugga**[2]

3 AUGUST 2018 • VOL 361 ISSUE 6401 **451**

SCIENCE sciencemag.org



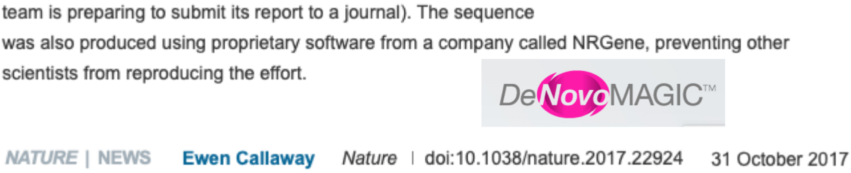
1. Introgression from the *Triticum* ancestor into the *Ae. mutica* ancestor
2. Origin of the D clade by hybridization between the A clade ancestor and the *Ae. mutica* ancestor
3. Introgression of the *Sitopsis* ancestor by the *Ae. speltoides* ancestor
4. Complex gene flows during the divergence of *Ae. caudata* and *Ae. umbellulata* probably involving at three events (or more)

Pervasive hybridizations in the history of wheat relatives

Sylvain Glemin, Celine Scornavacca, Jacques Dainat, Concetta Burgarella, Veronique Viader, Morgane Ardisson, Gautier Sarah, Sylvain Santoni, Jacques David, Vincent Ranwez

bioRxiv 300848; doi: https://doi.org/10.1101/300848

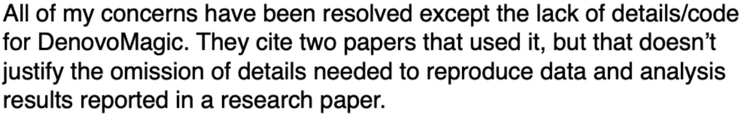
Why an open source pipeline?



team is preparing to submit its report to a journal). The sequence was also produced using proprietary software from a company called NRGene, preventing other scientists from reproducing the effort.

DeNovoMAGIC™

NATURE | NEWS Ewen Callaway *Nature* | doi:10.1038/nature.2017.22924 31 October 2017

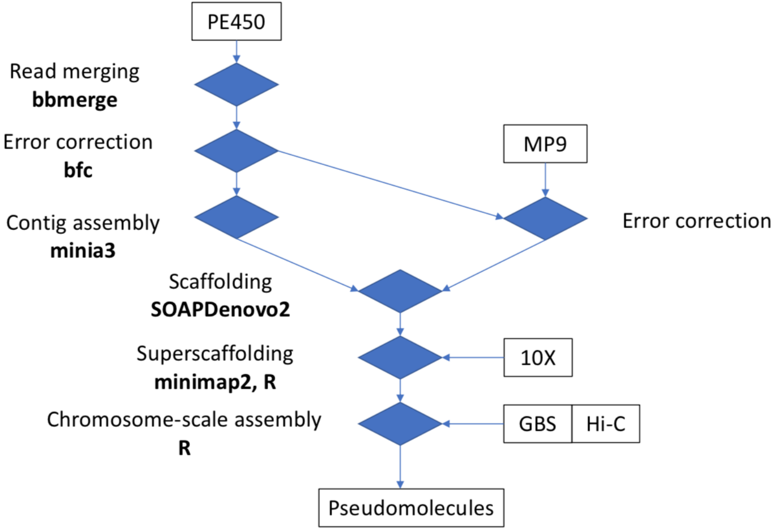


All of my concerns have been resolved except the lack of details/code for DenovoMagic. They cite two papers that used it, but that doesn't justify the omission of details needed to reproduce data and analysis results reported in a research paper.

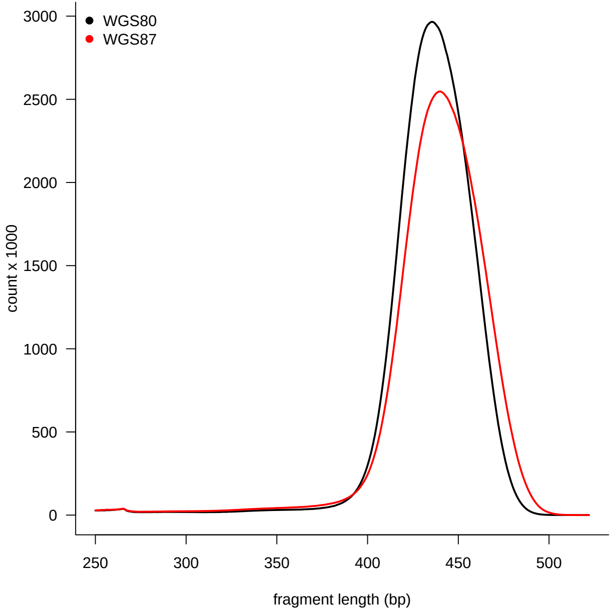
- Reviewer #3

Pipeline for chromosome-scale assembly in the Triticeae

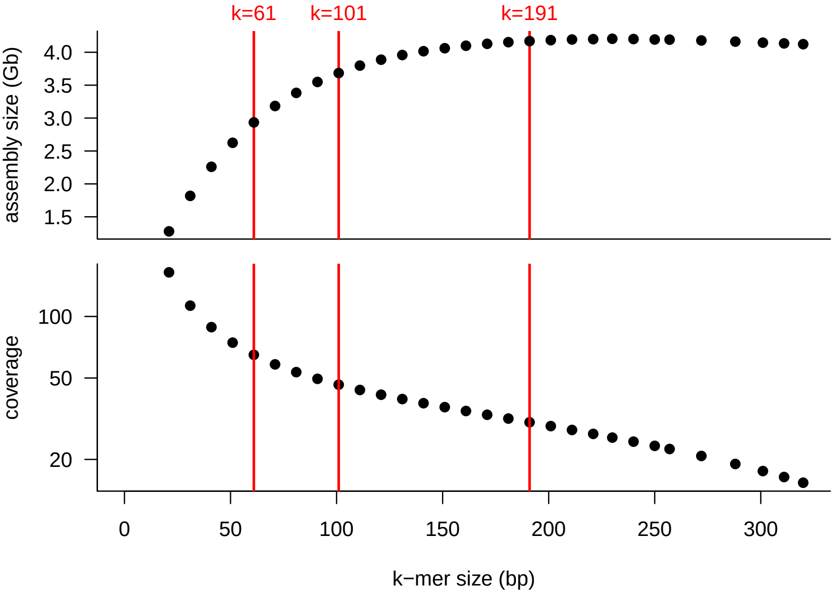
https://triticeae.bitbucket.io



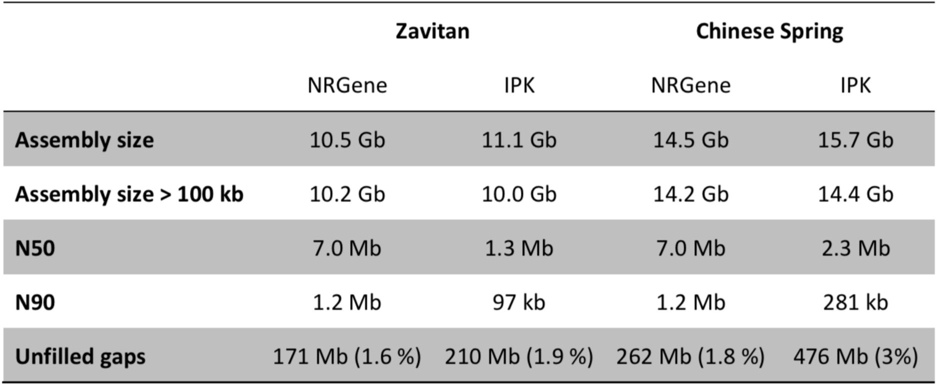
Importance of “long short reads”



Importance of “long short reads”

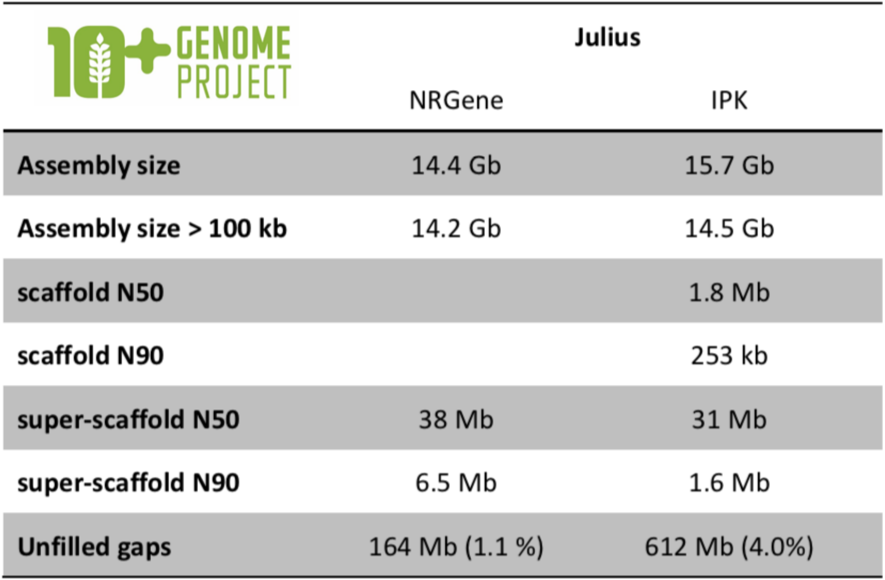


Assembly results with paired-end and mate-pair data



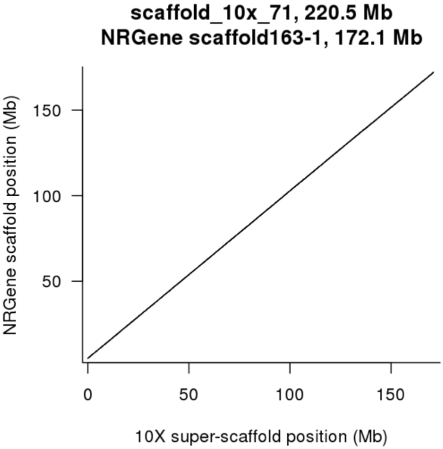
	Zavitan		Chinese Spring	
	NRGene	IPK	NRGene	IPK
Assembly size	10.5 Gb	11.1 Gb	14.5 Gb	15.7 Gb
Assembly size > 100 kb	10.2 Gb	10.0 Gb	14.2 Gb	14.4 Gb
N50	7.0 Mb	1.3 Mb	7.0 Mb	2.3 Mb
N90	1.2 Mb	97 kb	1.2 Mb	281 kb
Unfilled gaps	171 Mb (1.6 %)	210 Mb (1.9 %)	262 Mb (1.8 %)	476 Mb (3%)

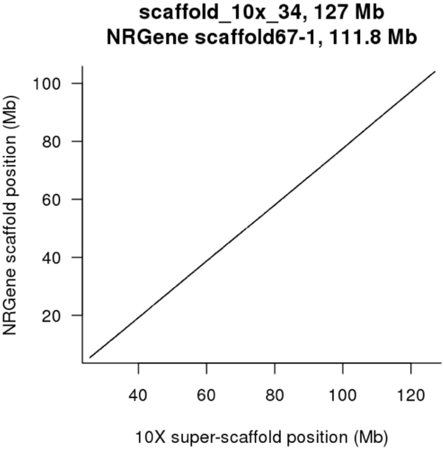
10X Chromium linked-reads improve assembly contiguity



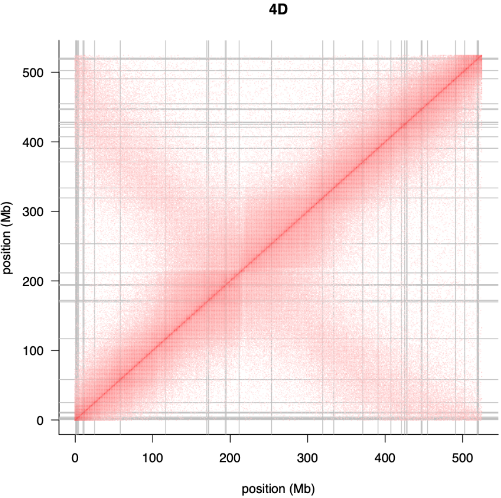
10+ GENOME PROJECT	Julius	
	NRGene	IPK
Assembly size	14.4 Gb	15.7 Gb
Assembly size > 100 kb	14.2 Gb	14.5 Gb
scaffold N50		1.8 Mb
scaffold N90		253 kb
super-scaffold N50	38 Mb	31 Mb
super-scaffold N90	6.5 Mb	1.6 Mb
Unfilled gaps	164 Mb (1.1 %)	612 Mb (4.0%)

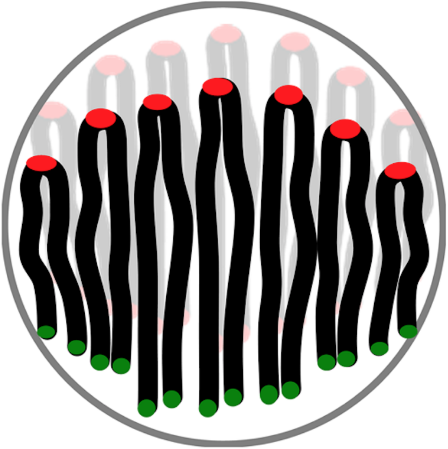
Collinearity to NRGene assembly



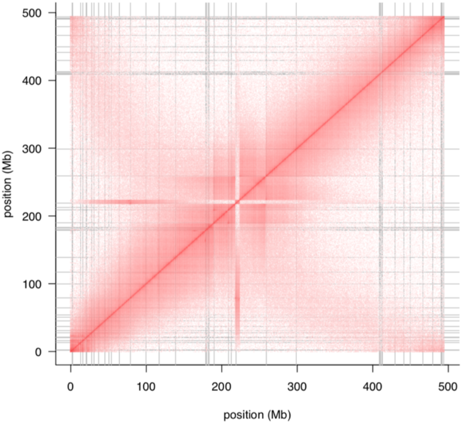


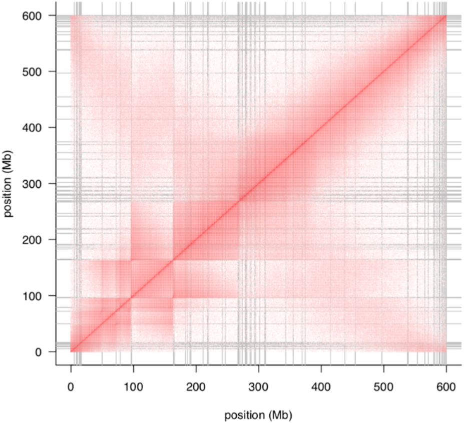
Hi-C for pseudomolecule construction



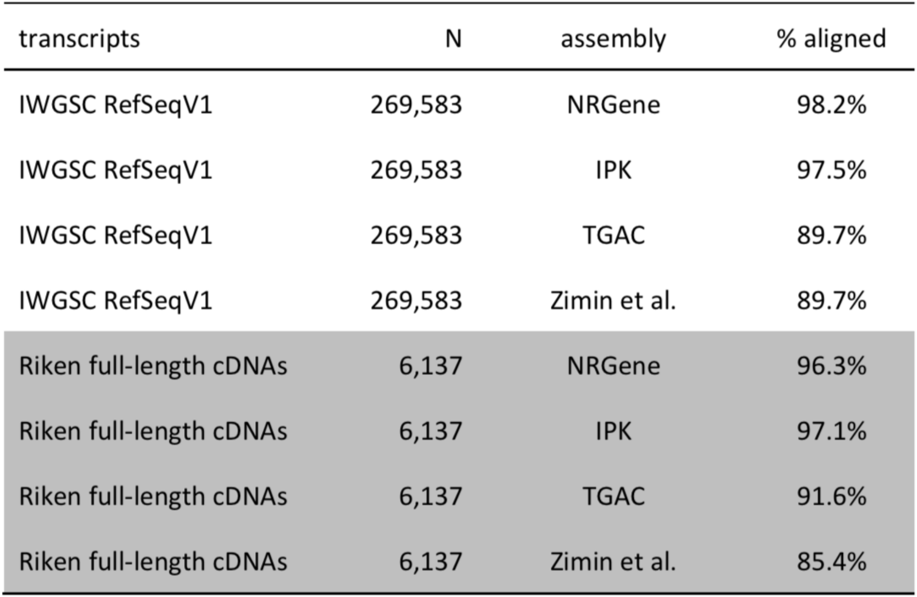


Hi-C for assembly validation and correction





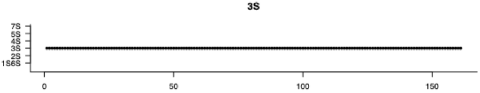
Gene space representation

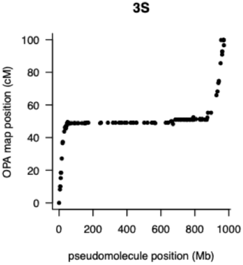


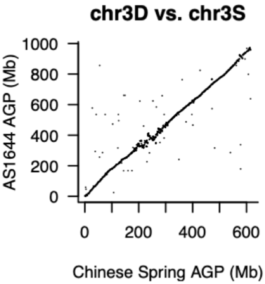
transcripts	N	assembly	% aligned
IWGSC RefSeqV1	269,583	NRGene	98.2%
IWGSC RefSeqV1	269,583	IPK	97.5%
IWGSC RefSeqV1	269,583	TGAC	89.7%
IWGSC RefSeqV1	269,583	Zimin et al.	89.7%
Riken full-length cDNAs	6,137	NRGene	96.3%
Riken full-length cDNAs	6,137	IPK	97.1%
Riken full-length cDNAs	6,137	TGAC	91.6%
Riken full-length cDNAs	6,137	Zimin et al.	85.4%

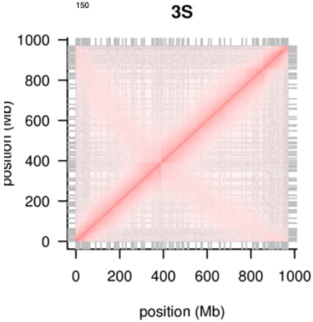
A reference genome for *Aegilops sharonensis*

assembly size	6.7 Gb
scaffold N50	12.3 Mb
scaffold N90	1.1 Mb
unfilled gaps	886 Mb (13 %)
pseudomolecule size	6.3 Gb (94 %)









Cost and timelines

- Cost for data collection per diploid genome: USD 50,000
- Cost for IT infrastructure: USD 50,000 (32 cores, 1 TB RAM, 3 TB disk space)
- Computationally skilled postdoc: USD 70,000 per year
- Time for computation: one month – six weeks per genomes
    - *Ae. sharonensis*: first dataset arrived on Oct 2, pseudomolecules finished on Dec 20

More on Triticeae genome assembly

- Updated barley reference genome and prospects of pan-genomics in barley

  Cécile Monat, Triticeae I, today, 11:20, Town & Country
  Martin Mascher, Sequencing Complex Genomes, tomorrow, 16:00, Golden Ballroom

- Chromosome-scale assembly of rye

  Tim Wallace, Triticeae I, today, 10:30, Town & Country

- Wheat 10+ Genomes Project

  Curtis Pozniak, Triticeae I, today, 10:55, Town & Country

Acknowledgements

Bioinformatics at IPK

IPK GATERSLEBEN

Uwe Scholz
Heiko Miehe
Jens Bauernfeind
Thomas Münch
Sebastian Fricke

Leibniz Association

Barley and wheat genomics at IPK

Nils Stein
Cécile Monat
Sudharsan Padmarasu
Axel Himmelbach
Ines Walde
Manuela Knauft

Federal Ministry of Education and Research

PLANT 2030

Federal Ministry of Food and Agriculture

10X data

10+ GENOME PROJECT

Curtis Pozniak
Jennifer Ens

University of Saskatchewan

Aegilops sharonensis genomics

Brande Wulff
Guotai Yu
Burkhard Steuernagel
Jonathan Jones
Raz Avni
Hanan Sela
Anna Minz
Amir Sharon
Istvan Molnar
Jaroslav Dolezel

John Innes Centre
Unlocking Nature's Diversity

2BLADES FOUNDATION

TEL AVIV UNIVERSITY

Institute of Experimental Botany of the AS CR, v. v. i.

Come visit IPK!

GRC2019

Gatersleben Research Conference - Applied Bioinformatics for Crops

18 - 20 March, 2019

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben // DE

Follow **@GRC2019**

https://meetings.ipk-gatersleben.de/grc2019-abc/

- Image-Based Data Analyses & Data Visualization
- Distributed Computing, Tools and Infrastructures
- Systems Biology & Modeling
- Biodiversity & Information Systems
- Breeding Informatics

Anne-Françoise Adam-Blondon
INRA, Versailles- Grignon, FR

Andrea Bräutigam
Computational Biology, University Bielefeld, DE

Oliver Ebenhöh
Heinrich-Heine-University, Düsseldorf, DE

Richard Finkers
Wageningen University & Research, NL

Malia Gehan
Danforth Center, St. Louis, USA

Björn Grüning
Bioinformatics, University Freiburg, DE

Barend Mons
GO FAIR International Support and Coordination Office, Leiden, NL

Zoran Nikoloski
MPI of Molecular Plant Physiology, Golm, DE

Kelly Robbins
Cornell University, USA

Sotirios Tsaftaris
University of Edinburgh, UK

Input data

1. PE450, PCR free, 2 × 250 bp, $\geq$70x coverage
2. (PE800, PCR free, 2 × 150 bp, $\geq$30x coverage)
3. MP (2-4kb), (5-7 kb), 8-10 kb, Nextera, 2 × 150 bp, each $\geq$30x coverage
4. Two (one) 10X Chromium libraries sequenced to 30x read coverage (2 × 150 bp reads)
5. 1 Hi-C library, 200 M – 400 M read pairs
6. Low density genetic map (GBS / SNP chip)