Chapter 8 Class Notes – DNA Sequence Assembly

8.1. The Biological Problem: here we consider DNA sequence assembly via both the top-down (or map-based) approach (such as originally used in the Human Genome Project) and the bottom-up approach. In the latter case, a small-insert library is used to assemble the restriction map of a larger region by detecting overlaps of inserts to build contigs (here for a sequence rather than a restriction map). "The process of producing the sequence of a DNA segment (perhaps a genome) from a large number of randomly chosen sequence reads derived from it is called shotgun sequencing." Our focus here: Whole-genome shotgun (WGS) sequencing.

8.2. Reading DNA: Older versions have been replaced by the Sanger dideoxy sequencing method, "which employs the counterintuitive approach of analysis by synthesis." "Rather than breaking down a duplex molecule into fragments to be sequenced, we start with a single-strand molecule and manufacture fragments whose sizes depend upon the actual DNA sequence." Check out:

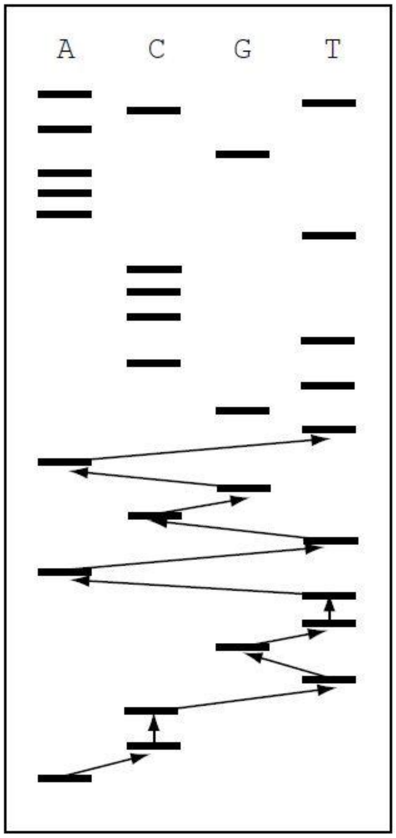
https://www.youtube.com/watch?v=bEFLBf5WEtc

In our text, the process is demonstrated with the following figure (Fig. 8.1 on p.199) for the "A" sequencing. The 12-nt primer is on the left in grey, followed by the 8 cases of "early stops" by the dideoxy process. The sequence is then determined by running through a gel: "Polyacrylamide gel electrophoresis physically sorts molecules by size"; see image below at bottom.

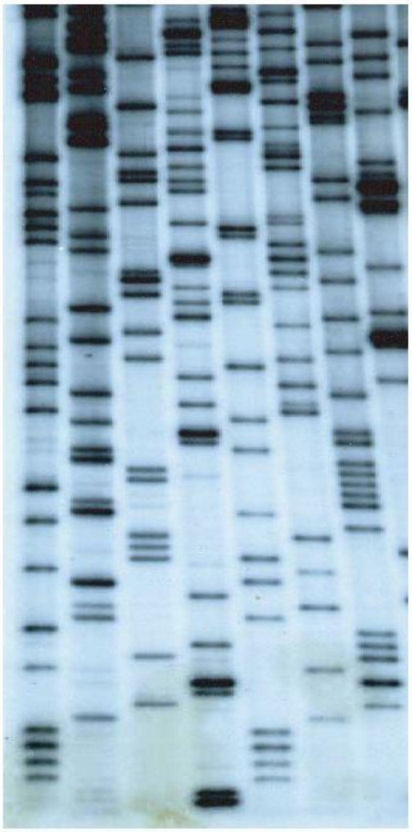
C.

```
5'catgacgatcggtttA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttAtcgAtgtctgggtAA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttAtcgAtgtctgggtAAAgActA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttAtcgAtgtctgggtAAAgA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttAtcgA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttAtcgAtgtctgggtA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'

5'catgacgatcggtttAcctgttAtcgAtgtctgggtAAA*
3'gtactgctagccaaaTggacaaTagcTacagacccaTTTcTgaTcagg…5'
```

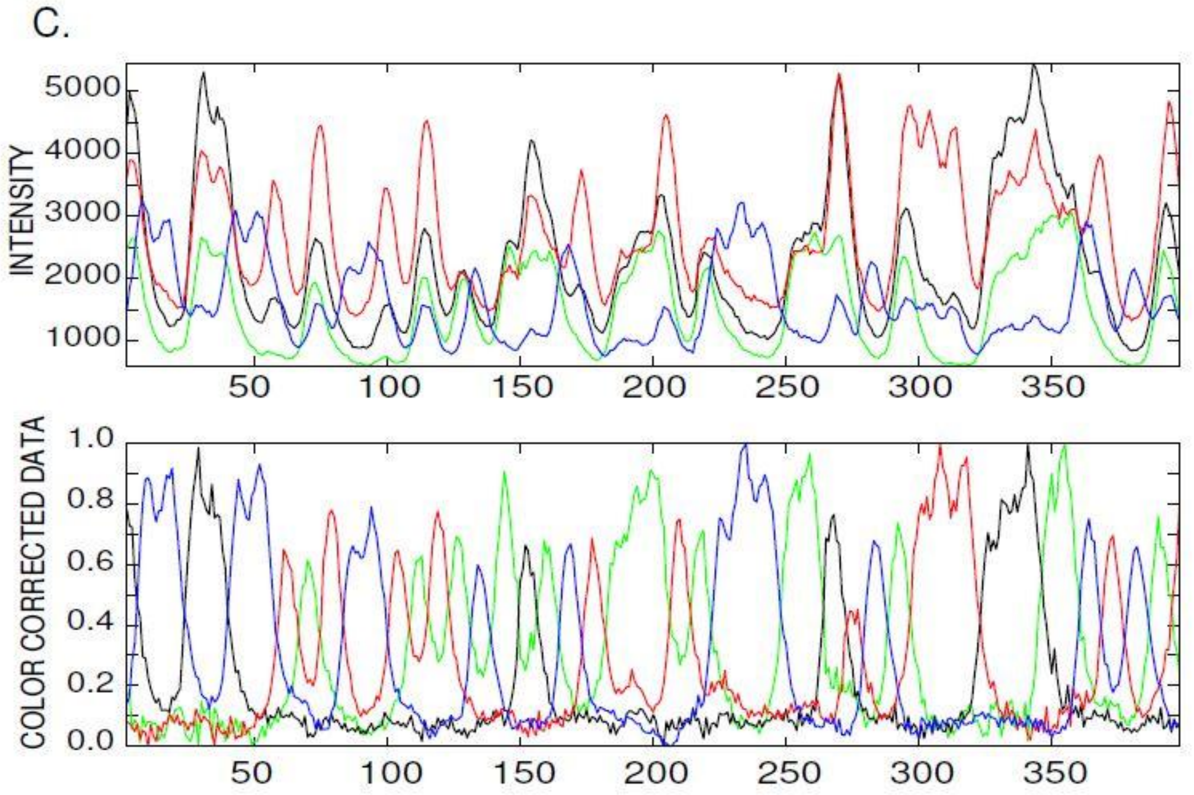
A.



B.



More recently, “automated capillary sequencers” are being used with four different fluorescent dyes, yielding images such as (with A, C, G and T):

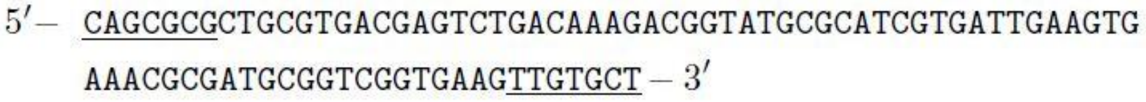


In the above, going from the top image to the bottom one is done by “color-correcting”. These plots are produced as in this video:

http://www.youtube.com/watch?v=AV35C36bBto

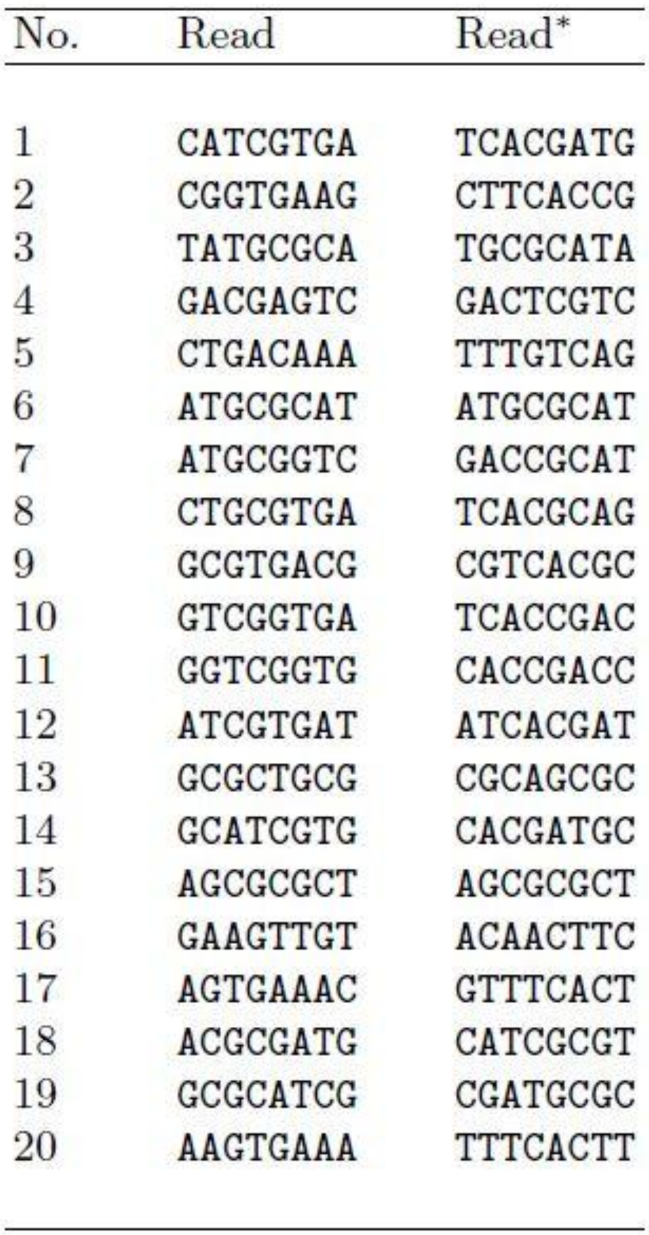
8.3. The Three-Step Method: Overlap, Layout, and Multiple Alignment – regarding alignments using shotgun sequencing, recall that it “is a mistake to think the entire target sequence will be determined even if the assembly is perfect” and we must keep in mind the statistical distribution of oceans and islands (p.112). Here, we’ll discuss the 3 steps in shotgun sequencing: (1) pairwise comparison, (2) layout, and (3) multiple alignment. Note that for reads *r* and *s*, we consider: *r* vs *s*, *r* vs *s**, *r** vs *s*, and *r** vs *s**.

Here’s an illustration (from p.205) where reads are 8 long:

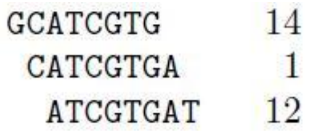


5′– CAGCGCGCTGCGTGACGAGTCTGACAAAGACGGTATGCGCATCGTGATTGAAGTG
AAACGCGATGCGGTCGGTGAAGTTGTGCT – 3′

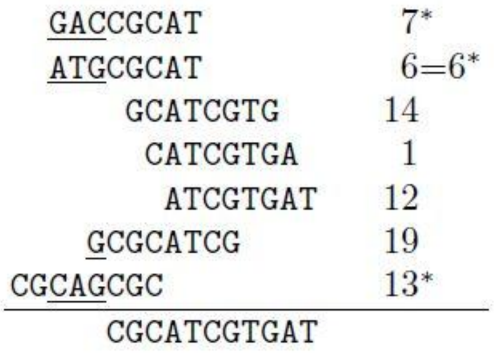
Table below left gives the 20 reads & their reversals (all 5’→ 3’)



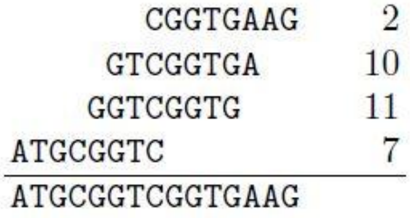
No.	Read	Read*
1	CATCGTGA	TCACGATG
2	CGGTGAAG	CTTCACCG
3	TATGCGCA	TGCGCATA
4	GACGAGTC	GACTCGTC
5	CTGACAAA	TTTGTCAG
6	ATGCGCAT	ATGCGCAT
7	ATGCGGTC	GACCGCAT
8	CTGCGTGA	TCACGCAG
9	GCGTGACG	CGTCACGC
10	GTCGGTGA	TCACCGAC
11	GGTCGGTG	CACCGACC
12	ATCGTGAT	ATCACGAT
13	GCGCTGCG	CGCAGCGC
14	GCATCGTG	CACGATGC
15	AGCGCGCT	AGCGCGCT
16	GAAGTTGT	ACAACTTC
17	AGTGAAAC	GTTTCACT
18	ACGCGATG	CATCGCGT
19	GCGCATCG	CGATGCGC
20	AAGTGAAA	TTTCACTT



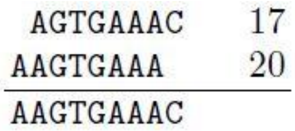
```
GCATCGTG     14
 CATCGTGA     1
  ATCGTGAT   12
```



```
  GACCGCAT        7*
  ATGCGCAT        6=6*
      GCATCGTG    14
       CATCGTGA    1
        ATCGTGAT  12
    GCGCATCG      19
CGCAGCGC          13*
-------------------------
     CGCATCGTGAT
```



```
       CGGTGAAG    2
     GTCGGTGA     10
    GGTCGGTG      11
ATGCGGTC           7
---------------------
ATGCGGTCGGTGAAG
```



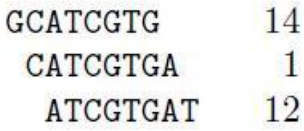
```
 AGTGAAAC   17
AAGTGAAA    20
---------------
AAGTGAAAC
```

The reads above are a subset of all possible reads, and were chosen at random locations in the sequence and in random orientations. We present below the 20×20 overlap matrix that indicates which fragments or their complements overlap by ≥ 3.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
*					3						7		7					5	
	*								6	5									3
		*			7								3					5	
			*					4											
				*															
3		7			*								4				3	6	
3						*			3	4									
							*					5							
								*				3					3		
			3						*	7									
										*									
											*		6					4	
		4										*		5				3	
					4	4							*					6	
												5		*					
															*				
																*			7
					3			4									*		
					6							4						*	
																			*

To understand the table, the '3' in the (7,1) entry corresponds to the read of 1 and the read* for 7, and the overlap is `CAT`; the '7' in the (1,12) entry corresponds to the reads of 1 and 12, and the overlaps is `ATCGTGA`.

We then start with the (1,14) and (1,12) high-scoring ‘7’ entries, and build the arrangement in the previous table at top right, reproduced here:



```
GCATCGTG      14
 CATCGTGA      1
  ATCGTGAT    12
```

Including other matches (namely 19, 6, 7* and 13*) gives the next entry and the table (above right) and the ultimate alignment: CGCATCGTGAT. In the original sequence (which normally we wouldn’t know and are trying to “estimate”), this is the word starting at position 31 (not counting the original 7-base left end). Continuing the process gives us alignments ATGCGGTCGGTGAAG starting at position 56 and AAGTGAAAC starting at position 44.

Clearly only part of the sequence is determined from these data, but this is to be expected since the coverage is only $c = \frac{20\times 8}{70} = 2.286$, and so $f_c = 1 - e^{-2.286} = 0.8983$. “At a coverage of 10 we would with reasonable probability recover most of the 70bp.”

8.4. High-Throughput Genome Sequencing – “the key underlying determinant is that we can obtain high-quality continuous sequence reads of up to 500 to 800 bases with current technology.”

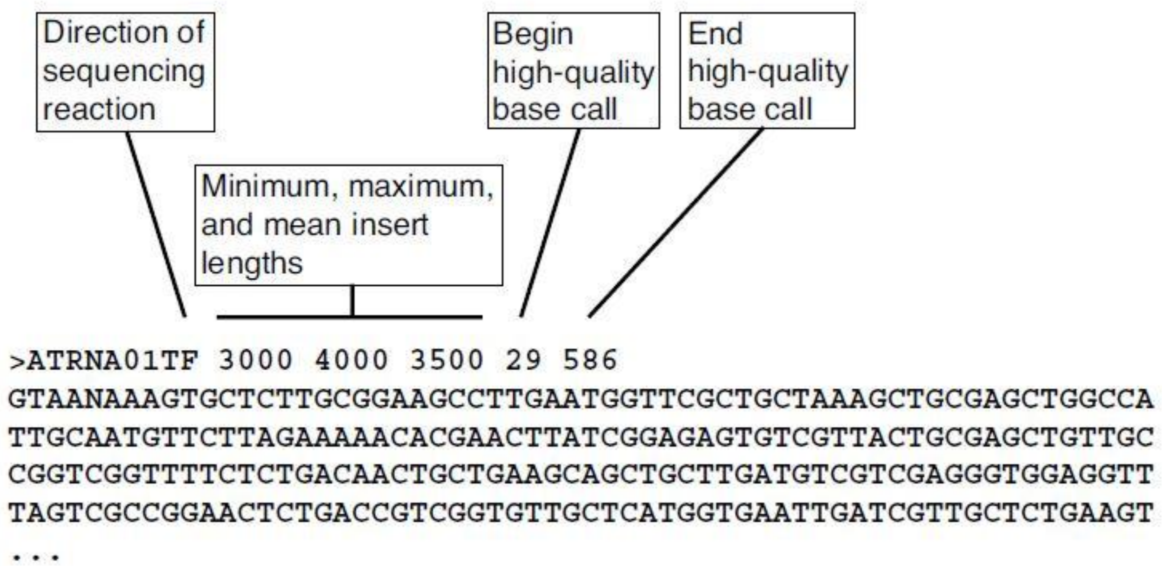
Big problem: eukaryotic genomes contain repeated sequences that may be longer than the average sequence read; possible solution: employ difference sizes of cloned inserts.

8.4.1. Computational Tools: base calling is the process of identifying which base corresponds to each position in a sequence read. One example of base-calling software is called Phred (http://www.phrap.org/phredphrapconsed.html); the trace

processing steps are: (a) idealized predicted peak locations, (b) observed peaks above a threshold, (c) matching observed and predicted peaks, and (d) accounting for missing peaks.. "A very important feature of Phred is that it associates with each base a probability p that the base call is in error. The probability p depends upon things such as peak spacings, peak resolution, and areas of uncalled peaks. The quality of each base call is described by the quality score Q, which is defined as $Q = -10\log_{10} p$."

The next step is sequence assembly; the inputs are:

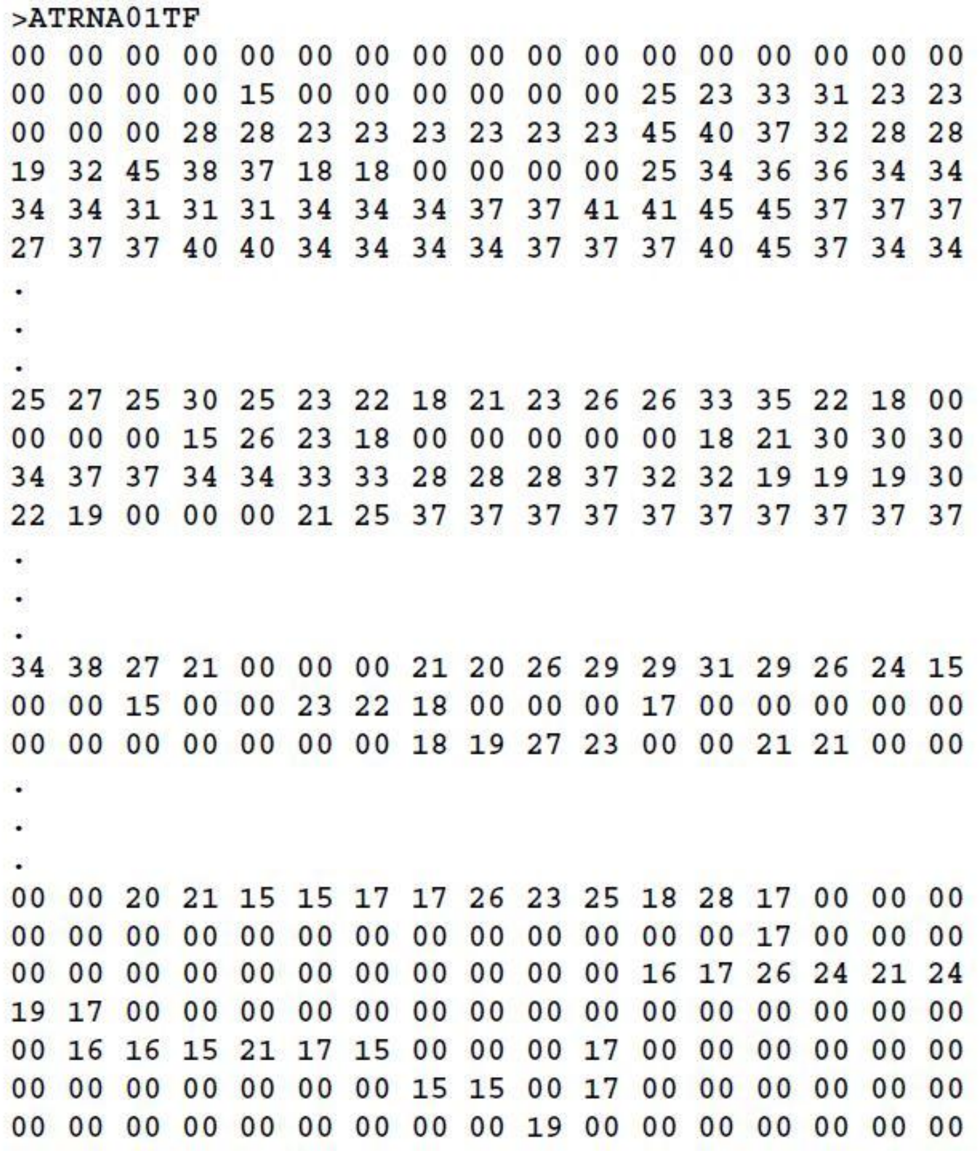
A. Sequence file


Direction of sequencing reaction
Begin high-quality base call
End high-quality base call
Minimum, maximum, and mean insert lengths

```
>ATRNA01TF 3000 4000 3500 29 586
GTAANAAAGTGCTCTTGCGGAAGCCTTGAATGGTTCGCTGCTAAAGCTGCGAGCTGGCCA
TTGCAATGTTCTTAGAAAAACACGAACTTATCGGAGAGTGTCGTTACTGCGAGCTGTTGC
CGGTCGGTTTTCTCTGACAACTGCTGAAGCAGCTGCTTGATGTCGTCGAGGGTGGAGGTT
TAGTCGCCGGAACTCTGACCGTCGGTGTTGCTCATGGTGAATTGATCGTTGCTCTGAAGT
...
```

Quality file (B) is on next page; note that "00" for Q corresponds to a p-value of 1 (i.e., a very unreliable base call). These inputs are used in the TIGR Assembler (The Institute for Genome Research, now merged into J. Craig Venter Institute & with other associations, http://www.jcvi.org/cms/home/). The ultimate output of the assembler is the consensus assembled sequence with the reliability score at each position.

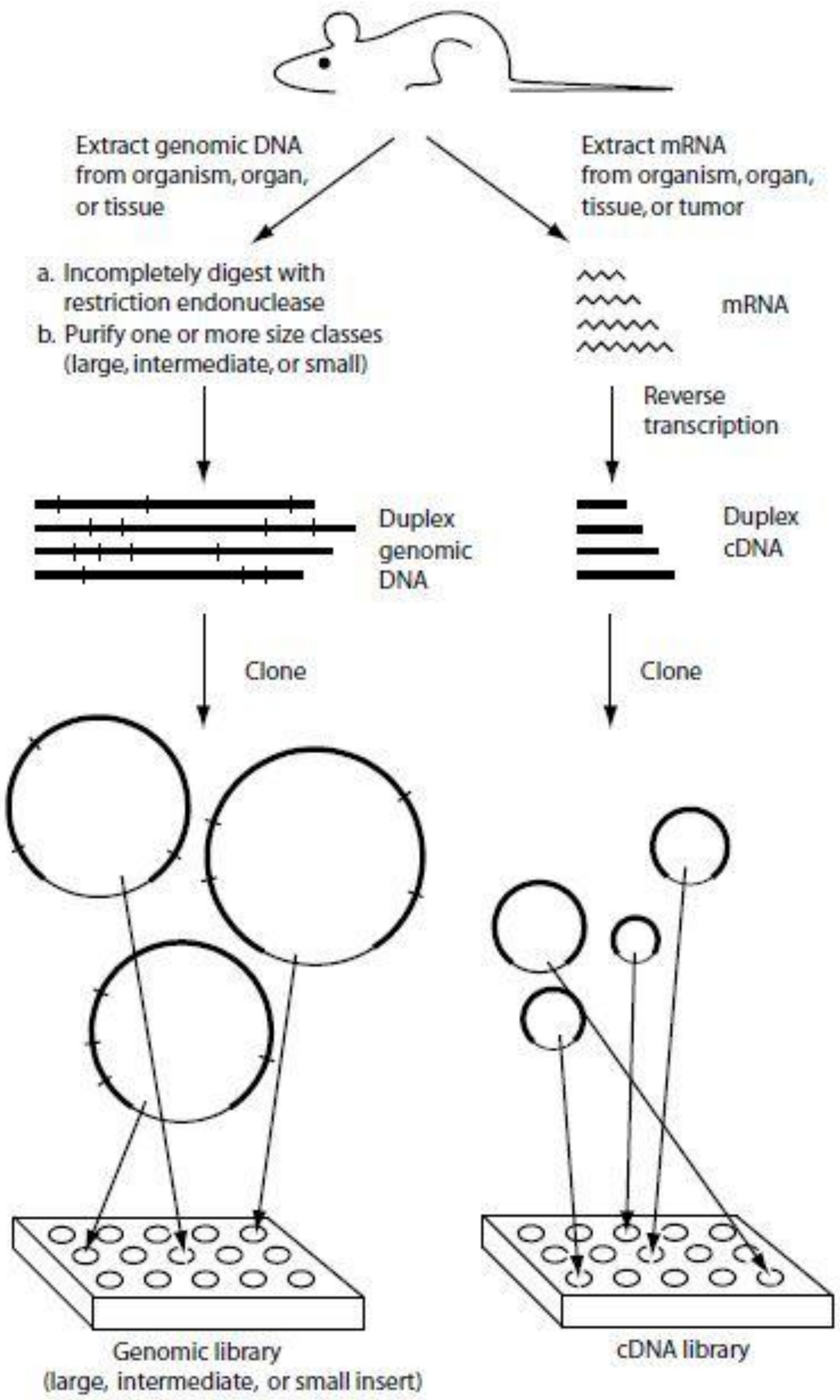
B. Quality file



```
>ATRNA01TF
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 15 00 00 00 00 00 00 25 23 33 31 23 23
00 00 00 28 28 23 23 23 23 23 23 45 40 37 32 28 28
19 32 45 38 37 18 18 00 00 00 00 25 34 36 36 34 34
34 34 31 31 31 34 34 34 37 37 41 41 45 45 37 37 37
27 37 37 40 40 34 34 34 34 37 37 37 40 45 37 34 34
.
.
.
25 27 25 30 25 23 22 18 21 23 26 26 33 35 22 18 00
00 00 00 15 26 23 18 00 00 00 00 00 18 21 30 30 30
34 37 37 34 34 33 33 28 28 28 37 32 32 19 19 19 30
22 19 00 00 00 21 25 37 37 37 37 37 37 37 37 37 37
.
.
.
34 38 27 21 00 00 00 21 20 26 29 29 31 29 26 24 15
00 00 15 00 00 23 22 18 00 00 00 17 00 00 00 00 00
00 00 00 00 00 00 00 18 19 27 23 00 00 21 21 00 00
.
.
.
00 00 20 21 15 15 17 17 26 23 25 18 28 17 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 17 00 00 00
00 00 00 00 00 00 00 00 00 00 00 16 17 26 24 21 24
19 17 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 16 16 15 21 17 15 00 00 00 17 00 00 00 00 00 00
00 00 00 00 00 00 00 15 15 00 17 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 19 00 00 00 00 00 00 00
```

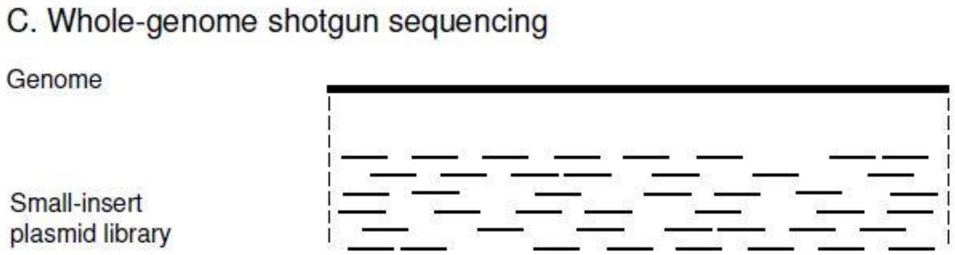
“In 2001, the human genome sequence assembly required 20,000 hours of CPU time and 500GB of storage, with the use of forty, four-processor machines, each having 4GB of RAM, running in parallel. Half of this time was employed in computing the overlaps between reads.”

8.4.2. Genome-Sequencing Strategies: recall cloning from Chap.1:


Extract genomic DNA from organism, organ, or tissue
Extract mRNA from organism, organ, tissue, or tumor
a. Incompletely digest with restriction endonuclease
b. Purify one or more size classes (large, intermediate, or small)
mRNA
Reverse transcription
Duplex genomic DNA
Duplex cDNA
Clone
Clone
Genomic library (large, intermediate, or small insert)
cDNA library

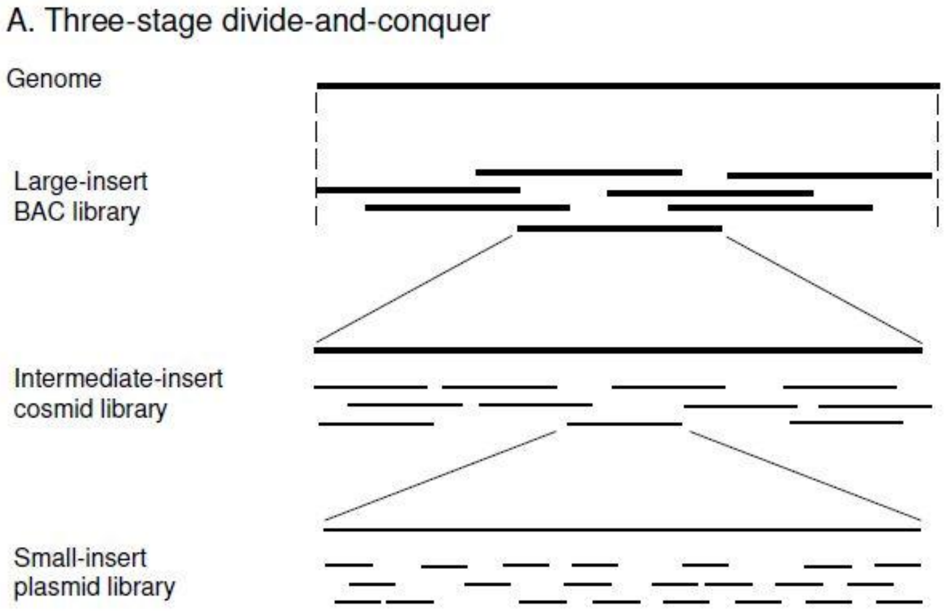
So the DNA (double-stranded genomic DNA or single-strand cDNA) is inserted into plasmid DNA (taken from bacteria) and a sequence primer is added – then called a recombinant plasmid. The plasmid is put in water with bacteria and heated so the plasmid enters into the bacteria. The goal is to then grow the bacteria (i.e., makes copies or clones). Aside: bacteriophage lambda is a virus which infects a bacterium.

All genome sequencing strategies require cloning and shotgun sequencing: "the differences in strategies lie in the use of clone mapping, and the point at which random shotgun sequencing is initiated." "Three particular strategies are: the clone-by-clone shotgun approach, shotgun sequencing of BACs joined into a minimum tiling path by sequence-tagged connectors, and whole-genome shotgun (WGS) assembly."


C. Whole-genome shotgun sequencing
Genome
Small-insert plasmid library

The (last) WGS assembly approach (illustrated above) was first used for viral genomes (such as the cauliflower mosaic virus: 8031bp, and bacteriophage lambda: 48502bp); randomly selected small-insert clones are used. Appears not feasible for larger genomes due to long repeats and lack of computing power. Note that here sequence coverage is the average number of times any given genomic base is represented in sequence reads.

The top-down approach is illustrated next – it was thought that this 'divide-and-conquer' strategy would be useful for the Human Genome Project:

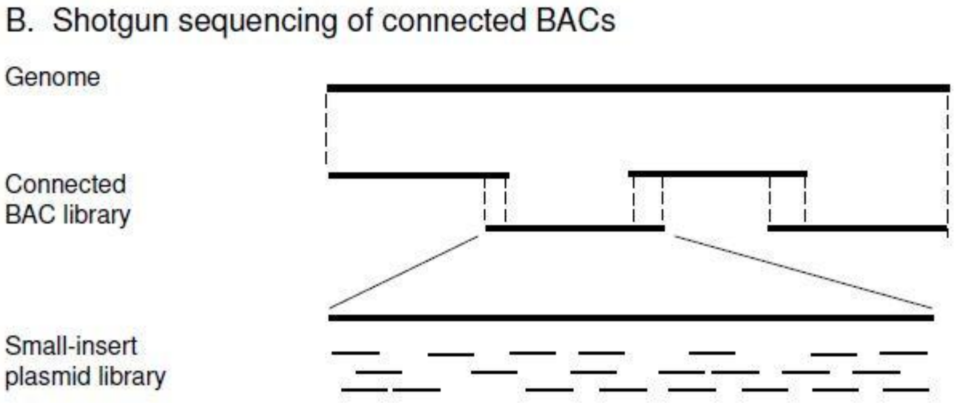

A. Three-stage divide-and-conquer
Genome
Large-insert BAC library
Intermediate-insert cosmid library
Small-insert plasmid library

The differences above (plasmids at bottom, cosmids in center, BACs at top) correspond to size: plasmids (Lederberg, 1952) are vectors akin to viruses and can accept DNA inserts of up to about 15kb; cosmids (Collins & Hohn, 1978) are a type of plasmids but they contain cos sequences and can accept inserts up to about 45-50kb, and BACs up to 300kb; the following link may be helpful: http://en.wikipedia.org/wiki/Genomic_library

This 3-stage method includes: construct a high-resolution genetic map, a low resolution physical map from large-insert clones (BACs), and a high resolution physical map bases on cosmids.

A third choice ("B" in the diagram) is a hybrid (intermediate approach); the intermediate insert cosmid step is omitted by

using large-insert BAC clones and sequence-tagged connectors (note the overlap at the ends below); PCR reactions are used to amplify unique sequences at the ends of a BAC. Then, each BAC clone (insert size ~150kb) is subjected to random shotgun sequencing.


B. Shotgun sequencing of connected BACs
Genome
Connected BAC library
Small-insert plasmid library

8.4.3. Whole-Genome Shotgun Sequencing of Eukaryotic Genomes: WGS not practical for some eukaryotes; for example, "the human genome consists of about 45% repeated sequences, with over a million copies of Alu elements alone."

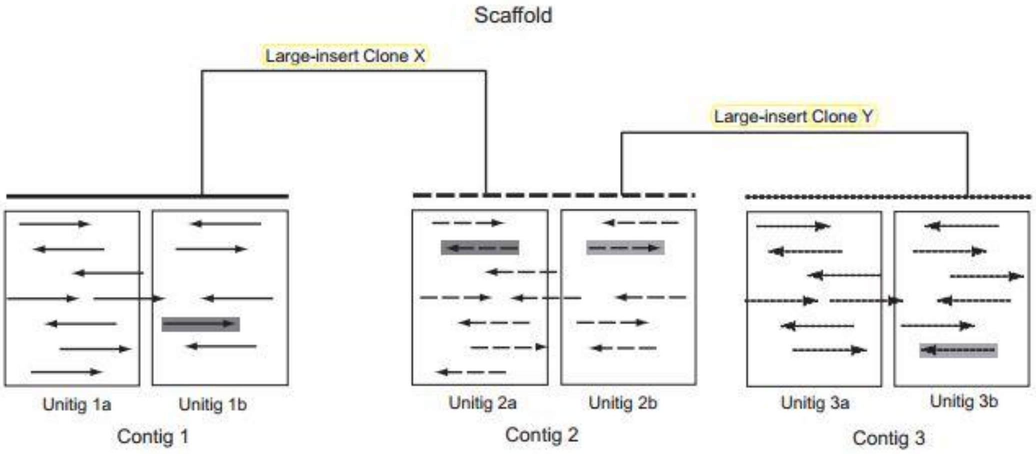


Fig. 8.7. Anatomy of a scaffold. Sequence reads are indicated by lines with arrowheads. Based upon pairwise comparisons, sequence reads can be assembled into small contigs called *unitigs*, which can further be merged into larger *contigs* based upon sequence reads present as mate pairs on small- and intermediate-insert clones. Contigs are completely spanned by a DNA sequence, although not all regions may have the same depth of coverage. Contigs can be further grouped to form a *scaffold*. Even though the regions between these contigs may not be represented by a DNA sequence, large-insert clones X and Y whose mate-pair reads lie in different contigs (shaded boxes) allow those contigs to be correctly positioned and oriented.