Assembly of large genomes from short WGS reads

Aleksey Zimin, Guillaume Marcais,
Daniela Puiu, Tanja Magoc,
Michael Roberts, Elliot Winston,
Steven Salzburg and James Yorke

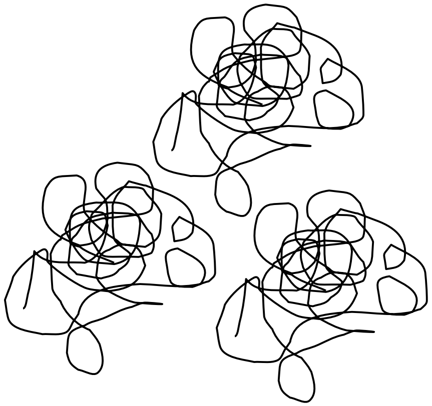
Johns Hopkins University, Baltimore
University of Maryland, College Park

USDA

PineRefSeq
http://pinegenome.org/pinerefseq/

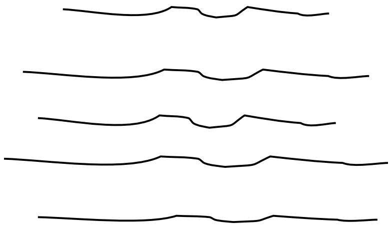
Whole Genome Shotgun reads

Multiple copies of DNA



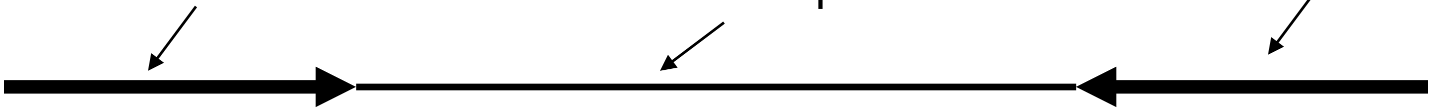
Shred
&
Size
select
→

Fragments of
150 - 200,000 bases



Sequence
the ends
→

Pairs of Reads of 150 – 500 bases each



CAAGCTGAT...

Unknown sequence

...GTTTGGAAC

Recently developed assemblers for NGS data

- MSR-CA
  - handles 454, Illumina, and Sanger reads
- Allpaths-LG
- SOAPdenovo
- Velvet
- ABySS
- Contrail
- SGA (also overlap-based)

Two assembly approaches

- Overlap-Layout-Consensus (OLC)
  - Used by most assemblers for previous generation (Sanger) sequencing
  - Celera Assembler, PCAP, Phusion, Arachne, etc
- Graph
  - Used by most assemblers for Illumina data
  - SOAPdenovo, Allpaths-LG, Velvet, Abyss, etc
- We use a combined approach that combines the benefits of both OLC and Graph in our MSR-CA assembler

Assembly approaches: OLC

- OLC: Overlap-Layout-Consensus
  - Compute overlaps of reads

```
AGTGATTAGATGATACTAGA
        ||||||| ||||
        GATGATAGTAGAGGATAGATTTA
```

  - Create layout of *contigs* from overlapping reads

```
AGTGATTAGATGATAGTAGA
       AGATGATACTAGAGATAGATAGACC
            ATAGTAGAGATAGATAGACCACTCATCATAC
```

  - Create consensus sequence of contigs

```
AGTGATTAGATGATAGTAGAGATAGATAGACCACTCATCATAC
```

5 billion reads?

... that’s 12.5 quadrillion overlaps

at 1 million overlaps/second, that would take 400 years

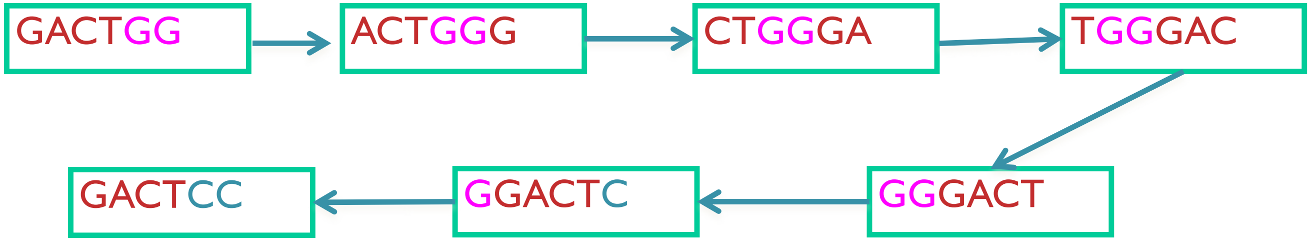
Two assembly approaches

- Overlap-Layout-Consensus (OLC)
  - Used by most assemblers for previous generation (Sanger) sequencing
  - Celera Assembler, PCAP, Phusion, Arachne, etc
- DeBruijn Graph
  - Used by most assemblers for Illumina data
  - SOAPdenovo, Allpaths-LG, Velvet, Abyss, etc
- We use a combined approach that combines the benefits of both OLC and Graph in our MSR-CA assembler

De Bruijn Graph Strategy

- Find all k-mers, build a graph
  - Every k-mer is a node
  - Two nodes are linked with an edge if they share k-1 mer

GACTGGGACTCC



- An assembly is a path through the graph that visits each edge at least once
- We can only roughly estimate the graph of the genome from reads due to sequencing errors and lack of coverage

Two assembly approaches

- Overlap-Layout-Consensus (OLC)
  - Used by most assemblers for previous generation (Sanger) sequencing
  - Celera Assembler, PCAP, Phusion, Arachne, etc
- Graph
  - Used by most assemblers for Illumina data
  - SOAPdenovo, Allpaths-LG, Velvet, Abyss, etc
- We propose to use a combined approach that combines the benefits of both OLC and Graph in our MSR-CA assembler

Benefits and drawbacks of OLC and Graph

- Benefits of OLC
  - Can deal with variable length reads and reads from different sequencing platforms
  - Overlaps can be long and thus more reliable
  - Overlaps do not have to be exact
  - Can resolve repeats of up to read size
- Drawbacks of OLC
  - Computationally intensive, number of overlaps grows quickly with the number of reads and coverage

- **Benefits of Graph**
  - **Computationally efficient**
- **Drawbacks of Graph**
  - **Errors in the reads create spurious branches in the graph – requires error correction**
  - **Max. size of k-mer is limited by the shortest read size**
  - **All overlaps in the graph are exact overlaps of k-1 bases**
  - **Repeats of longer than k-bases cannot be resolved**

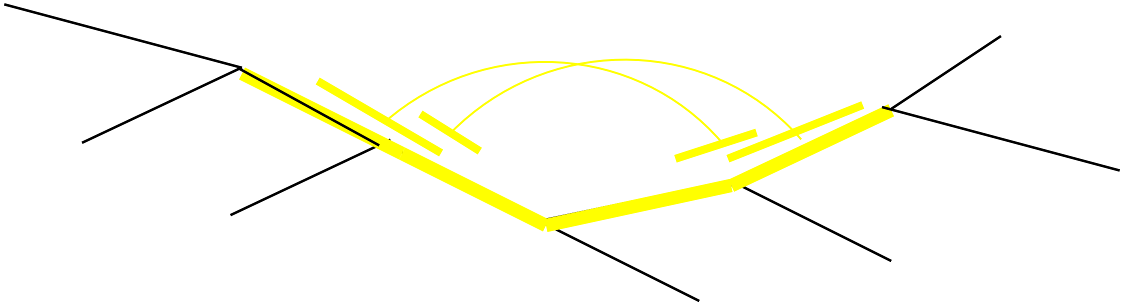
MSR-CA combines benefits of OLC and Graph

- Benefits of OLC
  - Can deal with variable length reads and reads from different sequencing platforms
  - Overlaps can be long and thus more reliable
  - Overlaps do not have to be exact
  - Can resolve repeats of up to read size
- Drawbacks of OLC
  - Computationally intensive, number of overlaps grows quickly with the number of reads and coverage

- **Benefits of Graph**
  - **Computationally efficient**
- Drawbacks of Graph
  - Errors in the reads create spurious branches in the graph – requires error correction
  - Max. size of k-mer is limited by the shortest read size
  - All overlaps in the graph are exact overlaps of k-1 bases
  - Repeats of longer than k-bases cannot be resolved

MSR-CA strategy

- Error correct Illumina reads
- Create a deBruijn graph: each k-mer is unique in the graph



  - Many pairs will extend to the same Super-read
  - Replace the pairs by the corresponding Super-reads
- Assemble the Super-reads with an OLC assembler (CABOG) using additional long mate pairs for linking the contigs

MSR-CA design

- Efficient multi-threaded code
- Designed to handle data sets with up to 12B reads
- Development is aimed at WGS assembly of the 24Gb Loblolly Pine genome on a computer with 48 cores and 512Gb of RAM in 1-2 months
- Current version 1.4 and being continually improved

Results on a pool of 500 Pine fosmids

	Sequence in assembly, bp	N50 contig size, bp	N50 scaffold size	Number of scaffolds>30 kb
Allpaths-LG	14,050,574	10,324	26,298	248
SOAPdenovo	13,470,572	1,632	33,389	322
MSR-CA	14,604,209	7,640	22,740	218

Notes: N50 computed from estimated 19.25Mb total sequence, each fosmid was estimated at 38Kb, assemblies used short pairs and jumping library pairs

Acknowledgements

- Funding agencies
  - USDA
  - NIH
- JCVI
  - Jason Miller
  - Brian Walenz
- UMD
  - Carl Kingsford