

# CS262 Computational Genomics

## Lecture 12: Fragment Assembly II

Lecturer: Serafim Batzoglou

Scribe: Harry Robertson

February 23, 2009

### Fragment Assembly

As we saw in Lecture 10, efficient DNA Fragment Assembly is performed in four steps:

1. Find overlapping reads
2. Merge reads into contigs
3. Link contigs into supercontigs
4. Derive consensus sequence

Steps 1 and 2, and the associated vocabulary, have already been covered in the lecture notes for Lecture 10.

At present we shall continue where we left off at Step 2, having obtained a contig graph (see Fig. 1) where reads have been merged up to the boundaries of repeats.

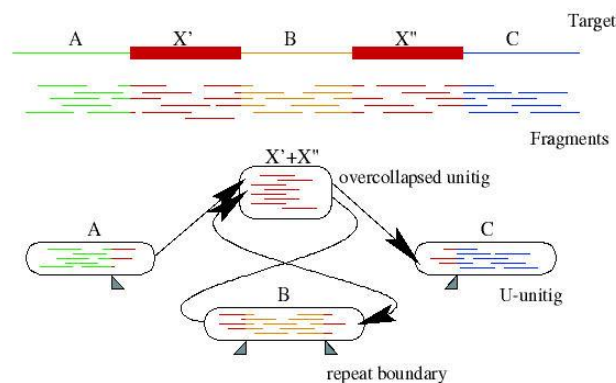


Fig 1. Contig graph structure

### Aside on repeats

It is clear that repeats are generally problematic for our algorithm.

However, in two cases we can easily detect them and treat them appropriately:

- if they are either shorter than the read length (because reads that span the repeat will disambiguate them)
- if they have a higher rate of pair diffs than our sequencing error rate (because in that case our algorithm will directly be able to differentiate between them)

In other words, we can decrease the rate of problematic repeats for our algorithm by increasing read length and/or decreasing sequencing error rate.

In practice, the error correction performed by our algorithm corrects up to 98% of single-letter sequencing errors.

### Step 3 – Link contigs into supercontigs

#### a) Identify unique contigs

Once we have our contig graph from Step 2, we need to determine which of the contigs are “unique” contigs, and which are “overcollapsed” (repeat) contigs.

We can do this using two criteria:

- Density of reads (e.g. a high density indicates an overcollapsed contig) (see example in Fig. 2)
- The number of incident edges in the contigs graph (e.g. if a contig has several incoming and outgoing edges then it is a repeat) (see Fig. 3)



Fig. 2: Unique contig (of left) vs. denser overcollapsed contig

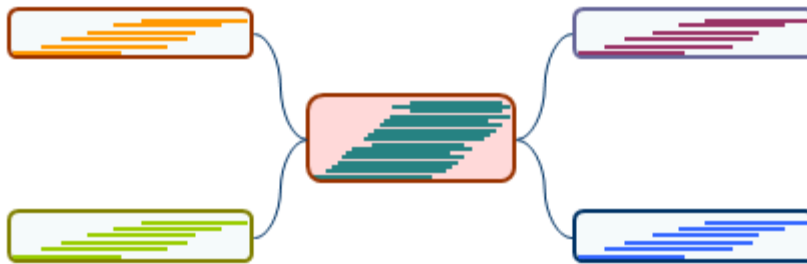


Fig. 3: Overcollapsed contig (in center) with several incident edges

### b) Chain together unique contigs

Once we have identified the unique contigs, we can use paired reads to find links between them (i.e. if the left read of a pair is at the end of contig A, the right read is at the start of contig B, we deduce that A is to the left of B in the contig chain). The resulting chain is made up of supercontigs (made up of contiguous unique contigs), and gaps which correspond to the repeat contigs (see Fig. 4).



Fig. 4: Unique contig chain

We can then fill in the gaps with the repeat contigs, using the contig graph from Step 2.

### Step 4 - Derive consensus sequence

We now have a chain of contigs, which are each made up of many overlapping reads. All that remains is to derive the final sequence of bases.

We choose each base by weighted voting among all the overlapping reads (see Fig. 5). Another possible approach is to take the maximum-“quality” base.



Fig. 5: Weighted voting to build consensus sequence

## Some assemblers

Some assemblers that are used in practice are:

- PHRAP
  - Early assembler, widely used, good model of read errors
  - Overlap  $O(n^2)$  -> layout (no mate pairs) -> consensus
- Celera
  - First assembler to handle large genomes (fly, human, mouse)
  - Overlap -> layout -> consensus
- Arachne
  - Public assembler (mouse, several fungi)
  - Overlap -> layout -> consensus
- Phusion
  - Overlap -> clustering -> PHRAP -> assemblage -> consensus
- Euler
  - Indexing -> Euler graph -> layout by picking paths -> consensus

## Quality of assemblies

### Paired-read-distance distribution

A major factor in assembly quality is the distribution of the distances between paired reads that we use. In practice we use paired-read-distances from 2k up to 150k: shorter distances allow us to efficiently link nearby contigs, while longer ones link far-off contigs which may be separated by very long repeats.

Reads with small (<20k) paired-read-distances typically use plasmid vectors, whereas very distant paired reads (150-200k) use BAC vectors.

### Mouse genome example

As a sample distribution, consider the case of the assembly of the mouse genome (see Table below).

Table 1 Distribution of sequence reads

Insert size (kb)*	Vector	Reads (millions)				Bases† (billions)		Sequence coverage‡		Physical coverage§
		All	Used	Paired	Assembled	Total	>Phred20	Total	>Phred20	
2	Plasmid	3.8	3.7	3.1	2.9	1.8	1.5	0.71	0.61	1.2
4	Plasmid	31.3	24.7	22.1	21.5	14.7	12.6	5.89	5.03	17.7
6	Plasmid	1.2	1.0	0.8	0.8	0.5	0.5	0.22	0.19	1.0
10	Plasmid	2.5	2.4	2.1	1.7	1.3	1.0	0.52	0.42	4.3
40	Fosmid	2.1	1.3	1.2	1.1	0.6	0.5	0.26	0.21	9.3
150-200	BAC	0.4	0.4	0.4	0.4	0.2	0.2	0.09	0.07	13.7
Other	Plasmid	0.07	0.05	0.03	0.04	0.03	0.03	0.01	0.01	0.02
Total		41.4	33.6	29.7	28.4	19.2	16.3	7.68	6.53	47.2

Centre	Reads (millions)				Bases† (billions)		Sequence coverage‡		Physical coverage§
	All	Used	Paired	Assembled	Total	>Phred20	Total	>Phred20	
Whitehead Institute	22.2	18.0	15.9	15.7	10.7	9.2	4.28	3.68	21.3
Washington University	11.5	8.3	7.5	7.1	4.7	3.9	1.87	1.57	5.9
Sanger Institute	6.7	6.3	5.4	4.7	3.3	2.7	1.31	1.09	5.3
University of Utah	0.6	0.6	0.5	0.5	0.3	0.3	0.13	0.11	1.0
The Institute for Genomic Research	0.5	0.4	0.4	0.4	0.2	0.2	0.09	0.08	13.7
Total	41.4	33.6	29.7	28.4	19.2	16.3	7.68	6.53	47.2

\* The approximate mean size of inserts of various libraries. Each library was individually tracked and evaluated. Insert sizes were intended to cover a narrow range as determined empirically against assembled sequence.

† Bases refers to the bases present in the used reads after trimming for quality.

‡ Sequence coverage estimated on the basis of all used reads after trimming for quality and a 2.5-Gb euchromatic genome. This excludes the heterochromatic portion, which contains extensive arrays of tandemly repeated sequence such as that found in the centromeres, rDNA satellites and the *Sp100-rs* array.

§ Physical coverage refers to the total cloned DNA in the paired reads.

|| Consists of a small number of unpaired reads and BAC-based reads used for methods development and consistency checks.

## N50 contig length

To quantify the distribution of contig lengths, a commonly used metric is the **N50 contig length**, which is defined as follows:

If we sort contigs from largest to smallest, and start covering the genome in that order, N50 is the length of the contig that just covers the 50th percentile.

Continuing with our mouse genome example, the Table below contains the N50 statistics for the assembly of the mouse genome:

Table 2 Basic statistics of the MGSCv3 assembly

Features	Number	N50 length (kb)*	Bases (Gb)	Bases plus gaps (Gb)	Percentage of genome
All anchored contigs†	176,471	25.9	2,372	2,372	94.9
All anchored supercontigs	377	18,600	2,372	2,477	99.1
All ultracontigs	88	50,600	2,372	2,493	99.7
Unanchored contigs‡	48,242	2.3	0.106	0.106	—
Largest 200 supercontigs	200	18,700	2,352	2,455	98.2
Largest 100 supercontigs	100	22,900	1,955	2,039	81.6

\* Not including gaps.

† Calculated on the basis of a 2.5-Gb euchromatic genome. Includes spanned gaps.

‡ The unanchored contigs, grouped into 44,166 unanchored supercontigs with an N50 value of 3.4 kb. The N50 value for all contigs is 24.8 kb, and for all supercontigs is 16,900 kb (excluding gaps). Inspection suggests that most of these unanchored contigs fall into gaps in the ultracontigs and are thus accounted for in the 'bases plus gaps' estimate.