DNSRA Project

De Novo Short Read Assembly Project

Leong Hon Wai & Melvin Zhang

- ❑ New project for CS5206
- ❑ Short Read Assembly is a hot research topic
- ❑ Data structures and algorithms challenges
- ❑ Fun and interesting

Based on GRP by Pramila (Thanks!!!)

De Novo Genome Assembly using Paired-End Short Reads

Pramila Ariyaratne

17th July 2009

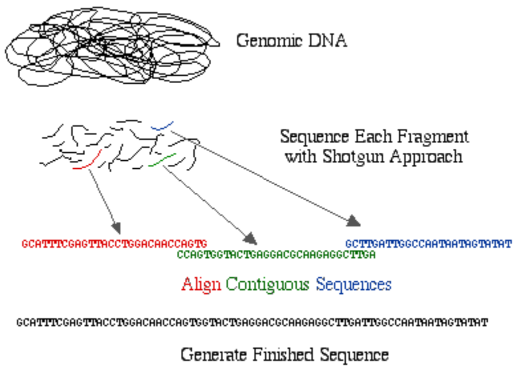
These notes are meant to give you a starting point for the project.

You are expected to do further reading on your own.

Motivation

- ❑ Complete genome sequence is essential to carry out various analysis.
- ❑ Sequencing a genome is not trivial
  - ❖ Chromosome length: up to ~250,000,000 bps
  - ❖ Longest sequence-able fragment: ~600 bps
- ❑ Therefore need for whole genome shotgun sequencing (WGSS).

WGSS overview


Genomic DNA
Sequence Each Fragment with Shotgun Approach
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence

Genomic DNA is sheared into small fragments.

Individual fragments are sequenced ('read')

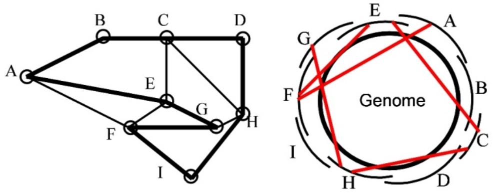
The reads are put together in dry lab to assemble target genome

Traditional method

- ❑ Sequenced using Sanger capillary sequencing
  - ❖ ~600bp length
  - ❖ ~10x coverage
- ❑ Assembly based on Overlap-Layout-Consensus approach.
- ❑ Based on *Overlap graph*
  - ❖ Each read formed a Node.
  - ❖ Edge exists between two nodes if the reads overlap.

Traditional method

- ❑ Overlap graph



- Red lines denote false overlaps.
- Thin edges denote false overlaps.

- ❑ Traverse the graph to find contiguous regions of target genome (*Contigs*)

Traditional method

- ❑ Very low throughput
- ❑ 384 sequences / day (0.4 million bps)
  - ❖ 10x coverage of human genome: ~30gbps

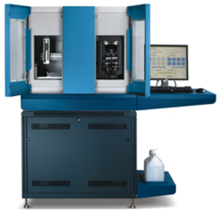
High-throughput sequencing

- ❑ … introduced in mid-2000s.
- ❑ Solutions by ABI SOLiD and Illumina Solexa.
- ❑ Characteristics:
  - ❖ Higher throughput
    - ◆ *1-4gbps / day*
  - ❖ Low cost / base pair.
  - ❖ Very short fragment length
    - ◆ *25 – 75bp*
  - ❖ High error rate
  - ❖ Inherent capability to do paired reads.

Next Generation Sequencing Machines

454 GS-SLX

ABI SOLID

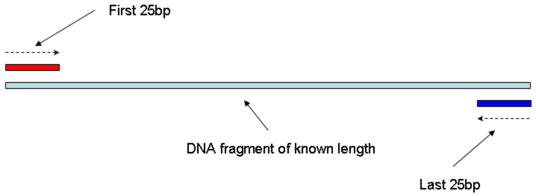


Solexa Genome Analyzer

- ❑ **Useful link about comparison and description**
  - ❖ **http://www.agencourt.com/services/nextgen/**
  - ❖ **http://www.genengnews.com/gen-articles/next-generation-sequencing-moves-to-next-next-level/3324/**

Paired reads (Mate pairs)

- ❑ **Paired-End sequencing (Mate pairs)**
  - ❖ **Sequence two ends of a fragment of known size.**



  - ❖ **Currently fragment length (insert size) can range from 200 bps – 10,000 bps**

High-throughput sequencing

- ❑ **Short read length = even short overlap**
- ❑ **Somewhat compensated by sequencing at higher coverage**
  - ❖ **Typically 80-100x coverage**
- ❑ **Large number of reads + short overlap + higher error rate ➔ Traditional Overlap-Layout-Consensus method impractical**
  - ❖ **Highly convoluted *Overlap graph***

Current approaches

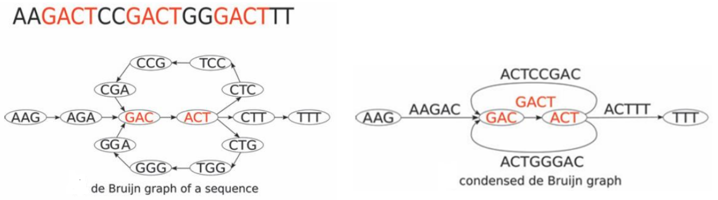
- ❑ **Most are based on Euler/de Bruijn graph method.**
- ❑ **Euler/de Bruijn graph method**
  - ❖ **Introduced as a alternative to traditional overlap graph.**
  - ❖ **More suited to short read assembly.**

De Bruijn graph

- ❑ For de Bruijn graph construction, all reads are broken in to overlapping subsequences on length *k* (*k-mer)*.
- ❑ Each *k*-1 subsequence represents a node in de Bruijn graph.
- ❑ A directed edge *e* exists between two nodes *a* and *b* iff there exists a *k-mer* such that its prefix = *a* and its suffix = *b.*

De Bruijn graph

- ❑ De Bruijn graph can be condensed by collapsing non-ambiguous paths.



de Bruijn graph of a sequence

condensed de Bruijn graph

- ❑ Ideally, find a Eulerian path in this graph which represents the genome.

Current approaches

- ❑ Velvet
- ❑ EULER-USR
- ❑ ALLPATHS

- ❑ Velvet and Euler USR are based on De Bruijn graph method, but differ in their error handling.

Velvet

- ❑ Currently the most popular approach for de novo assembly of short reads.
- ❑ Error handling done in de Bruijn graph
- ❑ Sequencing errors manifest as *tips* or *bubbles* in graph.
  - ❖ Tips: errors towards end of the read.
    - ◆ *Trim all tips shorter than* 2k *length.*
  - ❖ Bubbles: errors in middle of the read.
    - ◆ Tour bus *algorithm.*

Daniel Zerbino and Ewan Birney. Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. Genome Res. 18: 821-829. 2008

Velvet

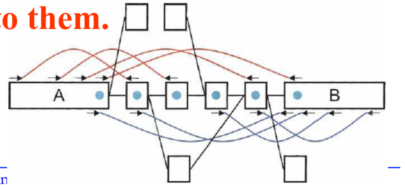
❑ Tour bus algorithm

- ❖ Start at arbitrary node and traverse breath first in *Dijkstra*-like algorithm.
- ❖ Distance metric such that paths with higher coverage are visited first.
- ❖ If a Node is visited twice backtrack both paths till common ancestor.
- ❖ If two paths (sequences) can be reconciled, merge them, giving priority to path with higher coverage.

Velvet

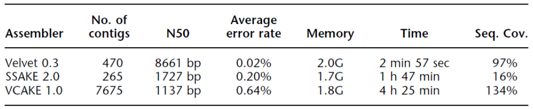
❑ *Breadcrumbs* algorithm

- ❖ Use paired data to resolve ambiguities.
- ❖ Mark all *long* nodes (Longer than *insert size*)
- ❖ Mark all other nodes connected to long nodes.
  - ◆ *Connected by >5 paired reads.*
- ❖ Find paths between two long nodes only via nodes connected to them.



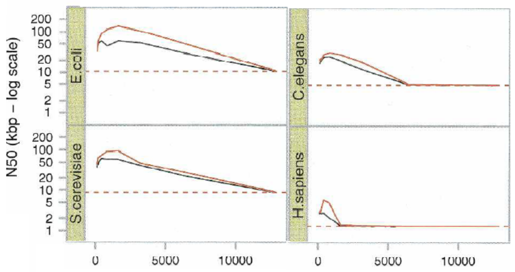
Velvet

❑ Results



Assembler	No. of contigs	N50	Average error rate	Memory	Time	Seq. Cov.
Velvet 0.3	470	8661 bp	0.02%	2.0G	2 min 57 sec	97%
SSAKE 2.0	265	1727 bp	0.20%	1.7G	1 h 47 min	16%
VCAKE 1.0	7675	1137 bp	0.64%	1.8G	4 h 25 min	134%

- Single tag data only
- *Streptococcus suis* on Solexa.



- Paired data/breadcrumbs
- Dotted: No breadcrumbs
- Red: Supercontigs
- Black: Contigs

Velvet

❑ Advantages

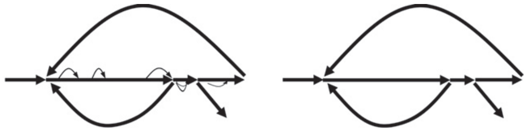
- ❖ Simple execution
- ❖ Extremely fast
- ❖ Moderate memory usage

❑ Disadvantages

- ❖ Error correction without localizing
- ❖ Paired data use only in latter part of execution
  - ◆ *Overly complicated de Bruijn graph.*
- ❖ Tour-bus likely to collapse large regions
- ❖ Currently *k* limited to 32bp. (on 64bit machines)

EULER-USR

- Use *Repeat graph* instead of De Bruijn graph
  - De Bruijn graph with bubbles and tips removed.
- Uses fact that ‘first part of a read sequence has less error’ to do handle errors.



de Bruijn graph of a genome

repeat graph of a genome

Mark J. Chaisson, Dumitru Brinza and Pavel A. Pevzner. De novo fragment assembly with short mate-paired reads: Does read length matter? Genome Res. 19:336-346. 2009

EULER-USR

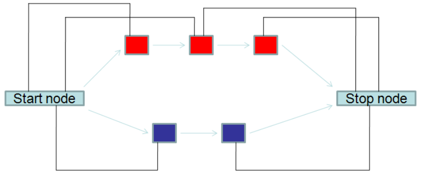
- Error correction of read prefixes
  - Given set of reads *R* and threshold *m*, a k-mer is *solid* if it occurs at least *m* times in *R.*
  - Use all read prefixes as R.
  - For each read prefix, check if all its k-mers are solid.
    - *If not, allow a few mutations to make it solid.*
  - Discard if cannot be made solid.

EULER-USR

- Error correction of read suffixes
  - Create repeat graph using assumed error-free prefixes.
  - Assume that suffix of a read is also prefix of another read.
    - *Therefore will be present in repeat graph.*
  - For each read, map the entire read to the repeat graph.
    - *Allow mismatches in suffix (error correction) if cannot be mapped.*
  - Rebuild repeat graph with error-free entire reads.

EULER-USR

- Use of paired data.
  - Fill gap between each paired tag to obtain sequence of size ‘*insert size + 2 x read length.*
  - Simple if there is only single path linking two tags
  - In case of multiple paths, use one with higher support.

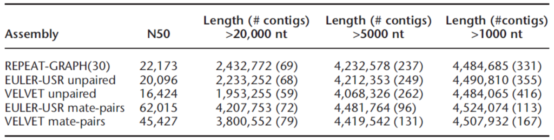


- Red: support 4
- Blue: support 2

  - Update repeat graph with complete sequences

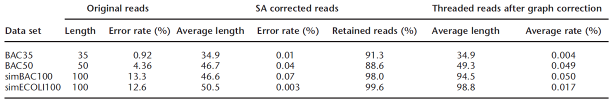
EULER-USR

❑ Results



Assembly	N50	Length (# contigs) >20,000 nt	Length (# contigs) >5000 nt	Length (#contigs) >1000 nt
REPEAT-GRAPH(30)	22,173	2,432,772 (69)	4,232,578 (237)	4,484,685 (331)
EULER-USR unpaired	20,096	2,233,252 (68)	4,212,353 (249)	4,490,810 (355)
VELVET unpaired	16,424	1,953,255 (59)	4,068,326 (262)	4,484,065 (416)
EULER-USR mate-pairs	62,015	4,207,753 (72)	4,481,764 (96)	4,524,074 (113)
VELVET mate-pairs	45,427	3,800,552 (79)	4,419,542 (131)	4,507,932 (167)

- Paired reads assembly
- *E. Coli* on Solexa.



	Original reads			SA corrected reads		Threaded reads after graph correction	
Data set	Length	Error rate (%)	Average length	Error rate (%)	Retained reads (%)	Average length	Average rate (%)
BAC35	35	0.92	34.9	0.01	91.3	34.9	0.004
BAC50	50	4.36	46.7	0.04	88.6	49.3	0.049
simBAC100	100	13.3	46.6	0.07	98.0	94.5	0.050
simECOLI100	100	12.6	50.5	0.003	99.6	98.8	0.017

- Error correction

EULER-USR

❑ Advantages

- ❖ Effective error correction.
- ❖ Clever use of prefix / suffix error rate difference.

❑ Disadvantages

- ❖ Error correction without localization.
- ❖ Use of paired end data is post processing step.

ALLPATHS

- ❑ Not based on Euler / de Bruijn graph approach.
- ❑ Use same solid *k*-mer error correction as EULER-USR. (without prefix/suffix differentiation)
- ❑ Builds unipath-graph (similar to repeat graph)
  - ❖ A linear section of the graph is referred to as a unipath.
- ❑ Localizes reads sequences before assembly.

Jonathan Butler, et al, "*ALLPATHS: De novo assembly of whole-genome shotgun microreads,*" Genome Research, (2008), 18: pp. 810-820.

ALLPATHS

❑ Read localization

- ❖ Select a unipath with 'normal' coverage
  - ◆*Avoid large repeat regions*
- ❖ All other unipaths and paired tags connected to this unipath is considered to be in its *neighborhood.*
- ❖ Assemble each neighborhood separately.

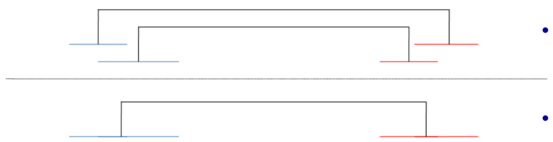
ALLPATHS

❑ Short fragment pair merger

- ❖ Similar to EULER-USR, fills the gap in between two paired reads.
- ❖ Builds a *local unipath graph.*
- ❖ Extend both ends (of all reads) based on the local unipath graph.
- ❖ For each pair, search for other pairs which overlap on both ends. Merge to obtain longer reads.

ALLPATHS

❑ Short fragment pair merger (cont.)



- Combine these..
- To obtain this.

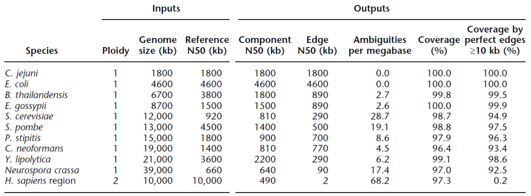
- ❖ Repeat the process for all pairs.

❑ Once sequence is complete, update the local unipath graph.

❑ Iteratively merge local unipath graphs to obtain a global unipath graph, representing the genome.

ALLPATHS

❑ Results



	Inputs			Outputs				
Species	Ploidy	Genome size (kb)	Reference N50 (kb)	Component N50 (kb)	Edge N50 (kb)	Ambiguities per megabase	Coverage (%)	Coverage by perfect edges ≥10 kb (%)
C. jejuni	1	1800	1800	1800	1800	0.0	100.0	100.0
E. coli	1	4600	4600	4600	4600	0.0	100.0	100.0
B. thailandensis	1	6700	3800	1800	890	2.7	99.8	99.5
E. gossypii	1	8700	1500	1500	890	2.6	100.0	99.9
S. cerevisiae	1	12,000	920	810	290	28.7	98.7	94.9
S. pombe	1	13,000	4500	1400	500	19.1	98.8	97.5
P. stipitis	1	15,000	1800	900	700	8.6	97.9	96.3
C. neoformans	1	19,000	1400	810	770	4.5	96.4	93.4
Y. lipolytica	1	21,000	3600	2200	290	6.2	99.1	98.6
Neurospora crassa	1	39,000	660	640	90	17.4	97.0	92.5
H. sapiens region	2	10,000	10,000	490	2	68.2	97.3	0.2

- Results on simulated data

- ❖ Our experiments showed ALLPATHS was very good in deciphering tandem repeat regions.

ALLPATHS

❑ Advantages

- ❖ Read localization
- ❖ Multi-CPU compatible
- ❖ Extremely good results.

❑ Disadvantages

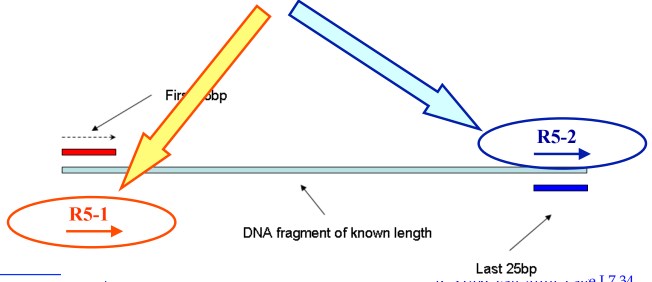
- ❖ Slow
- ❖ Very memory intensive
- ❖ Impractical assumptions on input data
  - ◆ *500bp +/- 5bp insert size*

Some remarks from Pramilae…

- □ Current existing method seems adequate
- □ Not necessarily exploiting paired data to fullest.
- □ Error correction steps are not localized.

For your Project (Simplifications)

- □ Only errors are *mutation* errors
  - ❖ No indels (insertions/deletions)
- □ No litigation errors for Paired-Ends
  - ❖ Paired-End reads are from *same* fragment
  - ❖ Both reads in the *same* direction



Project Milestones

- □ Three Milestones:
  - ❖ M1 : 10-Oct-2010
    - ◆ *Very simple model*
    - ◆ *Write simple program (know input/output, etc)*
  - ❖ M2 : 15-Oct-2010
    - ◆ *Your Preliminary Proposal*
  - ❖ M3 : 11-Nov-2010
    - ◆ *Final Deliverables*

Thank you.

Q & A

NUS National University of Singapore

School of Computing