Libraries and Prep for next generation sequencing

4/6/16

Questions from last class?

You have discovered a new species of fish on an expedition into a cave system in Mexico

- What do you want to learn about this organism?
- How would you go about learning those things?

https://www.wired.com/2016/03/dissecting-blind-cave-fish-walks-like-salamander/

Things we might want to know

- What other fish is it most closely related to?
- What gene expression changes allow it to grow in this unusual environment?
- What does it feed on in the environment?

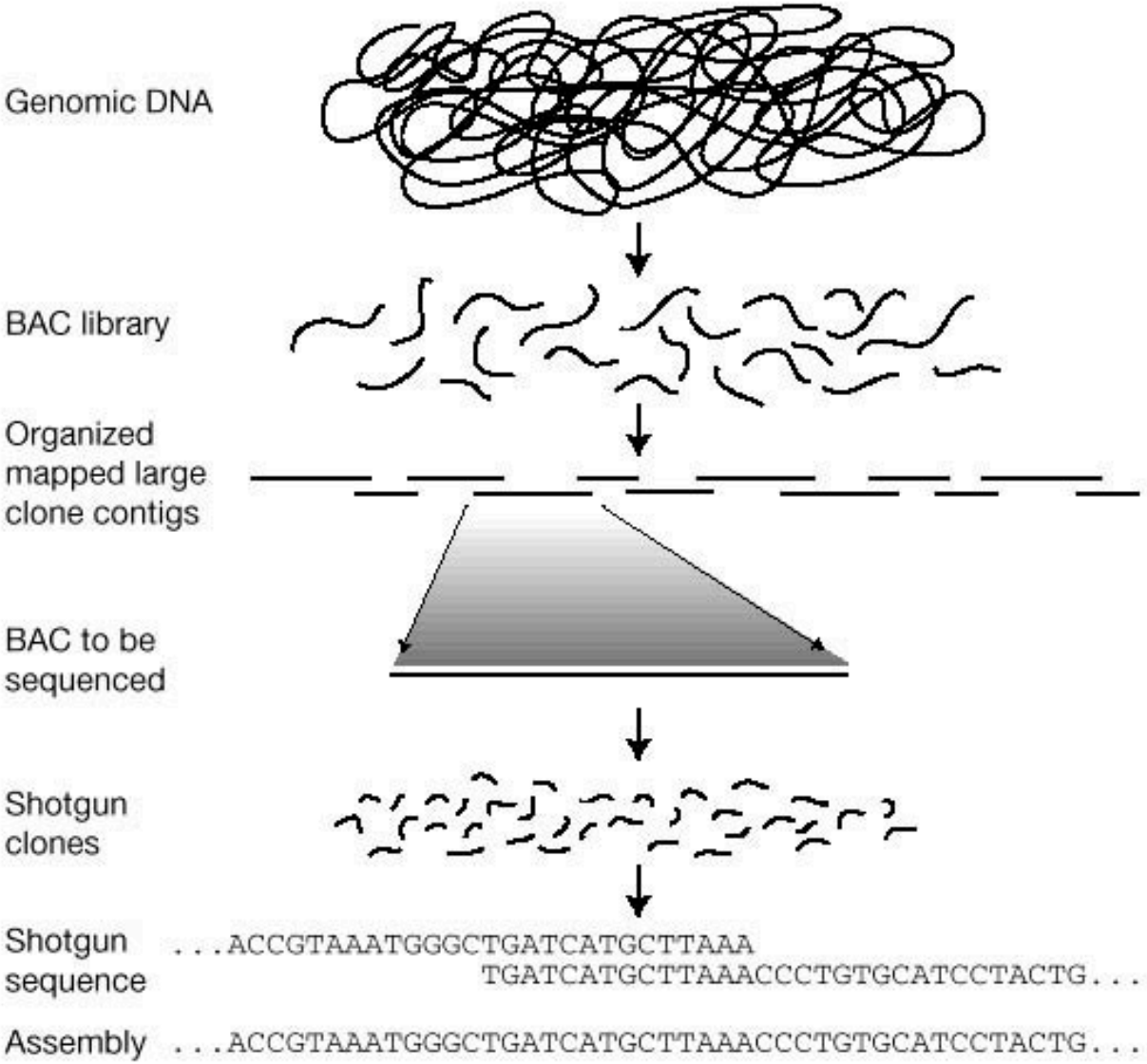
Similarity to other species - Phylogeny

- How would you address?

Sequence whole genome

- Haploid genomes of bony fish range from ~850Mbp to 3000Mbp (humans have ~3000Mbp genome)
- You determine your new species has ~2000Mbp.

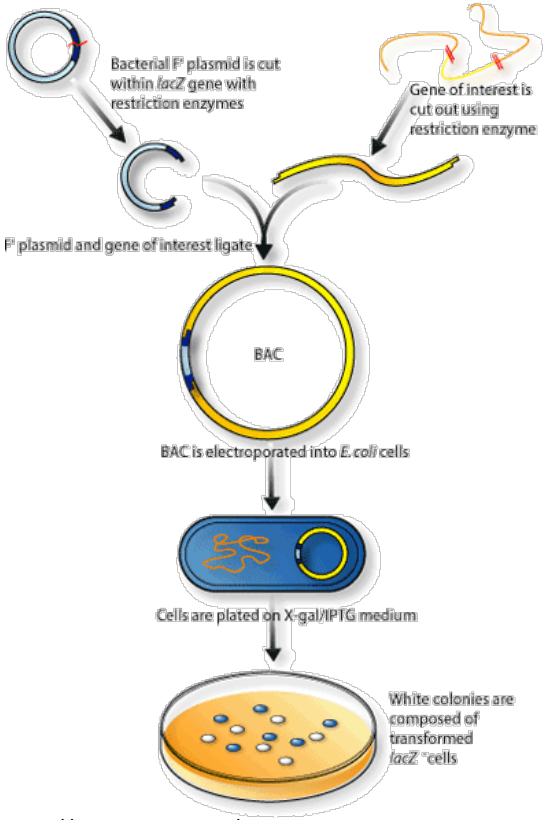
Hierarchical Shotgun Sequencing



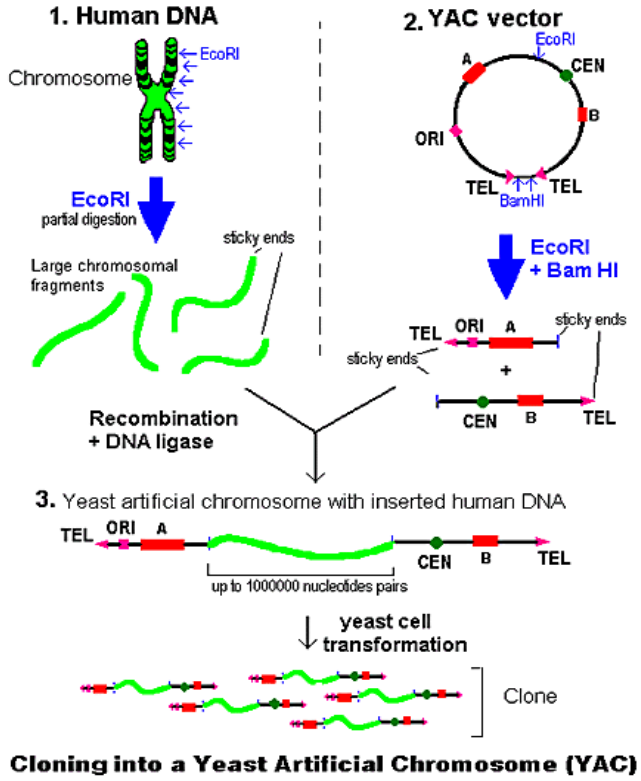
International Human Genome Sequencing Consortium; *Nature* 409, 860-921 (15 February 2001) | doi:10.1038/35057062

What is an artificial chromosome?

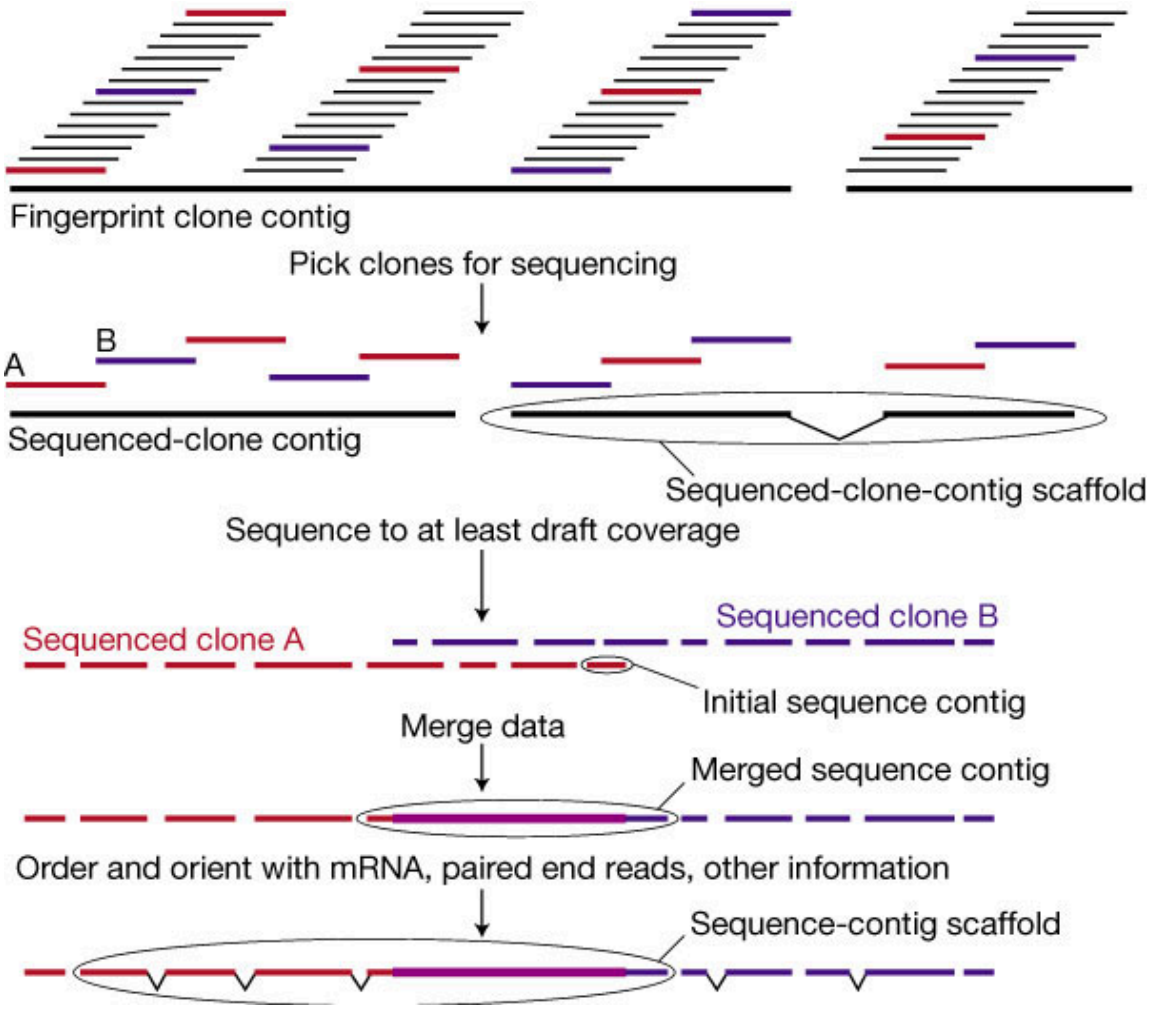
- What does a bacterial chromosome need to reproduce inside of a bacteria?
- What does a yeast chromosome need?



http://www.scq.ubc.ca/the-big-bad-bac-bacterial-artificial-chromosomes/



http://www.ebioworld.com/2011/08/yeast-artificial-chromosome-yac-vectors.html

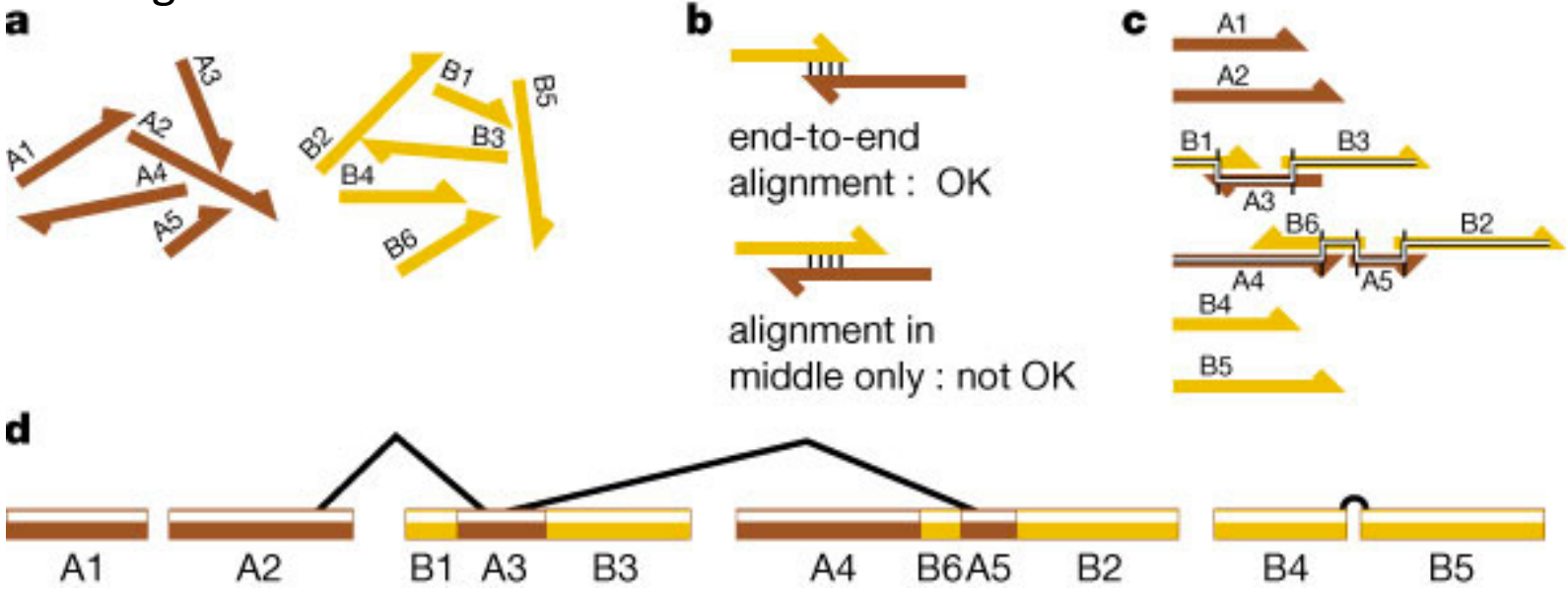


International Human Genome Sequencing Consortium; *Nature* 409, 860-921 (15 February 2001) | doi: 10.1038/35057062

Restriction digests of large scale clones to predict which clones overlap and give good coverage - fingerprinting

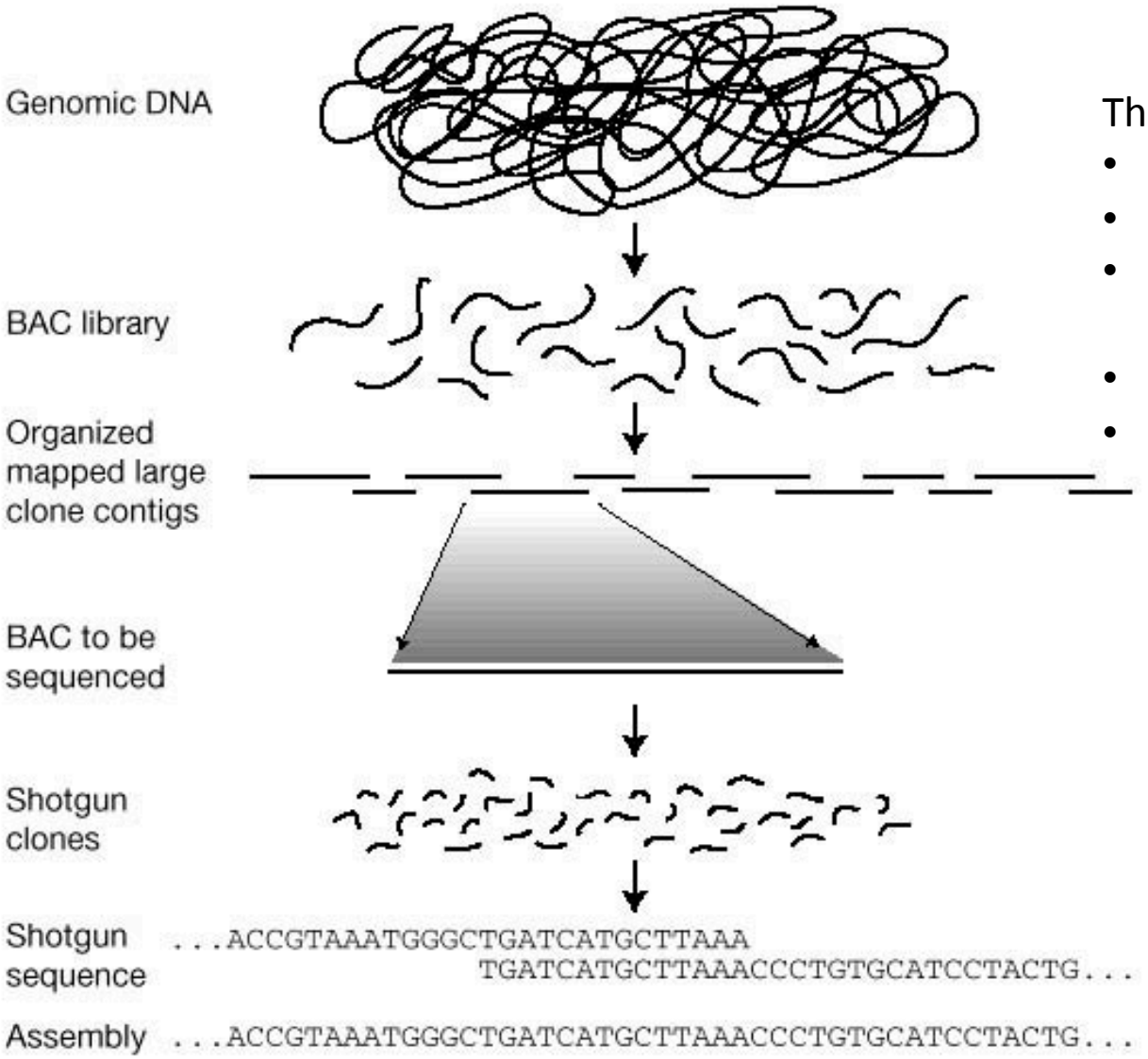
To figure out which contig follows the other isn't straightforward

Aligning the sequenced fragments requires LOTS of computation and some guidelines



International Human Genome Sequencing Consortium; *Nature* 409, 860-921 (15 February 2001) | doi: 10.1038/35057062

Hierarchical Shotgun Sequencing



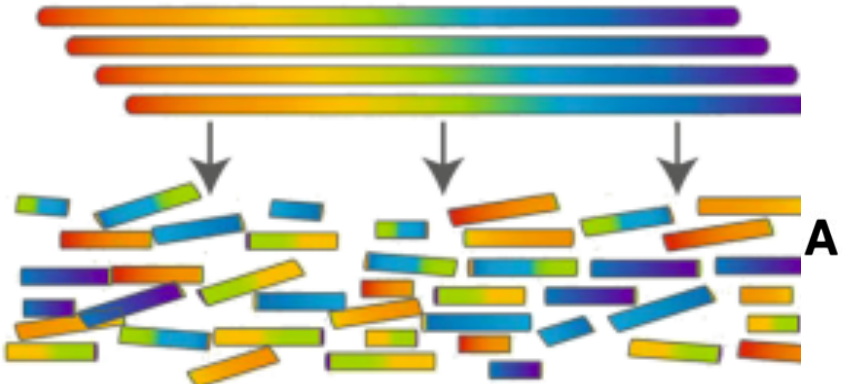
This requires lots of:

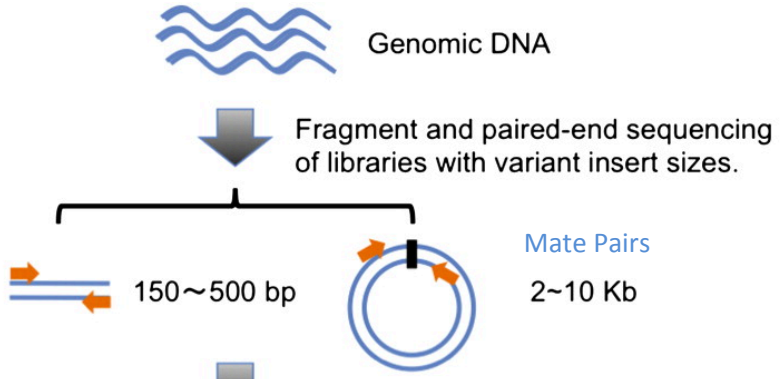
- Cloning
- Restriction enzyme digests
- Sequencing of individual clones
- TIME
- MONEY

International Human Genome Sequencing Consortium; *Nature* 409, 860-921 (15 February 2001) | doi:10.1038/35057062

Sanger Sequencing	Second Generation Sequencing
Long pieces 900-1500 bp	Short pieces 300-500 bp
One at a time	Many at a time
Need lots of DNA • Plasmid DNA 500ng • PCR products 10ng	Need less DNA • 1-1.5 µg (total)
Usually from a plasmid, BAC, or YAC	Directly from organism and fragmented into size appropriate for sequencing
Need known primer to start sequencing by synthesis (SBS)	Need known primer to start sequencing by synthesis (SBS)

Fragment the DNA in the genome





http://genome.cshlp.org/content/20/2/265/F2.large.jpg

Physical Shearing

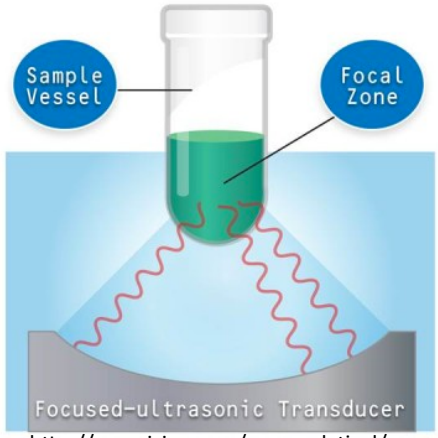
- Acoustic
- Hydrodynamic
- Sonication

Enzymatic Cutting

- DNAseI and other restriction and endonucleases
- Transposease

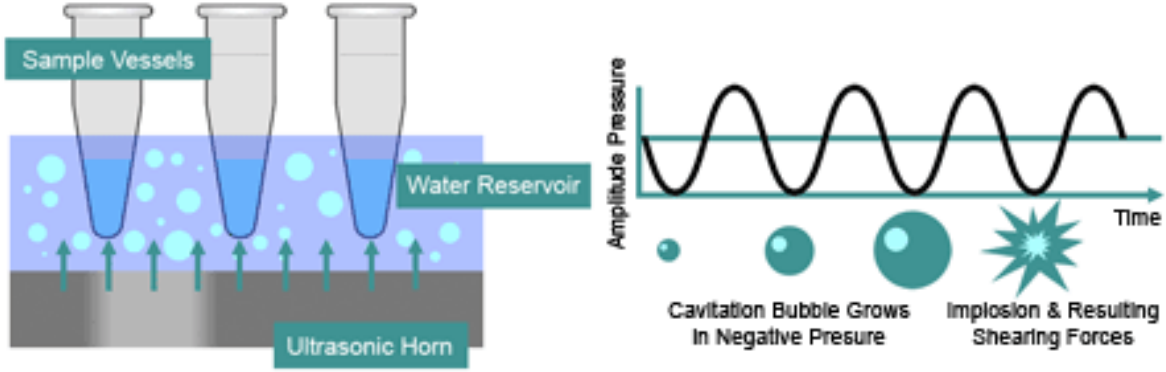
Chemical Fragmentation

- Heat and magnesium or zinc (only for RNA)



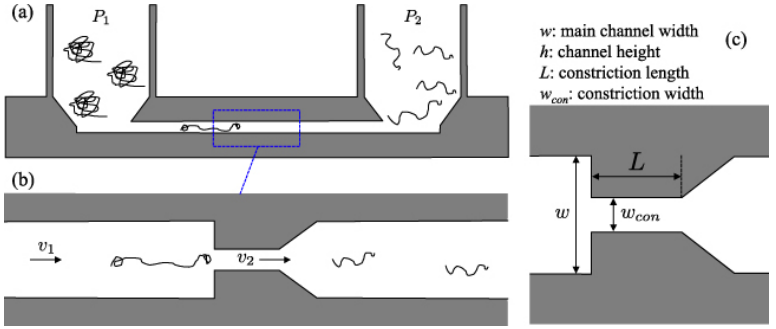
http://covarisinc.com/pre-analytical/afa-technology/

Sonication:
Energy from sound waves causes random breaks in the DNA backbone



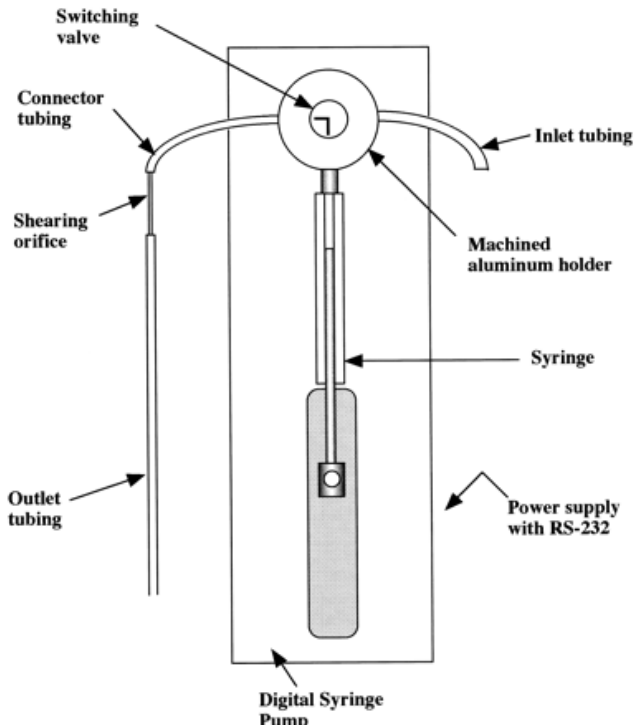
https://www.epigentek.com/catalog/sonication-devices-c-72_73.html?currency=USD&height=190&width=500&border=1&modal=true&random=1459971707966

Hydrodynamic Shearing: Force DNA through small opening – hydrodynamic forces cause shearing Can control size of piece by size of opening



L. Shui, et al; **Nanotechnology,** Volume 22, Number 49

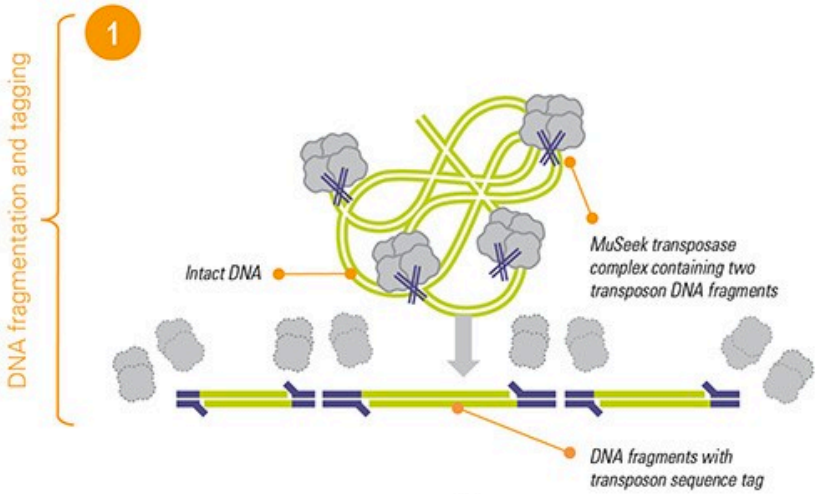
http://covarisinc.com/products/g-tube/



Thorstenson, YR, et al; **Genome Res**. 1998. 8: 848-855

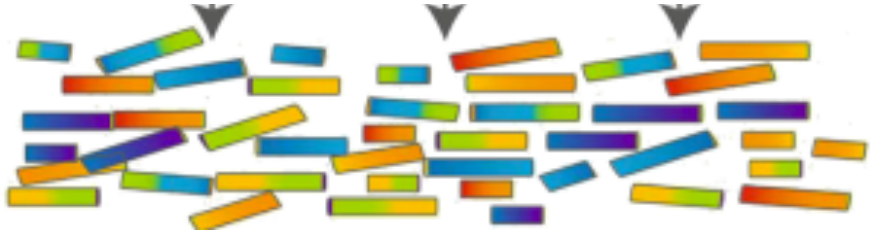
Enzymatic methods:

- Need to find ways to not bias sample to specific cutting locations
- Use a combination of enzymes that can nick/digest/cut to create random library
- Amount of time enzymes are allowed to digest usually determines the size of fragment



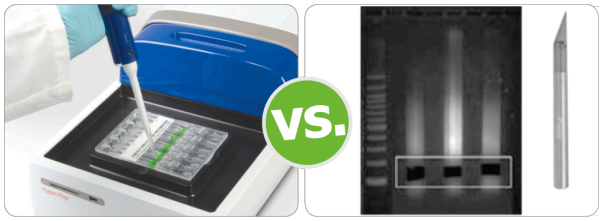
Transposeases:
Can add tag to the cut end
May introduce bias

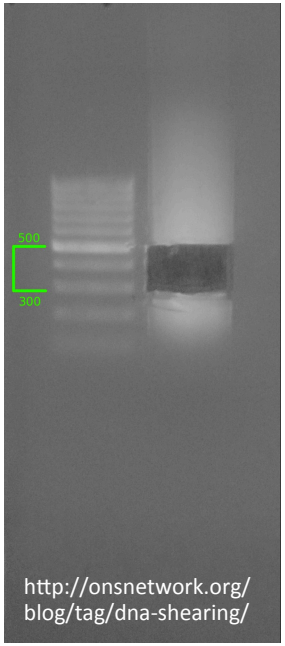
http://www.thermofisher.com/us/en/home/brands/thermo-scientific/molecular-biology/thermo-scientific-specialized-molecular-biology-applications/sequencing-thermo-scientific/next-generation-sequencing-library-preparation-thermo-scientific.html



Have lots of fragments:

Select for the size you want:
300bp in our case





We are planning SBS which means we need a primer.

Why is this problematic?

How can we solve this problem?

We are planning SBS which means we need a primer.

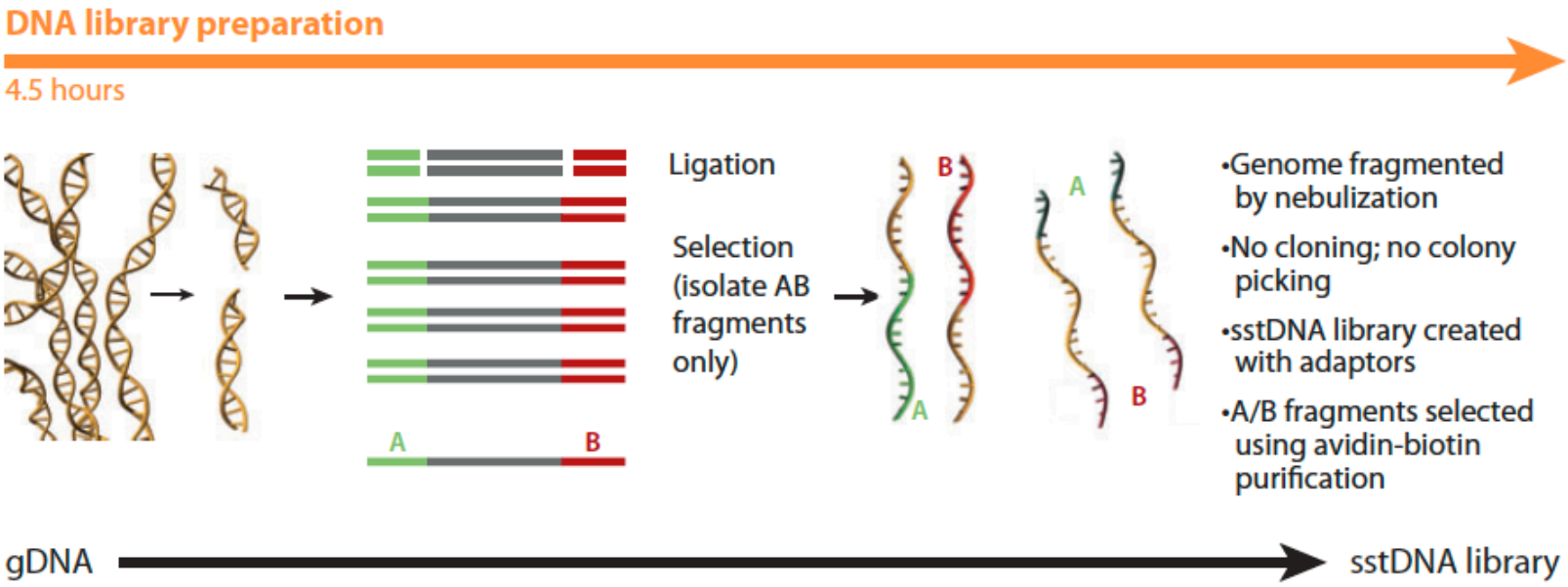
Why is this problematic?

We don’t know the sequence, can’t design a primer

How can we solve this problem?

We need to add known sequence - adaptors

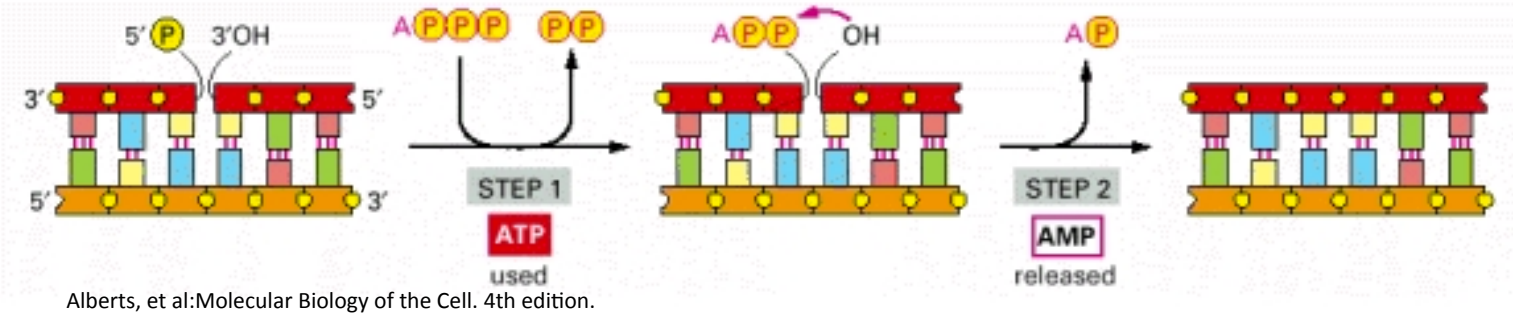
Adapt sequences with known sequences



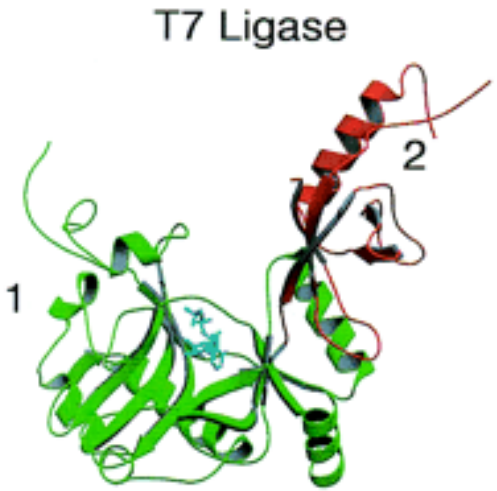
Adapters can also include a barcode sequence, so multiple libraries can be sequenced at the same time

Mardis, ER; Ann Rev Genom & Hum Gen

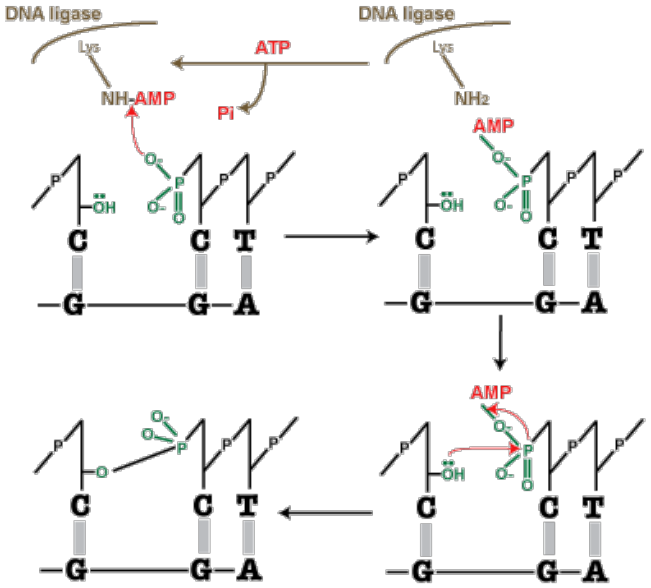
DNA ligase



Alberts, et al:Molecular Biology of the Cell. 4th edition.



http://nar.oxfordjournals.org/content/28/21/4051/F2.expansion



http://bitesizebio.com/10279/the-basics-how-does-dna-ligation-work/

Library Sizes for next-generation sequencing

- You can get ~300bp/read. How big does your library need to be?

Library calculations

Assume we want ~98% probability of seeing each base (f_c)

L length of the read, ***G*** length of genome

For any given sequence in the genome, the probability that it is in a particular read is the fraction of the genome contained in the read.

$$L/G$$

The probability that any given base is not in that read is

$$1-L/G$$

If you have a large pool of independent reads (N), the probability of any given base not being in any of the reads is

$$(1-L/G)^N$$

The probability that any given base is found in the sequencing data of all the selected reads (f_c) is

$$1-f_c=(1-L/G)^N$$

Therefore the number of reads needed to reach a specific base probability is

$$N=\log(1-f_c)/\log(1-L/G)$$

For our fish of 20,000 Mbp, how many reads of ~300 bp do we need to achieve 98% f_c?

Library calculations

Assume we want ~98% probability of seeing each base (f_c)

L length of the read, ***G*** length of genome

For any given sequence in the genome, the probability that it is in a particular read is the fraction of the genome contained in the read.

$$L/G$$

The probability that any given base is not in that read is

$$1-L/G$$

- ***L*** length of the read, ***G*** length of genome

If you have a large pool of independent reads (N), the probability of any given base not being in any of the reads is

$$(1-L/G)^N$$

The probability that any given base is found in the sequencing data of all the selected reads (f_c) is

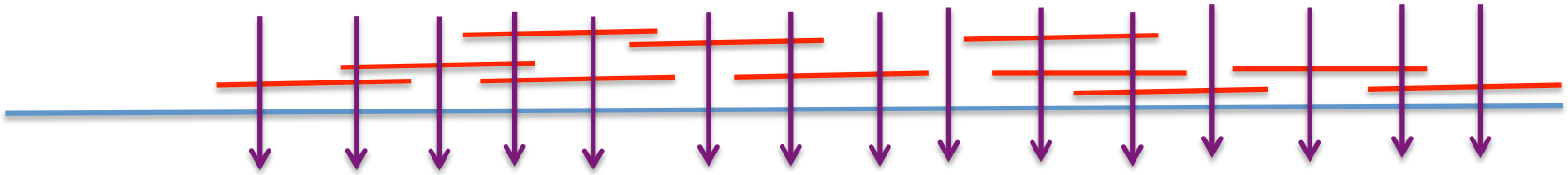
$$1-f_c=(1-L/G)^N$$

Therefore the number of reads needed to reach a specific base probability is

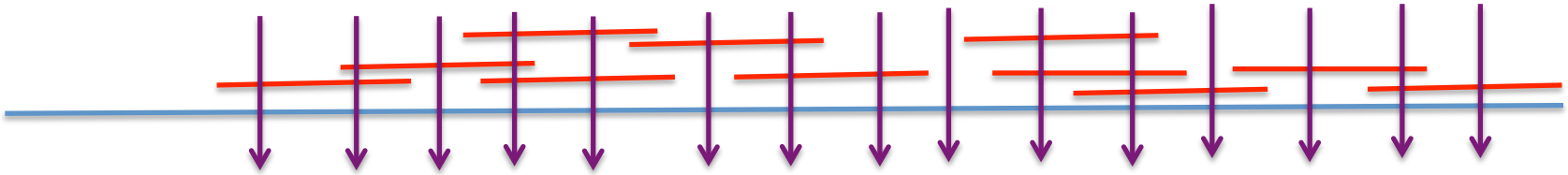
$$N=\log(1-f_c)/\log(1-L/G)$$

For our fish of 20,000 Mbp, how many reads of ~300 bp do we need to achieve 98% f_c?

2.6×10^8 reads → 260,000,000



What is the coverage of the locations indicated by the arrows? What is the average coverage for this given set of locations?



1 2 1 3 2 1 2 1 0 2 3 1 1 2 1

Average = 1.53 times each base is seen in all of the reads

How can we calculate the expected coverage from the number of reads?

$$c = NL/G$$

Given the data from before about our new fish, what coverage do we expect to have?

3.9

Library calculations

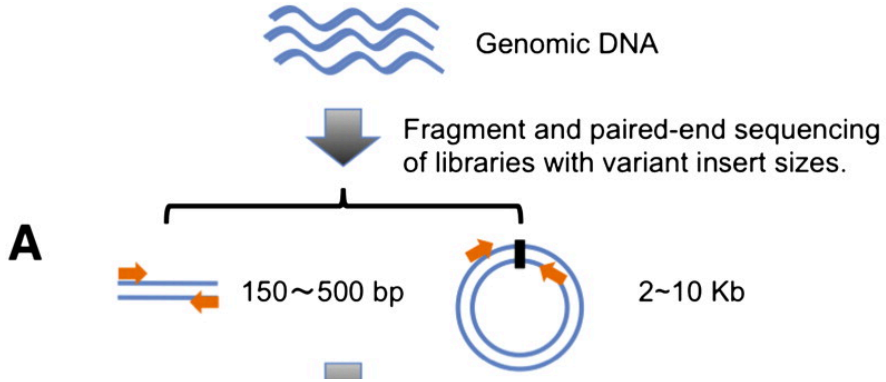
$$f_c = 1 - e^{-c}$$

We can determine the amount of coverage we need to achieve the probability of seeing each base

c	1	2	3	4	5
f_c	0.632	0.865	0.950	0.982	0.993

To get full genome need additional library

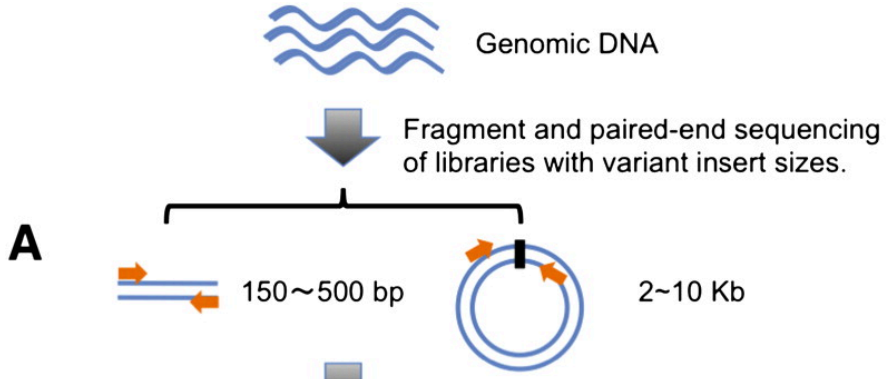
Will need a mate-pair library to allow us to join contigs and better place repetitive DNA



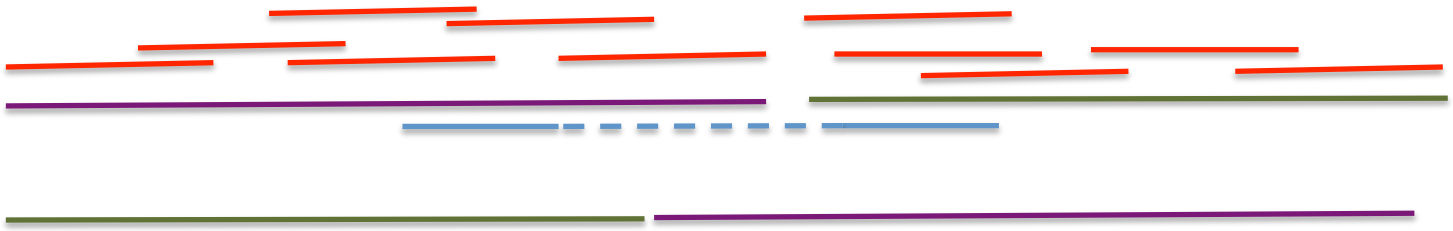
http://genome.cshlp.org/content/20/2/265/F2.large.jpg

To get full genome need additional library

Will need a mate-pair library to allow us to join contigs and better place repetitive DNA



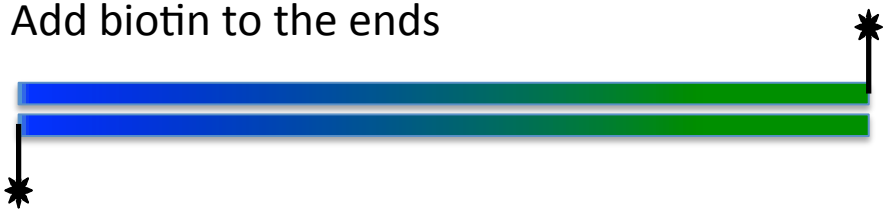
http://genome.cshlp.org/content/20/2/265/F2.large.jpg



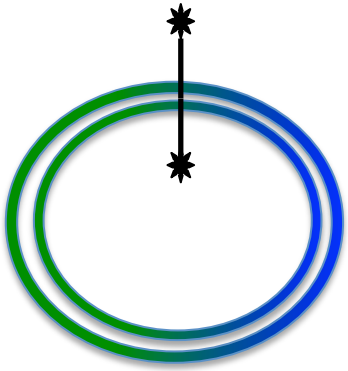
Fragment DNA - longer than size of read

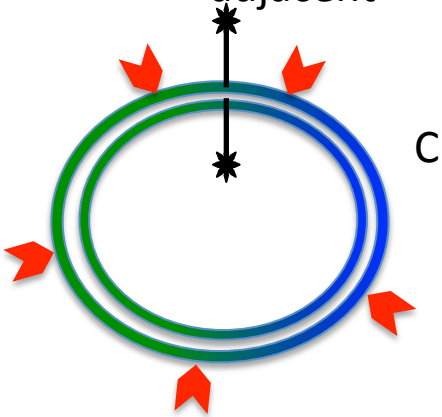
1000 bp

Add biotin to the ends

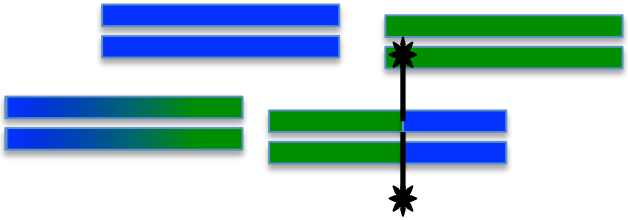


Circularize – the biotin containing ends are now adjacent

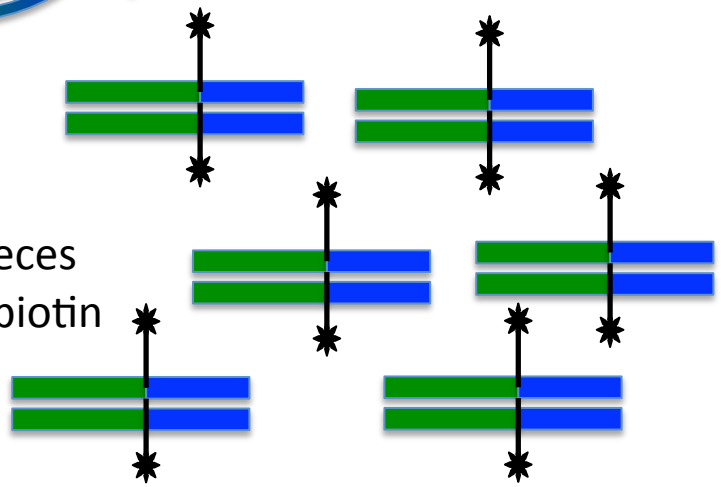


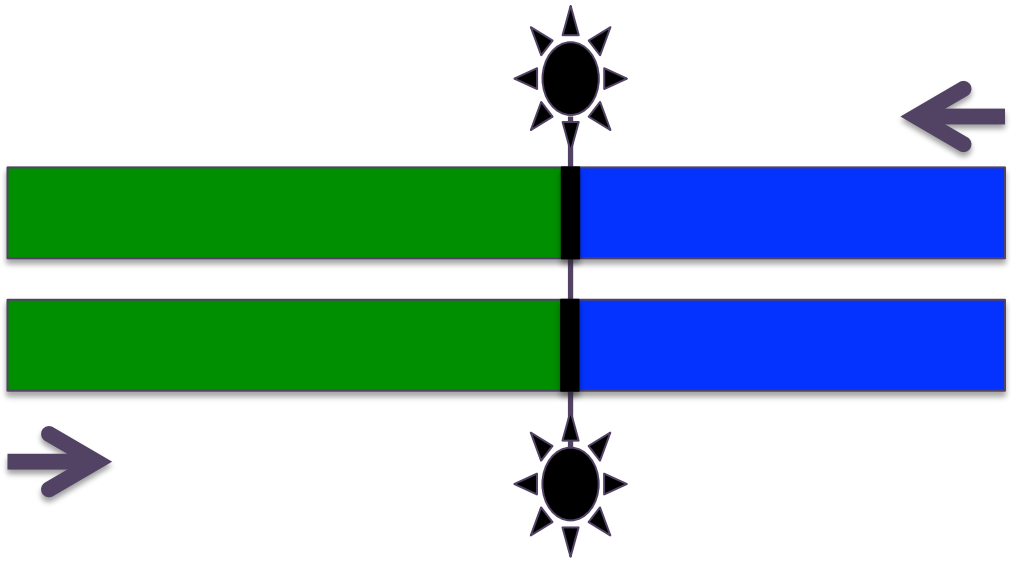


Cleave the circular DNA

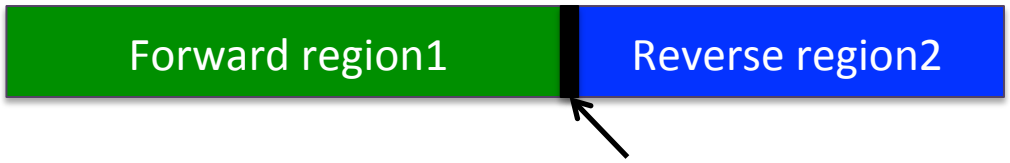


Isolate the cleaved pieces that have biotin

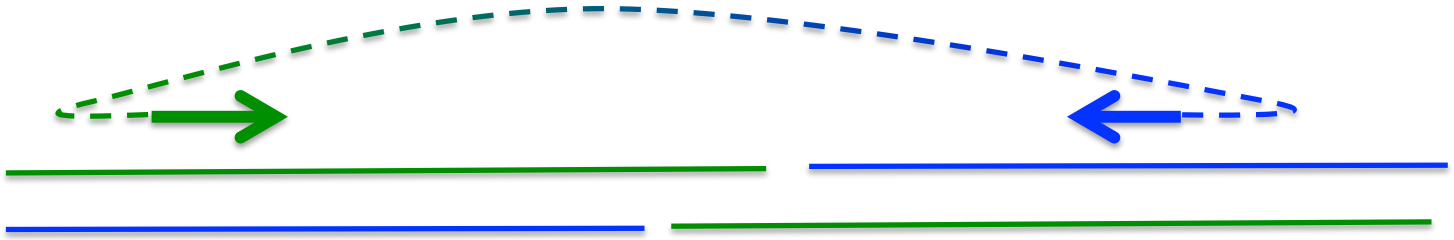


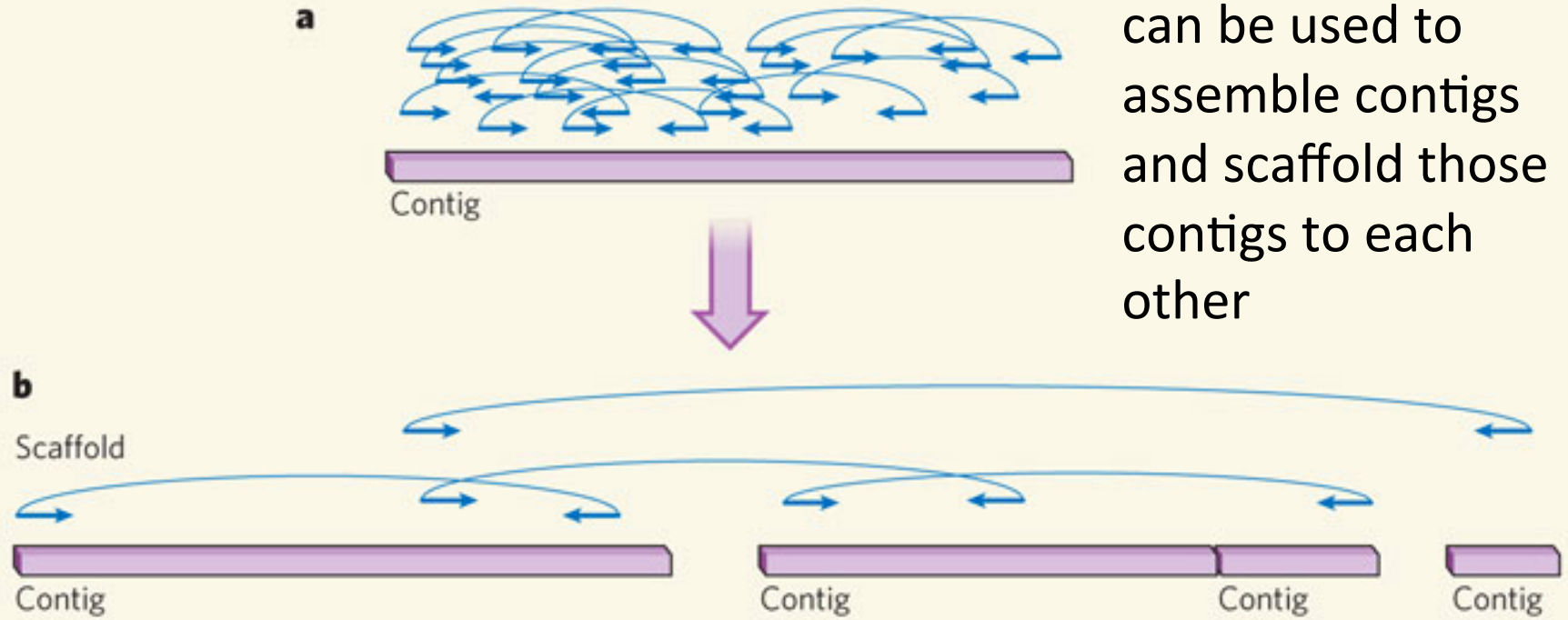


Sequence from either (or both) end



Conjoining sequence





Mate pairs of different lengths can be used to assemble contigs and scaffold those contigs to each other

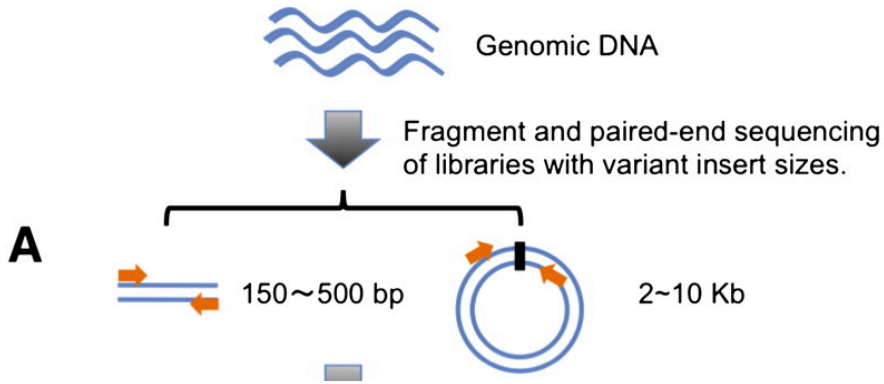
Kim C. Worley & Richard A. Gibbs; *Nature;* **463**, 303-304: doi:10.1038/463303a

How many of these will we need?

If assume 3Kb?

Need lower coverage, so assuming coverage of ~2

N=$2*10^6$



http://genome.cshlp.org/content/20/2/265/F2.large.jpg

$$c = NL / G$$

Other types of nucleic acid libraries

- cDNA libraries
- Amplified region libraries (16srRNA genes)
- Genomes from environmental samples (multiple organisms at once)

Questions about library creation?

Alignments

- Why do we care about aligning sequences?
- What kinds of things can we align with each other?

Judging an alignment

How can we tell if an alignment is good?

```
AGCTAGCT  AGCTAGCT  AGCTAGCT AGCTAGCT
ACCTTTCC  AGGTAGCT  ACGGATTA AGC-AGCT
```

Which of these is better? How do you know?

Scoring alignments

Look for the max score (can do min if you prefer, computationally equivalent)

- Match (m) = 1, mismatch (x) =-2, gap (g)= -5

AGCTAGCT	AGCTAGCT	AGCTAGCT	AGCTAGCT
ACCTTTCC	AGGTAGCT	ACGGATTA	AGC-AGCT
mxmmxxmx	*mmxmmmmm*	*mxxxmxxx*	*mmmgmmmm*
*S=4*m+4*x =-4*	*S=7*m+1*x =5*	*S=2*m+6*x =-10*	*S=7*m+1*g =2*

How do we do this in practice?

Algorithm for Computing Edit Distance

If we have two strings or sequences:

$a=a_1a_2a_3a_4\ldots a_n$

$b=b_1b_2b_4b_4\ldots b_m$

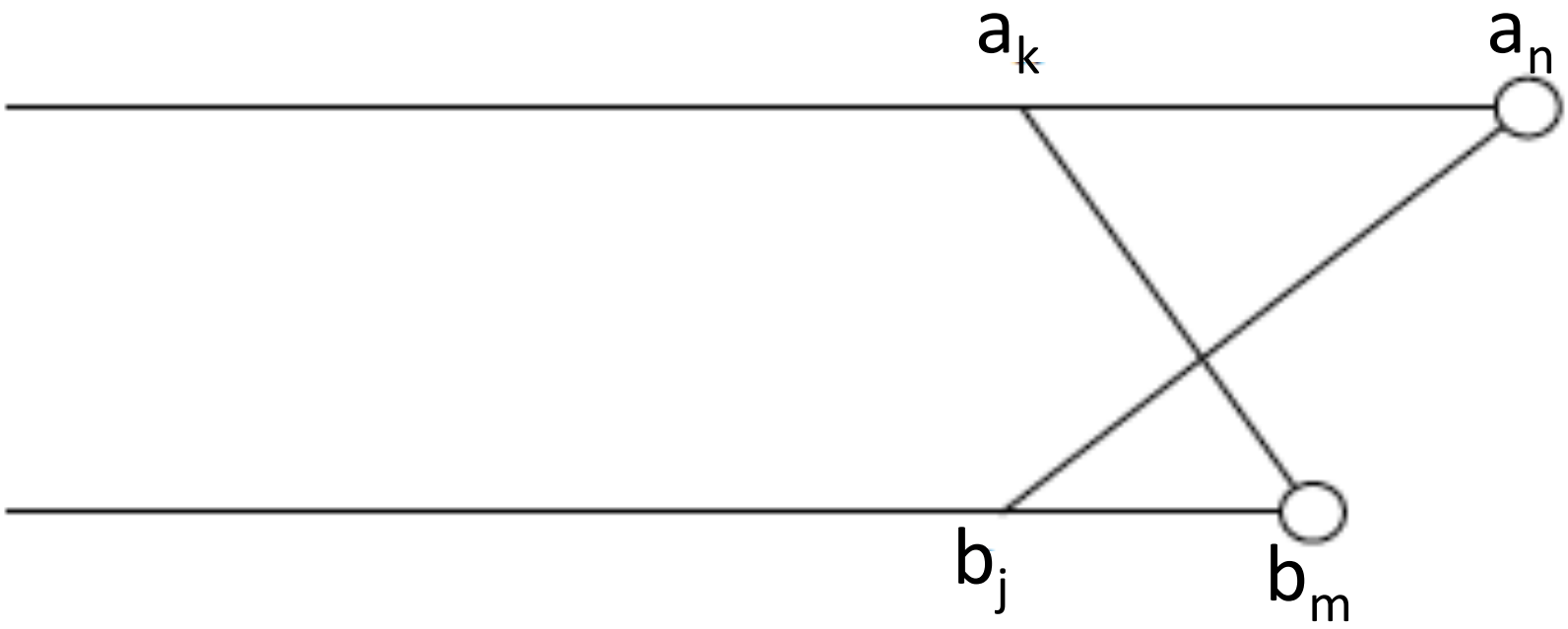
If we consider the last letter of each, they can either

(1) align with one another with a match or a mismatch,

(2) a_n doesn't align to anything,

(3) or b_m doesn't align to anything

a_n and b_m can't align with a different letter as it would violate the no crossing rule of alignments



Dynamic Programming

Combination of recursion and memoization

- Recursion is calling a function within itself
  - Need a known value to stop recursion and start building
- Memoization is recording the results at each step to allow for quick look up

		m	o	n	k	e	y
	0	1	2	3	4	5	6
m	1	0	1	2	3	4	5
o	2	1	0	1	2	3	4
n	3	2	1	0	1	2	3
e	4	3	2	1	1	1	2
y	5	4	3	2	2	2	1

Edit distance algorithm

3 possible outcomes become 3 cases of recurrence

$$S(i,j) = \max \begin{cases} \mathrm{c}(a_i, b_j) + S(i-1, j-1) \\ g + S(i-1, j) \\ g + S(i, j-1) \end{cases}$$

- $\mathrm{c}(a_i, b_j) + S(i-1, j-1)$: align a_i with b_j - match or mismatch
- $g + S(i-1, j)$: a_i is not matched, gap is added
- $g + S(i, j-1)$: b_j is not matched, gap is added

Don't know which one is the best, so we calculate them all and record the best value

Recursion stops when we reach S(i,0) or S(0,j). These are all gaps, so are equal to i/j*gap

	-	a_1	a_2	a_3	a_4	a_5
-	0	1g	2g	3g	4g	5g
b_1	1g					
b_2	2g		S(i-1,j-1)	S(i-1,j)		
b_3	3g		S(i-1,j)	S(i,j)		
b_4	4g					
b_5	5g					

Match = +1, mismatch =-1, gap = -2

	-	T	C	G	T	G
-	0	-2	-4	-6	-8	-10
A	-2	-1	-3	-5	-7	-9
T	-4	-1	-2	-4	-4	-6
C	-6	-3	0	S(i,j)		
G						
T						

Match = +1, mismatch =-1, gap = -2

	-	T	C	G	T	G
-	0	-2	-4	-6	-8	-10
A	-2	-1	-3	-5	-7	-9
T	-4	-1	-2	-4	-4	-6
C	-6	-3	0	-2	-4	-5
G	-8	-5	-2	1	-1	-3
T	-10	-7	-4	-1	2	0

What is the final alignment score?

EXERCISE

Calculate the alignment score for the following sequences. Use the following values. What is the best alignment?

Match =+1, Mismatch =-2, Gap =-5

GACGT ACGGT

Alignment score is -4

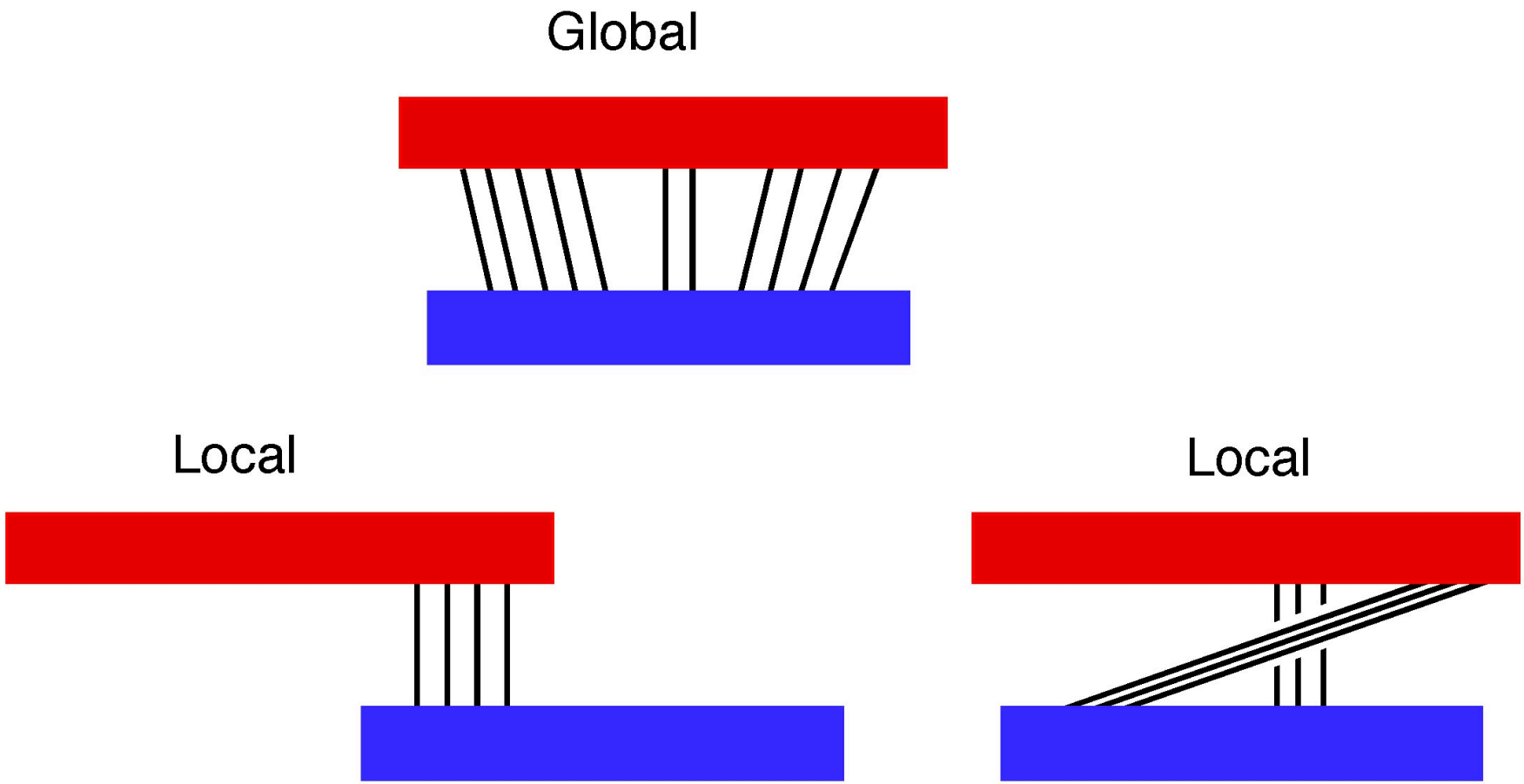
	-	G	A	C	G	T
-	0	-5	-10	-15	-20	-25
A	-5	-2	-4	-9	-14	-19
C	-10	-7	-4	-3	-8	-13
G	-15	-9	-9	-6	-2	-7
G	-20	-14	-11	-11	-5	-4
T	-25	-19	-16	-13	-10	-4

GACGT
ACGGT

mm*3+m*2
-2*3+2*1=-4

	-	G	A	C	G	T
-	0	-5	-10	-15	-20	-25
A	-5	-2	-4	-9	-14	-19
C	-10	-7	-4	-3	-8	-13
G	-15	-9	-9	-6	-2	-7
G	-20	-14	-11	-11	-5	-4
T	-25	-19	-16	-13	-10	-4

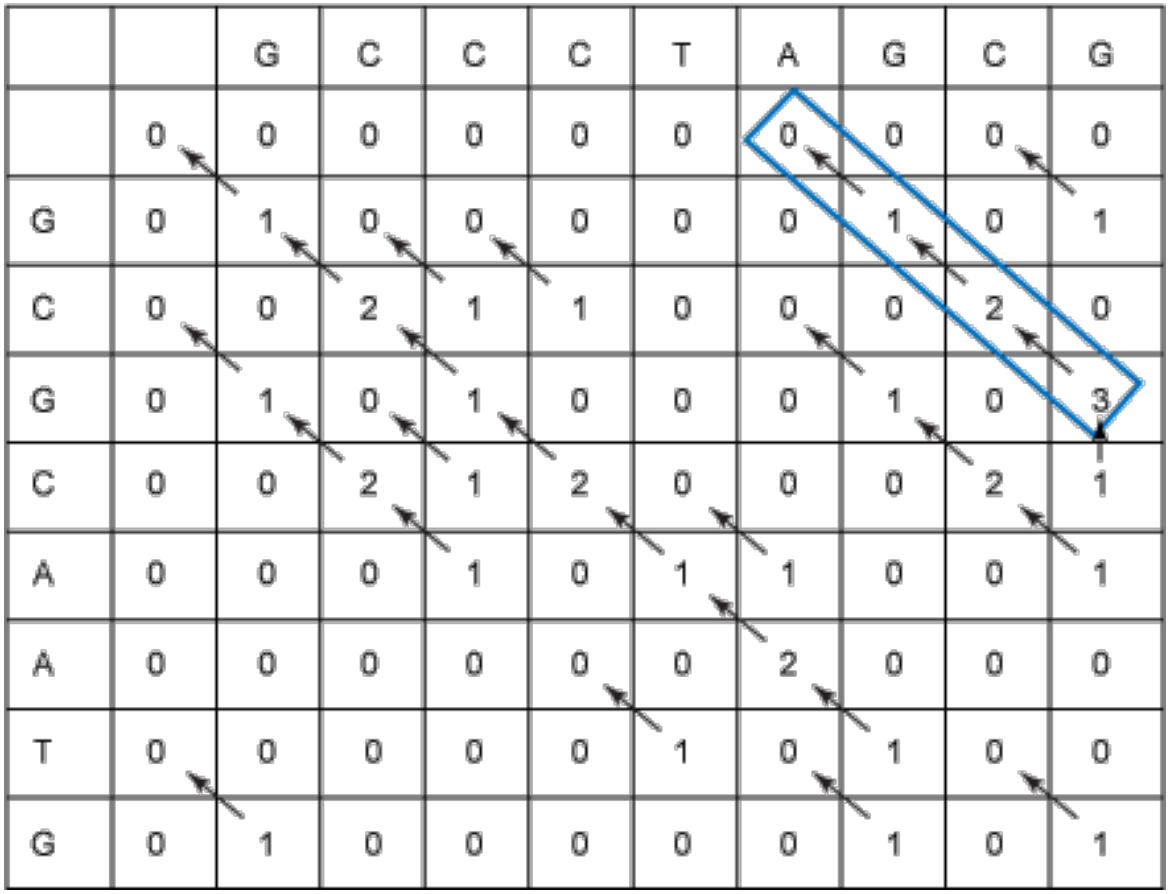
Global vs. Local Alignments



http://www.pitt.edu/~mcs2/teaching/biocomp/tutorials/global.html

To make local alignment:

- don't allow for negative numbers
- Start traceback at the highest value in table



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

http://www.ibm.com/developerworks/library/j-seqalign/

Why might you want a local alignment?

Current scoring matrix

All mutations have the same value

	A	G	T	C
A	1	-2	-2	-2
G	-2	1	-2	-2
T	-2	-2	1	-2
C	-2	-2	-2	1

Another scoring matrix

C→ T
Most common mutation, weight less

Pyrimidine to purine might be less damaging to RNA, weight less

	A	G	T	C
A	1	-1.5	-2	-2
G	-1.5	1	-2	-2
T	-2	-2	1	-1
C	-2	-2	-1	1

Most commonly used in protein alignments
Some amino acids are more different than others
BLOSUM62 matrix usually used

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4