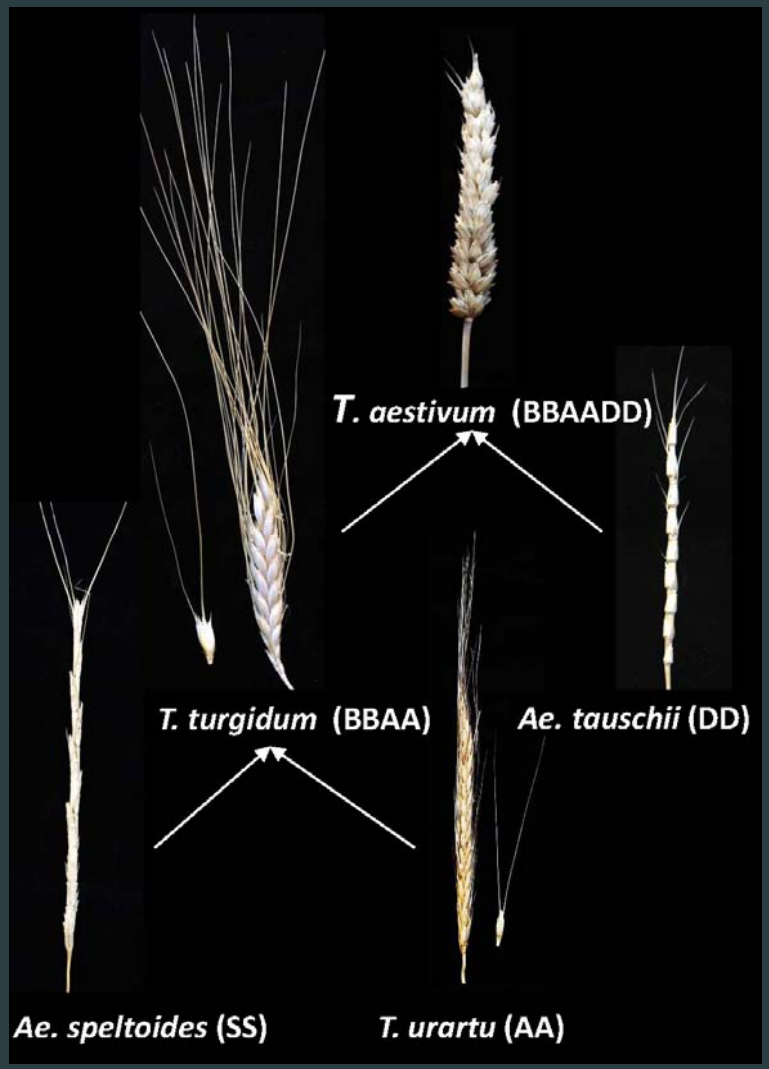
HYBRID ASSEMBLY OF THE ANCESTRAL WHEAT Ae. tauschii 4.25Gb GENOME

Aleksey Zimin*, Guillaume Marcais, Daniela Puiu, Steven Salzberg and James Yorke, et al.

University of Maryland, Johns Hopkins University

Aegilops tauschii is one of the progenitors of common wheat



- *Aegilops tauschii* **sequence provides a reference for study of polyploid genome evolution by facilitating comparison of the wheat D-genome and** *Ae. tauschii* **genomic sequences**

Dominant Sequencing technologies

- Illumina
  - ☺ Cheap: as low as $5K for de novo mammalian genome
  - ☺ Accurate – 1-2% error rate
  - ☹ Only 150 to 350 bp reads, almost all paired
- PacBio
  - ☺ Inexpensive: $100k+ for a mammalian genome
  - ☹ High error rate ~15%
  - ☹ Random sequence insertions, chimeric reads
  - ☺ We get on average ~10000 bp reads

Data for A. tauschii assembly

WGS Illumina data and Pacbio

- Data used for the hybrid assembly:
  - 62x WGS coverage by 2x250bp Illumina Paired end reads
    - 450bp fragment size
  - 35x WGS P6C4 Pacbio reads, ~10Kb N50 size

Definitions

Read, super-read, mega-read

- *Read*
  - a fragment of genome sequence read by a sequencing machine
  - 100-250bp long for Illumina sequencing
- *Super-read*
  - a synthetic read produced by extending Illumina read(s) by using k-mer graph;
  - typically several reads extend to the same super-read
  - 400-2000 bp average length
- *Mega-read*
  - a synthetic read produced by merging super-reads with exact sequence overlaps guided by a template long read (Pacbio read)
  - 5000-8000 bp average length

Advantages of our hybrid approach

- We aim at producing long near-perfect "mega-reads" from the Pacbio SMRT reads
- We pre-process the Illumina reads to form Super-reads:
  - *Much longer – average >500bp*
  - *2-3x overlapping genome coverage*
  - *Exact k-overlaps (usually min 69+bp) known*
- Require only ~10-20x coverage of Pacbio reads and 50x-100x Illumina 150-250bp reads
- Can preserve and resolve haplotype information based on the accurate Illumina and long Pacbio reads
- Relatively inexpensive computationally (less than 1 month on 48-core 512Gb computer for 3Gb genome).

Preliminary Assembly results

	The First Preliminary Result	My Goal
Assembled sequence	**3.98Gbp**	**4.2Gbp**
N50 contig size	**242Kbp**	**~400Kbp**
N50 scaffold size	**252Kbp**	

In every project we always run assembly more than once to achieve the best result

Overview of the MaSuRCA mega-reads technique

Efficient error correction of the PacBio reads

- We use every single PacBio read as a *template*
- Find the best tiling of each Pacbio read with the super reads to produce accurate *mega-reads*
- Assemble the mega-reads + the other data with Celera Assembler 8.3
- Optional: Post-process to resolve (unzip) haplotypes

MaSuRCA mega-reads 3.2.x

- The mega-reads technique was developed by our group at the University of Maryland and added to the MaSurca assembler. Available now for beta-testing from us http://www.genome.umd.edu, by request.
- Design aimed at large genomes
- The latest version can assemble mammalian genome in ~1 month on one 48-core computer
- We are using it to assemble 22Gbp genome of Loblolly pine, 2.8Gbp cow genome, and 1Gb Manakin genome (in collaboration with Smithsonian Institution)

Mega-read sizes

Three hybrid data sets

Data set	Pacbio N50 sub-read size (Kb)	Mega-read N50 read size (Kb)
Ae. tauschii, 35x	**11.5**	**9.2**
S.cerevisiae (yeast), 20x	**11.6**	**9.3**
Drosophila pseudoobscura, 20x	**8.4**	**6.9**
Manacus vitellinus, 10x	**18.0**	**12.3**

- Total size of Pacbio reads was used for N50
- Mega-reads N50 length is up to 80% of the Pacbio N50 length

MaSuRCA performance on PacBio Hybrid data

Results on S. cerevisiae W303 data set

Assembler	Input data	Aligned NGA50 Contig Kb)	Structural mis-assemblies
REFERENCE	40x 454 Sequencing	924	N/A
HGAP	237x PacBio	809	4
ECTools	20x PacBio + 100x MiSeq	401	9
PacBioToCA	20x PacBio + 100x MiSeq	320	15
MaSuRCA Mega-reads	20x PacBio + 100x MiSeq	804	3

MaSuRCA performance on PacBio Hybrid data

Faux haplotype experiment – in development

- We created faux haplotype data set of 100x Illumina + 20x Pacbio for the yeast (S. cerevisiae)
  - Create modified version of the finished sequence by introducing SNPs
  - Split Illumina and Pacbio reads into 2 groups and introduce SNPs into one group based on alignment to the finished sequence (where the matches are)
  - Assemble, and then separate (unzip) the haplotypes
  - Use for development and validation

Assembler	Haplotype difference rate	Aligned NGA50 Contig Kb)
REFERENCE	N/A	924
MaSuRCA Mega-reads	N/A	793
MaSuRCA Mega-reads Haplotype resolved	1%	268
MaSuRCA Mega-reads Haplotype resolved	0.1%	250

Super-reads

A key idea used in the assembly

- Based on the observation that most of the sequence in genomes is *locally* unique – branches are relatively rare
- Consider 10-mers (we use much longer k of course):

AGCTGACTGACTGGTAACAA
AGCTGACTGA
GCTGACTGAC

- **The idea is to make reads longer instead of breaking them into k-mers.**

Super reads

Extending a read to become a super-read

- Consider a read – can its ends be extended uniquely?

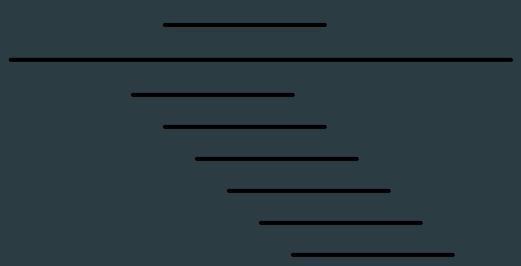
ACTGACCAGATGACCATGACAGATACATGGT

extend 5 GACTGACTGG **CTGACTGGTA 10 stop**

CTGACTGGTC 2

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:

read R extended to super read S (red)



Many *other reads extend to the same S as well*

Super reads

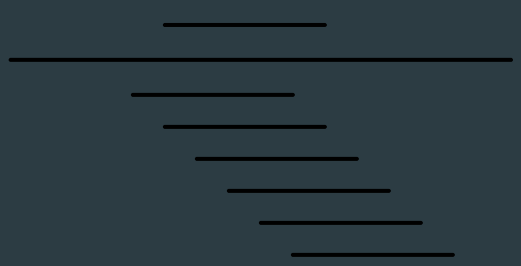
Extending a read to become a super-read

- Consider a read – can its ends be extended uniquely?

```
          GACTGACCAGATGACCATGACAGATACATGGT
extend  5 GACTGACTGG                      CTGACTGGTA 10 stop
                                          CTGACTGGTC  2
```

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:

read R extended to super read S (red)



Many *other reads extend to the same S as well*

Super reads

We can keep Extending on the left

- Consider a read

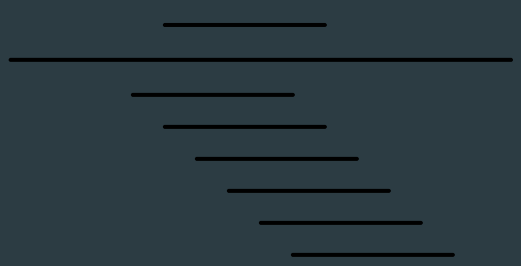
CGACTGACCAGATGACCATGACAGATACATGGT *stop*

extend 5 GACTGACTGG **CTGACTGGTA 10 stop**

extend 3 CGACTGACTG **CTGACTGGTC 2**

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:

read R extended to super read S (red)



Many *other reads extend to the same S as well*

Super reads

We can keep Extending on the left

- Consider a read

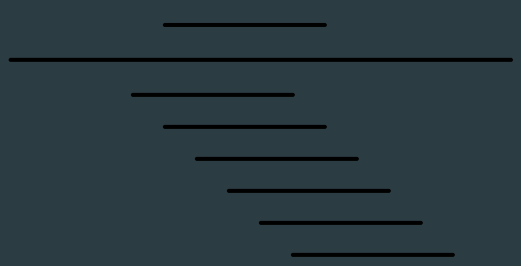
CGACTGACCAGATGACCATGACAGATACATGGT *stop*

extend 5 GACTGACTGG **CTGACTGGTA 10 stop**

extend 3 CGACTGACTG **CTGACTGGTC 2**

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:

read R extended to super read S (red)



Many *other reads extend to the same S as well*

Super reads

Extend, stopping at the next branch (or where there is no data)

- Consider a read

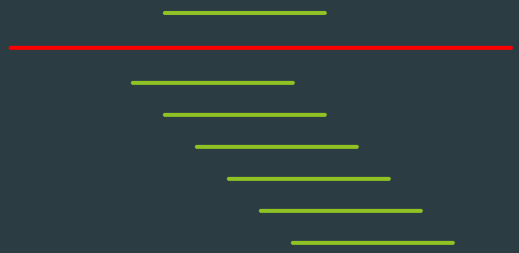
CGACTGACCAGATGACCATGACAGATACATGGT *stop*

extend 5 GACTGACTGG **CTGACTGGTA 10 stop**

extend 3 CGACTGACTG **CTGACTGGTC 2**

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:

read R extended to super read S (red)



Many *other reads extend to the same S as well*

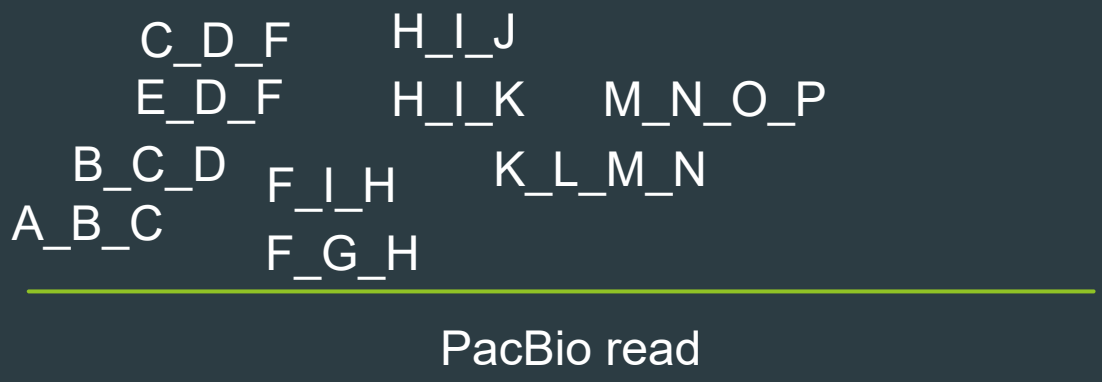
Overview of the mega-reads technique

Efficient error correction of the PacBio reads

- We use every single PacBio read as a *template*
  - Map (approximately) super-reads to PacBio reads
  - Exact overlaps between the super reads confirmed by mapping = *proper* overlaps
  - Mega-read is a *properly* overlapping contig of super-reads that matches the PacBio read best
  - If more than one mega-read tiling the pacbio read then join (or not) with pacbio sequence
- Assemble the corrected reads + the other data with Celera Assembler 8.3
- Post-process to resolve (unzip) haplotypes

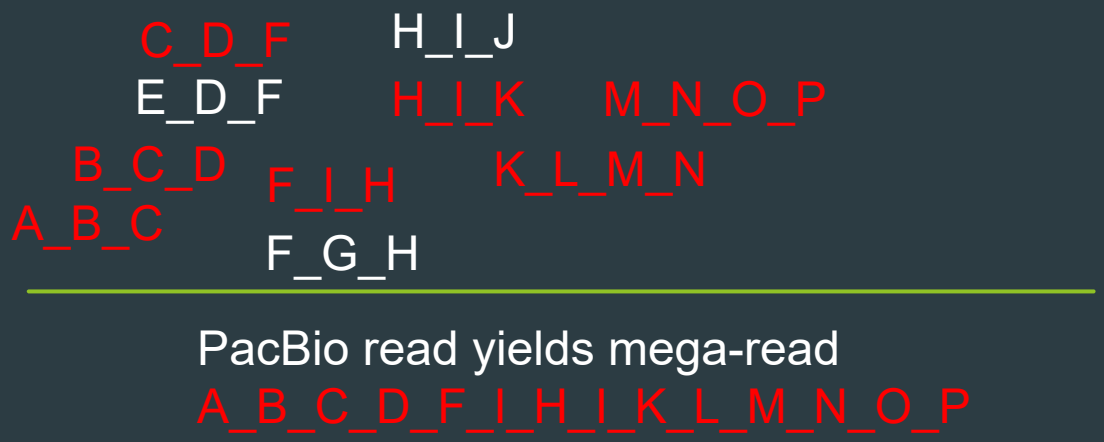
Overview of the mega-reads technique

Matching super reads, letters represent segments of k-mer graph



Overview of the mega-reads technique

Matching super reads, letters represent segments of k-mer graph, path indicated in red



More than one mega-read tiling a Pacbio read

- Some Pacbio reads yield more than one covering mega-read with a gap in corrected coverage
  - Insertions in Pacbio reads
  - Chimeric Pacbio reads
  - Repeats
  - Missing Illumina coverage
- We use super reads to decide whether we can use raw Pacbio sequence to join the covering mega-reads
- Each join must be in 2+ reads -- same flanking super-reads

Progress towards WGS PacBio/Illumina hybrid assembly of the Chinese Spring genome

- Doubled haploid Chinese Spring wheat (accession Dv418)
- 33X wheat genome equivalents of PacBio WGS long and superlong reads (Dv418)
- 50X wheat genome equivalents of Hiseq 3000 150bp pairend reads (Dv418)
- 50X wheat genome equivalents of Hiseq 2500 250bp pairend reads (Dv418)

Summary

- Mega-reads benefit from the accuracy of Illumina and the length of the Pacbio reads
- The goal is to assemble haplotype-resolved mammalian-size genome in 2-3 weeks on a single 64-core computer
- Up to 30Gbp genome on a computer with 1Tb of RAM
- MaSuRCA 3.2.x to be released in 2016

Acknowledgements

- Funding agencies
  - USDA
  - NIH
  - NSF