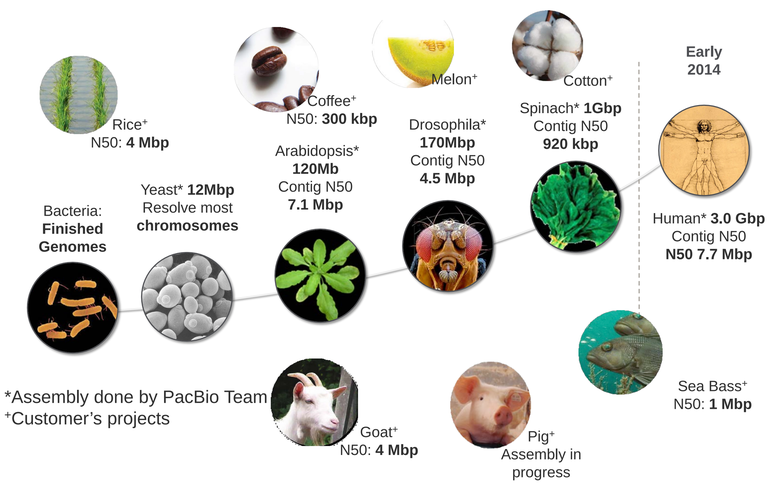
Toward Comprehensive Genomics Analysis With De Novo Assembly

Jason Chin, Paul Peluso, David Rank
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

Whole genome sequencing can provide comprehensive information important for determining the biochemical and genetic nature of all elements inside a genome. The high-quality genome references produced from past genome projects and advances in short-read sequencing technologies have enabled quick and cheap analysis for simple variants. However even with the focus on genome-wide resequencing for SNPs, the heritability of more than 50% of human diseases remains elusive. For non-human organisms, high-contiguity references are deficient, limiting the analysis of genomic features. The long and unbiased reads from single molecule, real-time (SMRT®) Sequencing and new *de novo* assembly approaches have demonstrated the ability to detect more complicated variants and chromosome-level phasing. Moreover, with the recent advance of bioinformatics algorithms and tools, the computation tasks for completing high-quality de novo assembly of large genomes becomes feasible with commodity hardware. Ongoing development in sequencing technologies and bioinformatics will likely lead to routine generation of high-quality reference assemblies in the future. We discuss the current state of art and the challenges in bioinformatics toward such a goal. More specifically, explicit examples of pragmatic computational requirements for assembling mammalian-size genomes and algorithms suitable for processing diploid genomes are discussed.

Progress of Large Genome Assembly with Only PacBio® Long Reads



*Assembly done by PacBio Team
+Customer's projects

"Cheaper sequencing, but at what cost scientifically?"

The Resurgence of Reference Quality Genome Sequence
Michael Schatz

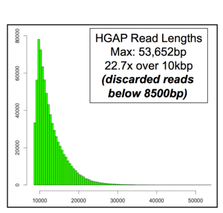
Jan 13, 2015
PAG XXIII

@mike_schatz / #PAGXXIII

Source: http://schatzlab.cshl.edu/presentations/2015.01.13.PAG.Resurgence%20of%20Ref%20Quality%20Genomes.pdf

O. sativa pv Indica (IR64)
Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19 kbp
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18 kbp
HGAP + CA 22.7x @ 10kbp	4.0 Mbp
Nipponbare BAC-by-BAC Assembly	5.1 Mbp



BAC-by-BAC Assembly
N50=5.1Mbp
cost > $100M

Short Read Assembly,
N50=20kbp
cost ~$10k

Simple Theorem for Perfect Assembly

Gene Myers, ISMB 2014 Keynote talk

https://dazzlerblog.wordpress.com/2014/05/15/on-perfect-assembly/

Theorem: Perfect assembly possible
if and only if

a) errors are random
b) sampling is Poisson
c) reads long enough to solve repeats.

Note: low error rate not needed

Better Algorithm Efficiency

Early 2014 → Mid 2014

BLASR
400,000 cpu hours

MHAP
(Celera® Assembler 8.2)
80,000 cpu hours
(high sensitivity setting)

New Results
Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing
Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James Drake, Jane M Landolin, Adam M Phillippy
doi: http://dx.doi.org/10.1101/008003

daligner
20,000 cpu hours

Efficient Local Alignment Discovery amongst Noisy Long Reads
Gene Myers*
MPI for Molecular Cell Biology and Genetics, 01307 Dresden, Germany
myers@mpi-cbg.de
Algorithms in Bioinformatics,
14th International Workshop, WABI 2014

Need Google® scale computation

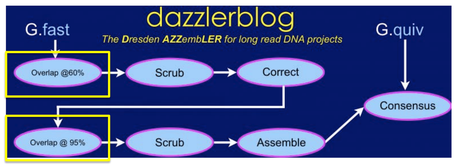
Can be done with a homebrew 5 to 7 node cluster. Google compute not needed.

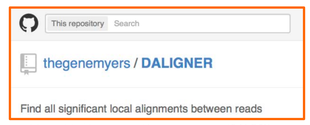
Dazzler And Falcon: Open Source Assembler

Gene Myers' team is working a new assembler "DAZZLER" ("The Dresden AZZembLER").

Currently the "daligner" code for overlapping reads is released on GitHub.

Check http://dazzlerblog.wordpress.com





https://github.com/thegenemyers/DALIGNER

Late 2014

Integrate "daligner" from DAZZLER as the "read overlap engine" to "Falcon" Assembler (v0.2)

Modules inside Falcon*:
- Workflow management
- Error correction
- Overlap to graph
- Graph to contigs

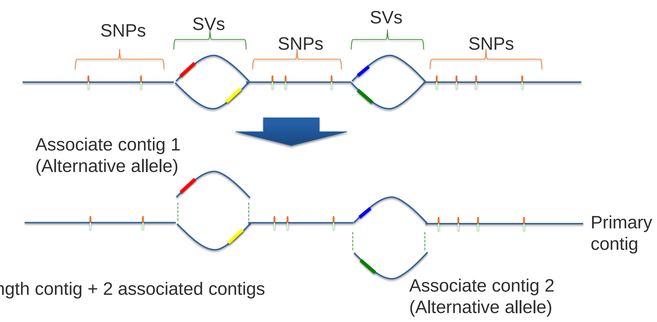
New CHM1 Assembly Statistics done with end-to-end Falcon workflow with Daligner:

#Seqs	5,528
Mean	509,822
Max	35,424,115
n50	**5,460,023**
Total	**2,818,296,359**

* For all different ways combining different modules for assembly (e.g. MHAP, HGAP.3, CA, etc.), please consult with Jason Chin or Richard Hall.

Get Falcon
https://github.com/PacificBiosciences/FALCON

Diploid Aware Contig Layout Rule

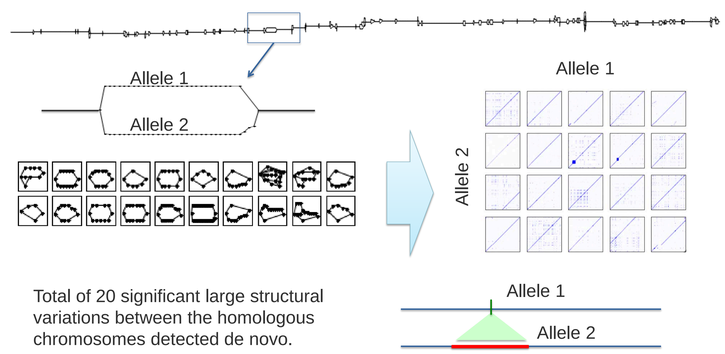


1 full length contig + 2 associated contigs

Keep the long-range information while maintaining the relations of the alternative alleles.

Assembly Graph of Diploid Human MHC Region

The assembly contig string graph contains many "bubbles"

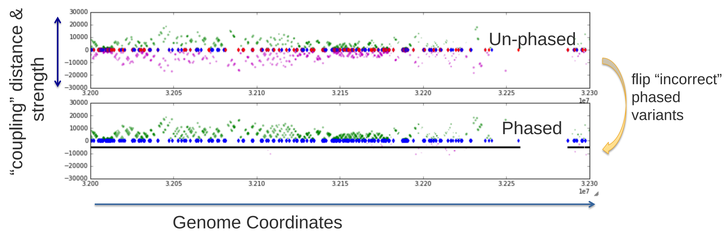


Total of 20 significant large structural variations between the homologous chromosomes detected de novo.

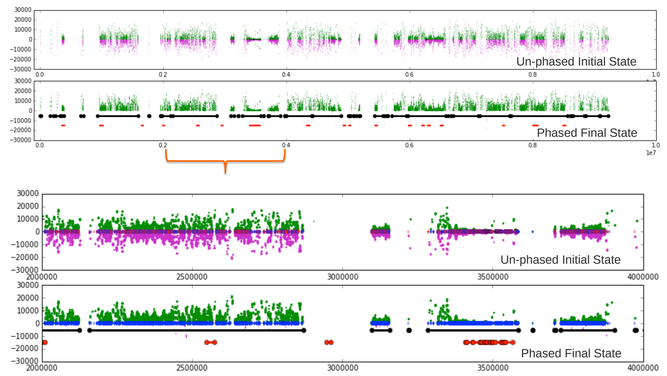
Phasing SNPs

Call het-SNPs directly from BAM output from BLASR

An "Ising model" (a model widely used in studying statistical physics) inspired greedy algorithm for phasing variants



Phasing Through a 9 Mb Contig



Toward Comprehensive Genomics Analysis

- Graph representation of reference genomes and assemblies will be essential.
- New algorithm and software tool development needed, e.g., more efficient haplotype re-construction
- Some other lower cost options include
  - Lower coverage assembly: cost vs. quality analysis
  - Incorporated other long-range information: optical mapping, Hi-C, genetic mapping
- Vision for scaling up post-assembly analysis
  - Crowd sourcing infrastructure for examining / annotating / correcting genome assemblies
  - Building Tools for large-scale comparative genomics with *de novo* assemblies

Acknowledgements.
- The authors like to thank Gene Myers, Adam Phillippy, Serge Koren, Mike Schatz, Aaron Klammer, Jim Drake and Lawrence Hon for suggestions and discussions during the development work for Falcon.
- The authors like to thank Richard Hall and Kathryn Keho for comments on improving the poster content.