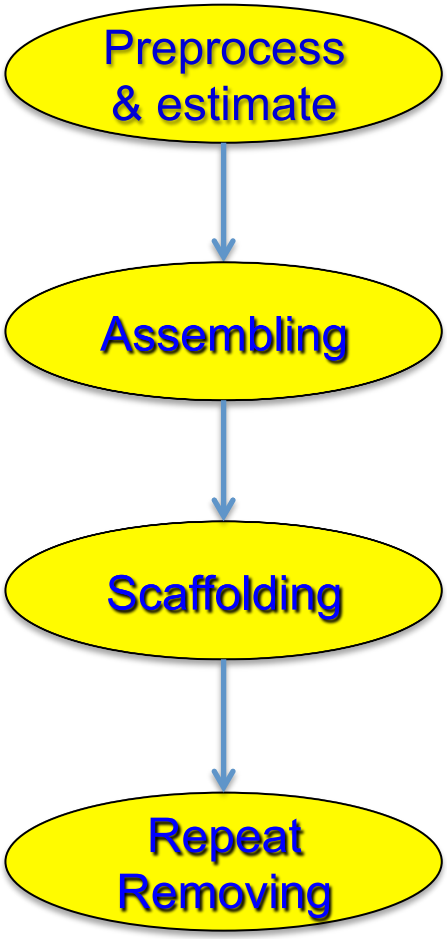
Next-generation sequencing

Lecture 6

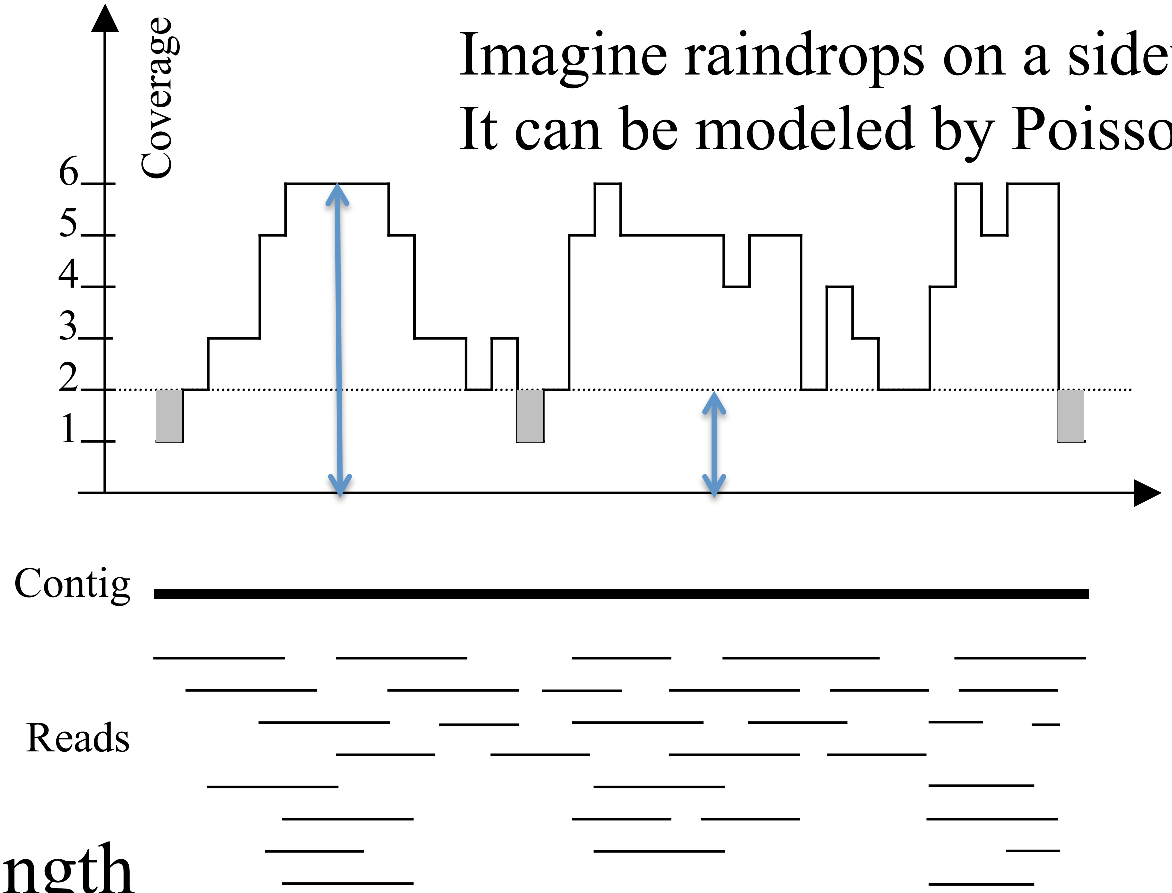
Assembly Pipeline



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

Why?

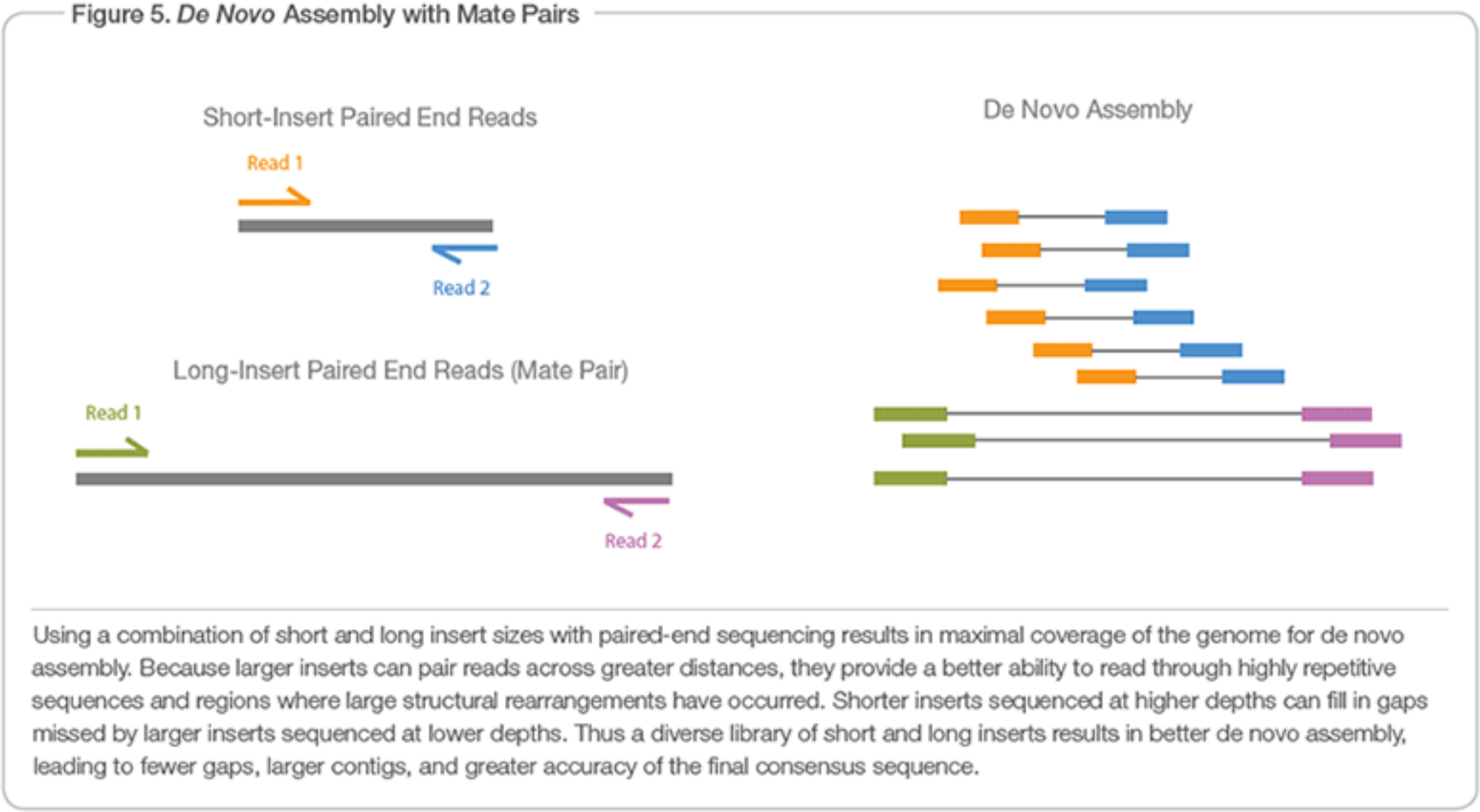
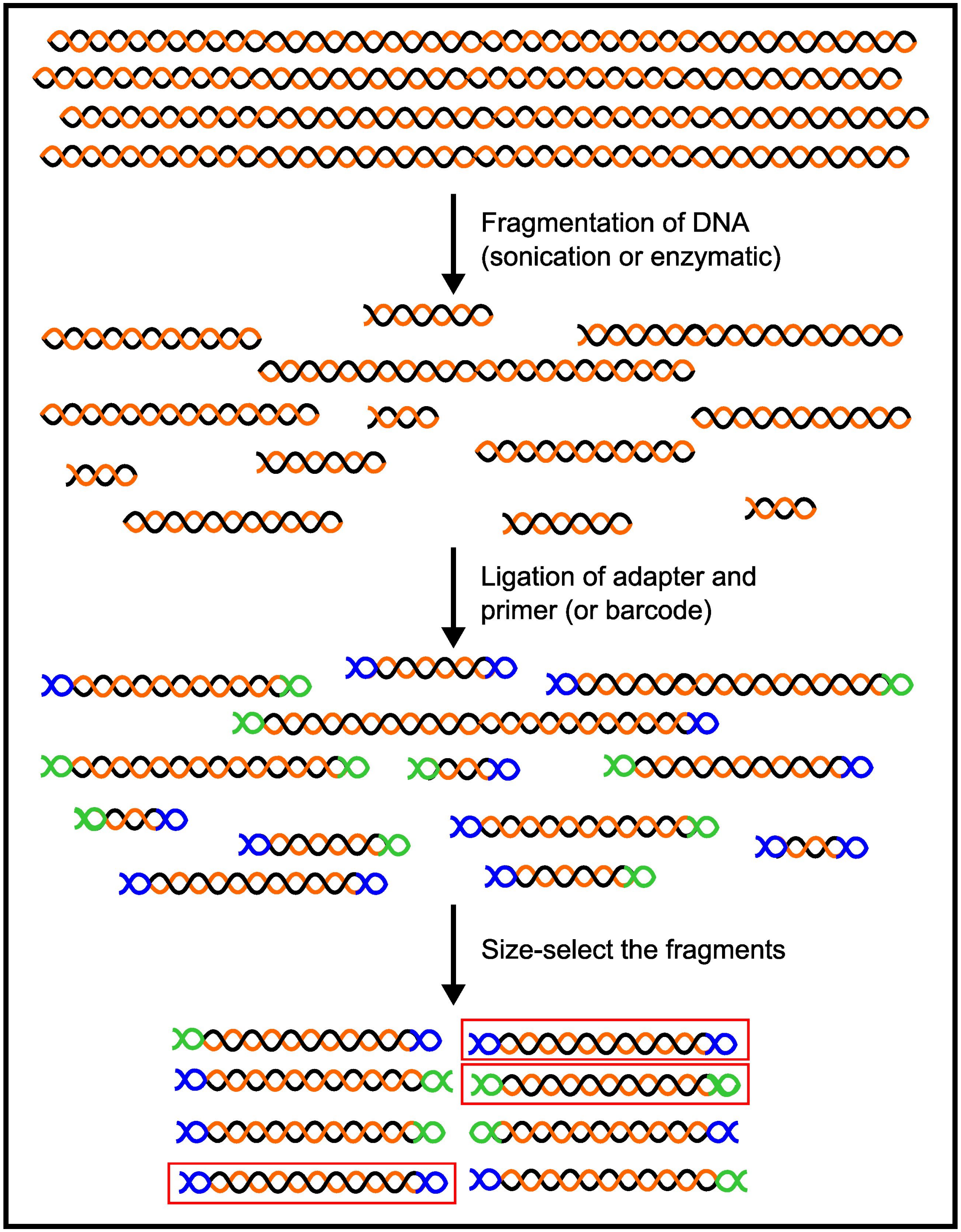


Figure 5. *De Novo* Assembly with Mate Pairs

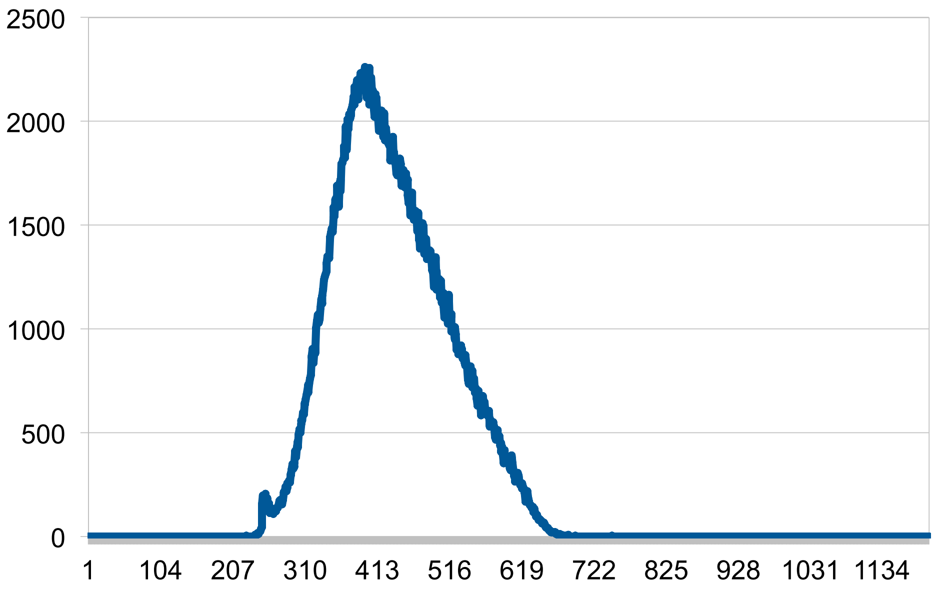
Short-Insert Paired End Reads
Read 1
Read 2
Long-Insert Paired End Reads (Mate Pair)
Read 1
Read 2
De Novo Assembly

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.



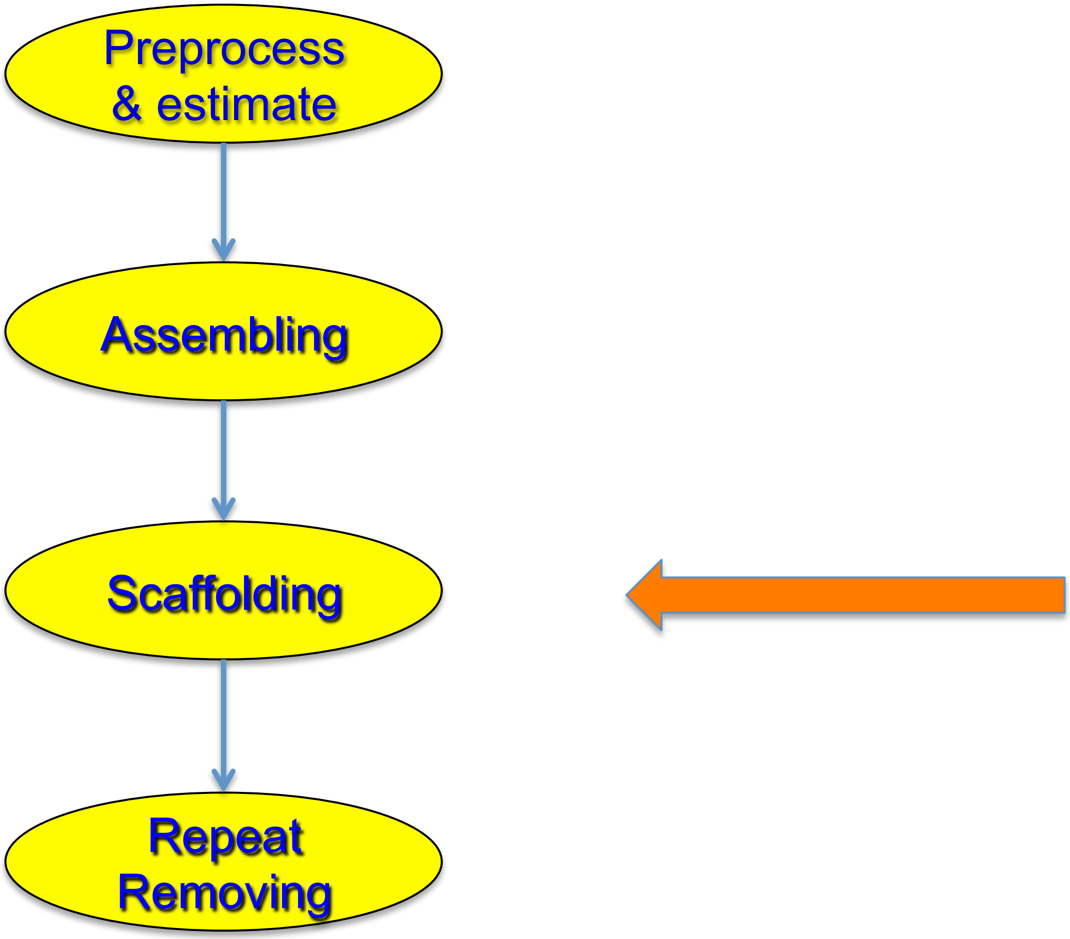
1. fragmenting the DNA (sonication, nebulization, or shearing)
2. DNA repair and end polishing (blunt end, phosphorylated end that is ready for ligation)
3. platform-specific adaptor ligation.
4. Size-selection

Distribution of distances between two paired-end reads



Library size = 400 bp

Assembly Pipeline



Preprocess & estimate

Assembling

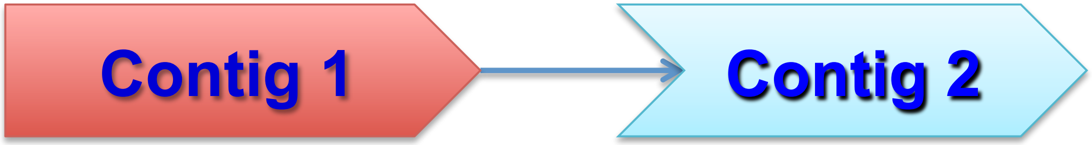
Scaffolding

Repeat Removing

Strain	coverage	# of Reads used	Longest contig	N50	# of contigs	Contigs >500/1000	# of Used contigs
980	40	915274/924368	1578387	1578387	30	20/17	17
982	40	829927/846400	44096	8775	713	608/492	
983	30	681053/696114	26649	6090	938	799/608	
985	30	738515/754370	53527	17916	398	336/287	
988	60	1494832/1509718	219011	87065	84	75/72	
030	60	1345113/1357034	2,004,569	2004569	15	10/8	9
033	34	777226/790462	1,353,777	520746	72	18/12	13
037	18	425061/429622	530,371	203421	30	23/22	23
038	38	846722/855856	1,478,783	477506	15	8/7	8
040	22	496806/502234	1,488,066	520651	35	18/12	12
041	27	656227/664846	1,085,840	905754	14	8/7	8
042	21	544711/554070	481,065	206399	960	34/29	28
043	25	635572/651446	1,092,671	1012472	1244	23/12	13

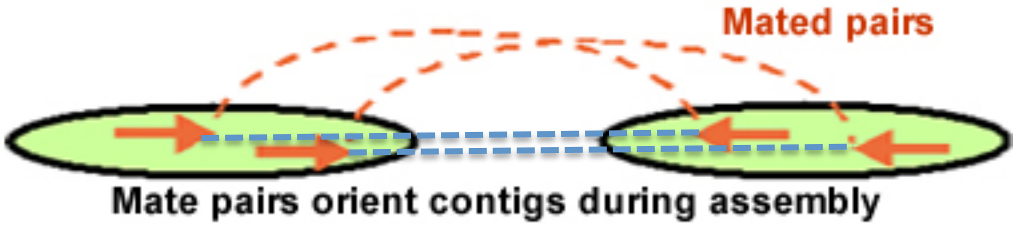
Scaffolding

- 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.

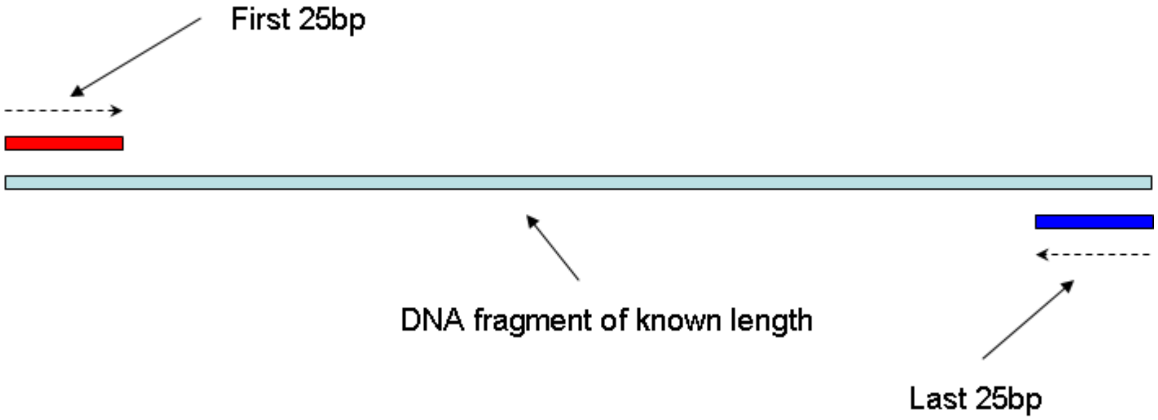


Scaffolding

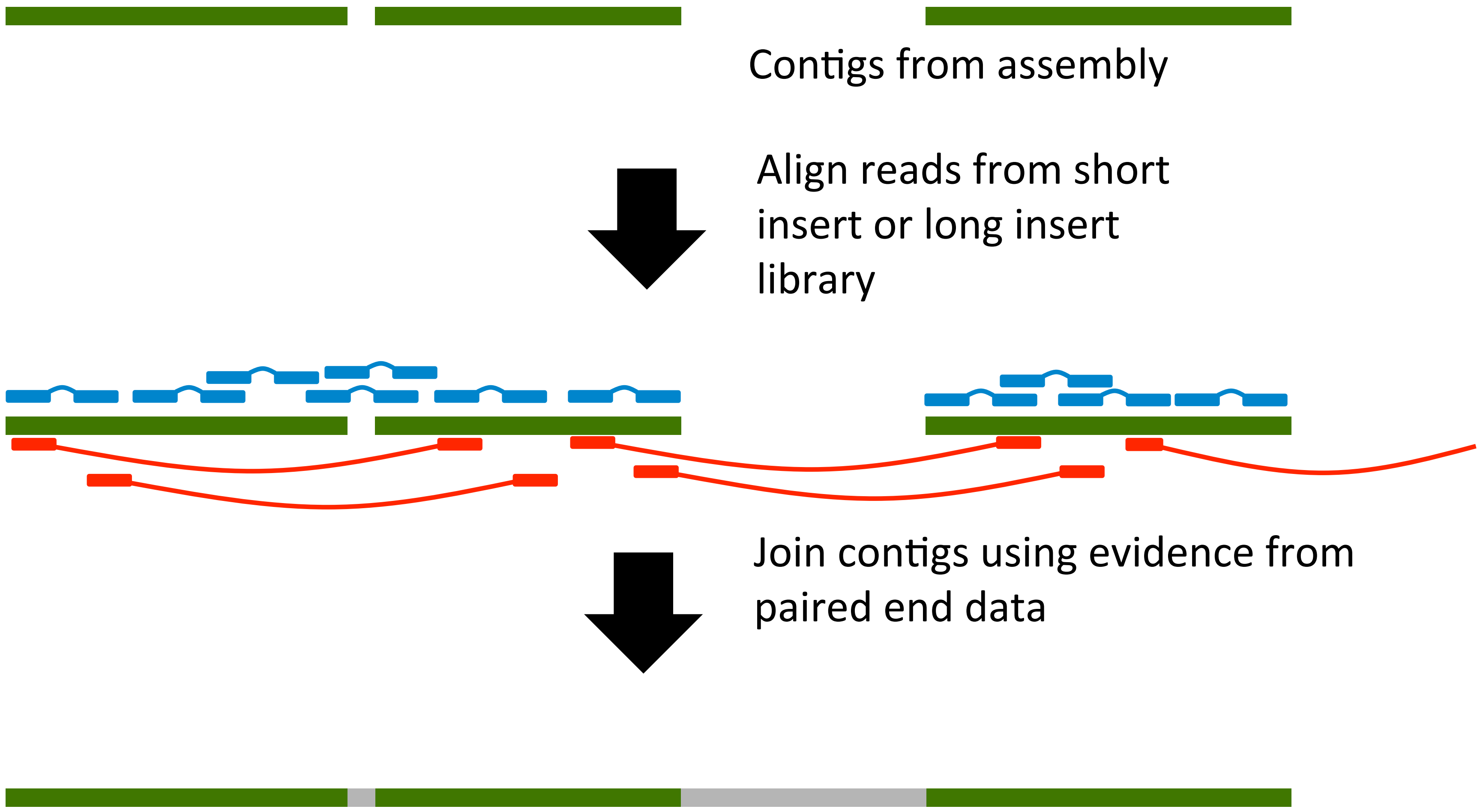
- Mate pairs , if in different contigs, have a chance of being neighbors.



Mate pairs orient contigs during assembly



Scaffolding



Contigs from assembly

Align reads from short insert or long insert library

Join contigs using evidence from paired end data

Scaffold

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/
- Volvet and Soap-denovo have buid-in scaffolding tools.

Additional techniques for orientation

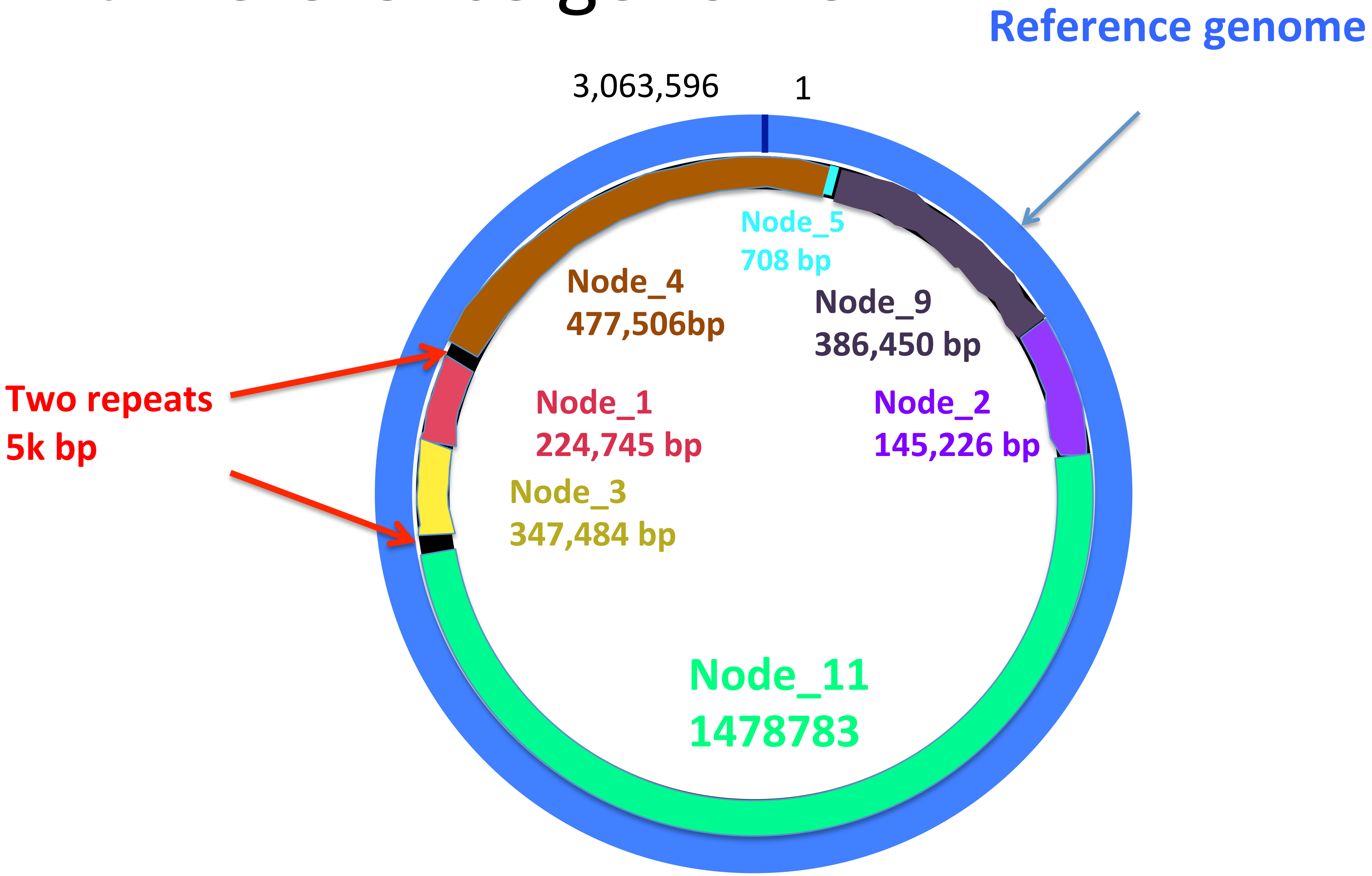
- **Physical mappin**g. Using information from Bacterial Artificial Chromosome (BAC)-based physical maps. Physical maps are built by clustering together of BACs sharing portions of a DNA "fingerprint," which is a pattern of DNA fragments of various sizes.
- Using **markers** along a DNA strand as independent 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

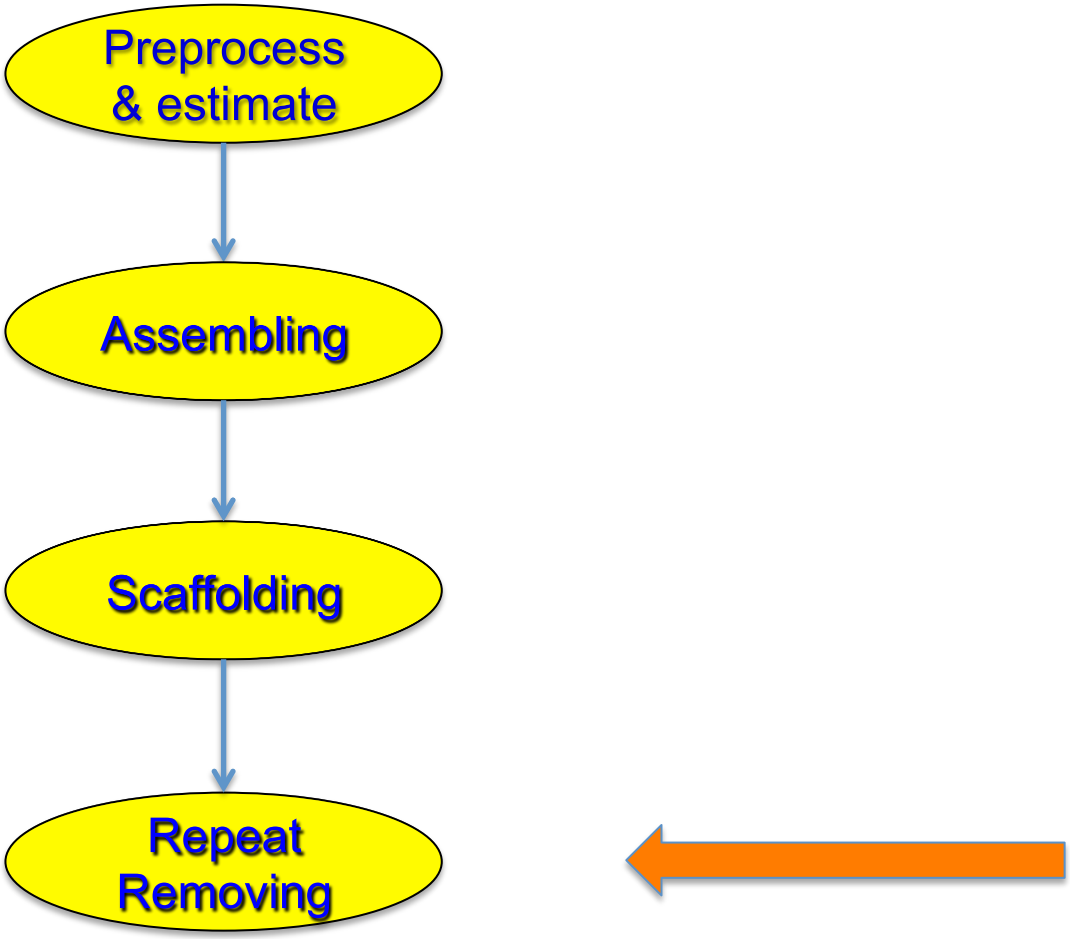
With reference genome



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.

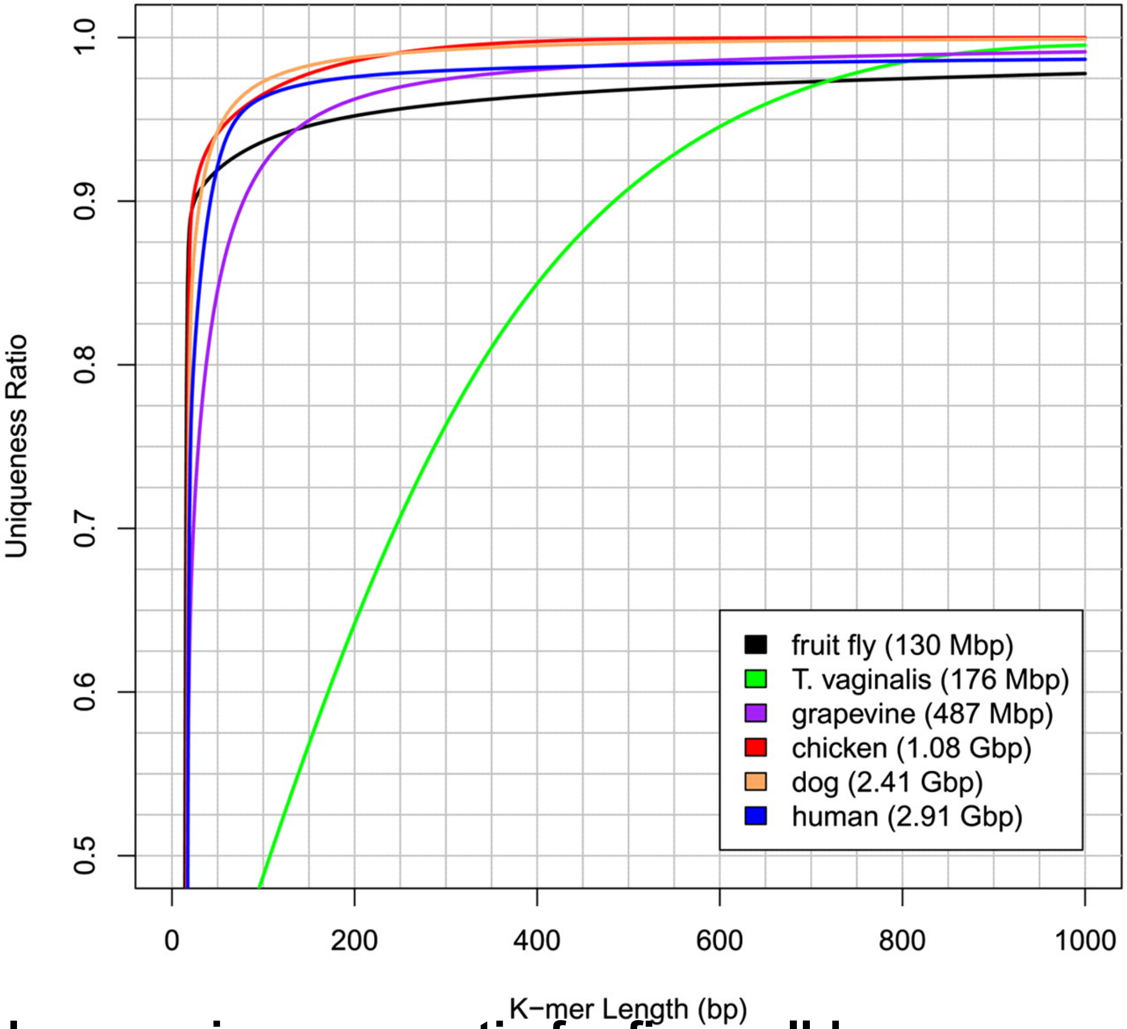
Assembly Pipeline



The variability in repetitiveness among species species.

The ratio == the percentage of the genome that is covered by unique sequences of length k or longer.

The figure shows how much of each genome would be covered by *k*-mers (reads) that occur exactly once.



The k-mer uniqueness ratio for five well-known organisms and one single-celled human parasite.

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

Repeat Control Issues

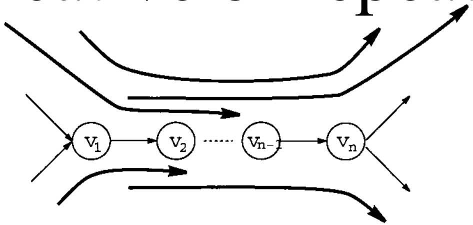
- Assembly programs should detect repeats in the assembly process and not after.
  - Incorrect genome reconstruction

- Assemblers should try to resolve correctly as many repeats as possible.
  - Avoid intensive human labor

Repeat Control – When? & How?

- **pre-assembly:** find fragments that belong to repeats
  - statistically (most existing assemblers)
  - repeat database (*RepeatMasker*)

- **during assembly:** detect "tangles" indicative of repeats (Pevzner, Tang, Waterman 2001)



- **post-assembly:** find repetitive regions and potential mis-assemblies.
  - *Reputer, RepeatMasker*
  - "unhappy" mate-pairs (too close, too far, mis-oriented)

Detecting repeats **pre-assembly:**

- Statistical methods
  - Assemblers assume that reads are sampled uniformly at random.
  - Significant deviations from average coverage flagged as repeats.
  - frequent k-mers are ignored
  - “arrival” rate of reads in contigs compared with theoretical value.

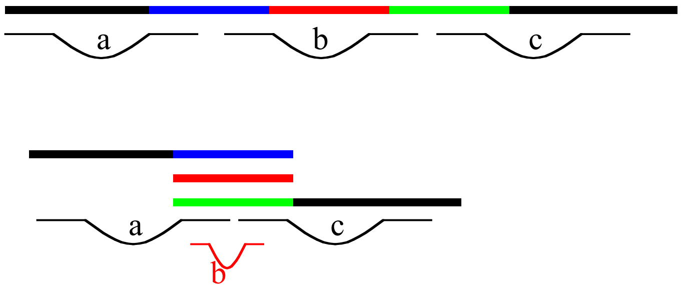
(e.g., 800 bp reads & 8x coverage - reads "arrive" every 100 bp)

Detecting repeats **during assembly**

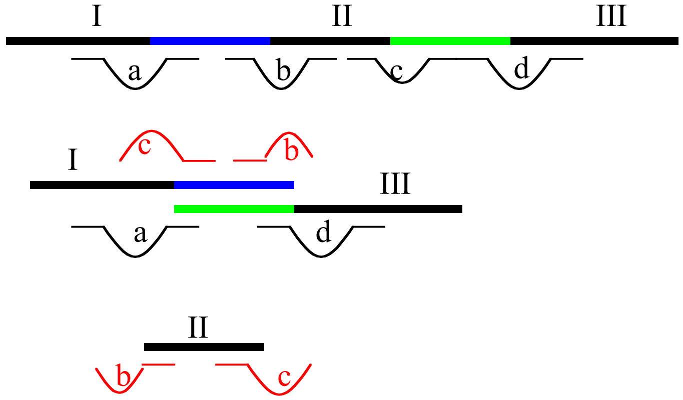
- Example: In Euler assembly program
  - Finds repeats by complex parts of the graph constructed during the assembly process.
  - Researchers look into these complex areas to try and resolve repeats.
  - Assemblers can use clone mate information to find incorrect assemblies. This is based on finding clone-mate pairs too close or too far from one another. (“unhappy” mate-pairs)

Detecting repeats **post-assembly:** Mis-assembled repeats

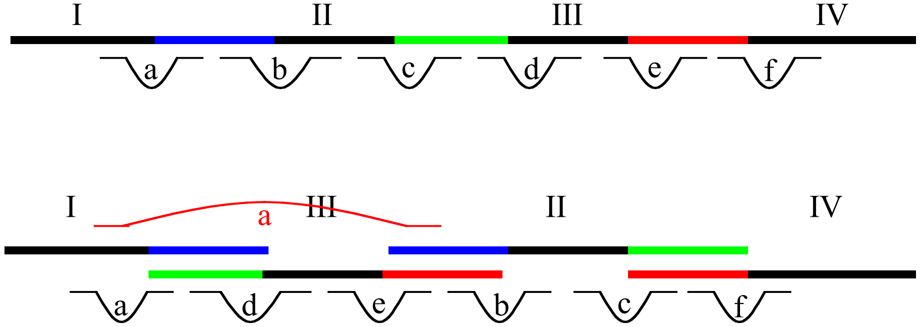
collapsed tandem



excision



rearrangement



Repeat resolution

- Assemblers deduce that areas covered by a large number of reads may show an over-collapsed repeat.
- Problems with this - samples are not uniformly distributed (for example, non-random libraries and poor clonability regions). leads to false positives.
- Repeats with low copy number are missed - leads to false negatives.

Repeat resolution

- Techniques for repairing sequencing errors during repeat resolution
  - find clusters of reads where the clusters share differences.
    - For example, four reads contain an A , four contain a B. it is likely that the first four reads are from one copy and the last four from a different one.
  - Drawbacks are if certain areas of the sequence have low coverage.
  - Difficult to separate from true polymorphism

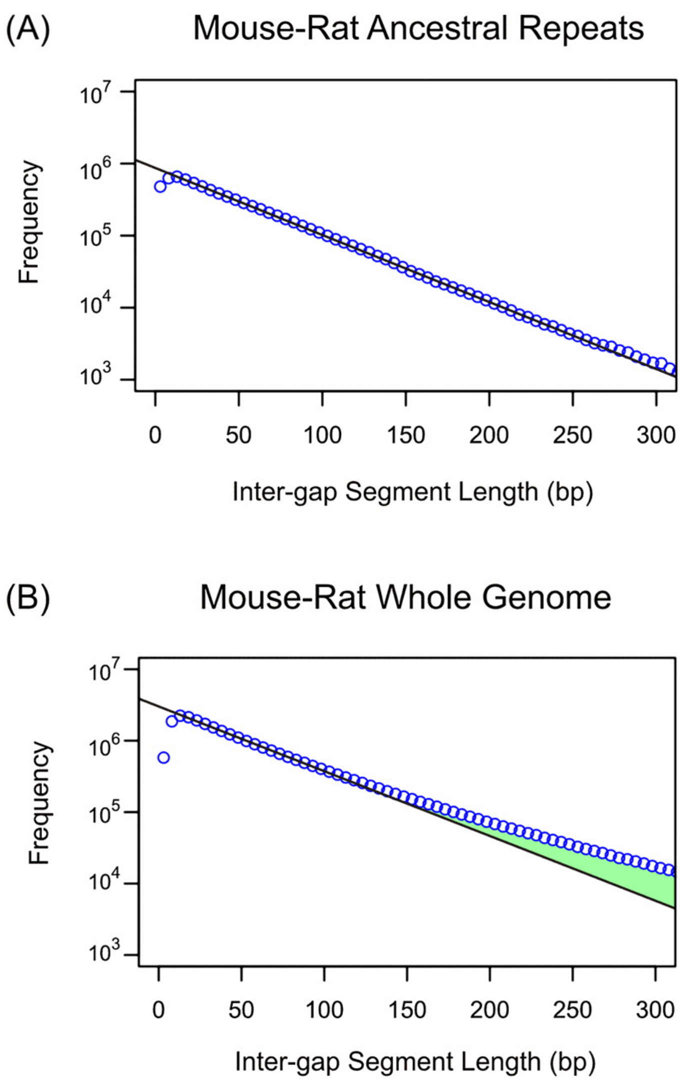
Assembled genome validation

- Quality at the nucleotide level for contigs can be used to detect fine-scale inaccuracies, such as substitution and indel errors.
- Method 1: Once assembled, a base is assigned a consensus quality score (CQS) depending on its read depth and the quality of each base contributing to that position. (Huang and Madan 1999, Genome Research, 9: 868–877).
- Method 2: A multiple sequence alignment of reads is constructed and a consensus sequence along with a quality value for each base is computed for each contig.

Assembled genome validation

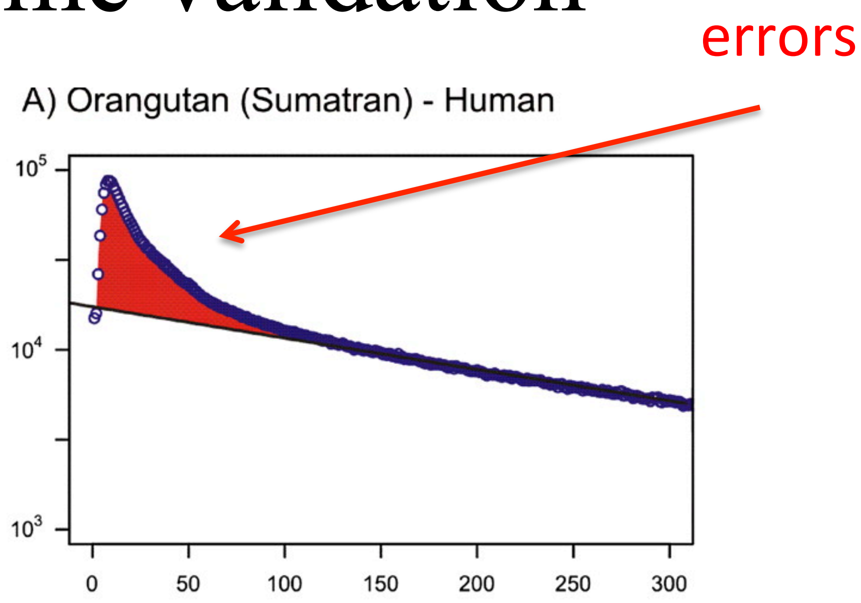
- Method 3: a statistical and comparative genomics method that quantifies the fine-scale quality of a genome assembly and that has the merit of being complementary to the aforementioned approaches.
- This approach estimates the abundance of indel errors between aligned genome pairs, by separating these from true evolutionary indels.
- indel mutations leave a precise and determinable fingerprint on the distribution of ungapped alignment block lengths. These block lengths, which represent distances between successive indel mutations are intergap segment (IGS) lengths.

Assembled genome validation



(A) Mouse-Rat Ancestral Repeats

(B) Mouse-Rat Whole Genome



errors

A) Orangutan (Sumatran) - Human

Under the **neutral indel model**, these inter-gap segment (IGS) lengths are expected to follow a geometric frequency distribution.

Meader et al., Genome Research, 2010, 20(5):675

Assembled genome validation

Compare with existing genes.

CEGMA: Core Eukaryotic Genes Mapping Approach

- Looks in your assembly for genes that should be there
- Usually best assembly have best CEGMA score

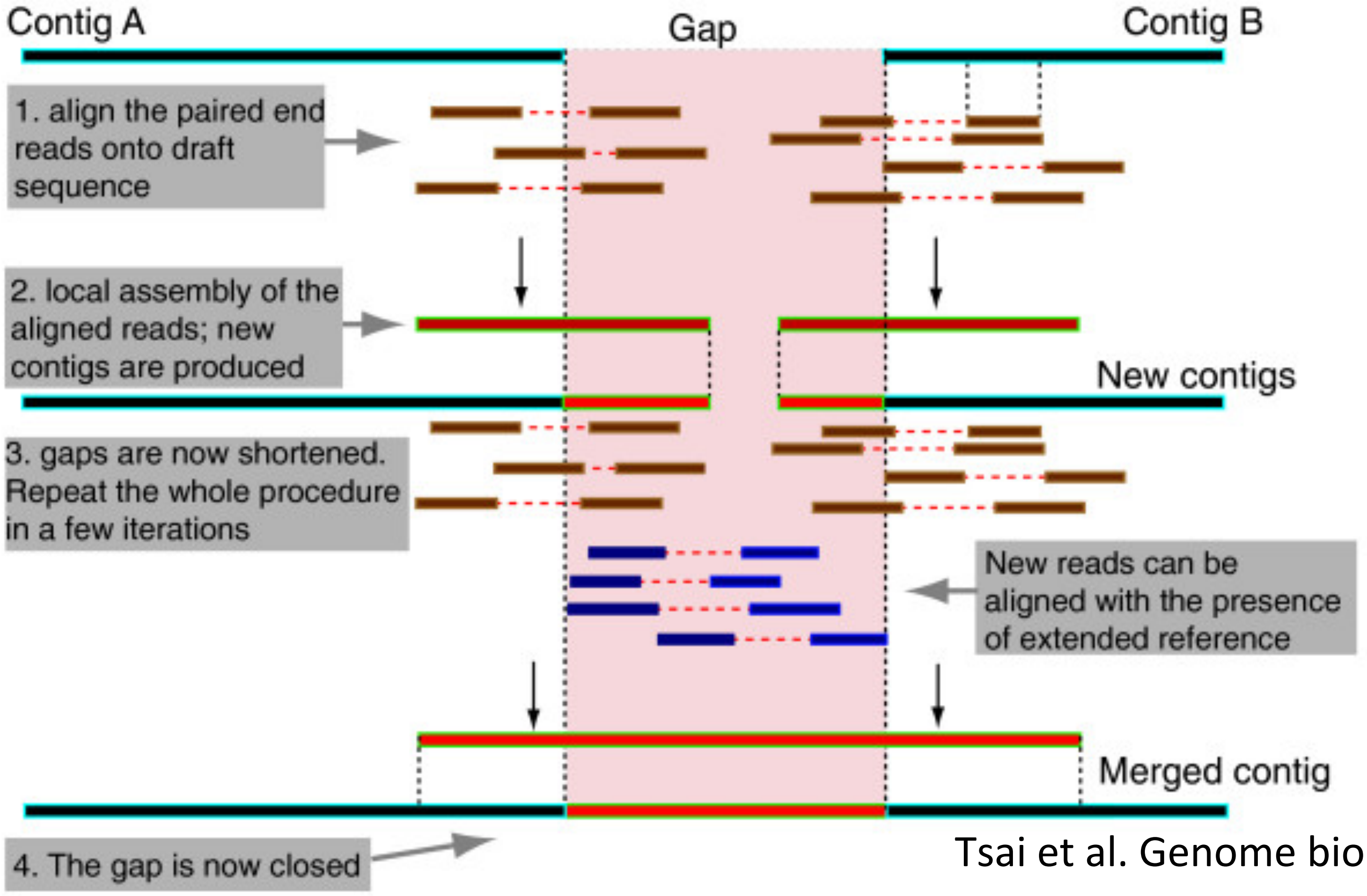
http://korflab.ucdavis.edu/datasets/cegma/

What makes an assembly good?

- High coverage: 50 to 300X
- Different but precise insert size libraries (Paired end from different library sizes will allow you to stitch across several repeat type.)
- Avoid large number of variant.
- Error Correction: Correct the read before assembly

What makes your assembly better?

IMAGE: Gap Filling. improve draft genome assemblies by aligning sequences against contig ends and performing local assemblies to produce gap-spanning contigs.


Contig A
Gap
Contig B
1. align the paired end reads onto draft sequence
2. local assembly of the aligned reads; new contigs are produced
New contigs
3. gaps are now shortened. Repeat the whole procedure in a few iterations
New reads can be aligned with the presence of extended reference
Merged contig
4. The gap is now closed

Tsai et al. Genome biology 2010

Gene annotation

- RAST http://rast.nmpdr.org/
- IGS Prokaryotic Analysis Engine Services(http://ae.igs.umaryland.edu/cgi/index.cgi)
- AGeS http://www.bhsai.org/ages.html
- BG7 http://bg7.ohnosequences.com/
- **Prokka** http://www.vicbioinformatics.com/software.prokka.shtml

Gene annotation

Prodigal (Hyatt 2010)	**gene prediction and Coding sequence (CDS)**
RNAmmer (Lagesen et al., 2007)	**rRNA genes**
Aragorn (Laslett et al, 2004)	**Transfer RNA genes**
SignalP (Petersen et al., 2011)	**Signal leader peptides**
Infernal (Kolbe and Eddy, 2011)	**Non-coding RNA**

Protein function annotation

Databases:

(1) Bacterial proteins in UniProt and RefSeq

(2) Protein domains in Pfam and TIGRFAMs

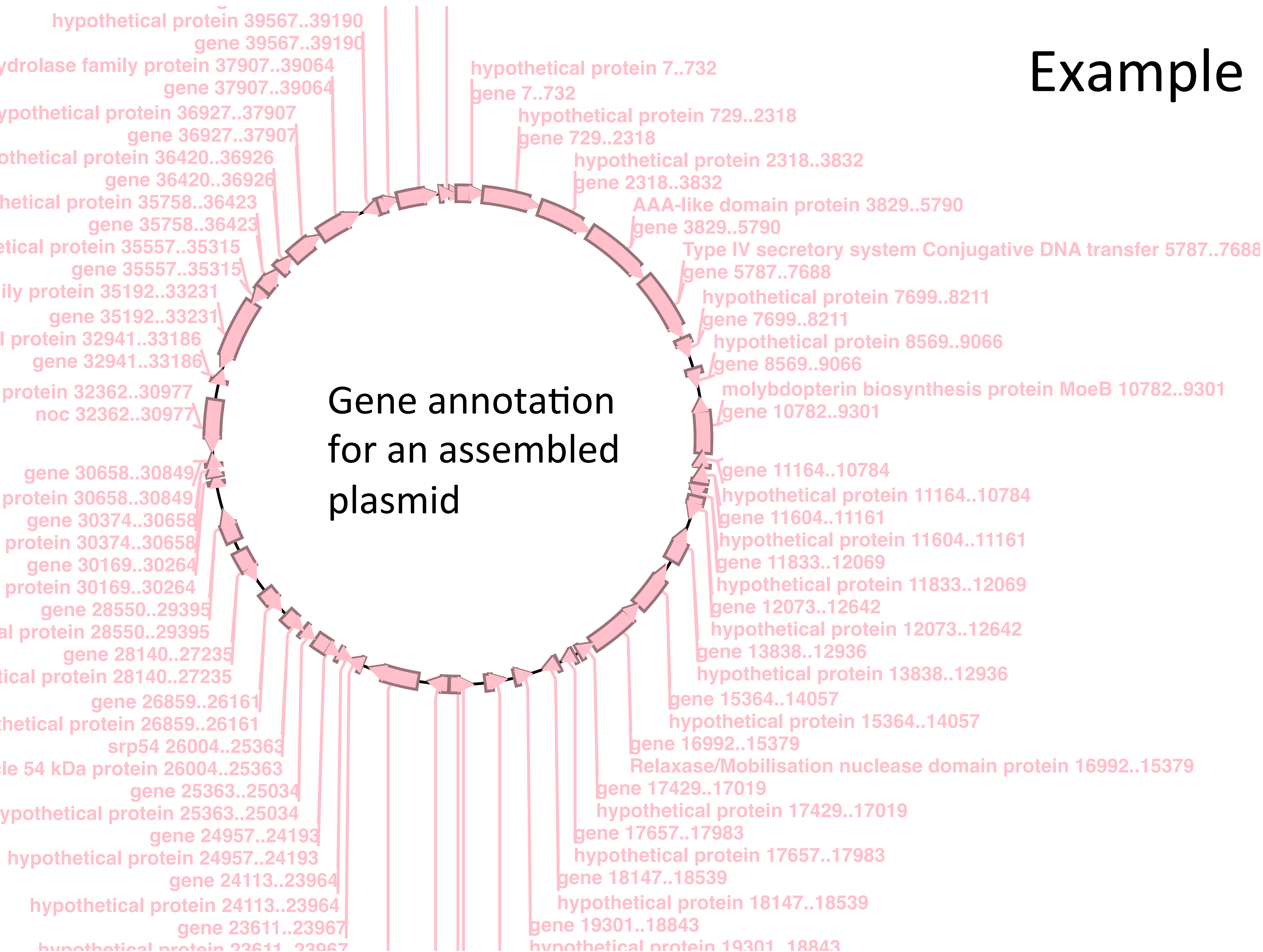
Searching tools:

(1) Blastp

(2) Hidden Markov Model (HMMER 3.0)

Strain	# of Used contigs	bases	genes	CDS	Misc_RNA	tRNA	tmRNA
980	13	3036852	2922	2862	9	50	1
030	8	3058742	2942	2883	8	50	1
033	11	3058510	2936	2882	8	45	1
037	23	3058614	2945	2887	9	48	1
038	8	3066896	2949	2892	8	48	1
040	13	3062944	2954	2896	8	49	1
041	8	3059145	2939	2886	8	44	1
043	13	3056218	2947	2891	8	47	1
Reference genome		3,063,006		2817		46	

Example


hypothetical protein 39567..39190
gene 39567..39190
ydrolase family protein 37907..39064
gene 37907..39064
ypothetical protein 36927..37907
gene 36927..37907
othetical protein 36420..36926
gene 36420..36926
hetical protein 35758..36423
gene 35758..36423
etical protein 35557..35315
gene 35557..35315
ily protein 35192..33231
gene 35192..33231
l protein 32941..33186
gene 32941..33186
protein 32362..30977
noc 32362..30977
gene 30658..30849
protein 30658..30849
gene 30374..30658
protein 30374..30658
gene 30169..30264
protein 30169..30264
gene 28550..29395
al protein 28550..29395
gene 28140..27235
tical protein 28140..27235
gene 26859..26161
hetical protein 26859..26161
srp54 26004..25363
le 54 kDa protein 26004..25363
gene 25363..25034
ypothetical protein 25363..25034
gene 24957..24193
hypothetical protein 24957..24193
gene 24113..23964
hypothetical protein 24113..23964
gene 23611..23967
hypothetical protein 23611..23967
hypothetical protein 7..732
gene 7..732
hypothetical protein 729..2318
gene 729..2318
hypothetical protein 2318..3832
gene 2318..3832
AAA-like domain protein 3829..5790
gene 3829..5790
Type IV secretory system Conjugative DNA transfer 5787..7688
gene 5787..7688
hypothetical protein 7699..8211
gene 7699..8211
hypothetical protein 8569..9066
gene 8569..9066
molybdopterin biosynthesis protein MoeB 10782..9301
gene 10782..9301
gene 11164..10784
hypothetical protein 11164..10784
gene 11604..11161
hypothetical protein 11604..11161
gene 11833..12069
hypothetical protein 11833..12069
gene 12073..12642
hypothetical protein 12073..12642
gene 13838..12936
hypothetical protein 13838..12936
gene 15364..14057
hypothetical protein 15364..14057
gene 16992..15379
Relaxase/Mobilisation nuclease domain protein 16992..15379
gene 17429..17019
hypothetical protein 17429..17019
gene 17657..17983
hypothetical protein 17657..17983
gene 18147..18539
hypothetical protein 18147..18539
gene 19301..18843
hypothetical protein 19301..18843

Gene annotation for an assembled plasmid

Discussion:
Virtual genome assembly

- Plant mitochondrion genome 500,000 bp DNA circular
- How can you get mitochondria DNA? What problems do we need to concern for this step?
- For DNA fragmenting, what sizes of DNA fragments will you use? A. 1Kbp, B. 5kbp, C. both
- Pair-ended or single ended?
- What depth do you sequence? how many lanes do you need if you use illumina hiseq 2000? or how many reads do you need to get?
- Which assembler will you use? Why?
- What computer do you used to do assemble? A. 4GB laptop B. 50GB workstation C. computer cluster in HCC
- According to your estimate, how long does it take for assemble? A. 30 minutes B.2 hours C. 12 hours D. 4 days
- What software do you used to do scaffold? how long does it take?
- What is longest gap in one scaffold? How do you fill gaps?
- How do you determine if your assembled genome is good enough?
- how do you annotate genes?

On Thursday