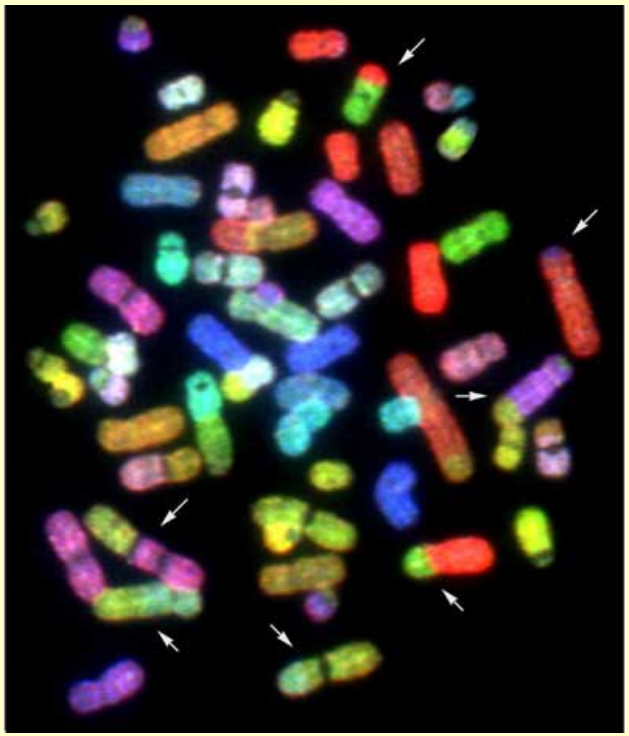
Sequencing the Human Genome

http://biochem158.stanford.edu/

Genomics, Bioinformatics & Medicine



Doug Brutlag
Professor Emeritus of Biochemistry & Medicine
Stanford University School of Medicine

Homework Assignments

- Homework is due midnight on the evening of the due date.
- If you are late you will lose 10% of the grade for each day the assignment is late
- Submit homework in an email or as an attachment in an email to brutlag@stanford.edu.
- Homework may be a Word, text, PDF, postscript, HTML or Google document.
- Always reference and quote copied material
  - Copying without quotes and references, even from Internet is plagiarism
  - Copying without quotes and references is also a violation of the Honor Code
- If you get less than 100% on an assignment you will have one week to submit a revised homework for full credit.

The Human Genome Project: Should we do it?

- Service, R. F. (2001). The human genome: Objection #1: big biology is bad biology. *Science, 291(5507), 1182.*
  - –Not hypothesis driven.
  - –Fishing expedition or stamp collecting.
  - –Eliminate funds from investigator initiated science.

- Vogel, G. (2001). The human genome: Objection #2: why sequence the junk? *Science, 291(5507), 1184.*
  - –Limit sequencing to 1.5% of genome that codes proteins.
  - –Do not sequence intergenic regions "genetic wastelands".
  - –Do not sequence repeated regions (telomeres and heterochromatin).

- Service, R. F. (2001). The human genome: Objection #3: impossible to do. *Science, 291(5507), 1186.*
  - –Technology of the time permitted 500 to 1,000 bp per day per person.
  - –Move from radioactively labeled sequencing to fluorescent sequencing permitted complete automation up to 1 gigabyte per year.

Chemical Structure of DNA

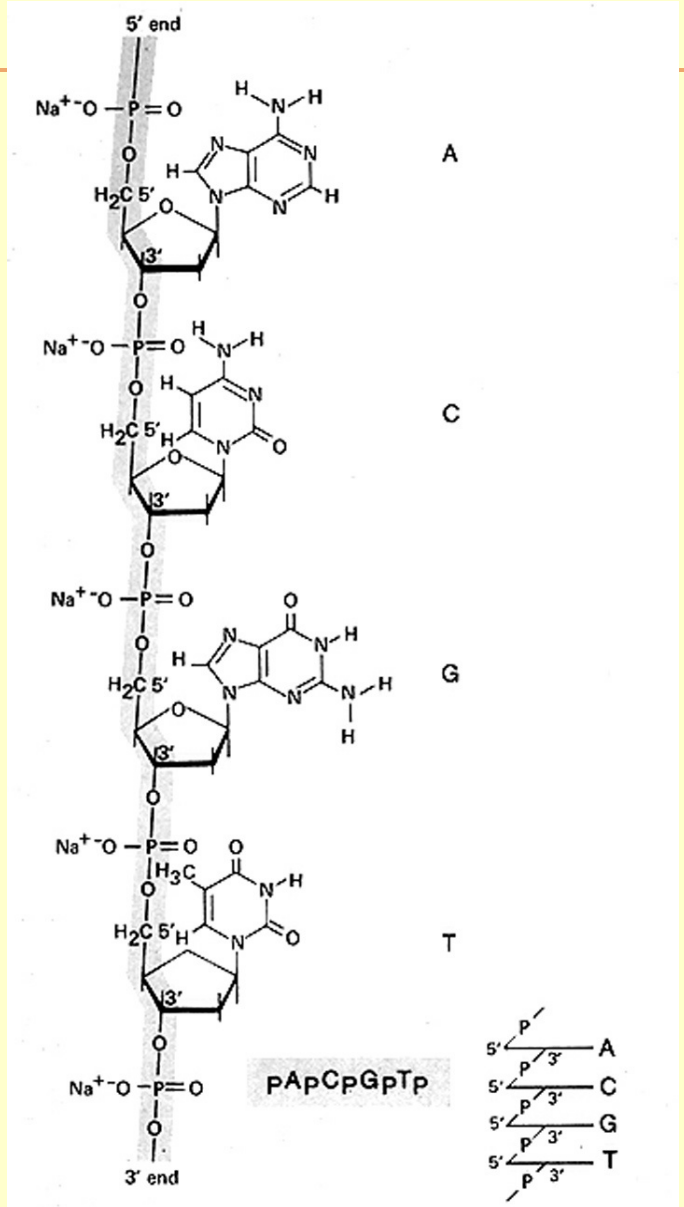
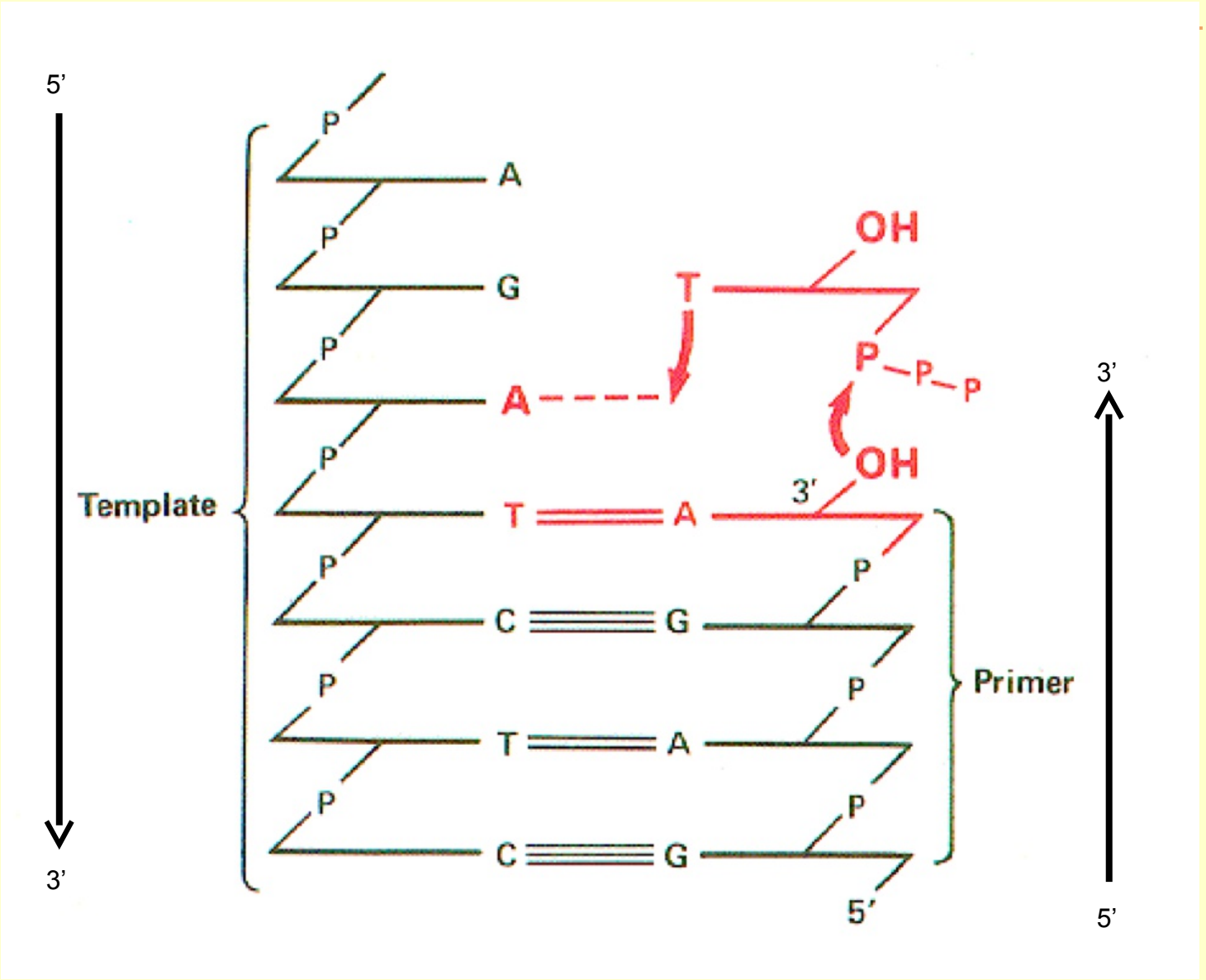
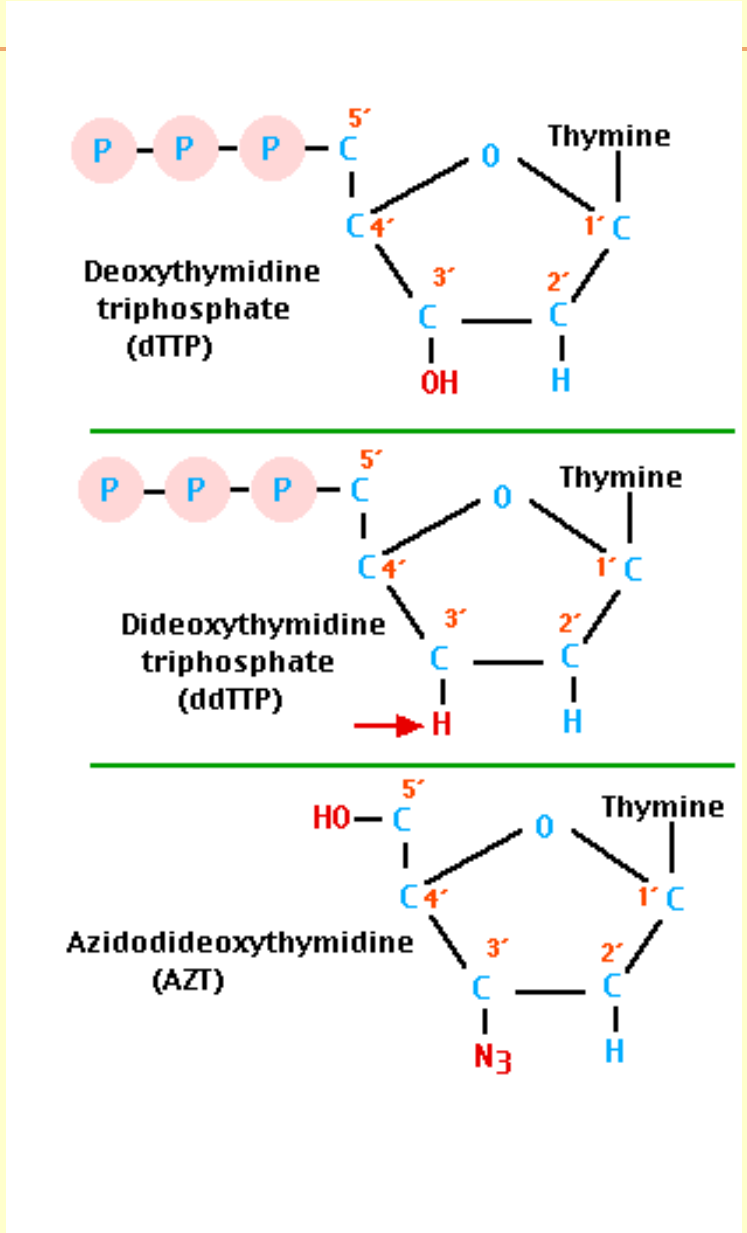


FIGURE 1-2
Segment of a polydeoxynucleotide as a sodium salt.

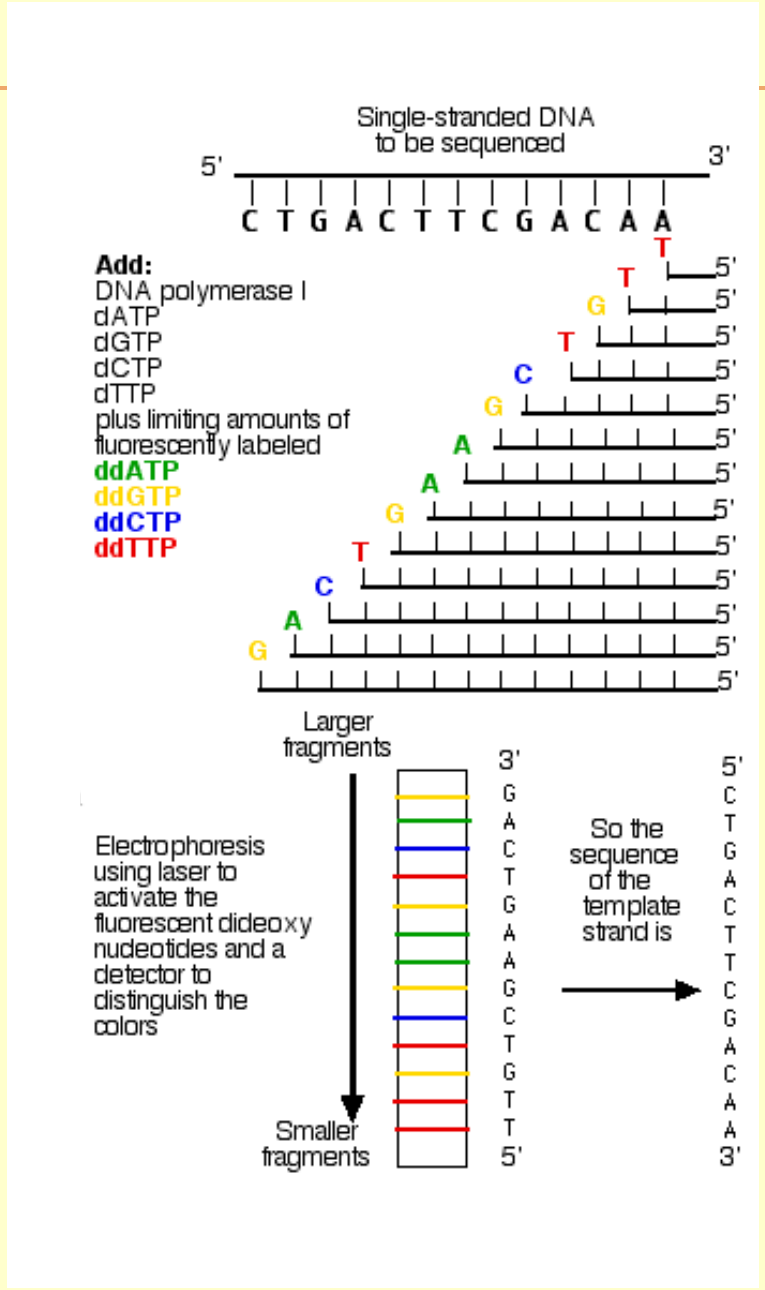
DNA Synthesis by DNA polymerases



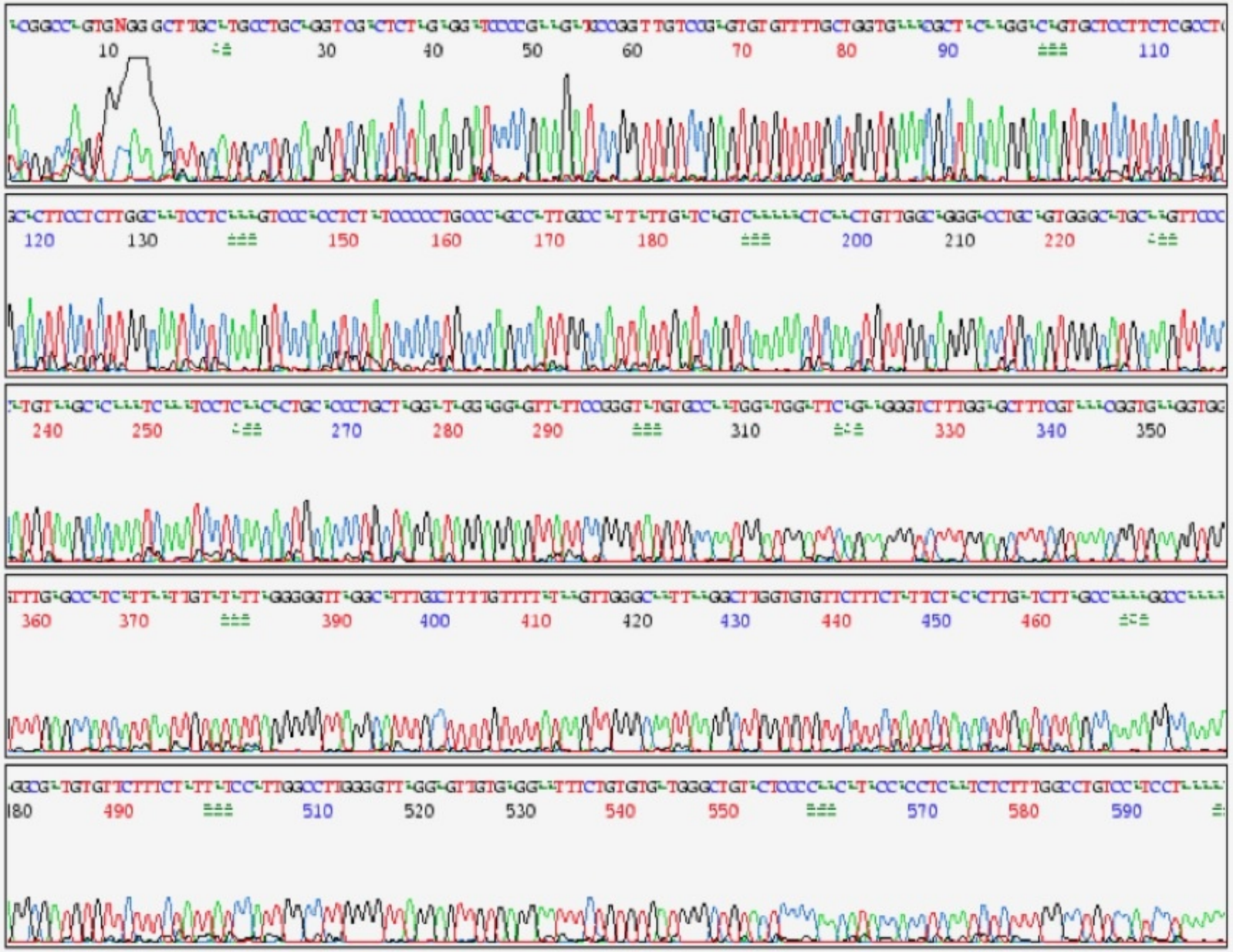
Sequencing using Chain terminators



DNA Sequencing by Chain Termination

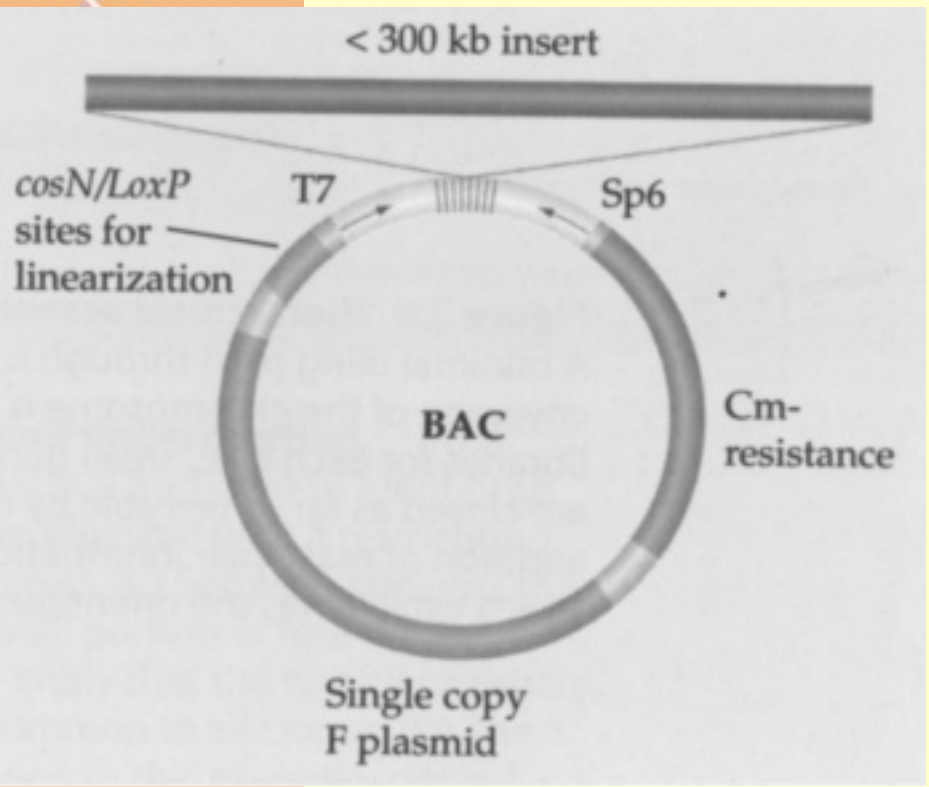


Fluorescent DNA Sequencing

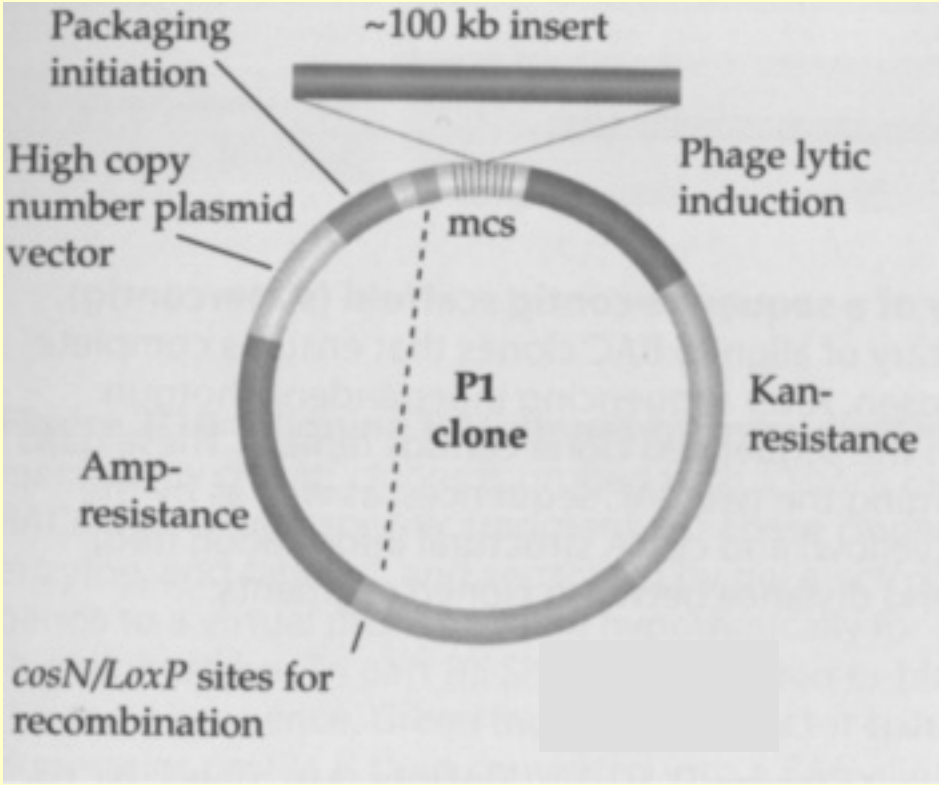


Bacterial Cloning Vectors Used in Genome Sequencing

BAC Vector

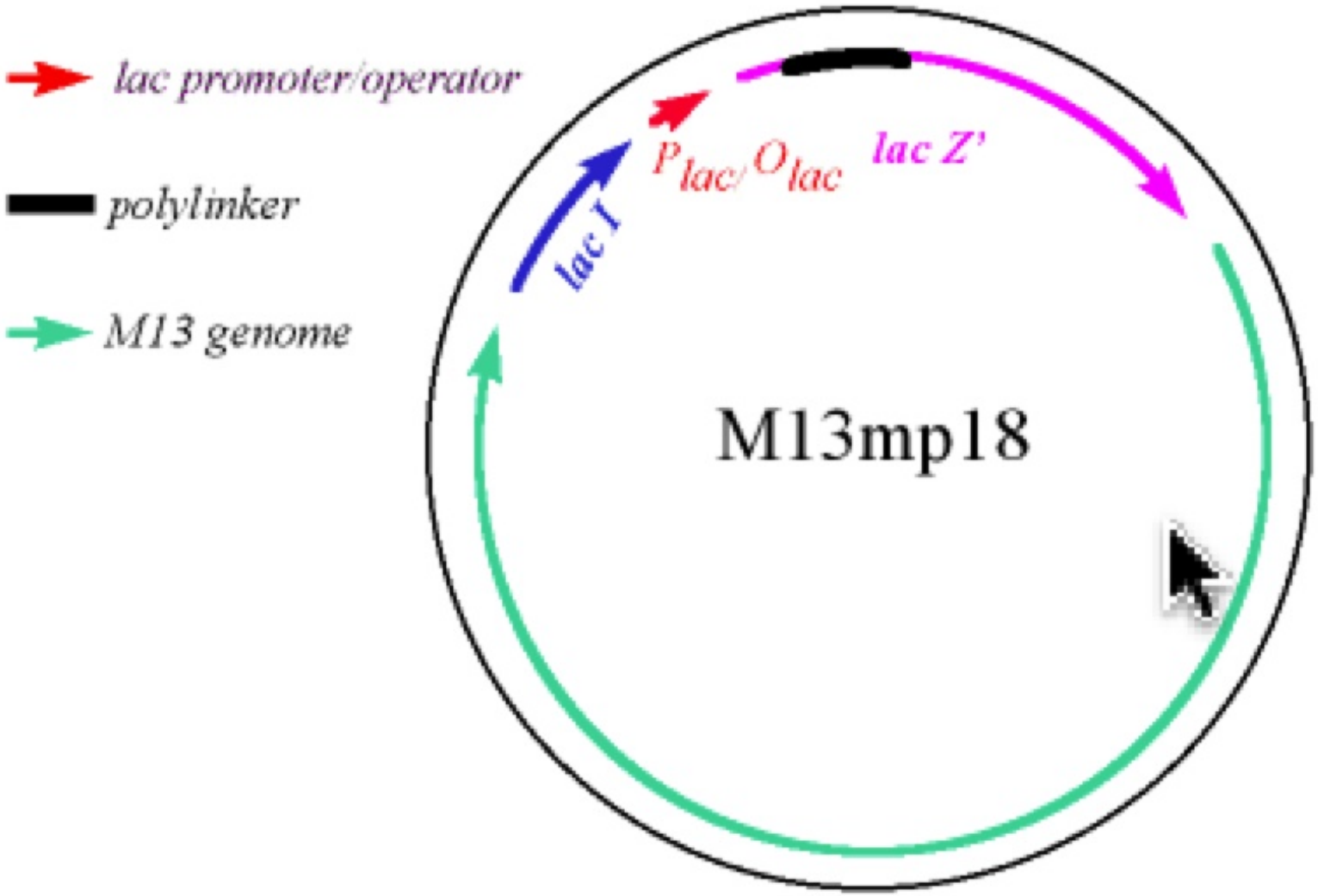


PAC Vector



M13 Sequencing Vector

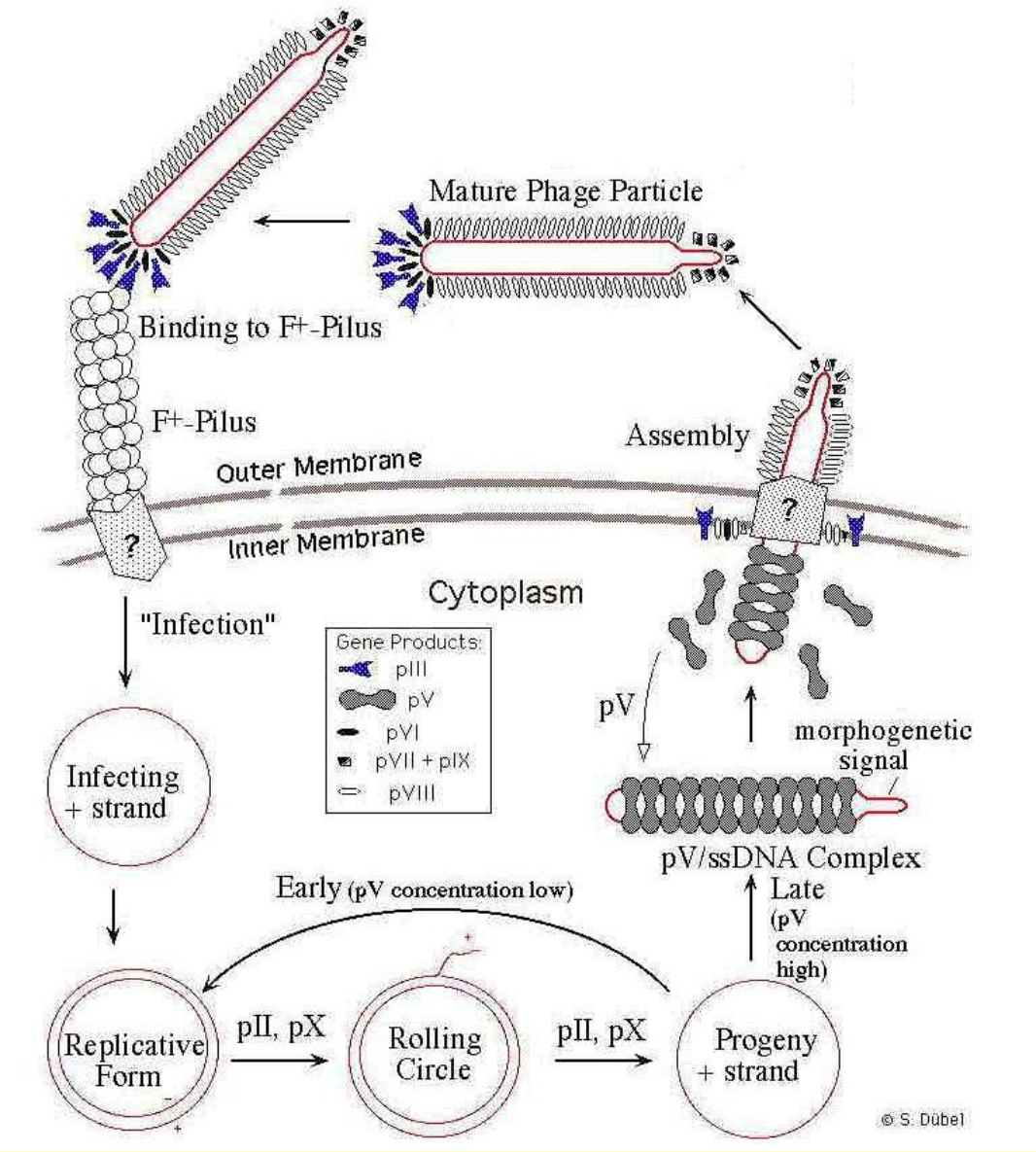
http://www.mikeblaber.org/oldwine/bch5425/lect33/lect33.htm



Michael Blaber@sb.fsu.edu

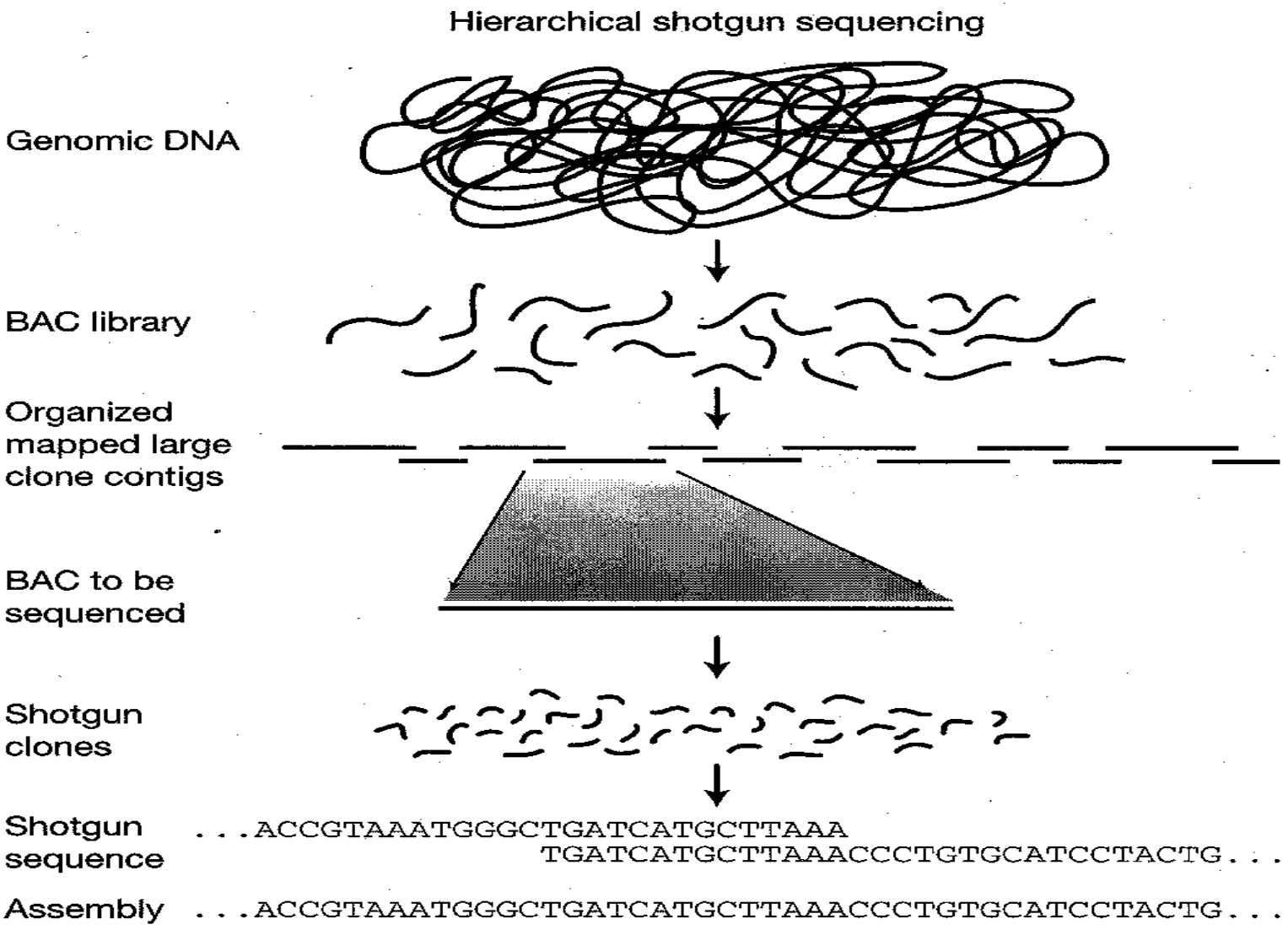
M13 Life Cycle

http://www.elec-intro.com/m13-cloning

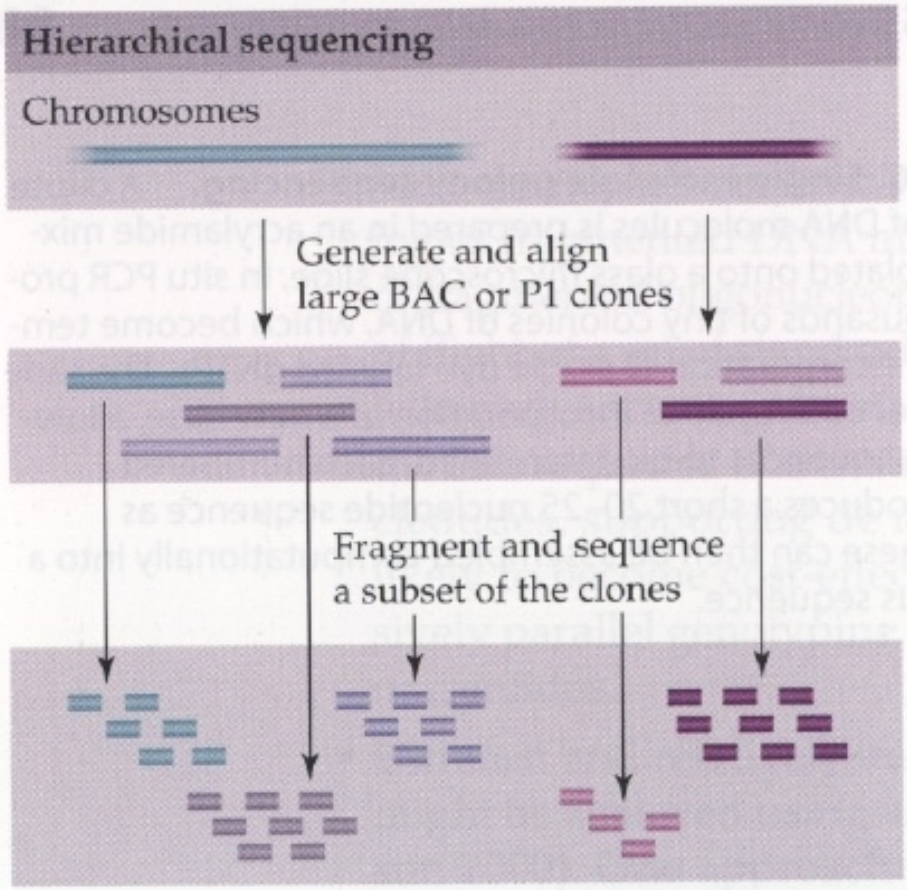


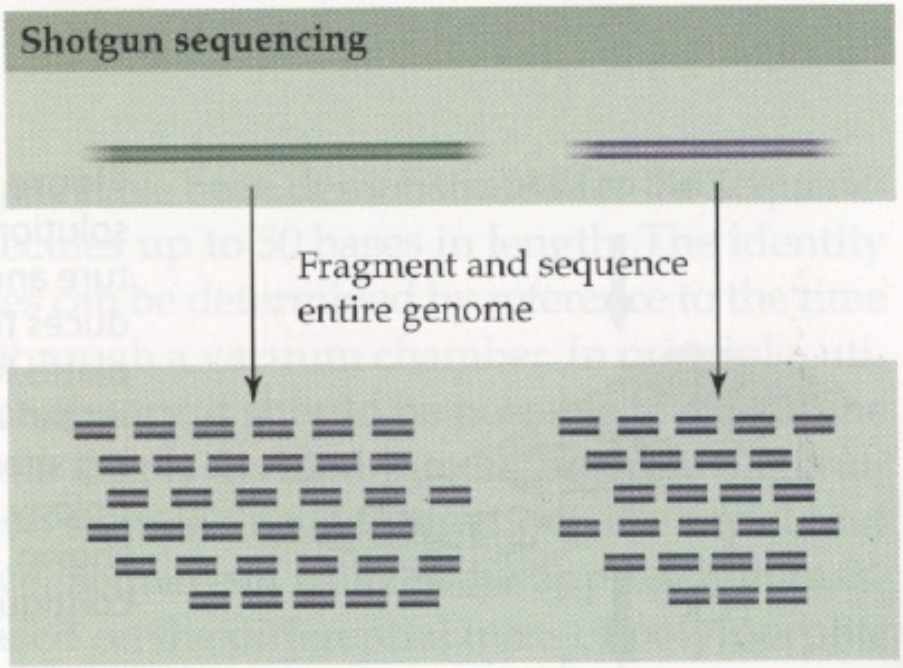
Public Human Genome Project Strategy

Published in Nature 15 February 2001



Hierarchical Sequencing Vs. Whole Genome Shotgun Sequencing



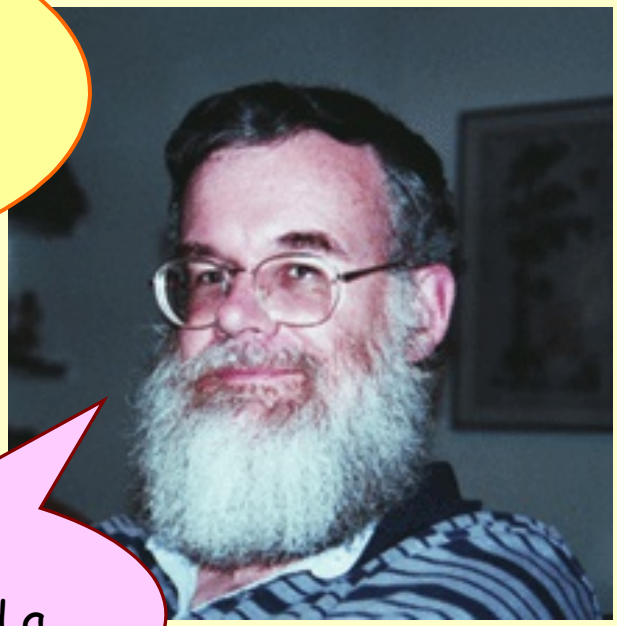


(from Gibson & Muse, A Primer of Genome Science)

Whole Genome Shotgun versus BAC Sequencing

1997

Let's sequence the human genome with the shotgun strategy



That is impossible, and a bad idea anyway

Phil Green

Gene Myers

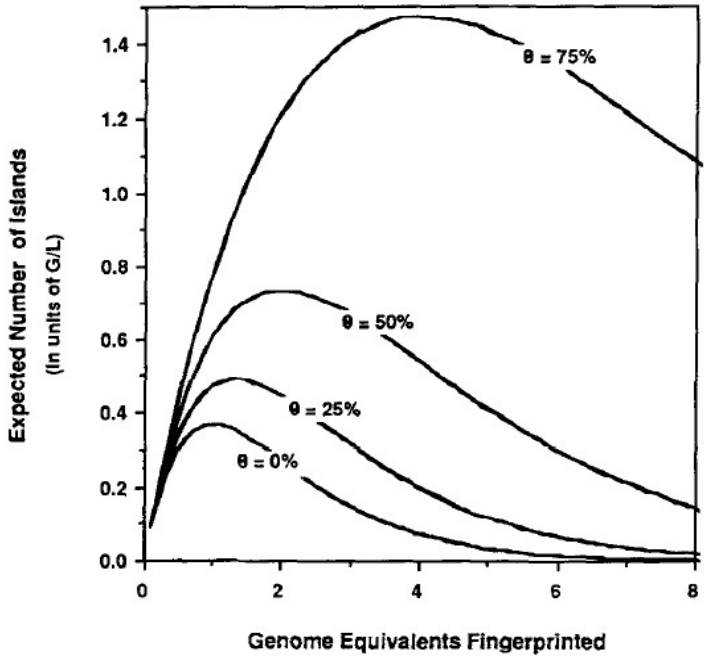
Thanks to Seraf ın Batzoglou

The Human Genome Project: How should we do it?

- Weber, J. L., & Myers, E. W. (1997). Human whole-genome shotgun sequencing *Genome Res, 7(5), 401-409.*
  - Use clone end sequencing generating mate-pairs
    - Referred to as double shotgun sequencing
  - Use multiple length clones 2 kb, 10 kb and 50 kb
  - Able to use long clones to leap over repeated regions
  - Clone length permits one to measure length of repeated regions.
  - Will find more polymorphisms (SNPs)
  - Costs less
  - Finishing easier
  - BAC clone artifacts
    - Differential amplification
    - BACs not stable in bacteria will be lost.
    - Repeated regions will recombine and be lost

- Green, P. (1997). Against a whole-genome shotgun. *Genome Res, 7 (5), 410-417.*
  - Preferred clone-by-clone BAC sequencing
  - Distributed versus monolithic organization
  - BACs linked to genetic maps
  - Costs less (sequence 4x human genome)
  - Finishing simplified and fewer gaps
  - Haplotyping automatic
  - Longer repeat regions lengths measured

Rate of Contig Formation
Lander & Waterman 1988

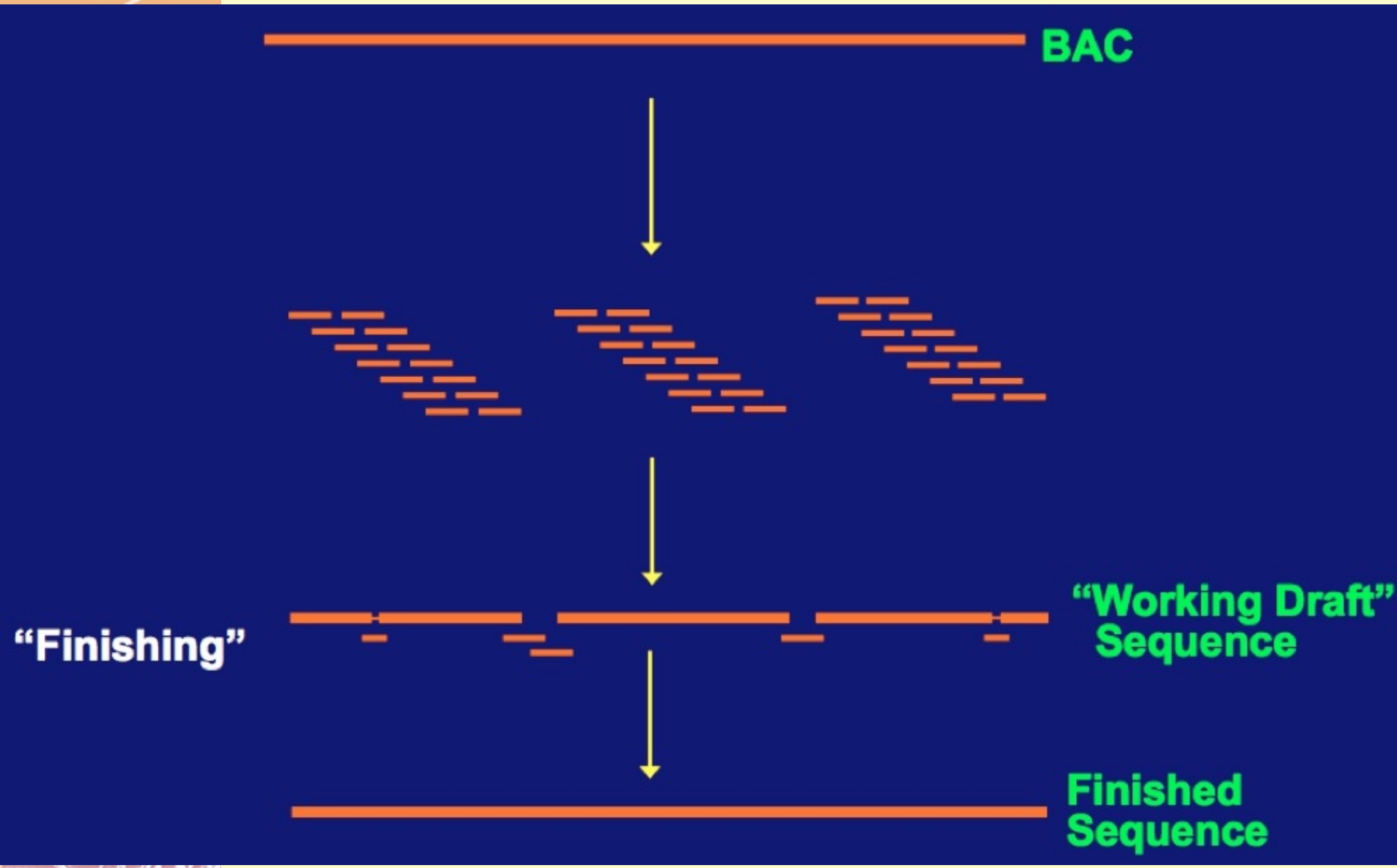


Approximate value of G/L

	Phage (15kb)	Cosmid (40kb)	Yeast (1Mb)
E. coli	267	100	4
S. cerevisiae	1333	500	20
C. elegans	5,667	2,125	85
Human	200,000	75,000	3,000

G = haploid genome length in bp;
L = length of clone insert in bp;
N = number of clones fingerprinted;
$\alpha = N/G$ = probability per base of starting a new clone;
T = amount of overlap in base pairs needed to detect overlap;
$\theta = T/L$;
c = redundancy of coverage $= LN/G$.

BAC Shotgun Sequencing Strategy



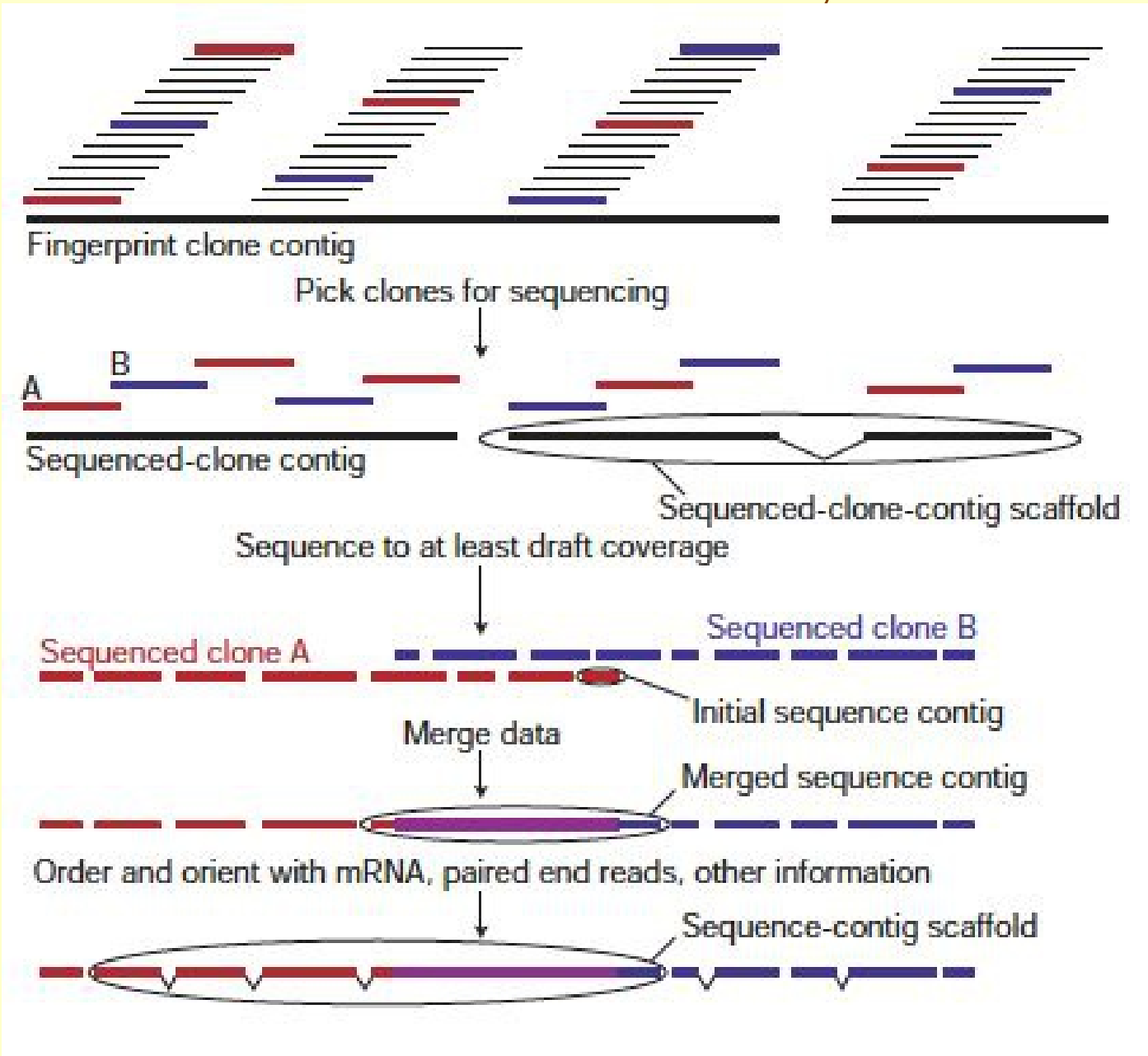
BAC and PAC Libraries in the Public Human Genome Project

Table 1 Key large-insert genome-wide libraries

Library name*	GenBank abbreviation	Vector type	Source DNA	Library segment or plate numbers	Enzyme digest	Average insert size (kb)	Total number of clones in library
Caltech B	CTB	BAC	987SK cells	All	*Hin*dIII	120	74,496
Caltech C	CTC	BAC	Human sperm	All	*Hin*dIII	125	263,040
Caltech D1 (CITB-H1)	CTD	BAC	Human sperm	All	*Hin*dIII	129	162,432
Caltech D2 (CITB-E1)		BAC	Human sperm	All			
				2,501–2,565	*Eco*RI	202	24,960
				2,566–2,671	*Eco*RI	182	46,326
				3,000–3,253	*Eco*RI	142	97,536
RPCI-1	RP1	PAC	Male, blood	All	*Mbo*I	110	115,200
RPCI-3	RP3	PAC	Male, blood	All	*Mbo*I	115	75,513
RPCI-4	RP4	PAC	Male, blood	All	*Mbo*I	116	105,251
RPCI-5	RP5	PAC	Male, blood	All	*Mbo*I	115	142,773
RPCI-11	RP11	BAC	Male, blood	All		178	543,797
				1	*Eco*RI	164	108,499
				2	*Eco*RI	168	109,496
				3	*Eco*RI	181	109,657
				4	*Eco*RI	183	109,382
				5	*Mbo*I	196	106,763
Total of top							1,482,502

https://bacpac.chori.org/

Public Genome Assembly Process



Total Genome Sequence Information 2001

Table 2 Total genome sequence from the collection of sequenced clones, by sequence status

Sequence status	Number of clones	Total clone length (Mb)	Average number of sequence reads per kb*	Average sequence depth†	Total amount of raw sequence (Mb)
Finished	8,277	897	20–25	8–12	9,085
Draft	18,969	3,097	12	4.5	13,395
Predraft	2,052	267	6	2.5	667
Total					23,147

*The average number of reads per kb was estimated based on information provided by each sequencing centre. This number differed among sequencing centres, based on the actual protocols used.

†The average depth in high quality bases (≥99% accuracy) was estimated from information provided by each sequencing centre. The average varies among the centres, and the number may vary considerably for clones with the same sequencing status. For draft clones in the public databases (keyword: HTGS_draft), the number can be computed from the quality scores listed in the database entry.

Whole Genome Shotgun Sequencing
Published in Science 16 February 2001

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

Whole Genome Sequencing Scaffolds

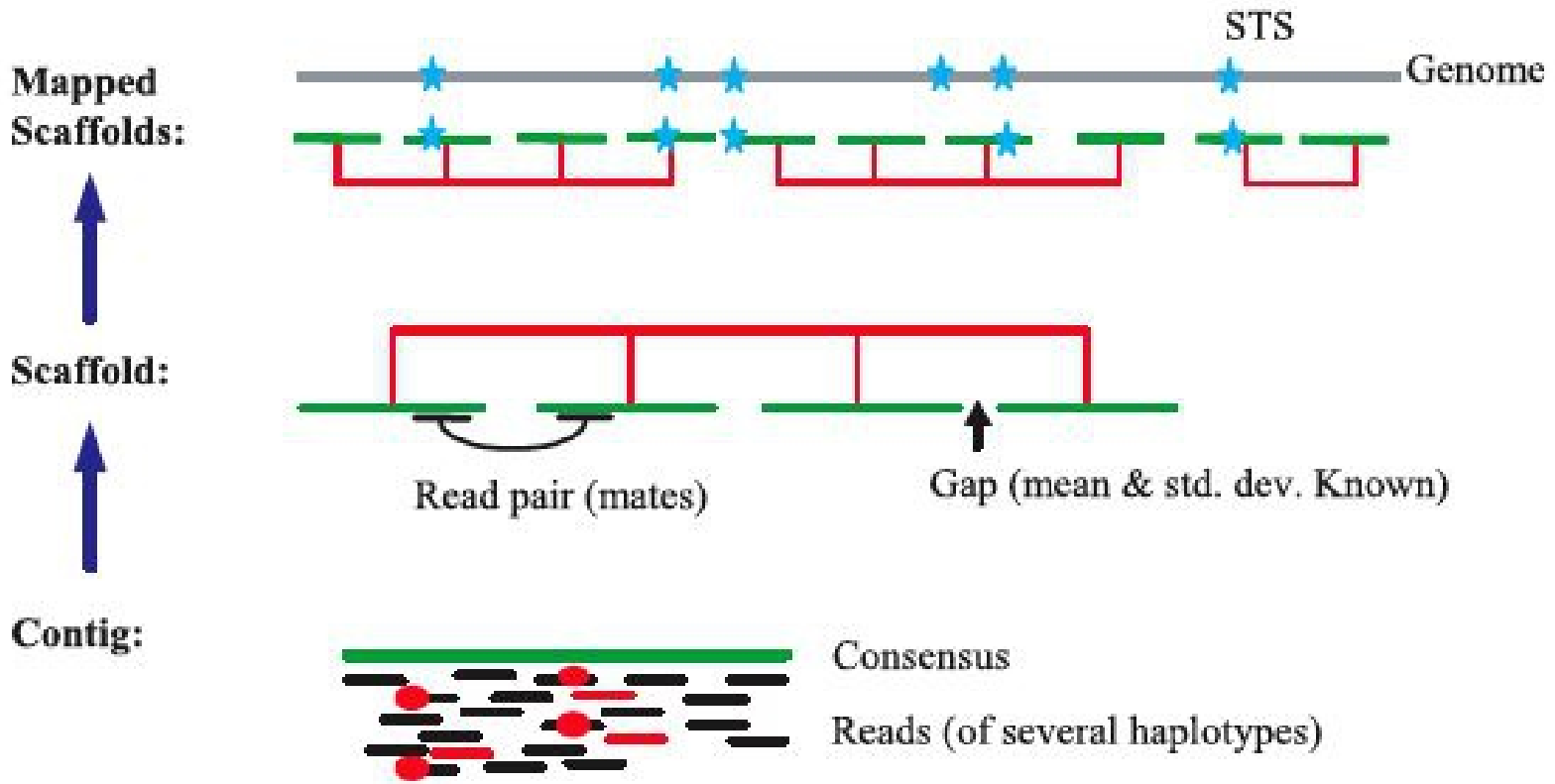
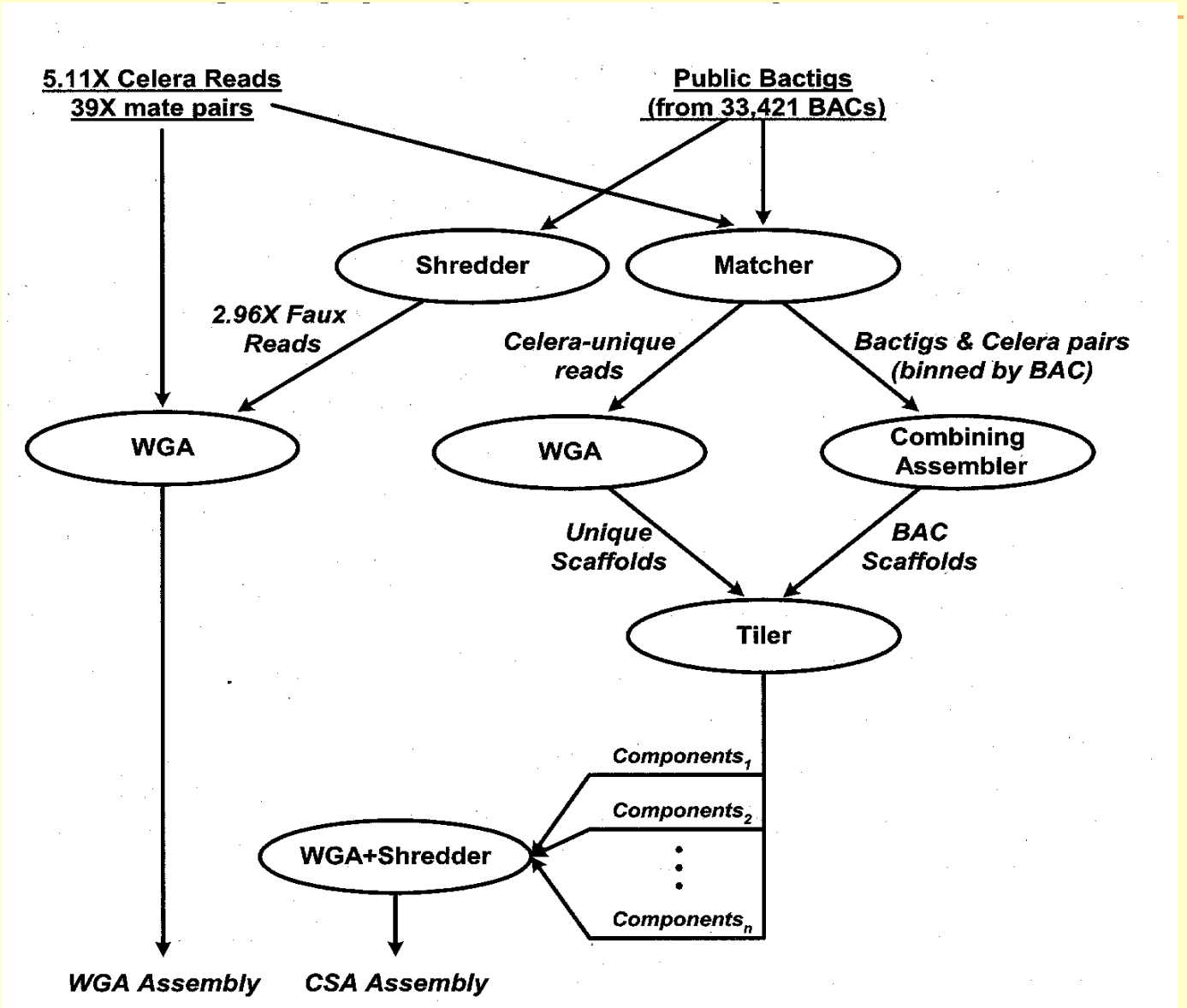


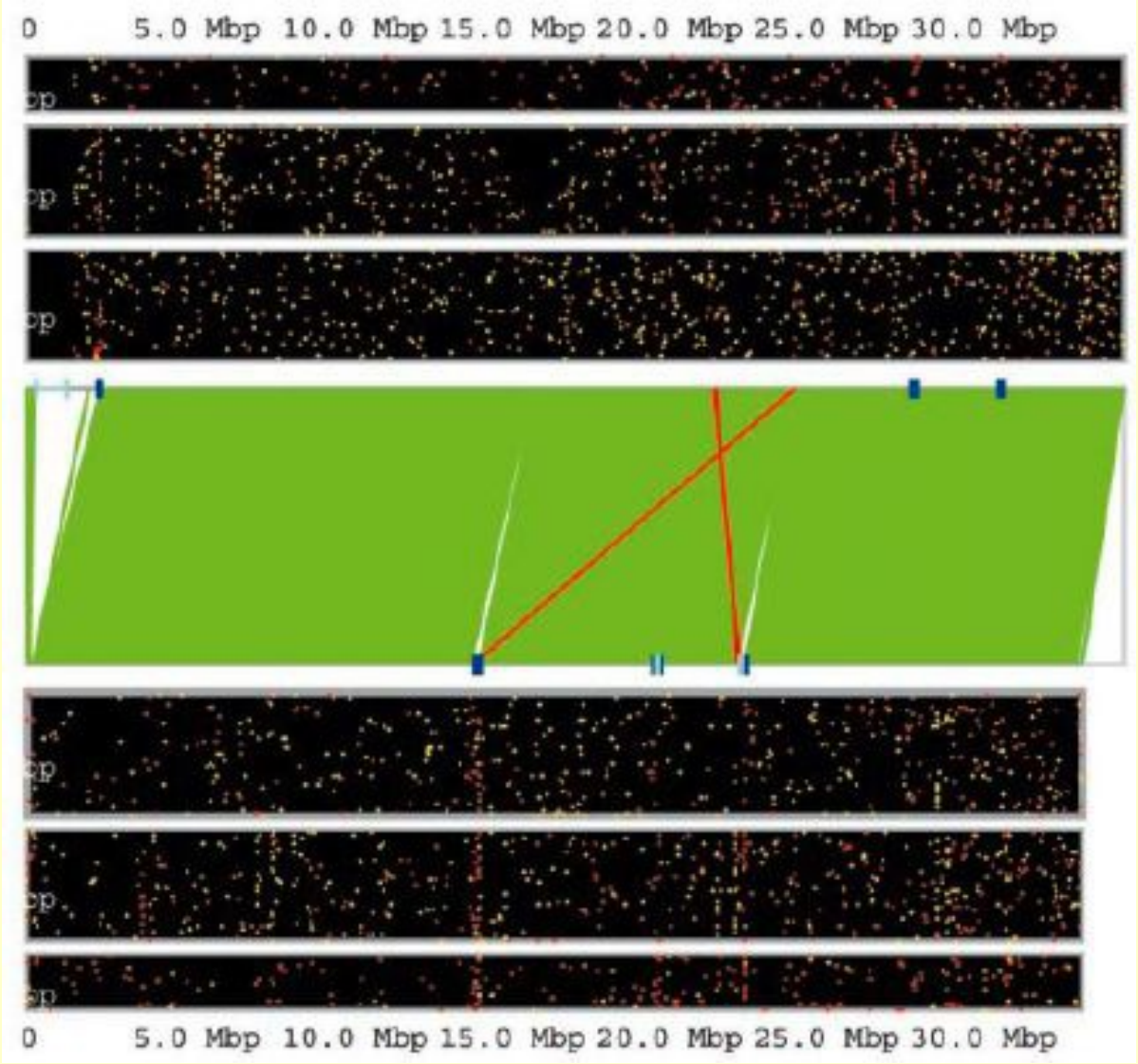
Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

Whole Genome Shotgun Assembler

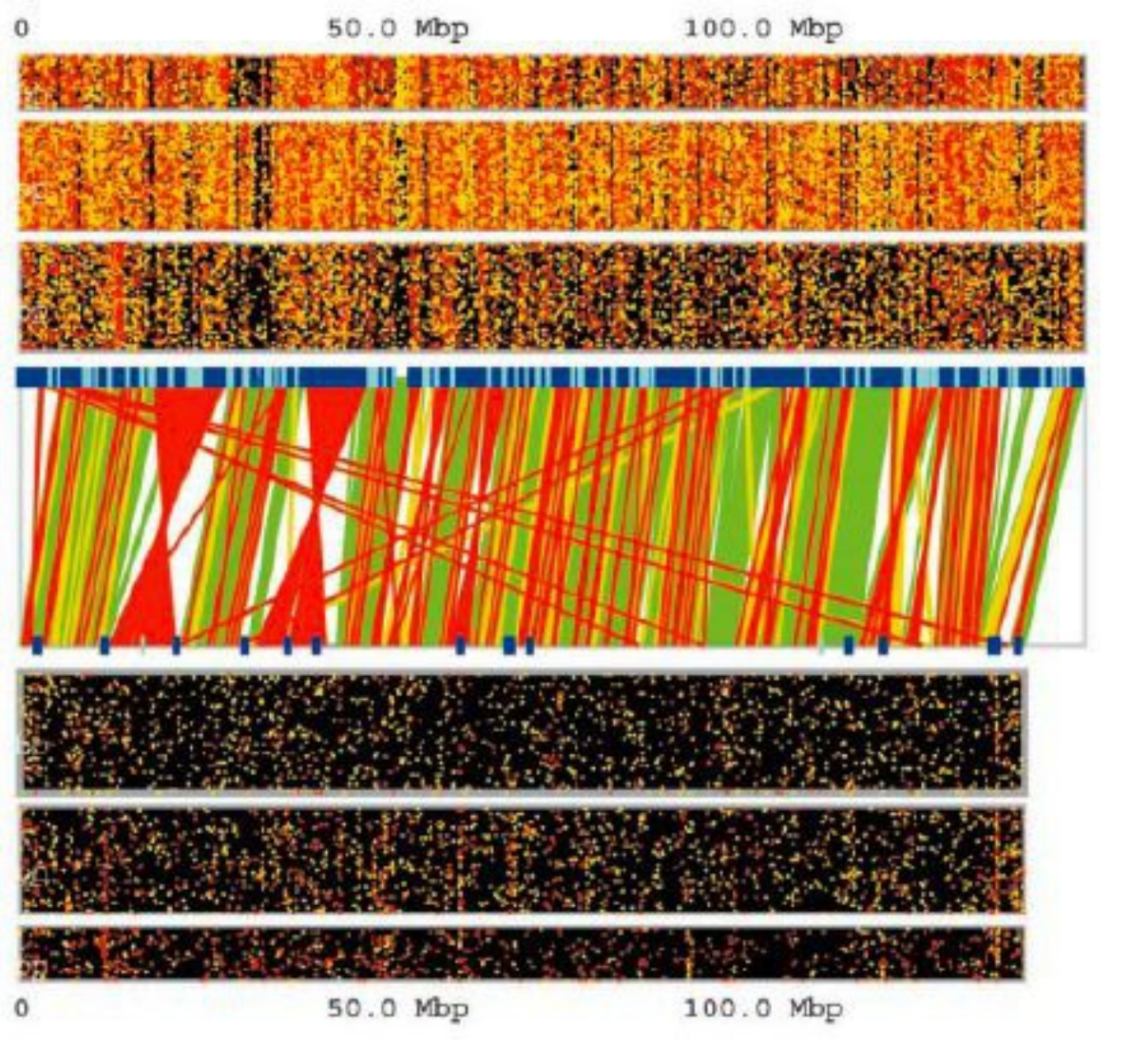
http://www.sciencemag.org/content/291/5507/1304.full



Chromosome 21:
Public vs Whole Genome Shotgun Assemblies



Chromosome 8:
Public vs Whole Genome Shotgun Assemblies



Comparing Chromosome 2 Sequence Ve

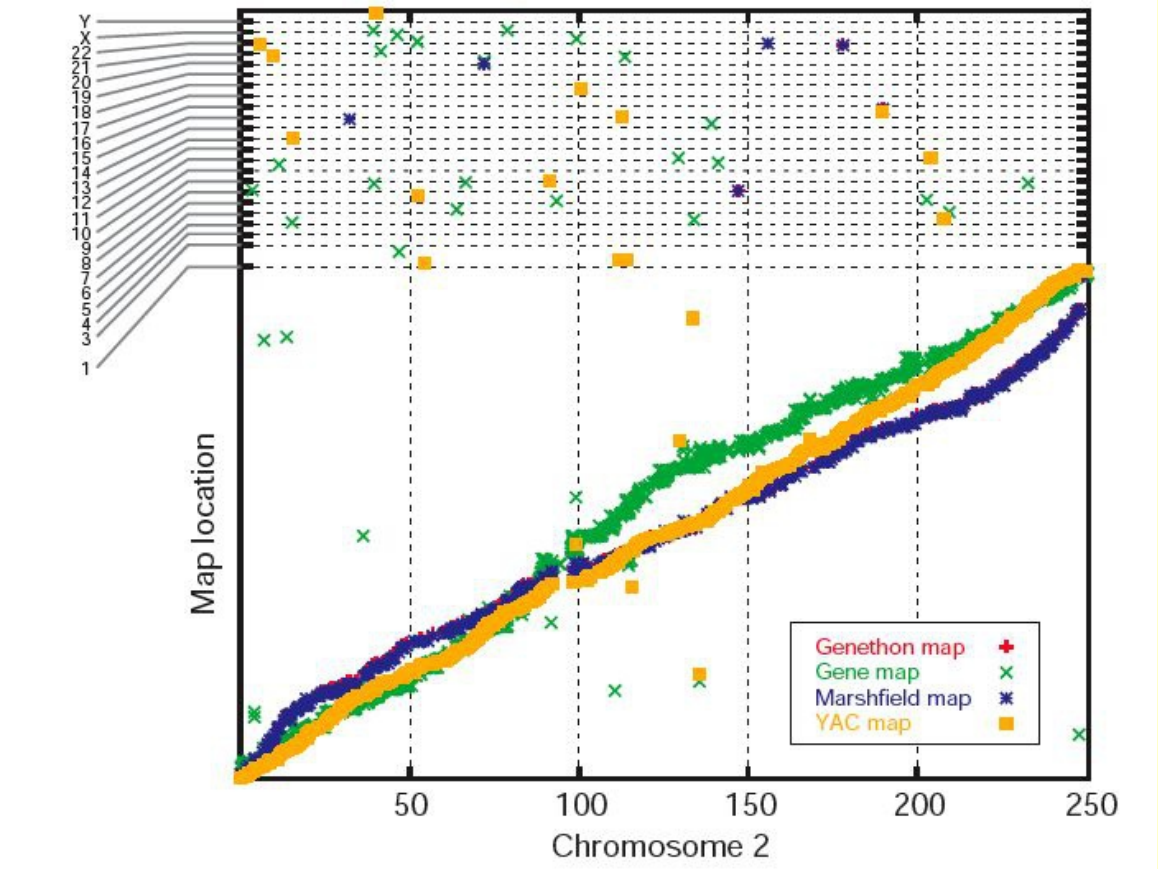
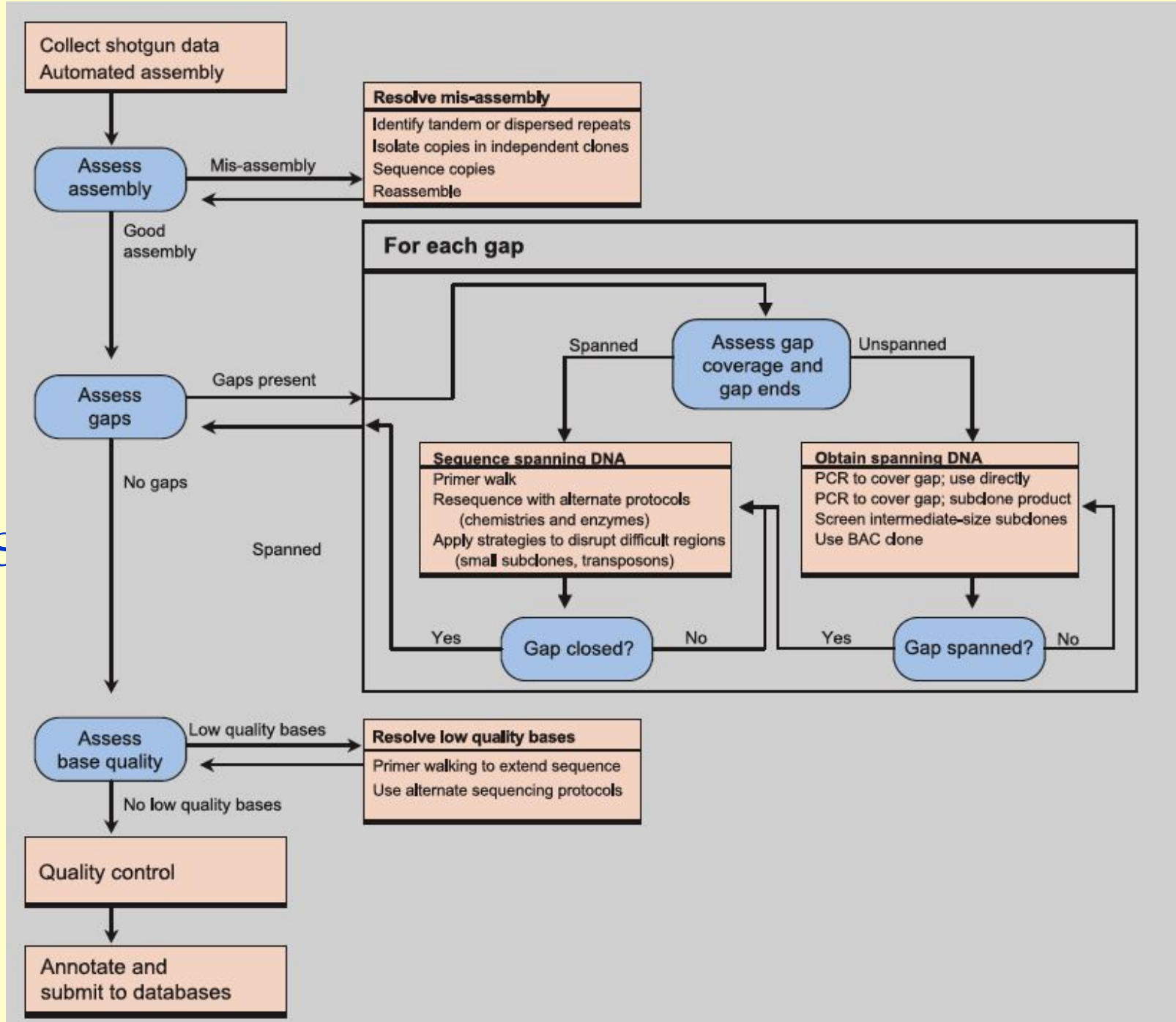


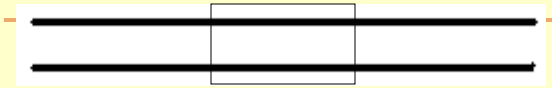
Figure 5 Positions of markers on previous maps of the genome (the Genethon[101] genetic map and Marshfield genetic map (http://research.marshfieldclinic.org/genetics/genotyping_service/mgsver2.htm), the GeneMap99 radiation hybrid map[100], and the Whitehead YAC and radiation hybrid map[29]) plotted against their derived position on the draft sequence for chromosome 2. The horizontal units are Mb but the vertical units of

Finishing S

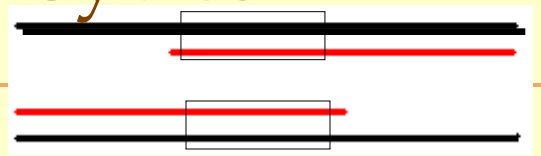

Collect shotgun data
Automated assembly
Assess assembly
Mis-assembly
Resolve mis-assembly
Identify tandem or dispersed repeats
Isolate copies in independent clones
Sequence copies
Reassemble
Good assembly
For each gap
Assess gaps
Gaps present
Spanned
Assess gap coverage and gap ends
Unspanned
Sequence spanning DNA
Primer walk
Resequence with alternate protocols (chemistries and enzymes)
Apply strategies to disrupt difficult regions (small subclones, transposons)
Obtain spanning DNA
PCR to cover gap; use directly
PCR to cover gap; subclone product
Screen intermediate-size subclones
Use BAC clone
Yes
Gap closed?
No
Yes
Gap spanned?
No
No gaps
Spanned
Assess base quality
Low quality bases
Resolve low quality bases
Primer walking to extend sequence
Use alternate sequencing protocols
No low quality bases
Quality control
Annotate and submit to databases

Polymerase Chain Reaction Overview: Exponential Amplification of DNA

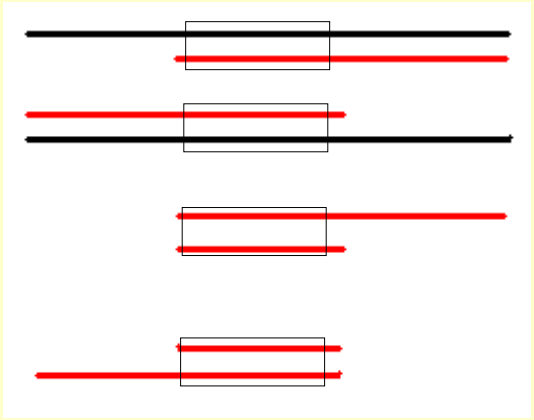
The First Three Cycles



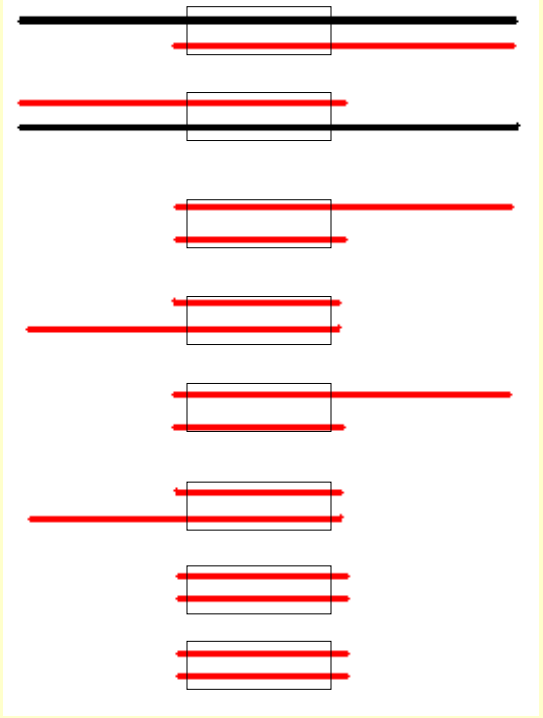
Original DNA



After Cycle 1



After Cycle 2



After Cycle 3

After N cycles, amount of target DNA is 2^N-2N

PCR Requirements

DNA

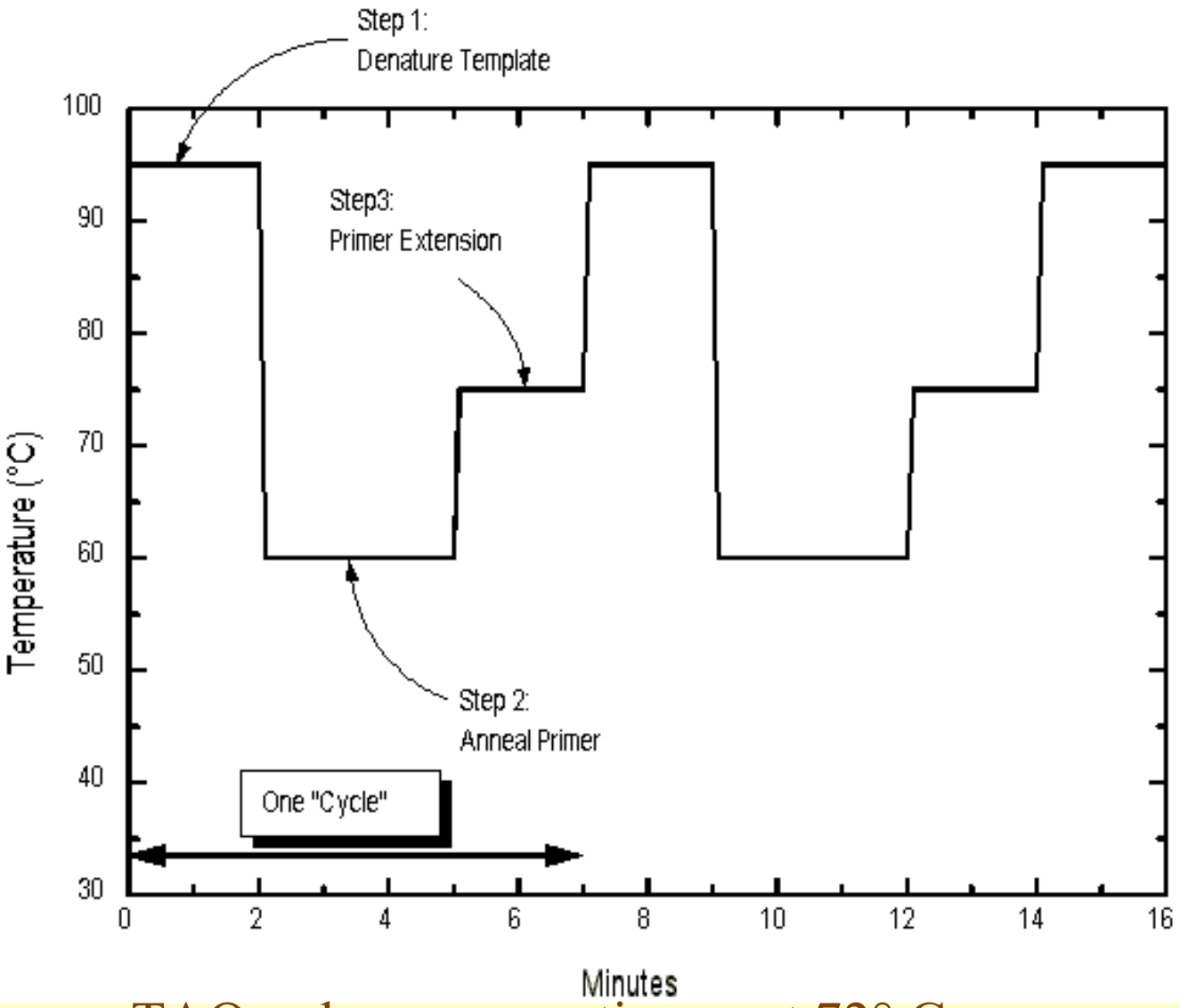
- Need to know at least the beginning and end of DNA sequence
- These flanking regions have to be unique to strand interested in amplifying
- Region of interest can be present in as little as one copy
- *Enough DNA in 0.1 microliter of human saliva to use PCR*

DNA Polymerase Enzyme

- DNA polymerase from *Thermus aquaticus*--Yellowstone
- Alternatives: *Thermococcus litoralis*, *Pyrococcus furiosus*

Thermocycler

Temperature Cycling



TAQ polymerase optimum at 72° C

PCR Applications

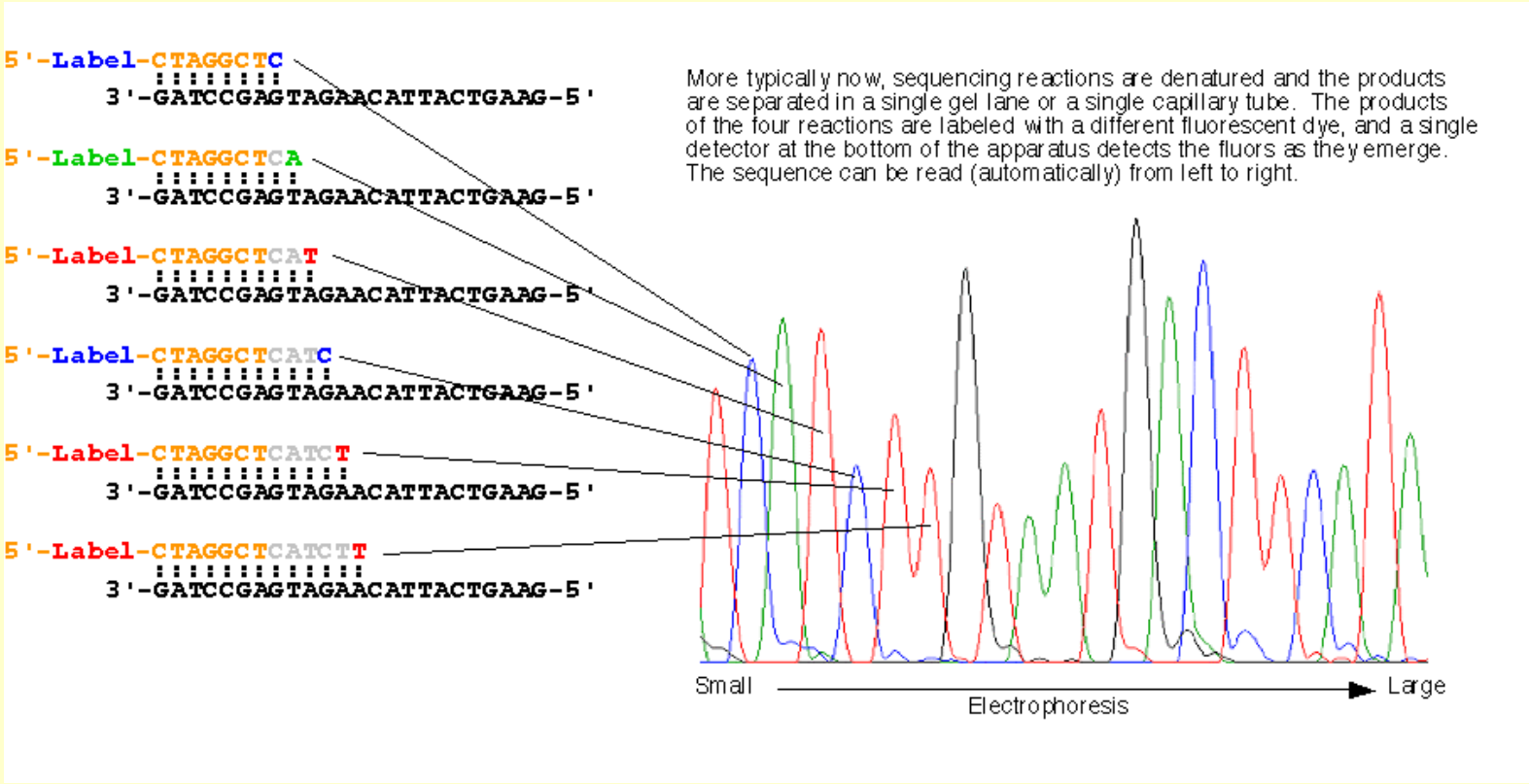
Forensics
- assessment/reassessment of crimes

Archaeology
- determine gene sequences of ancient organisms
- rethinking the past, human origins

Molecular Biology
- Cloning genes
- Sequencing genes
- Finishing genome sequences
- Amplification of DNA or RNA

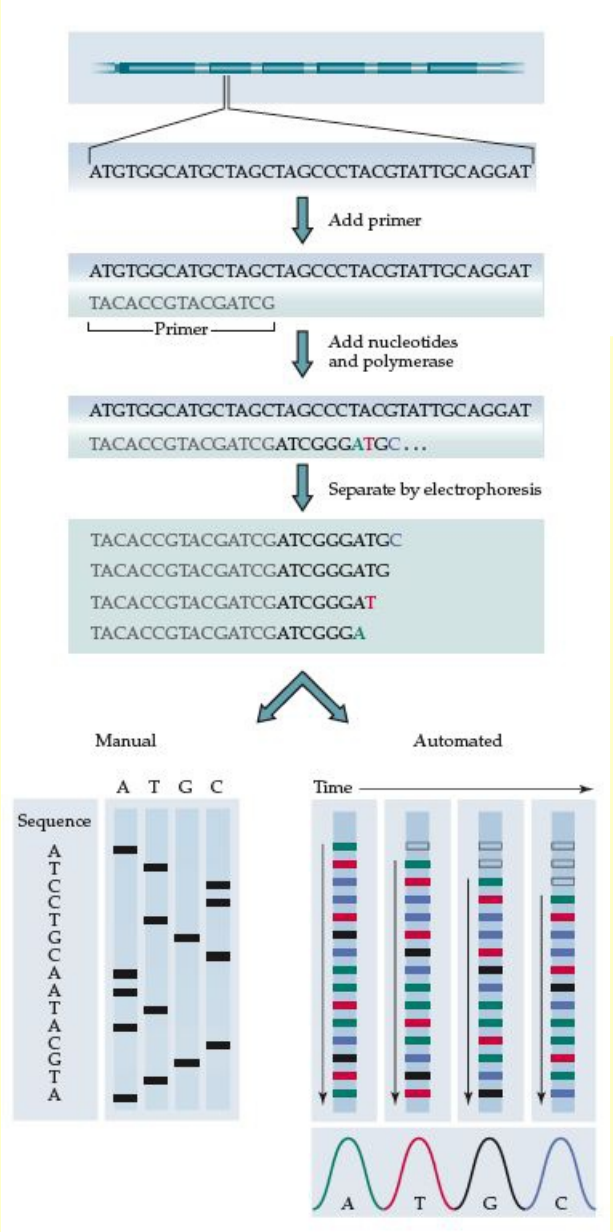
- Medicine
  - Diagnostics for inherited disease
  - Diagnostics for gene expression
  - Diagnostics for gene methylation

DNA Sequencing By Chain Termination

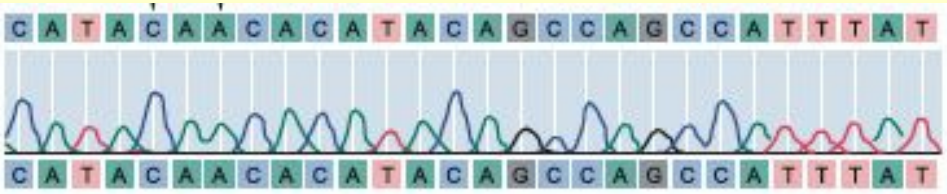


Sanger Sequencing Technology

(from Gibson & Muse, A Primer of Genome Science)



ABI Sequence Trace



Synteny Between Human and Mouse

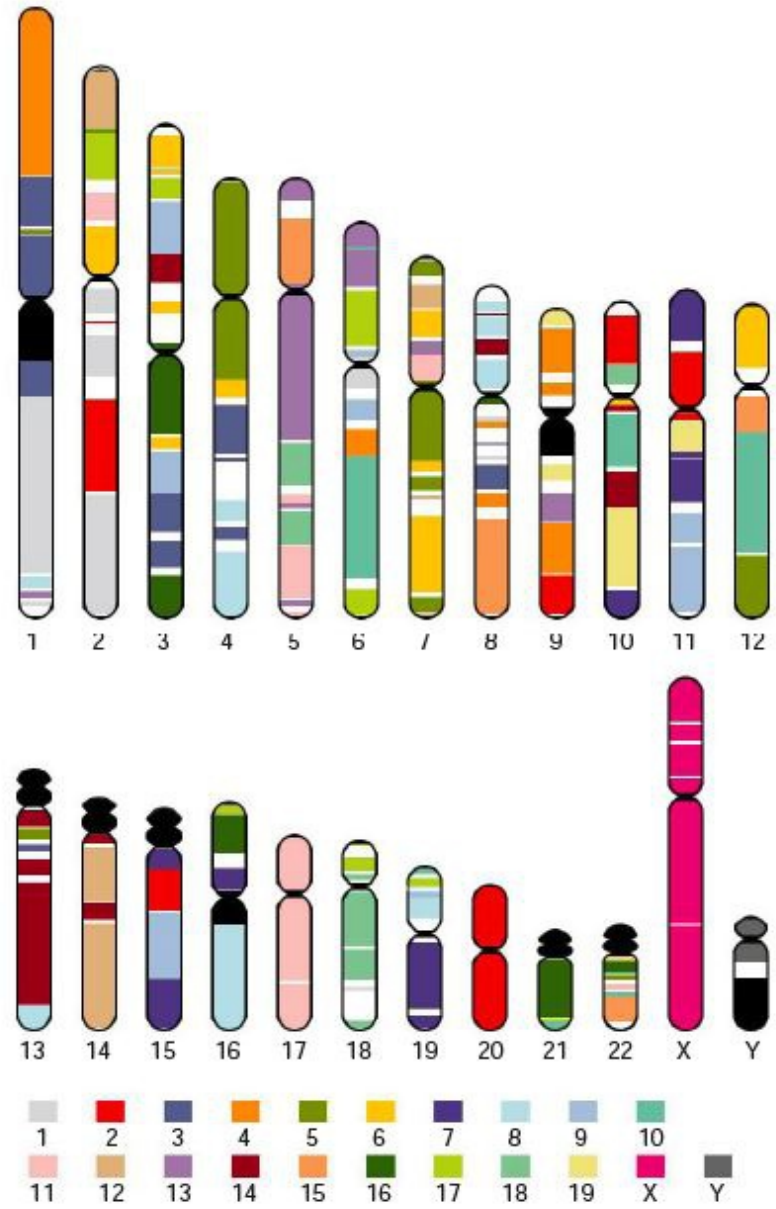


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.