Next-generation sequencing

Lecture 5

Assembly

- Assembly algorithms
  - Greedy algorithms (SSAKE, VCAKE)
  - Overlap Layout Consensus (Newbler, Mira)
  - De brujin graphs (Velvet, ABySS, Soap-denovo)

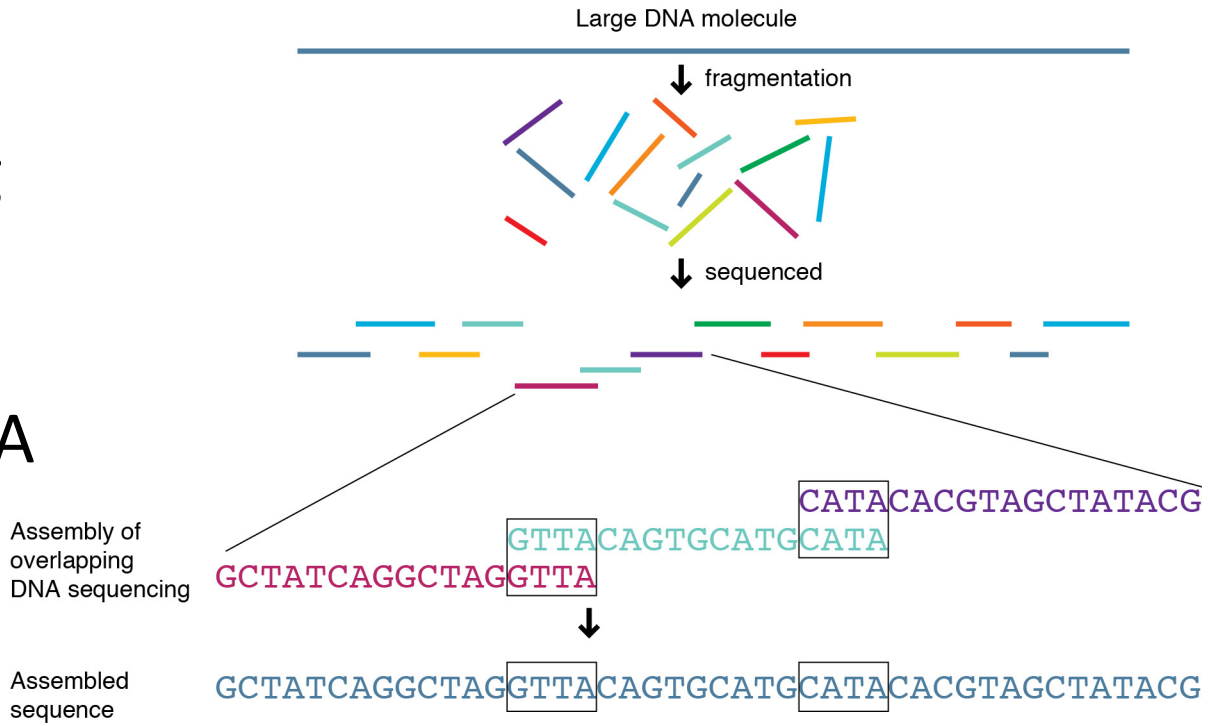
- *De novo* whole genome assembling strategies
- Mapping assembling strategies

De Novo sequencing

- New species/strains
- Challenge of assembly with short reads
  - 100x coverage of 3 GB genome with 100 bp reads= 3G fragments
  - Exponential problem for all-vs-all algorithm (overlap)
- Big problem with repeats
- Assemble contigs, fill gaps
- Paired-end reads are essential

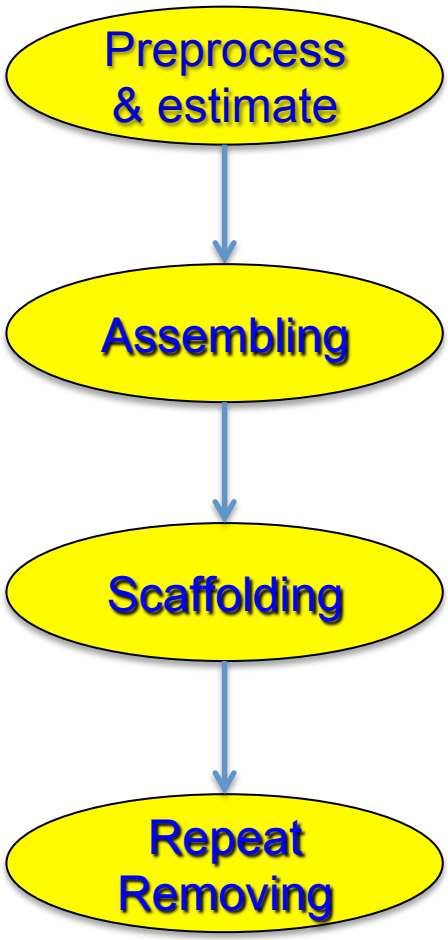
Shotgun Sequencing

Shotgun sequencing is a laboratory technique for determining the DNA sequence of an organism's genome.



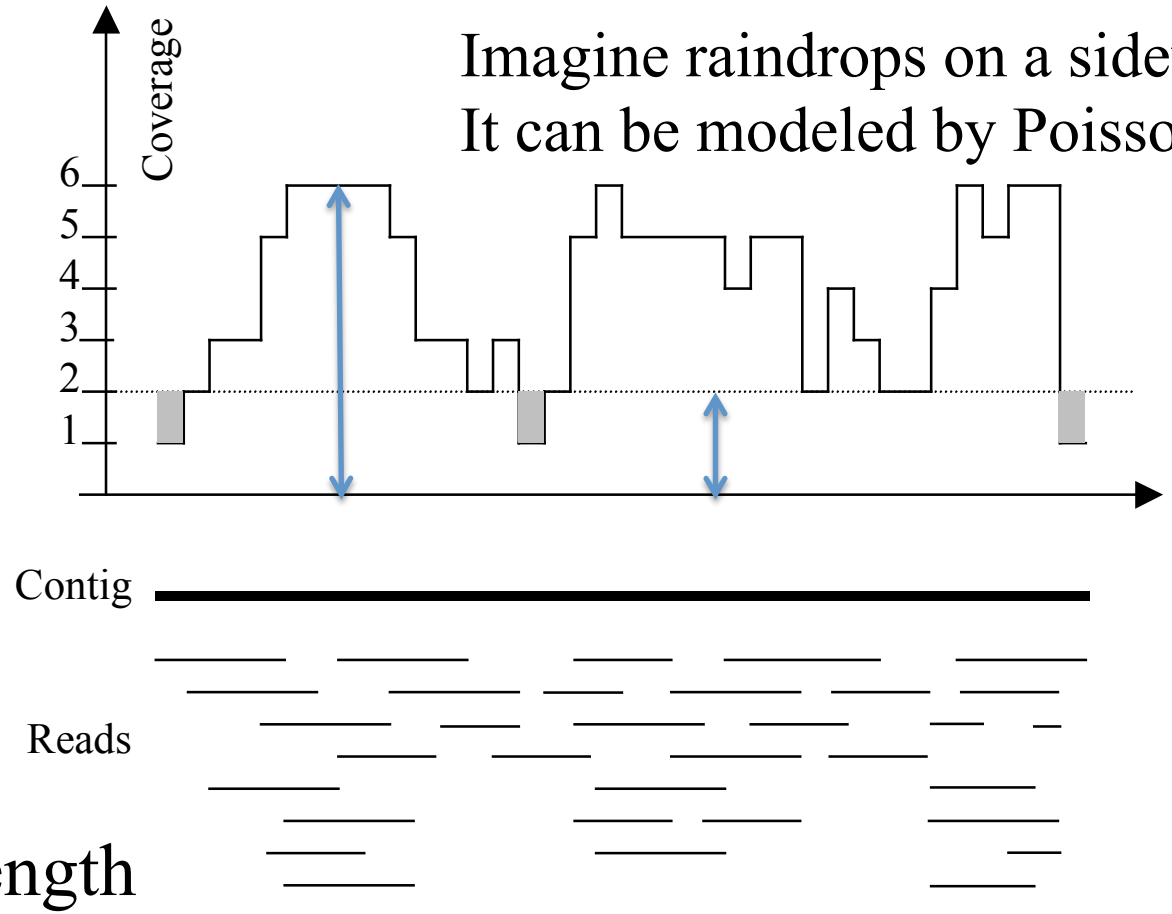
- Breaking the genome into a collection of small DNA fragments
- Sequencing.
- Reconstitute the genome.

Assembly Pipeline



Shotgun sequencing statistics

Typical contig coverage



Imagine raindrops on a sidewalk:
It can be modeled by Poisson distribution

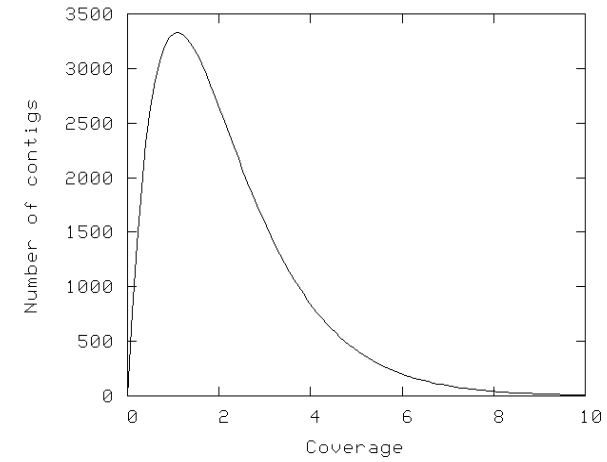
L = read length
G = genome size
N = number of reads
c = coverage= (NL / G)

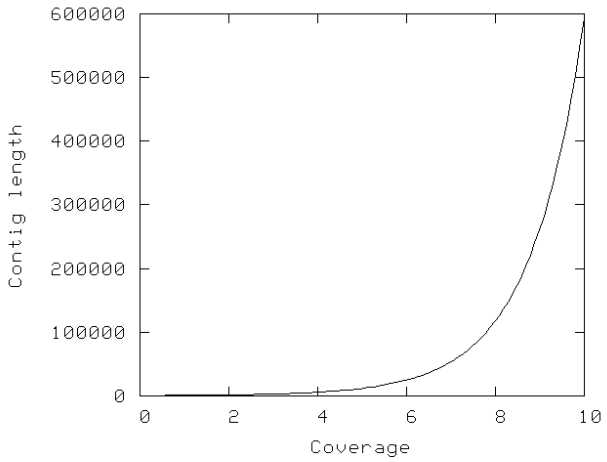
Average coverage

Lander-Waterman statistics

L = read length
G = genome size
N = number of reads
c = coverage =(NL / G)
T = minimum detectable overlap
$\sigma = 1 - T/L$

$E(\#\text{ of islands}) = Ne^{-c\sigma}$
$E(\text{island size}) = L((e^{c\sigma} - 1) / c + 1 - \sigma)$
contig = island with 2 or more reads





https://en.wikipedia.org/wiki/Michael_Waterman

Smith-Waterman algorithm for sequence comparison

Example

Genome size: 1 Mbp Read Length: 600

c	N	#islands	#contigs	bases not in any read	bases not in contigs
1	1,667	655	614	698	367,806
3	5,000	304	250	121	49,787
5	8,334	78	57	20	6,735
8	13,334	7	5	1	335

Experimental data

X coverage	# ctgs	% > 2X	avg ctg size (L-W)	max ctg size	# ORFs
1	284	54	1,234 (1,138)	3,337	526
3	597	67	1,794 (4,429)	9,589	1,092
5	548	79	2,495 (21,791)	17,977	1,398
8	495	85	3,294 (302,545)	64,307	1,762
complete	1	100	1.26 M	1.26 M	1,329

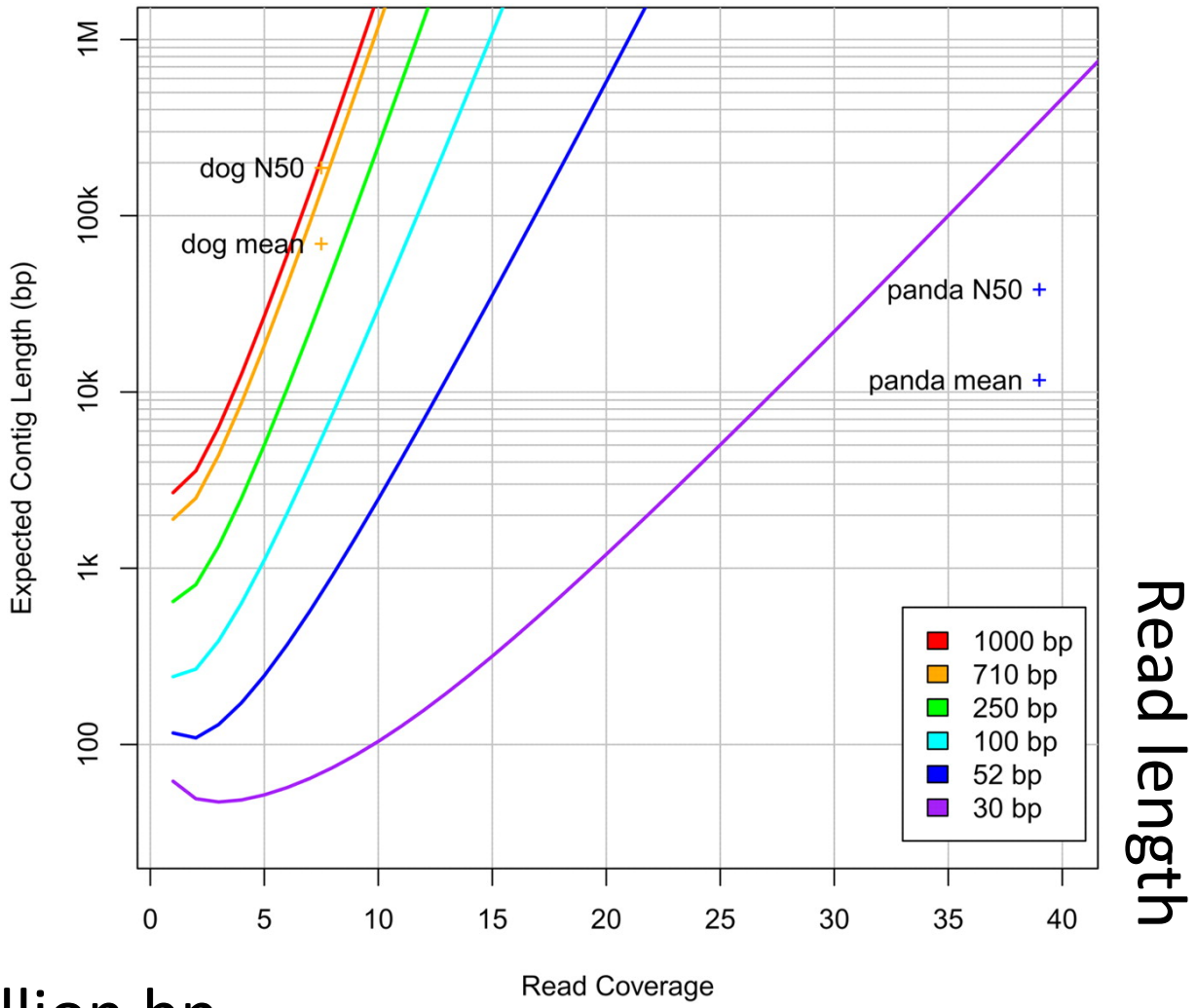
Numbers based on artificially chopping up the genome of *Wolbachia pipientis* dMel

Errors in Lander-Waterman Estimate

Lander-Waterman has limitations:

- repeats
- GC/AT rich regions
- other low complexity regions
- cloning biases in shotgun libraries

Expected average contig length for a range of different read lengths and coverage values.



Dog: 2.5 billion bp

Panda: 3 billion bp

Schatz M C et al. Genome Res. 2010;20:1165-1173

Organism/genome size	Assembler/status[a]	Input sequence						Contigs			
		Type	Pair size	Avgerage read (bp)	No. of reads	Read coverage[b]	Pair coverage[c]	No.	N50	Max	Total
Human (*H. sapiens*)/3.0 Gb	ABYSS published 2009	GA	210 bp	35–46	3.5 B	45×	120×	2.76 M	1.5 kb	18.8 kb	2.18 Gb
Grapevine (*V. vinifera*)/500 Mb	Myriad published 2007	Sanger	2–10 kb	579	5.95 M	6.9×	21×	58,611	18.2 kb	238 kb	531 Mb[d]
		Sanger	40 kb	460	144 k	0.13×	4.4×				
		Sanger	120 kb	369	68 k	0.02×	4.2×				
		454	None	169	12.5 M	4.2×	—				
Cucumber (*C. sativus*)/367 Mb	RePS2 published 2009	Sanger	2–6 kb	439	2.08 M	3.35×	9.9×	62,412	19,807	NR	226 Mb
		Sanger	40 kb	496	339 K	0.46×	16.7×				
		Sanger	140 kb	551	33.2 k	0.04×	5.6×				
		GA	200 bp	42	282 M	32.5×	76.8×	NR	2.6 kb	NR	204 Mb
		GA	400 bp	44	173 M	20.6×	94.4×				
		GA	2 kb	53	105 M	15.3×	286×	NR	12.5 kb	NR	190 Mb
Panda (*A. melanoleura*)/2.4 Gb	SOAPdenovo published 2010	GA	150	45	1.31 B	24.5×	43.3×	200,604	36,728	434,635	2.25 Gb
		GA	500	67	917 M	25.5×	90.2×				
		GA	2 kb	71	397 M	11.8×	192×				
		GA	5 kb	38	505 M	8.0×	533×				
		GA	10 kb	35	254 M	3.7×	571×				
Strawberry (*F. vesca*)/220 Mb	CABOG and Velvet announced	454	None	209	7.73 M	7.3×	—	16,487	28,072	215,349	202 Mb
		454	None	368	787 M	13.2×	—				
		454	2.5 kb	193	2.39 M	2.1×	6.9×				
		454	20 kb	236	1.58 M	1.7×	20×				
		GA	None	76	36 M	12.4×	—				
		SOLiD	2 kb	25	1.30 M	0.14×	6.4×				
Turkey (*M. gallopavo*)/1.1 Gb	CABOG announced	454	3 kb	180	6 M	1×	8×	128,271	12,594	90 kb	931 Mb
		454	20 kb	195	2 M	0.3×	18×				
		454	None	366	13 M	4×	—				
		GA	180 bp	74	200 M	13×	16×				
		GA	None	74	200 M	13×	—				

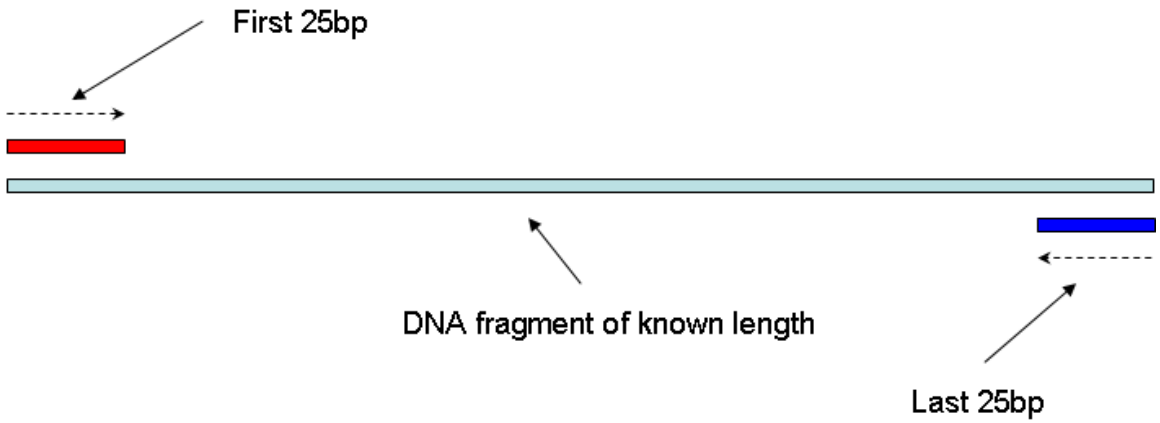
One more example

For yeast 12Mbp

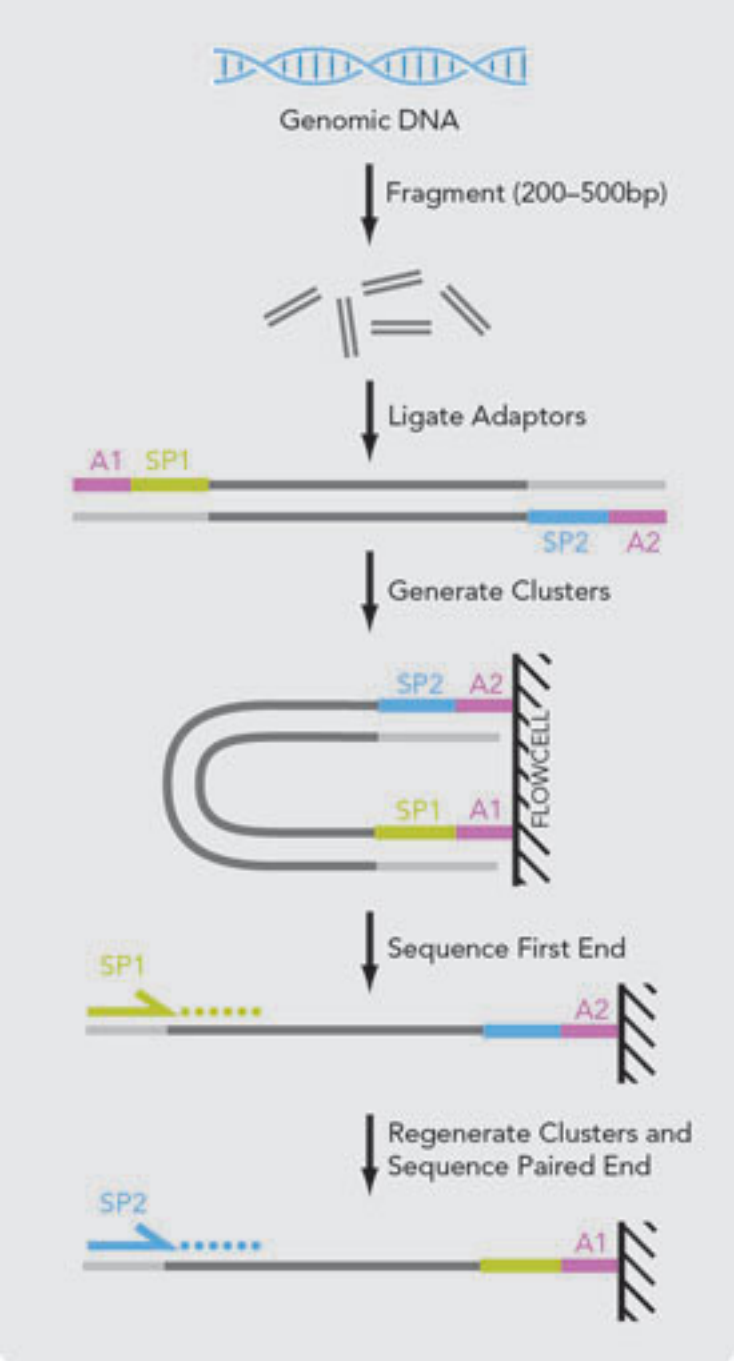
- read length: 200-400 bp
- coverage: 50X (how many reads do we need?)
- paired-end read insert size: 8kb (better to make multiple libraries with different insert sizes.)

Paired-end sequencing

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



  - Currently fragment length (insert size) can range from 200 bps – 10,000 bps
  - Paired-end sequencing is helpful for assembly and locating repeat. It also can detect rearrangements, including insertions and deletions (indels) and inversions.
  - As paired end reads are more likely to align to a reference, the quality of the entire data set improves



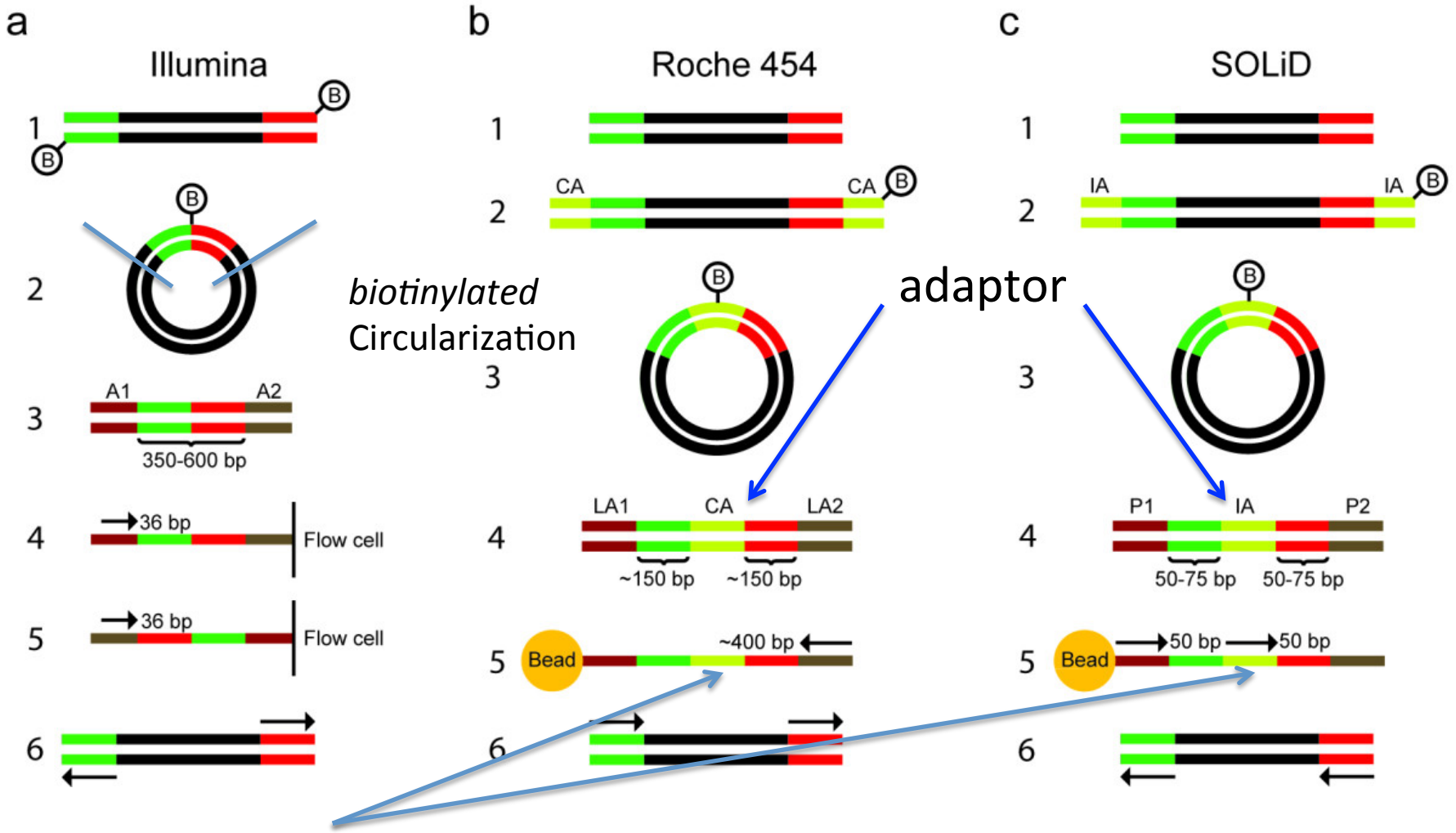
Paired-end sequencing by Illumina

Solid-phase amplification and Cyclic reversible termination

A simple modification to the standard single-read DNA library preparation.

Both the forward and reverse template strands of each cluster can be sequenced.

Mate-pair libraries



Use computer software to remove adaptor sequences

Berglund *et al. Investigative Genetics* 2011 **2**:23

difference

- Mate-pair is a specific type of library;
- paired-end is a type of sequencing
- mate-pair libraries require paired-end sequencing
- The decision to use mate-pair vs. standard libraries depends upon your application.
- Mate pair allows you to have your pairs be much farther apart, which can be more informative than the standard paired-end protocol.

Why?

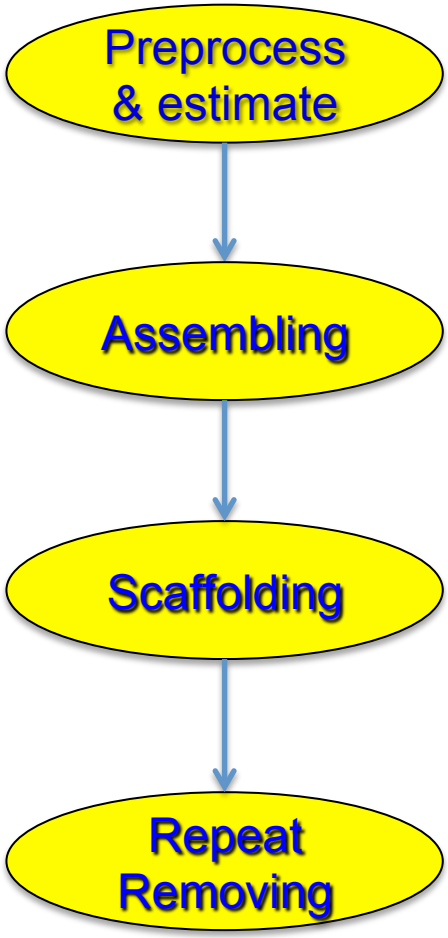
Figure 5. *De Novo* Assembly with Mate Pairs

Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

Strategies

- Coverage. The more, the better. De novo assembly, > 50x. But we usually want to have at least 300x.
- Multiple libraries with different insertion length. 800bp PE, 1kbp MP, 10k MP.

Assembly Pipeline



- **Velvet**: small genomes
- **ABySS**: large genome

Some issues

- For small genome, like bacteria, use Velvet.
- For large genome, use ABySS or Soap-denovo.
- For tools based on the De Bruijn graph, we need to find the optimal length of k-mer.
  - VelvetOptimiser (21-121bp)
  - http://dna.med.monash.edu.au/~torsten/velvet_advisor/
- Try different assemblers for a comparison

Assessing Assembly Quality

- Why do we need QC?
  - Misassembly correction is expensive
  - some assemblers have a simple quality-control method that does not capture larger errors
- Common measures of quality:
  - number and sizes of contigs (N50)
    - Assumption: few large contigs is better than many small contigs.
    - True because there are less gaps in the former, but, does not account for the possibility of misassemblies.
  - And more ..
  - Compare with a complete sequence

Assembly validation

N50 is the most commonly used metric:

Weighted median such as 50% of your assembly is contained in contigs with length >=N50

1. Make a list L of positive integers (contig lengths).
2. Create another list L', which is identical to L, except that every element n in L has been replaced with n copies of itself.
3. The median of L' is the N50 of L.

Assembly validation

For example:
L = {2, 2, 2, 3, 3, 4, 8, 8},
L'={2,2, 2,2, 2,2, 3,3,3, 3,3,3 ,4,4,4,4, 8,8,8,8,8,8,8,8,
8,8,8,8,8,8,8,8}

N50 of L is the median of L'.
N50=(4+8)/2 = 6.

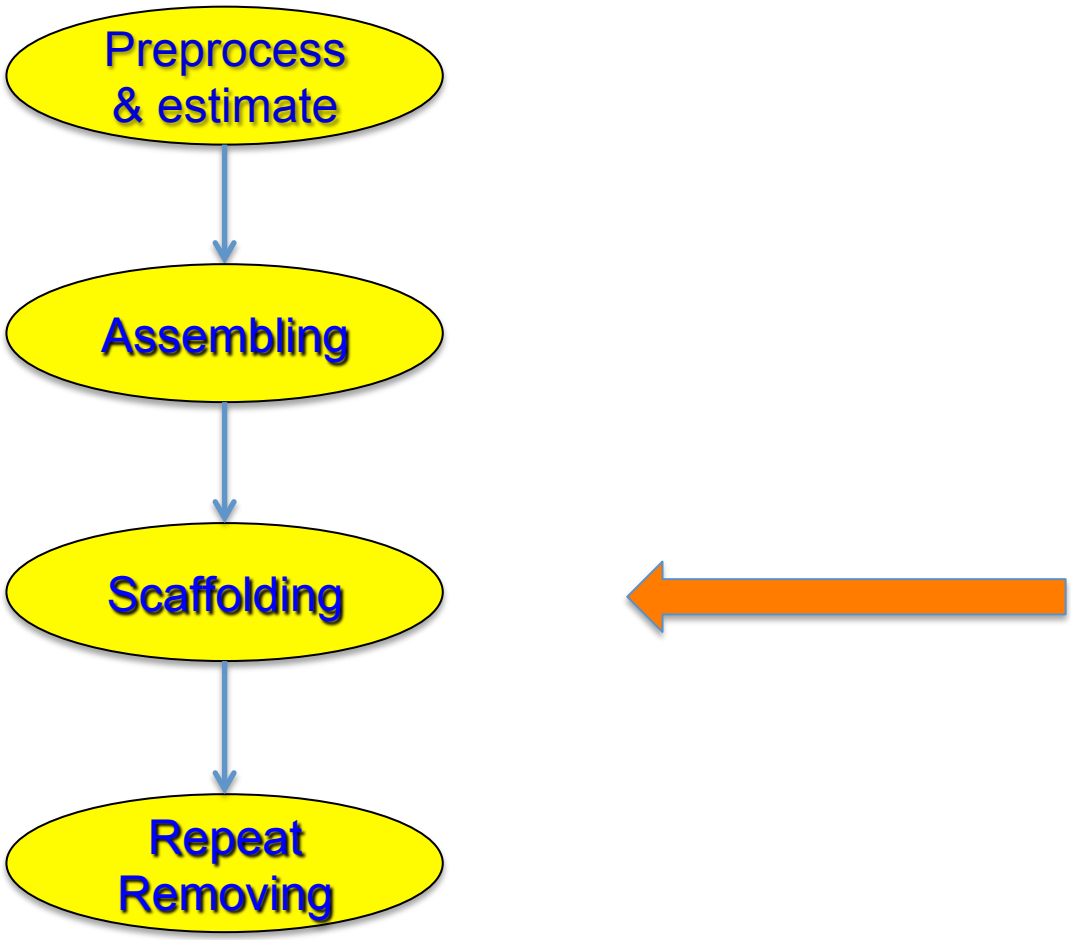
Assembly validation

While the N50 value thus quantifies the ability of the assembly algorithm to combine reads into large seamless blocks, it fails to capture all aspects of assembly quality.

For example, artificially high N50 values can be obtained by lowering thresholds for amalgamating smaller blocks of contiguous reads, resulting in misassembled contigs.

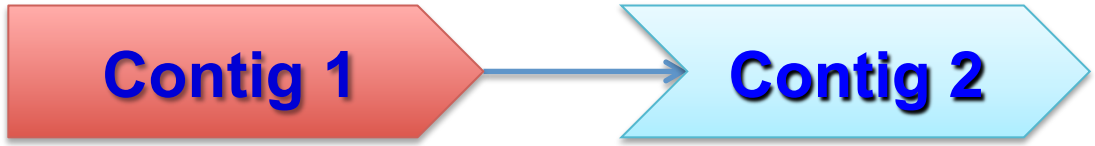
N50 values fail to reflect fine-scale inaccuracies, such as substitution and indel errors.

Assembly Pipeline



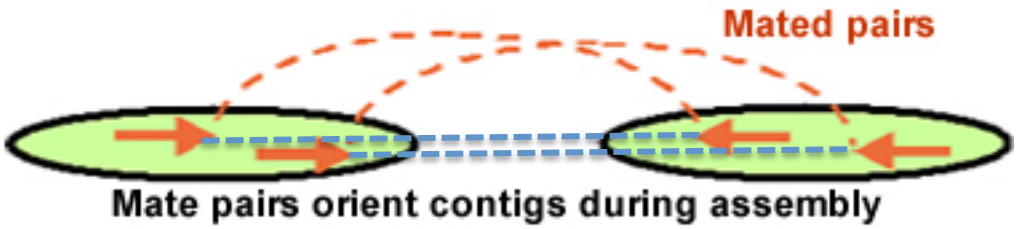
Scaffolding

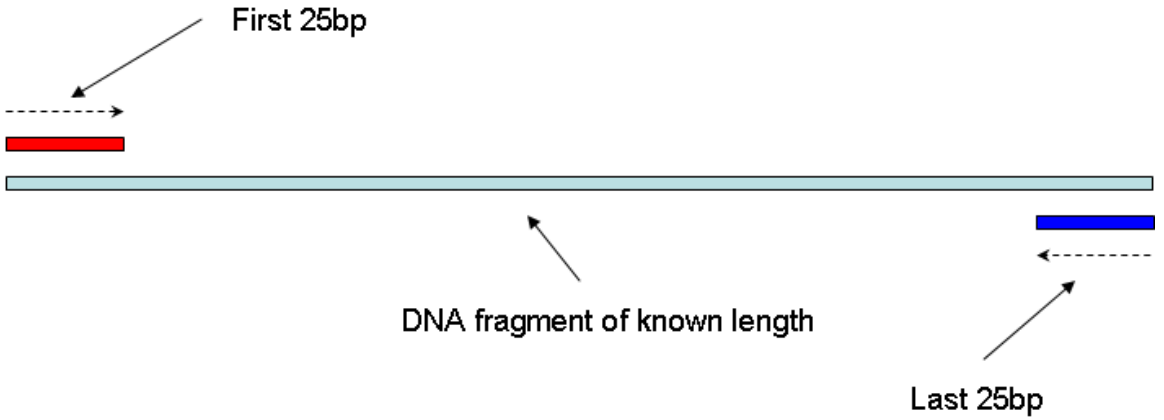
- Scaffolding groups contigs into subsets with known order and orientation.
- Nodes are contigs
- Directed edge is between two nodes if they are adjacent in the genome.



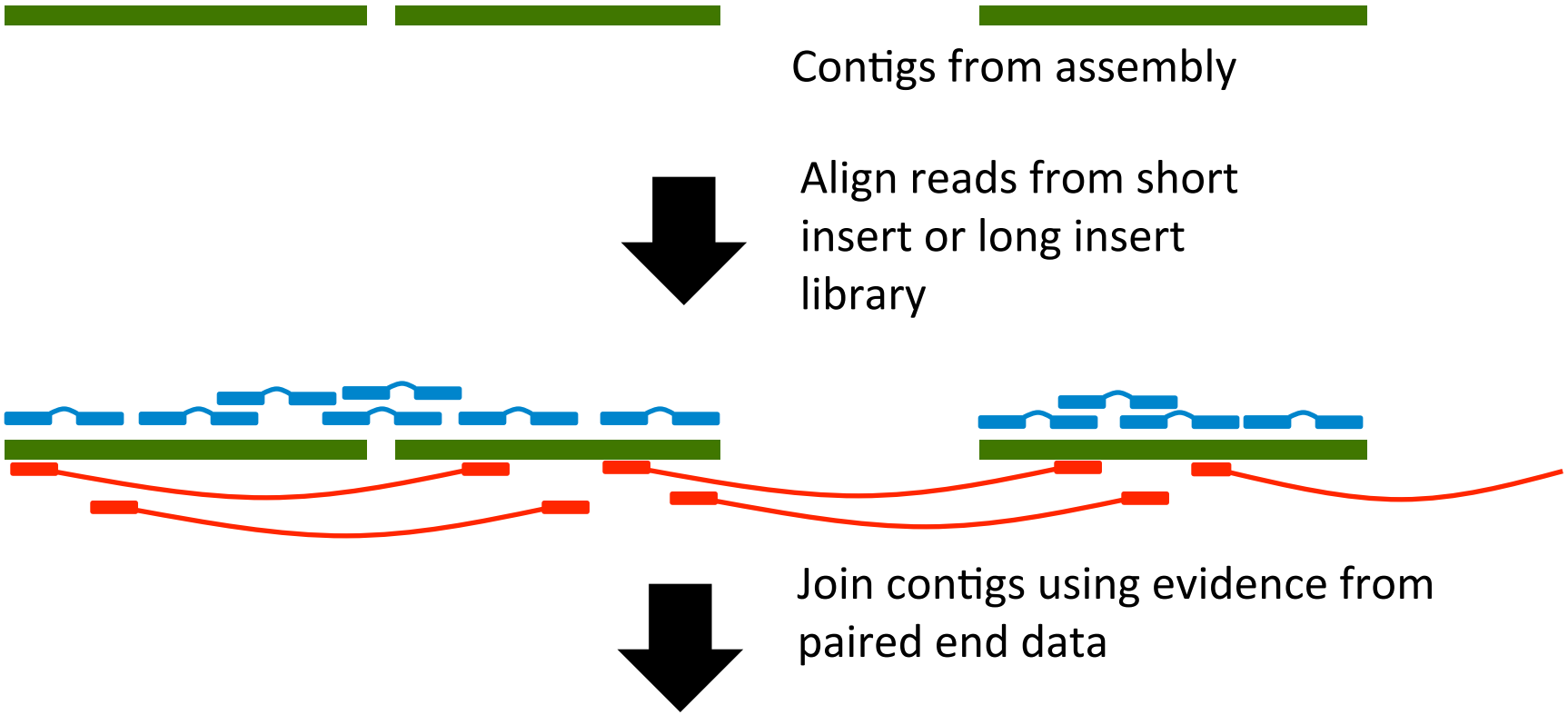
Scaffolding

- Mate pairs , if in different contigs, have a chance of being neighbors.





Scaffolding



Scaffolding Algorithm

- Find all connected components
- Find a consistent orientation for all nodes in the graph (all contigs).
  - Nodes (contigs) have two types of edges
    - Same orientation
    - Different orientation
  - Make sure linked contigs have consistent orientation.
  - Optimization problem – find the smallest number of edges to be removed so that all contigs have consistent orientation.
- Find the Hamiltonian path again.

Scaffolding software

- Some assembly software, such velvet, can do scaffolding as well.
- **Bambus** - http://www.cbcb.umd.edu/software/bambus
- **SSPACE** - http://www.baseclear.com/landingpages/basetools-a-wide-range-of-bioinformatics-solutions/sspacev12/
- **GRASS -** http://code.google.com/p/tud-scaffolding/

Additional techniques for orientation

- Physical mapping. Using information from Bacterial Artificial Chromosome (BAC)-based physical maps. Physical maps are built by clustering together of BACs sharing portions of a DNA "fingerprint," which is a pattern of DNA fragments of various sizes.
- Using markers along a DNA strand as independent information for scaffolding software. Markers are known sequences of nucleotides and tags. Markers are searched in the contigs.
- Using large scale maps of landmarks that lie along the the chromosomal DNA.

Scaffolding

- Additional information is also useful:
  - Sequences of closely related organisms are also used as scaffolding information.

    Example: aligning scaffolds of a mouse genome to the human genome

Scaffolding: Issues

- Errors in length of inserts (affecting distances between clone mates)
- Physical mapping is error prone.
- first builds a sequence based on linking information with high confidence, then factors in linking information with lower confidence.

Assembly Pipeline