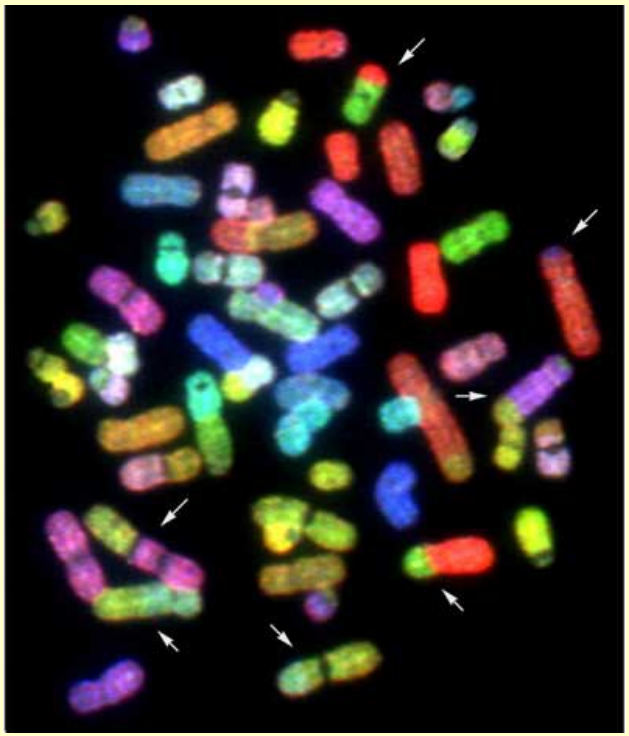
Sequencing the Human Genome

http://biochem118.stanford.edu/

Stanford Sophomore Seminar



Doug Brutlag, Professor Emeritus of
Biochemistry & Medicine (by courtesy)
Stanford University School of Medicine

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 bp per day per person.
  - Move from radioactively labeled sequencing to fluorescent sequencing permitted complete automation up to 1 gigabyte per year.

Covalent Structure of DNA

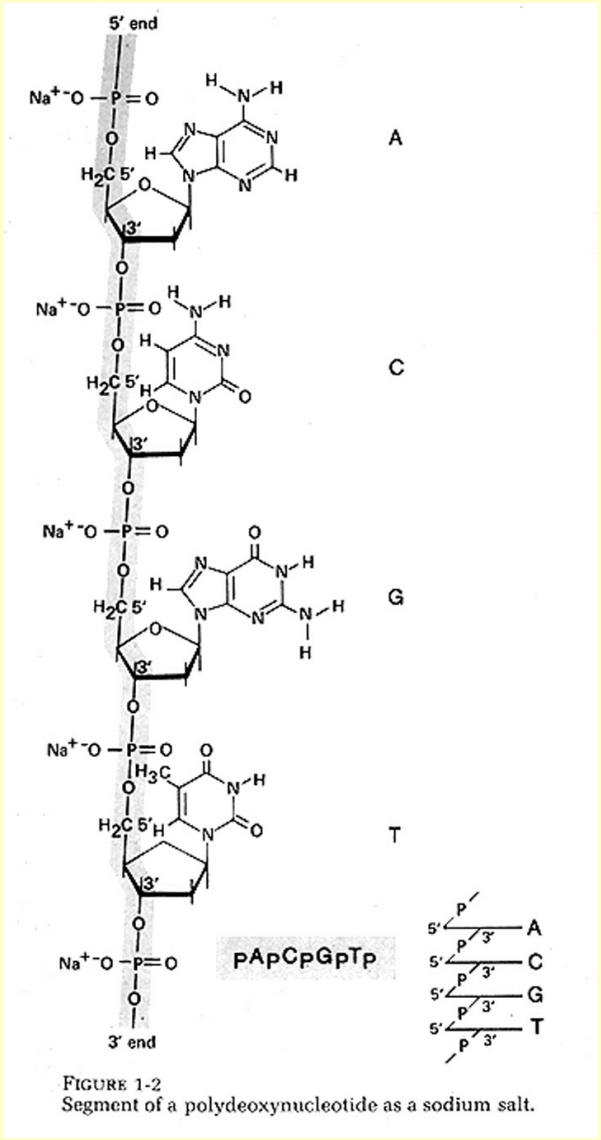
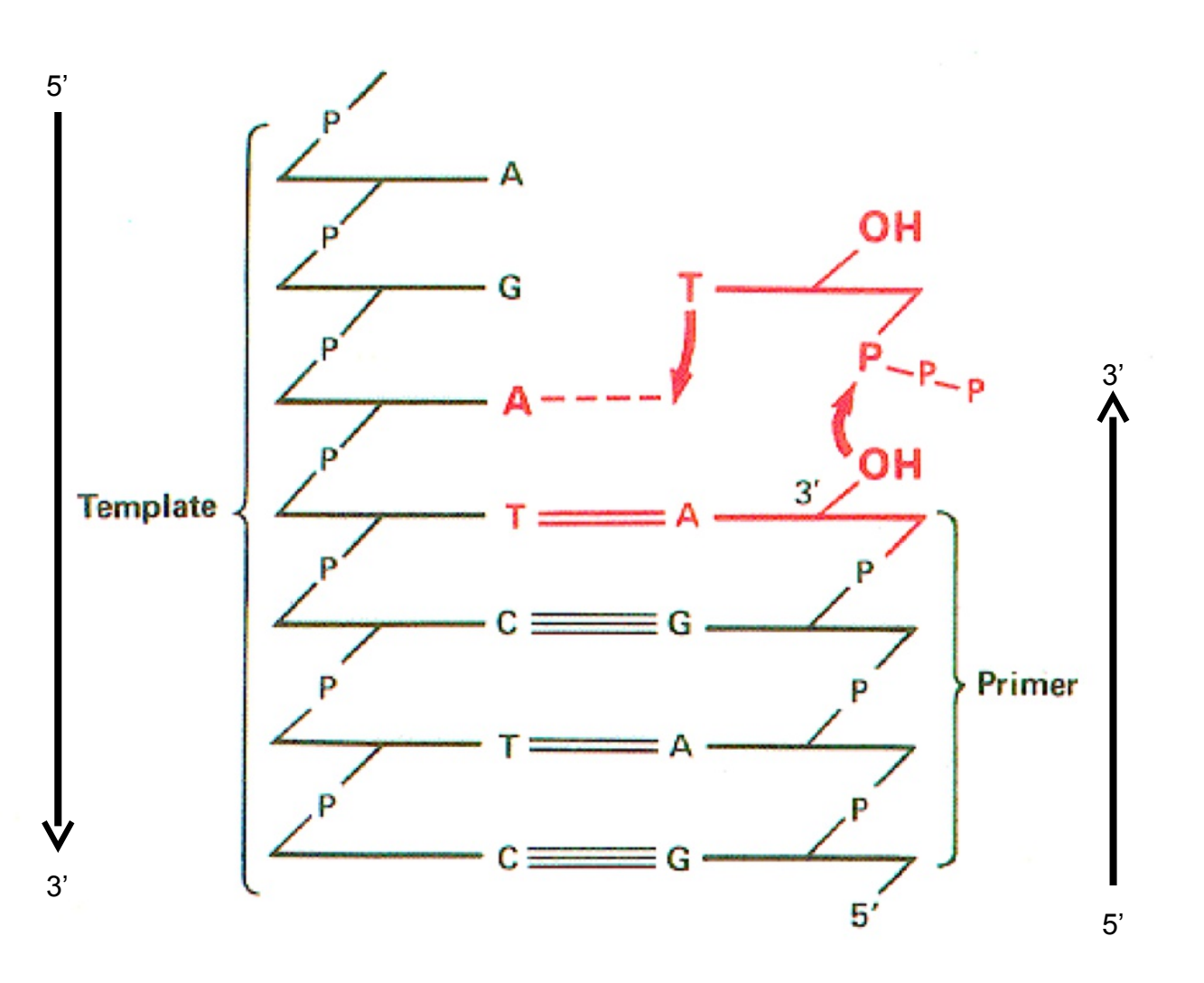
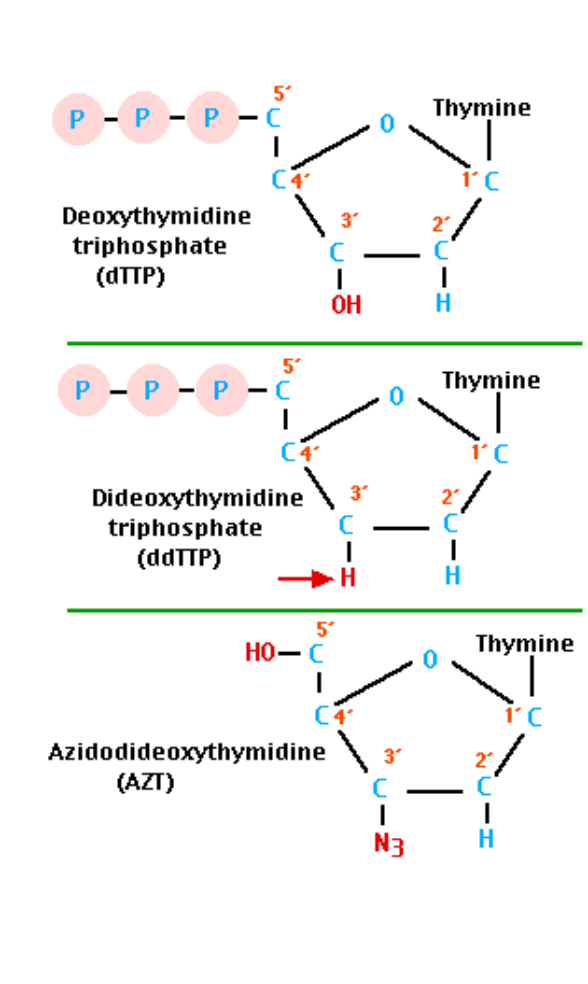


FIGURE 1-2
Segment of a polydeoxynucleotide as a sodium salt.

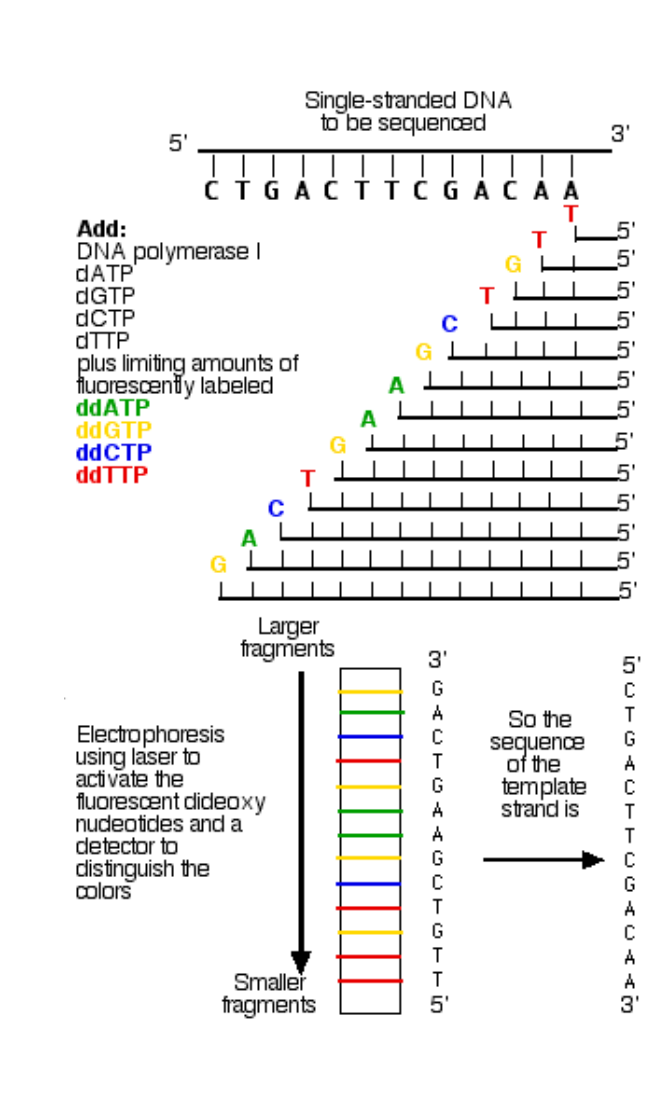
DNA Synthesis by DNA polymerases



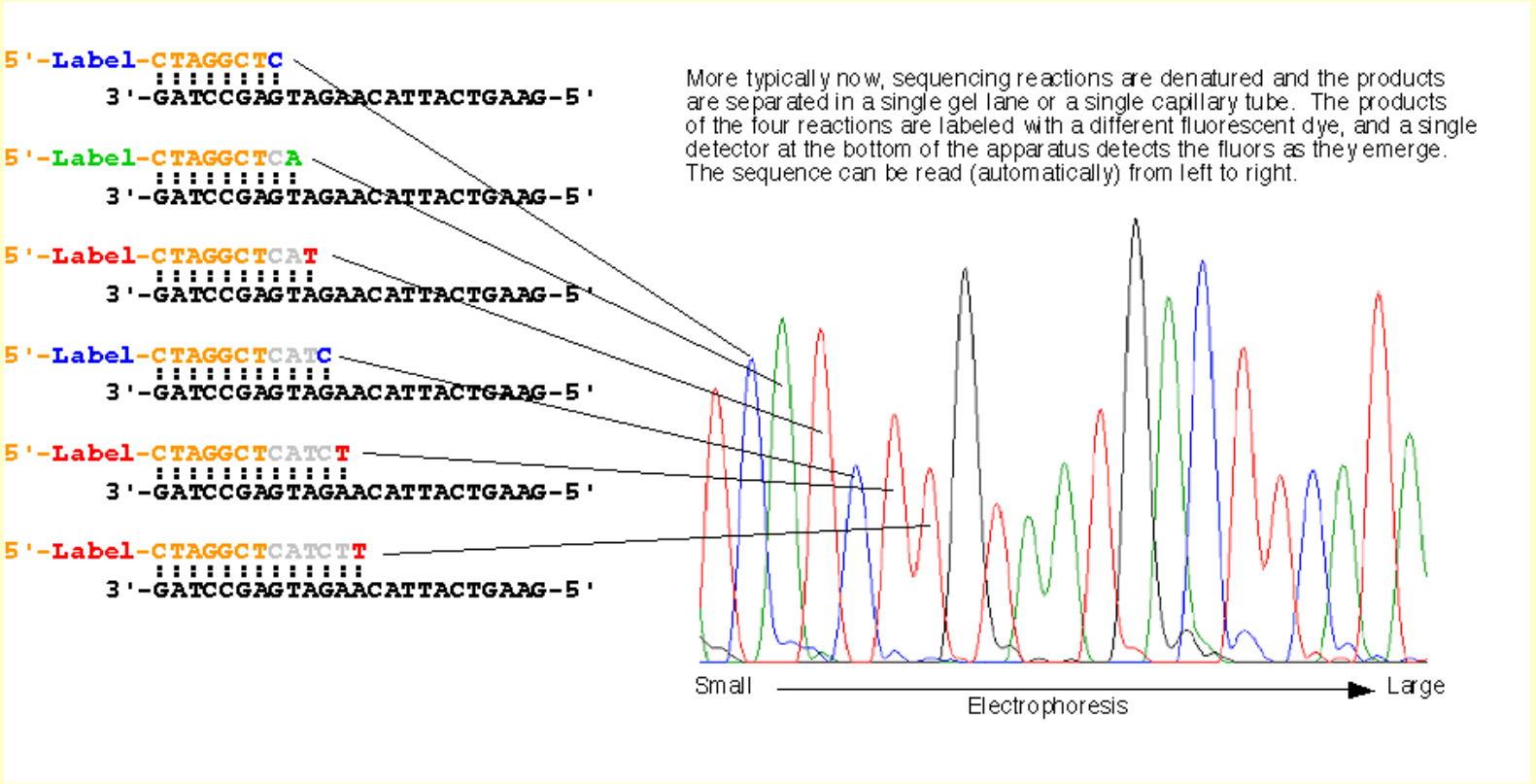
Sequencing using Chain terminators



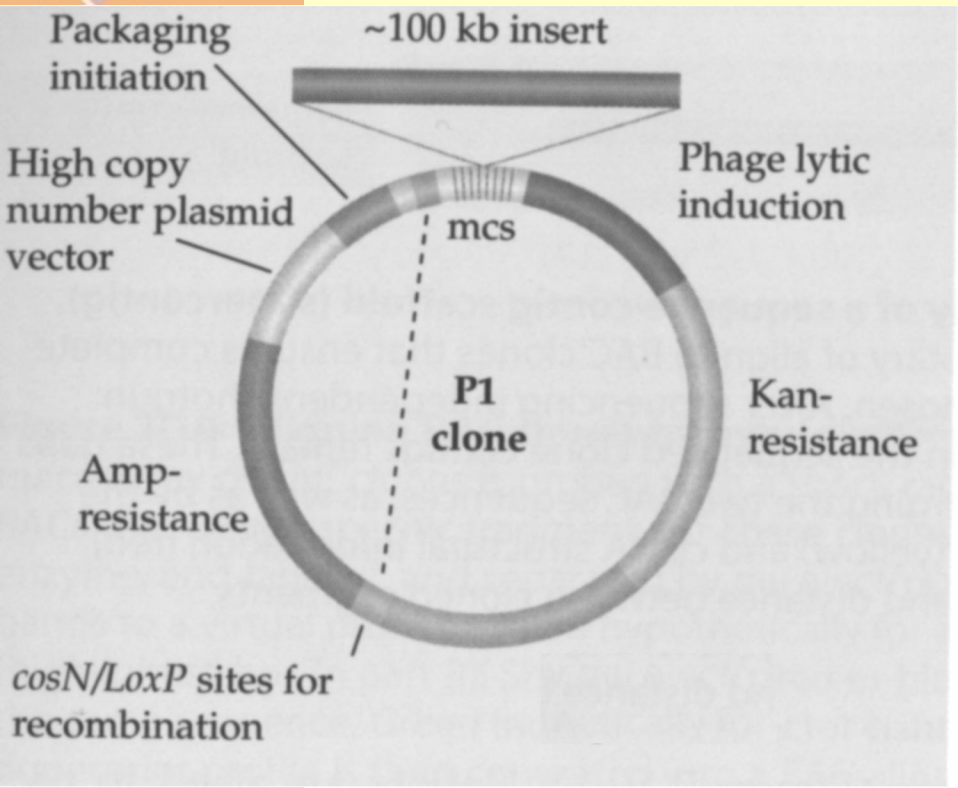
DNA Sequencing by Chain Termination

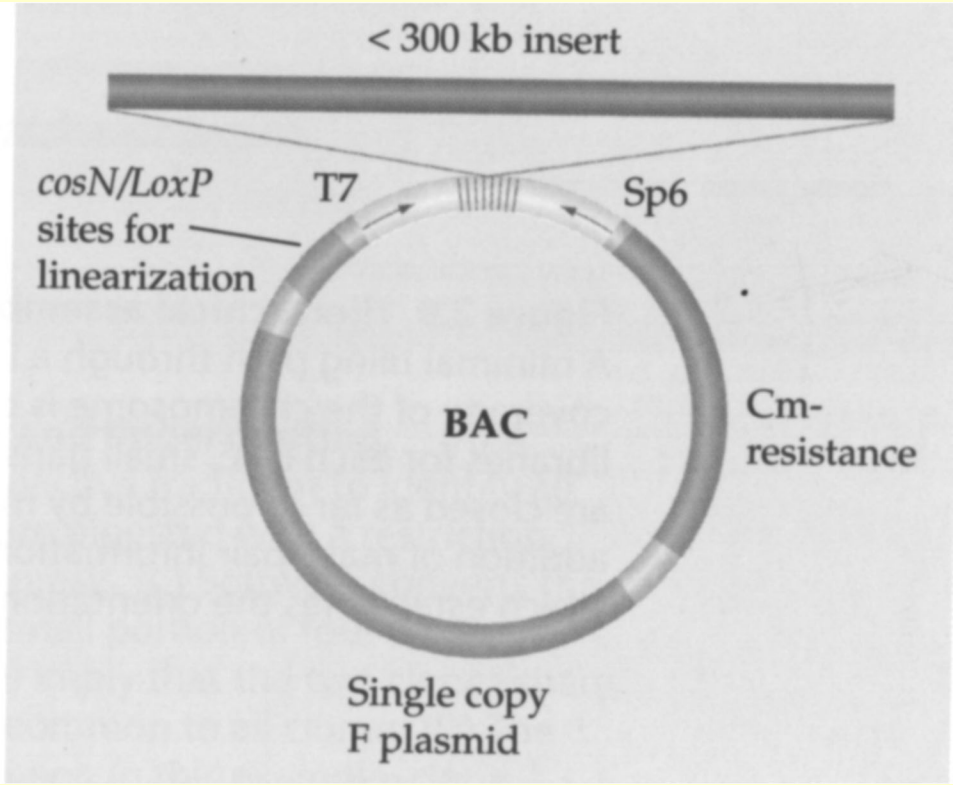


DNA Sequencing By Chain Termination



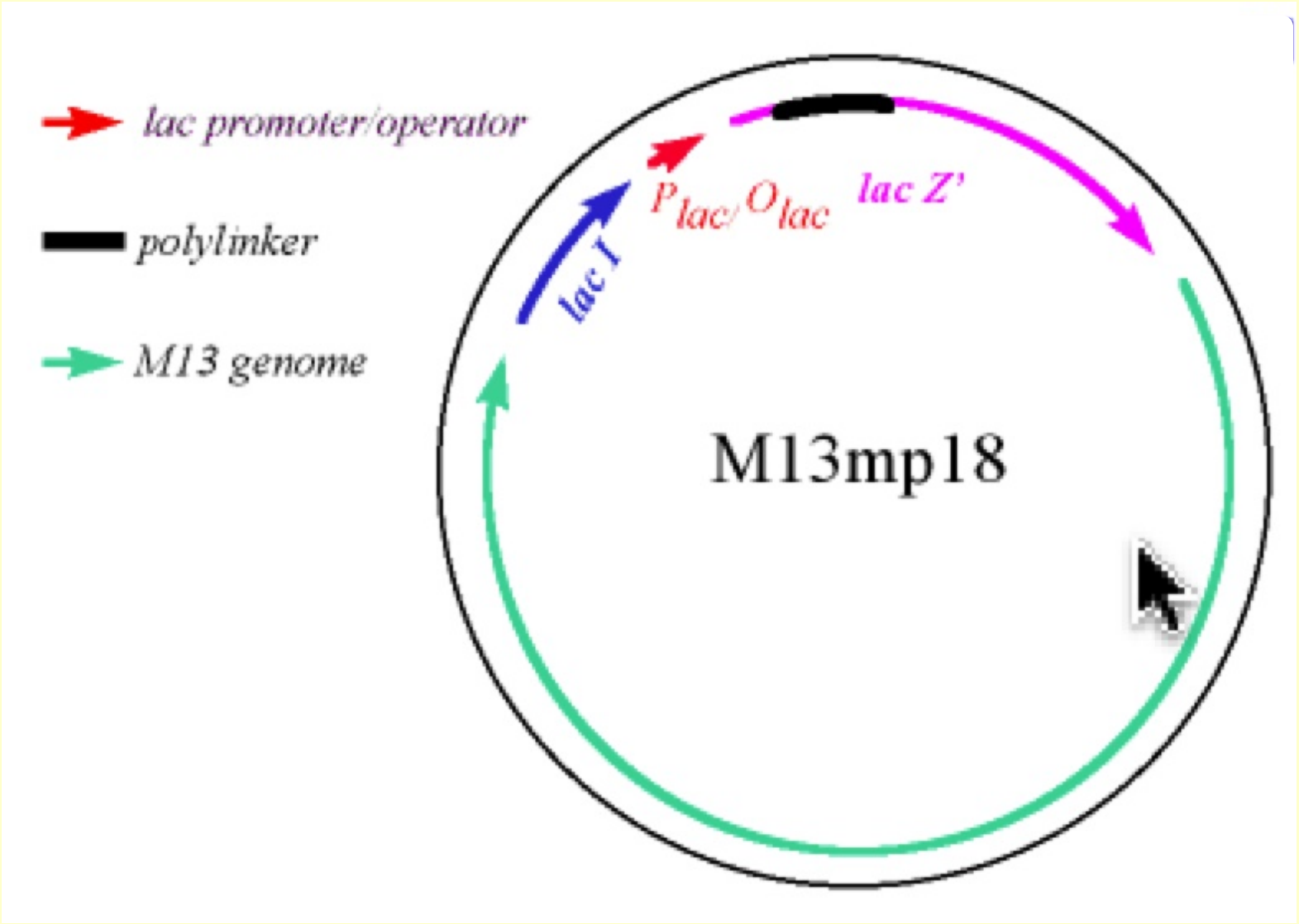
Cloning Vectors Used in Genome Sequencing





M13mp18 Sequencing Vector

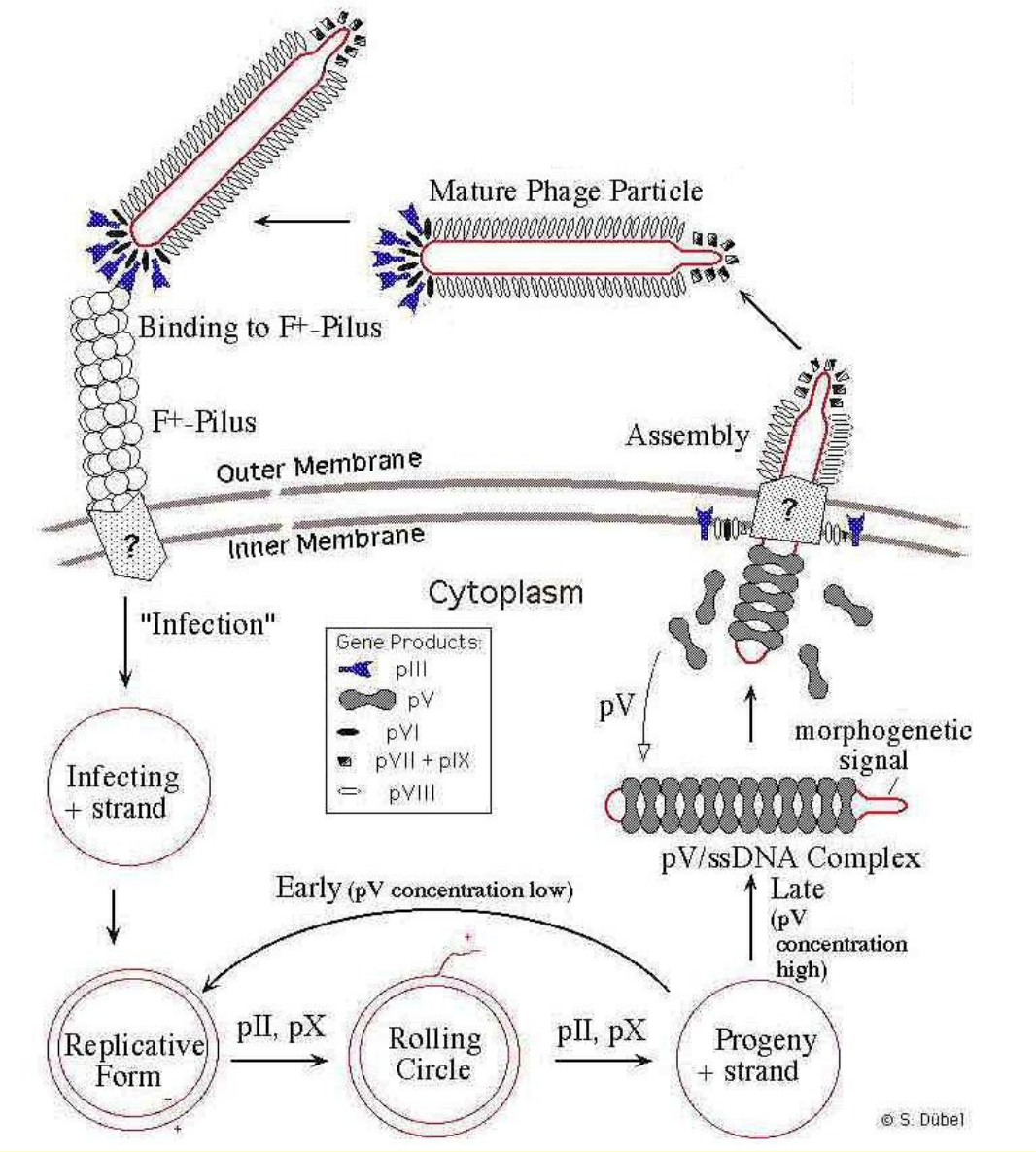
http://www.mikeblaber.org/oldwine/bch5425/lect33/lect33.htm



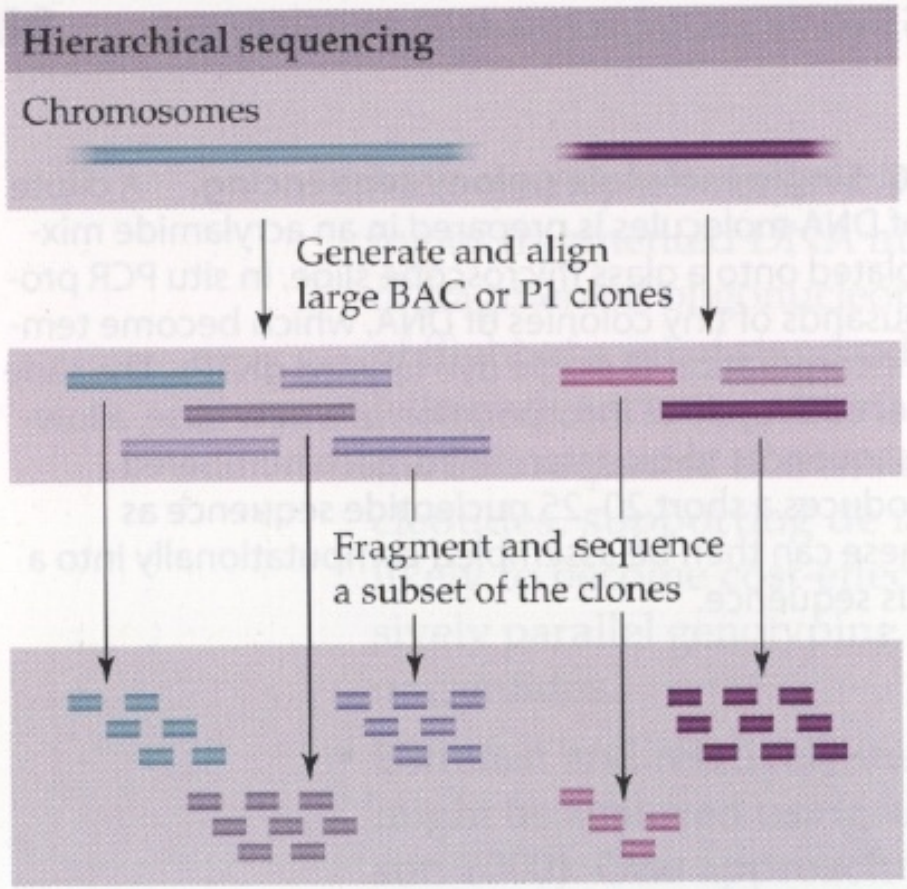
Michael Blaber@sb.fsu.edu

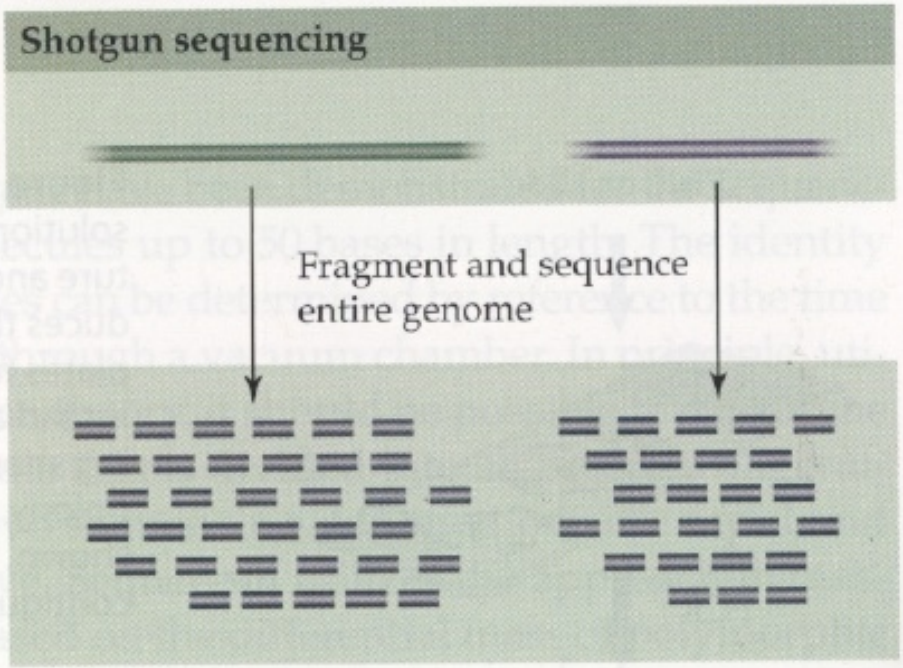
M13 Life Cycle

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



Hierarchical Sequencing Vs. Whole Genome Shotgun Sequencing





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

Any two humans differ at 0.1 % base positions in the genome

- There are 3 billion base pairs in the human genome
- Any two individuals differ at 3 million positions. One change every 1,000 bps.
- Most mutations are in non-essential regions
- Some cause different phenotypic traits (cultural and ethnic differences)
- Some are pathogenic (OMIM morbid mutations)
- Some are lethal (about 300 per person)

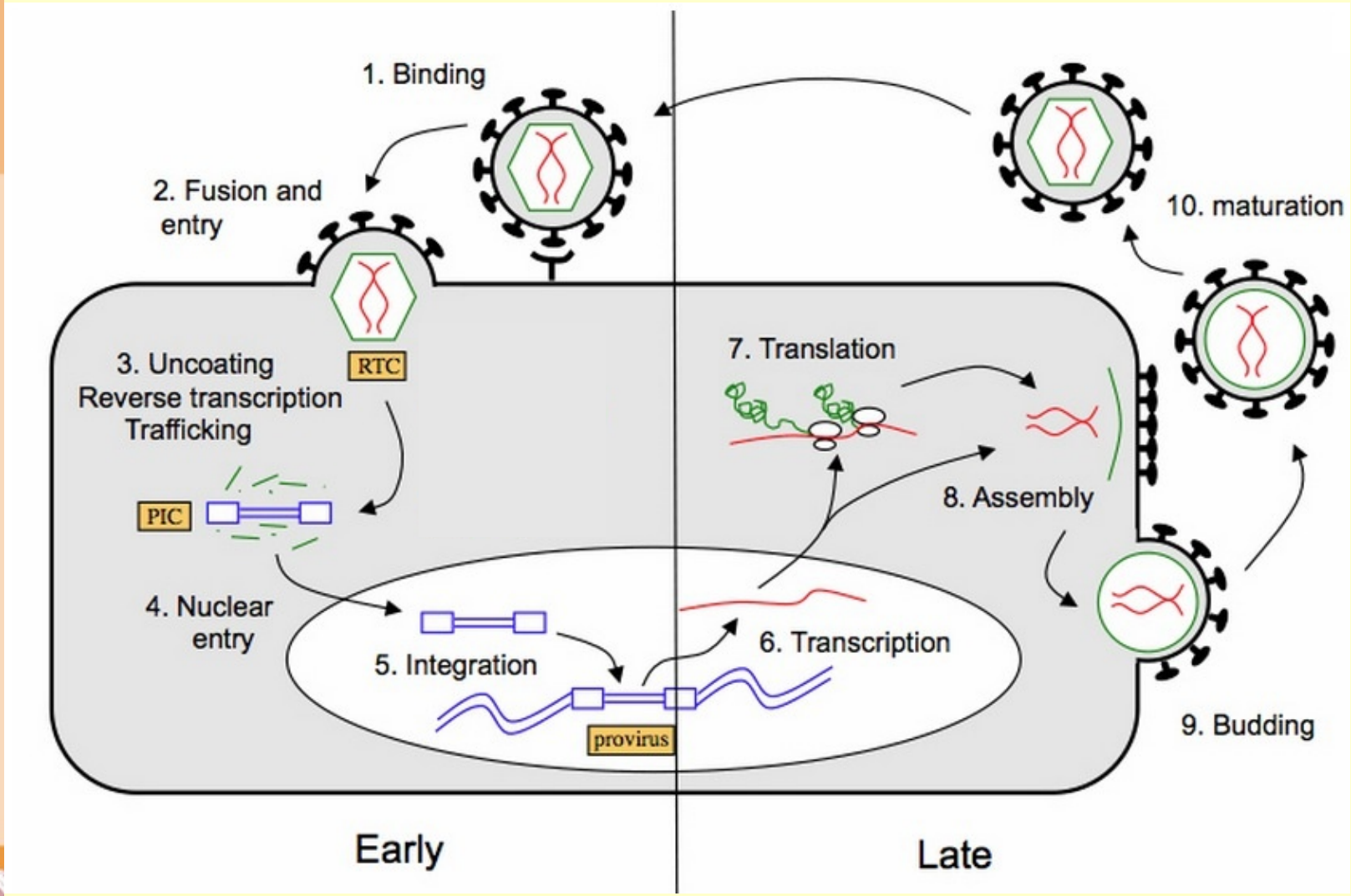
Repeated Elements in the Human Genome ERVs, LINES, SINES and ALUs

- ERVs-Endogenous Retroviruses
  - 10,000 base long RNA genome
  - Converted to DNA and integrate into genome with help of RNA reverse transcriptase and integrase enzymes and long tandem repeats (LTRs)
  - Transcribed into RNA and produce virus (example HIV)

Weiner Curr Opin Cell Biol, 2002 14 (3) 343-50

MRC National Institute for Medical Research

Retroviral Life Cycle



http://www.nimr.mrc.ac.uk/research/kate-bishop/

Repeated Elements in the Human Genome ERVs, LINES, SINES and ALUs

- ERVs-Endogenous Retroviruses
  - 10,000 base long RNA genome
  - Converted to DNA and integrate into genome with help of RNA reverse transcriptase and integrase enzymes and long tandem repeats (LTRs)
  - Transcribed into RNA and produce virus (HIV)
- LINES-Long Interspersed Nuclear Elements
  - About 868,000 in human genome
  - 6,500 base pairs long including LTRs
  - Encode reverse transcriptase and integrase
  - Copy-paste mechanism to insert elsewhere
- SINES-Short Interspersed Nuclear Elements
  - Millions in human genome
  - 100-400 bases long
  - Often contain RNA polymerase III promoters but no genes
- ALUs- The most common SINE
  - 1,500,000 copies = 11% of human genome
  - 350 base pairs in length
  - Contain an RNA Polymerase III promoter, Alu site
  - Appear to evolve from 7S RNA signal recognition particle

Weiner Curr Opin Cell Biol, 2002 14 (3) 343-50

Whole Genome Shotgun versus Bacterial Artificial Chromosome Sequencing

1997

Gene Myers

Thanks to Seraf ın Batzoglou

Whole Genome Shotgun versus Bacterial Artificial Chromosome Sequencing

1997

Let's sequence the human genome with the shotgun strategy

That is impossible, and a bad idea anyway

Gene Myers

Phil Green

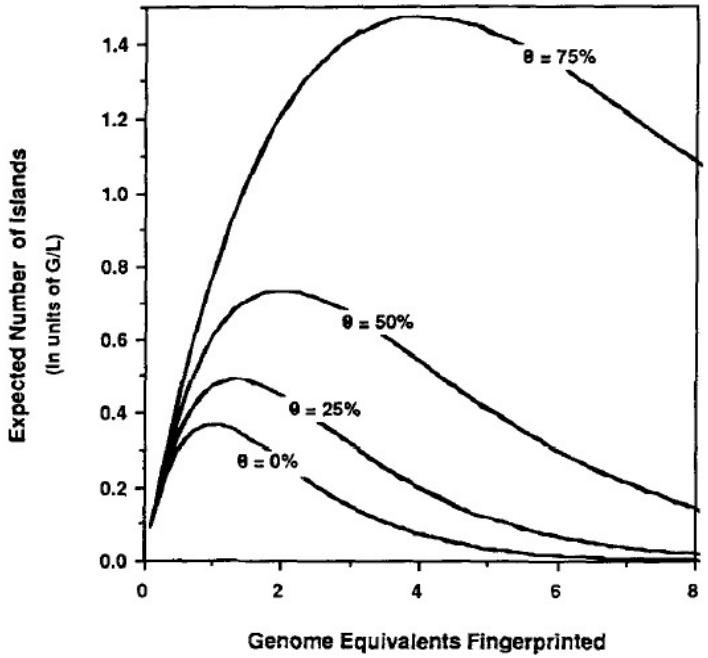
Thanks to Seraf in Batzoglou

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



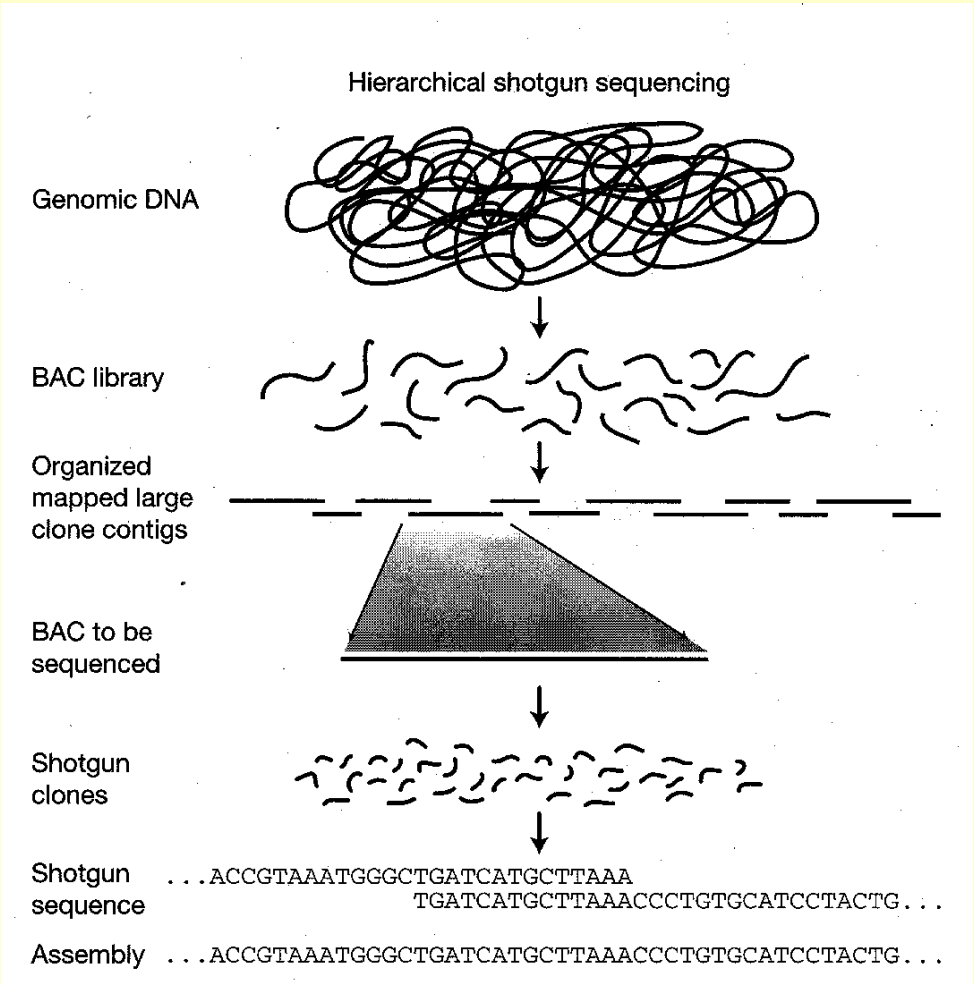
Approximate value of G/L

	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

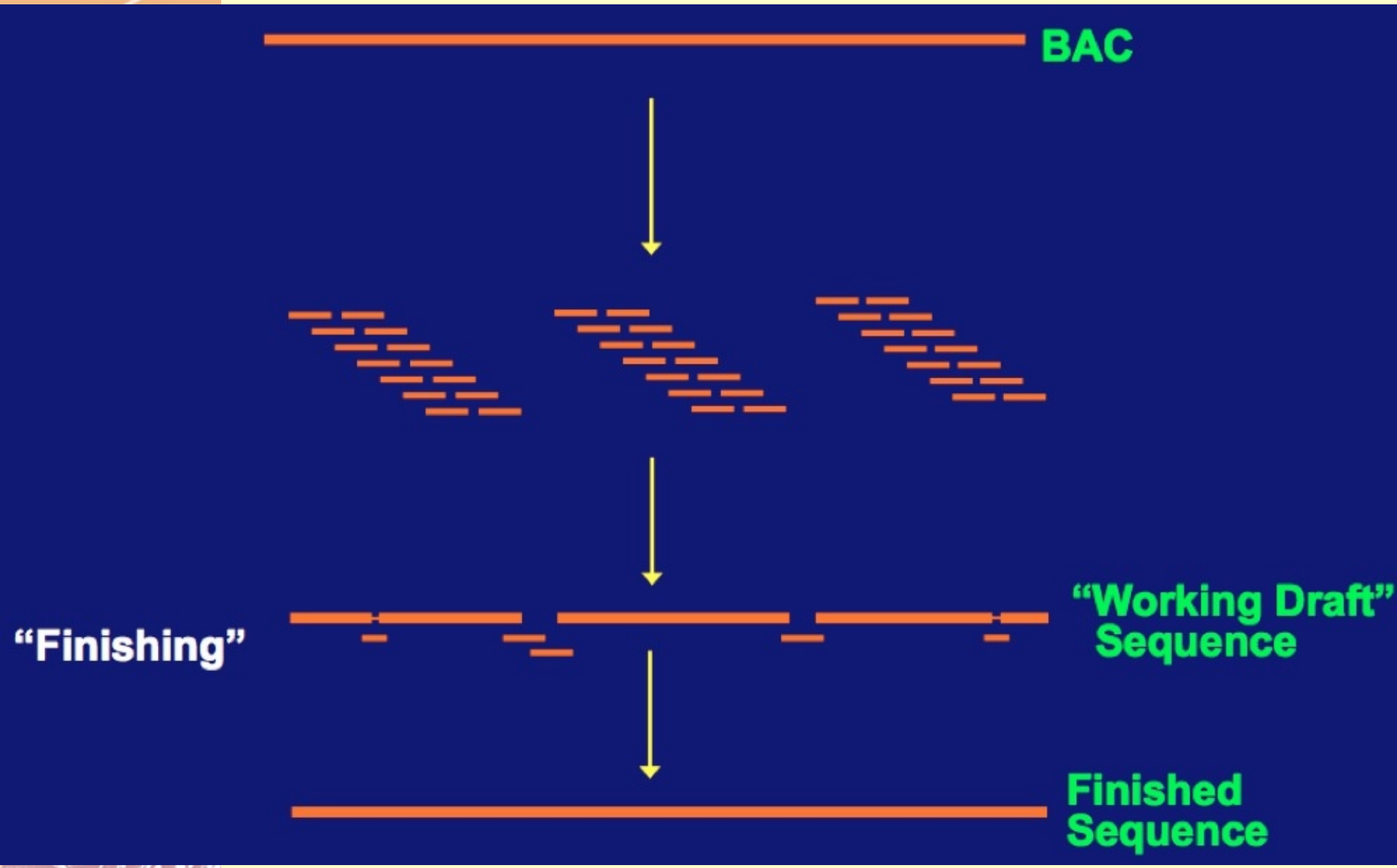
G = haploid genome length in bp;
L = length of clone insert in bp;
N = number of clones fingerprinted;
$\alpha = N/G$ = probability per base of starting a new clone;
T = amount of overlap in base pairs needed to detect overlap;
$\theta = T/L$;
c = redundancy of coverage $= LN/G$.

Public Human Genome Project Strategy

http://www.nhgri.nih.gov/



BAC Shotgun Sequencing Strategy

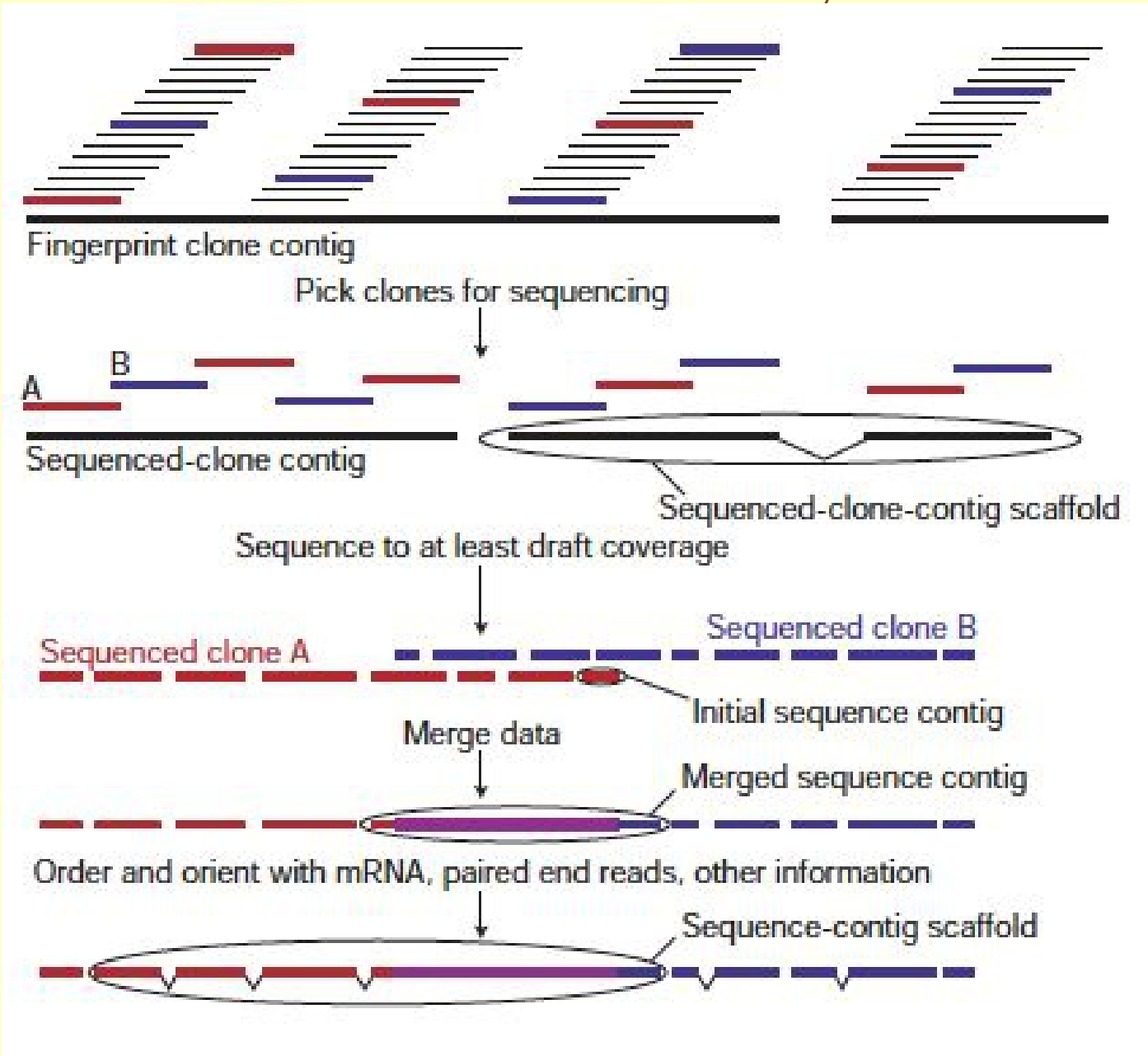


BAC and PAC Libraries in Public Human Genome Project

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

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.

Comparing Chromosome 2 Sequence Ve

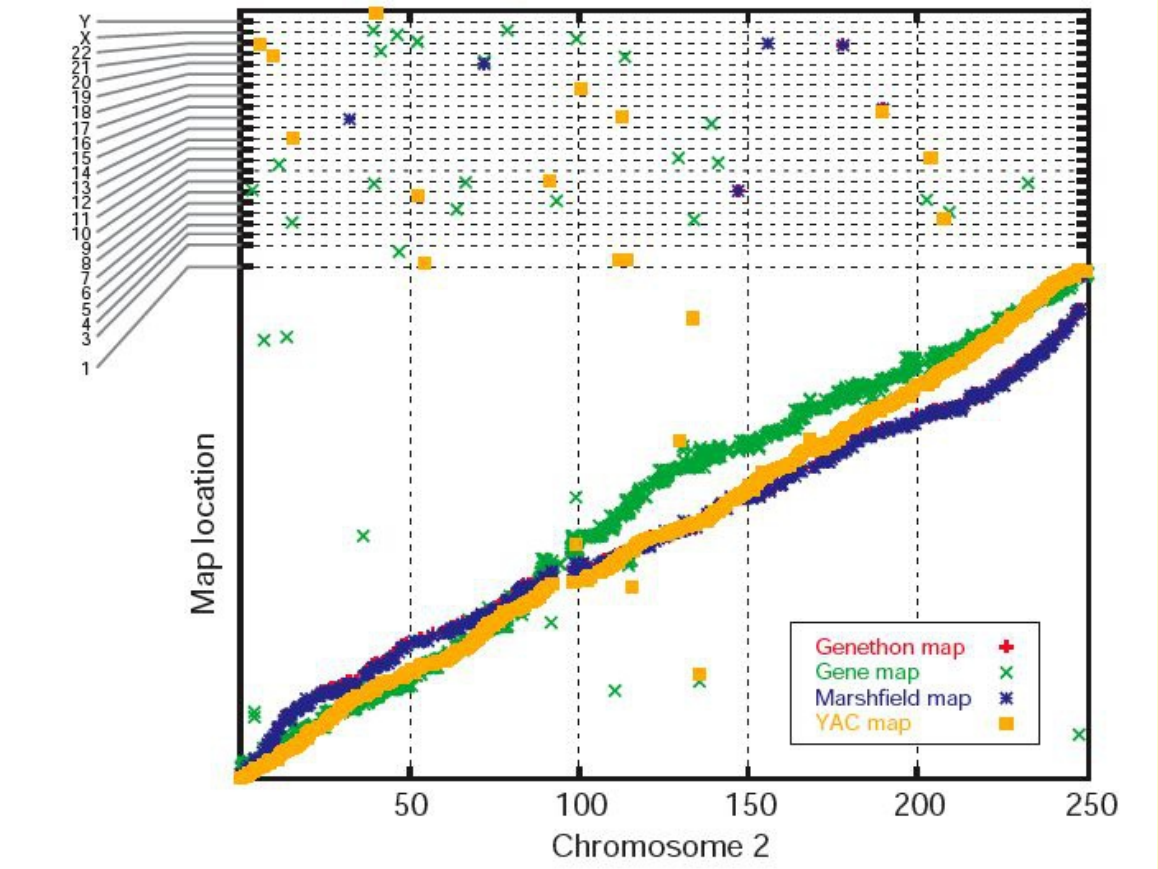


Figure 5 Positions of markers on previous maps of the genome (the Genethon[101] genetic map and Marshfield genetic map (http://research.marshfieldclinic.org/genetics/genotyping_service/mgsver2.htm), the GeneMap99 radiation hybrid map[100], and the Whitehead YAC and radiation hybrid map[29]) plotted against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

Synteny Between Human and Mouse

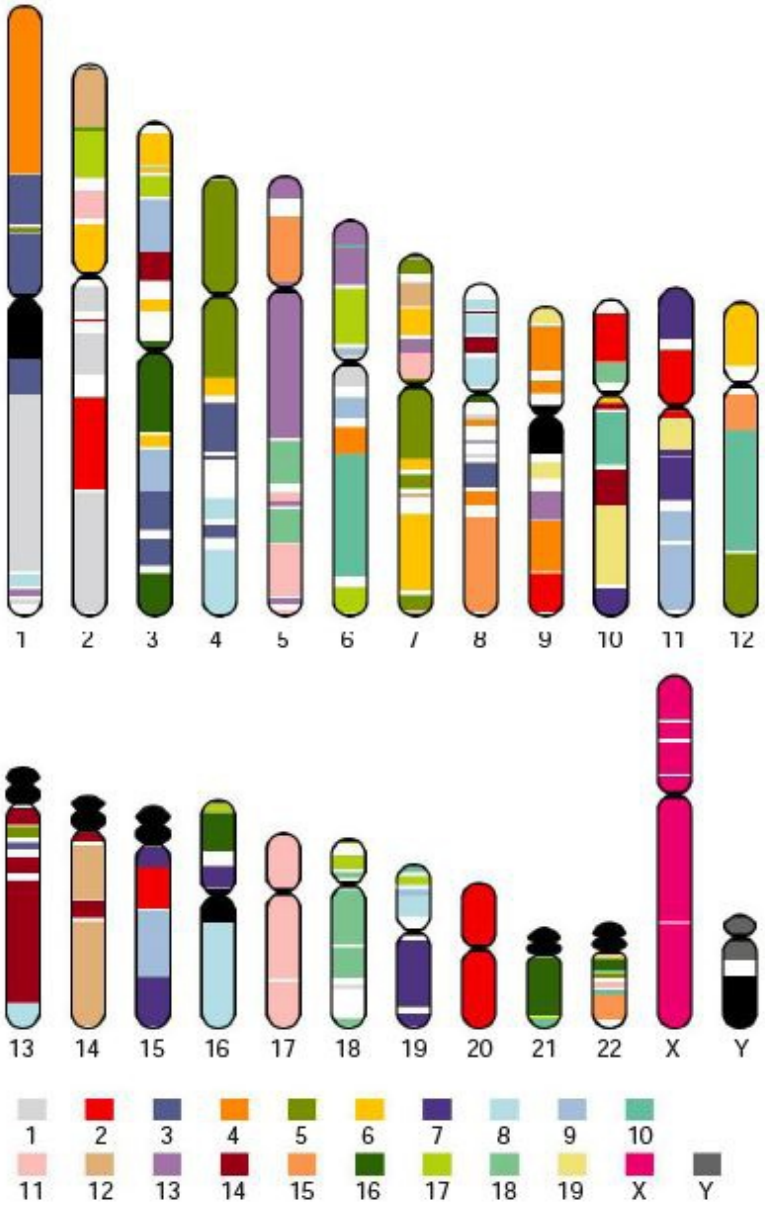


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes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

Celera Sequencing
http://www.celera.com/

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.

Celera Scaffolds

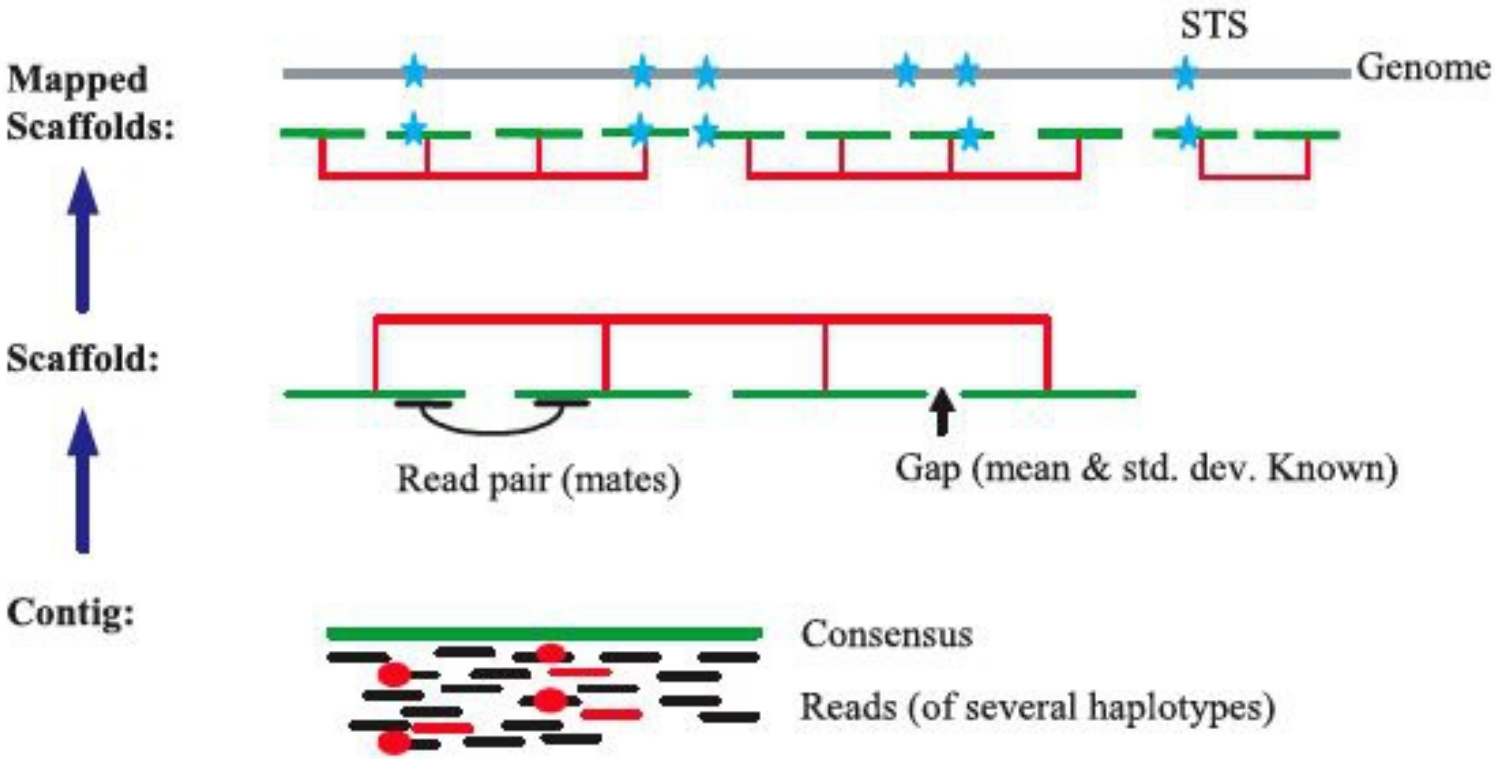
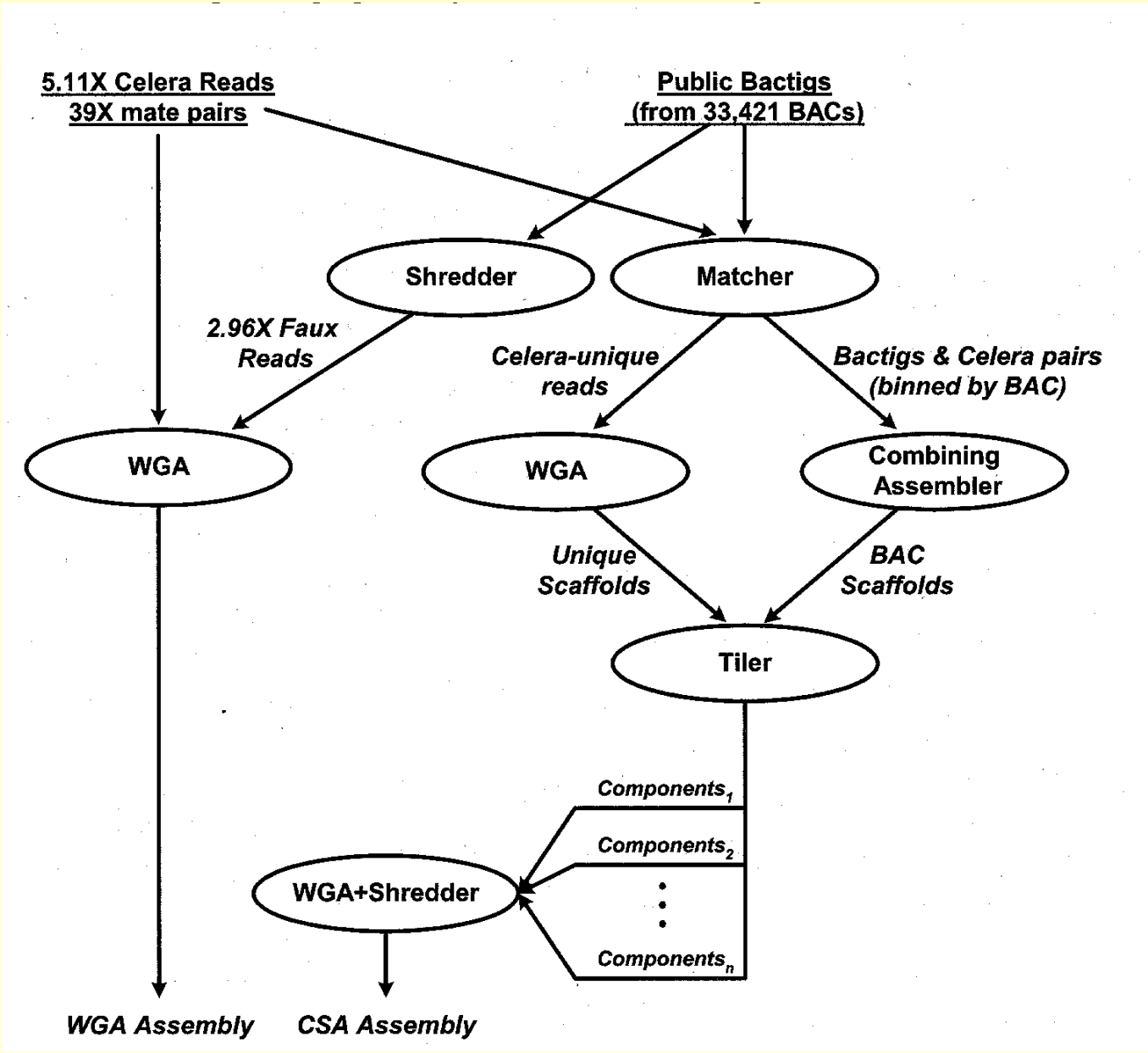


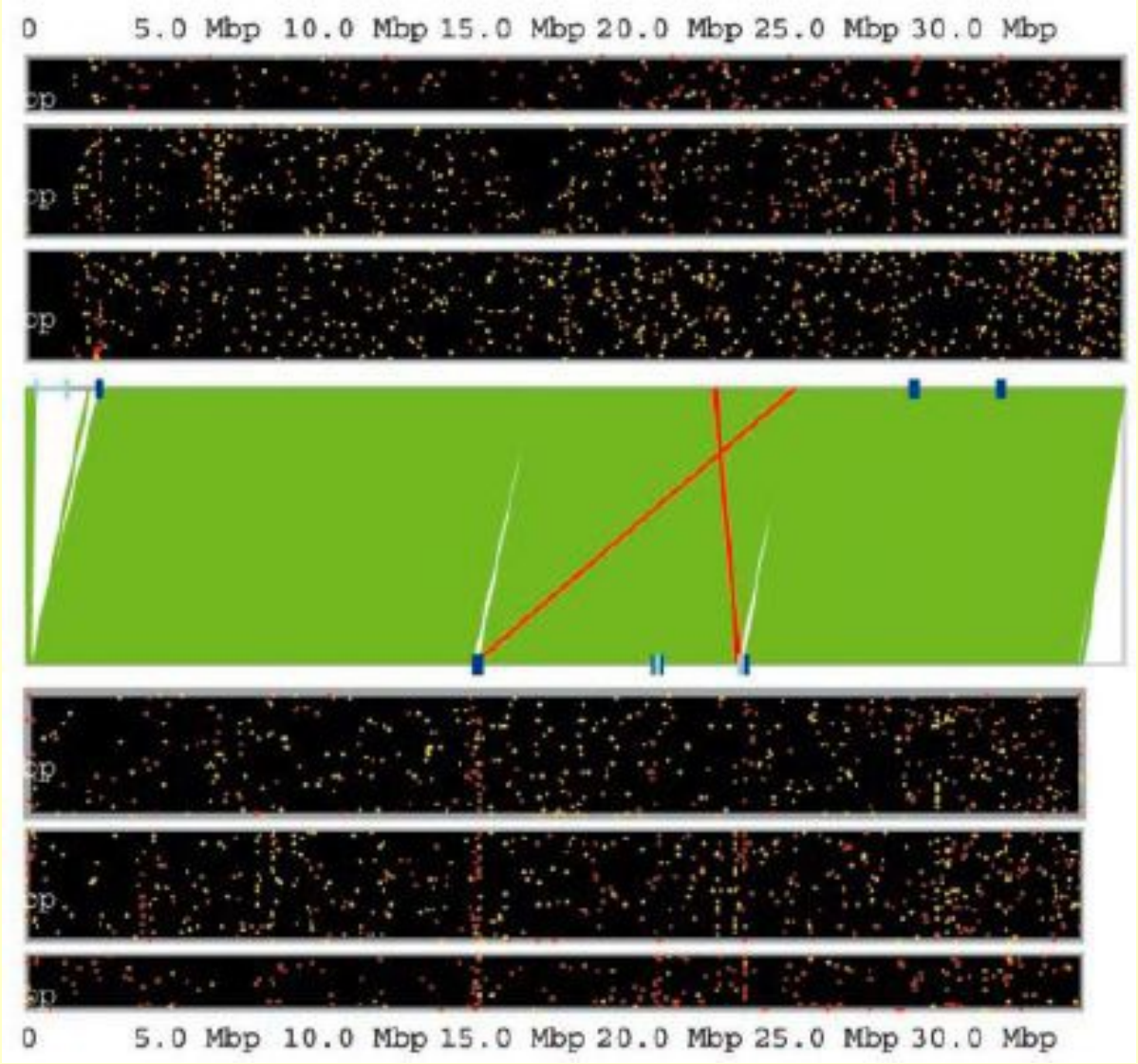
Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair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

Celera Assembler



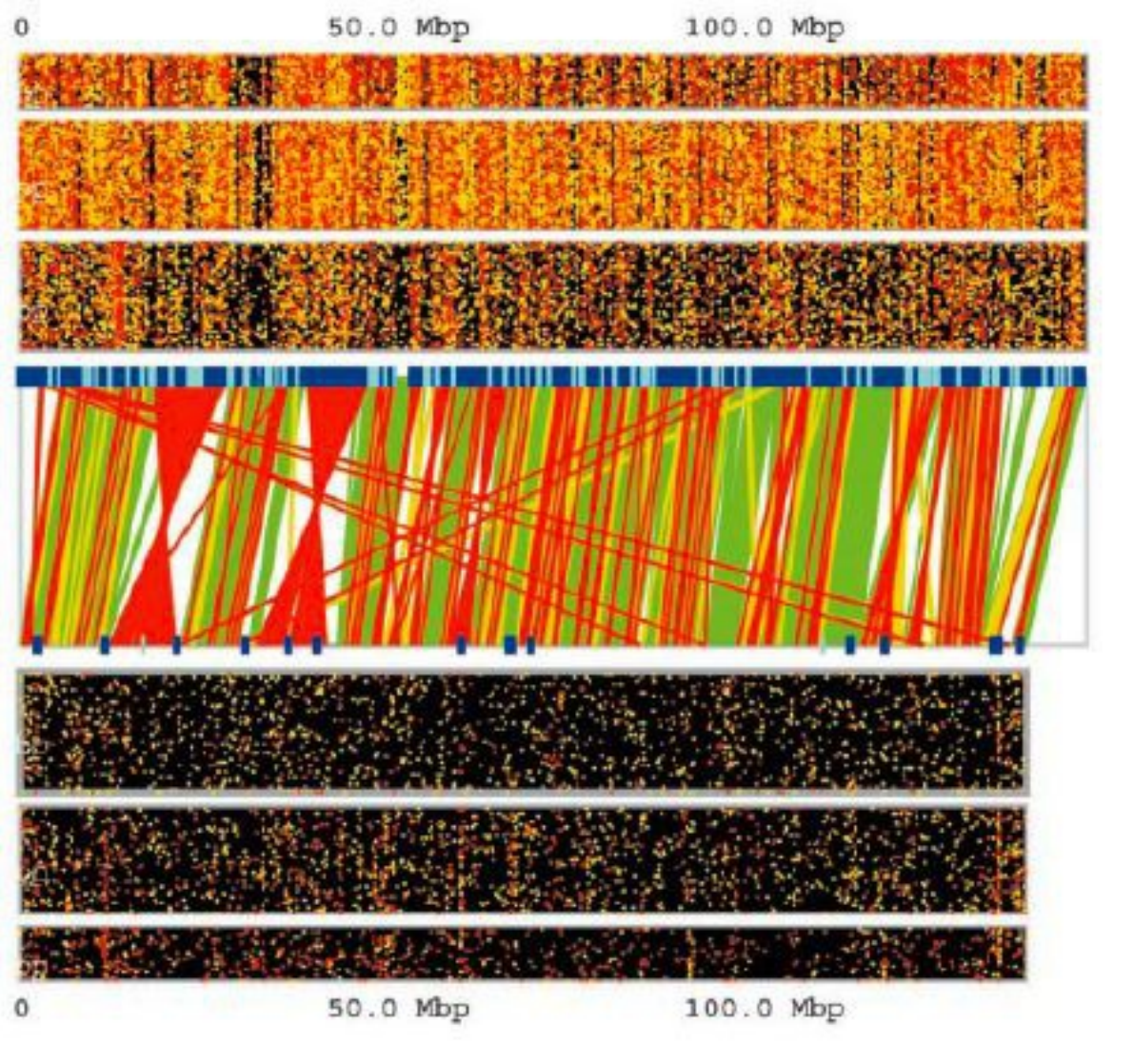
CELERA
an Applera Corporation Business

Chromosome 21: Public vs Celera Assemblies

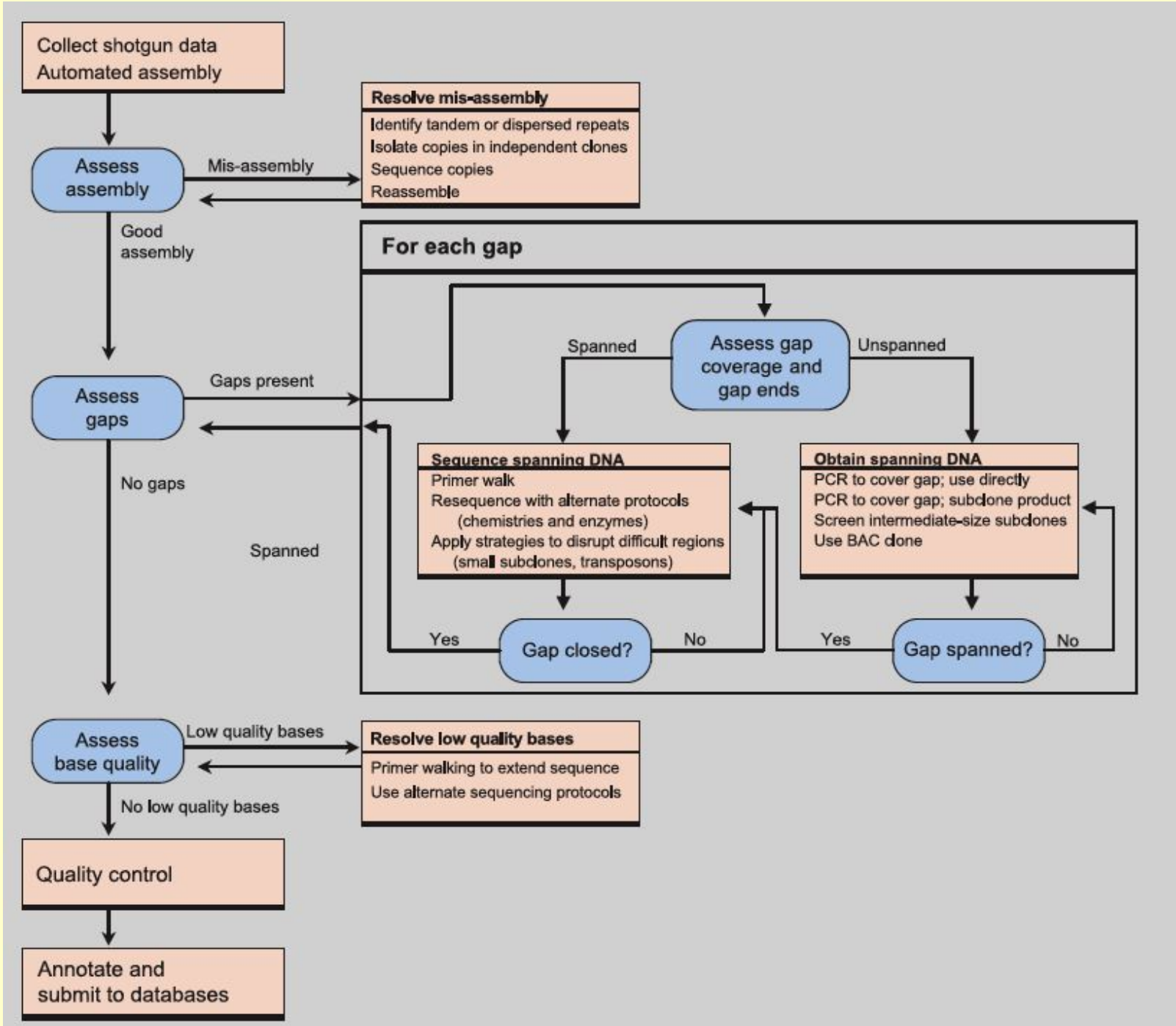


CELERA
an Applera Corporation Business

Chromosome 8: Public vs. Celera

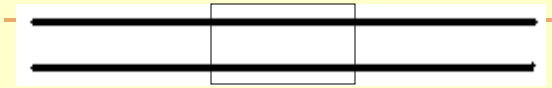


Finishing Strategy for the Public Genome Project

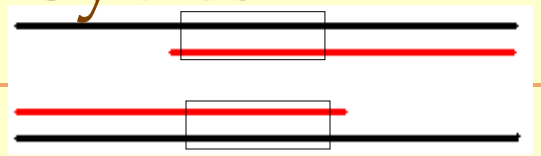


Polymerase Chain Reaction Overview: Exponential Amplification of DNA

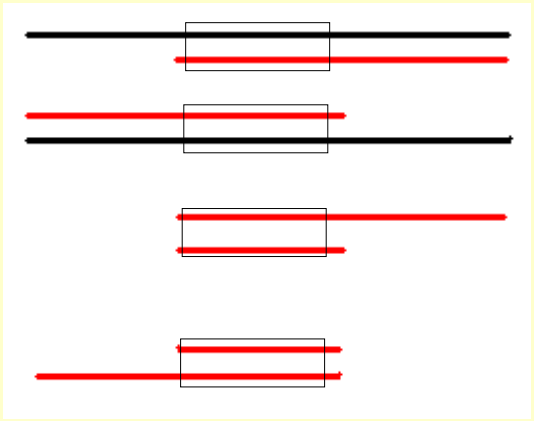
The First Three Cycles



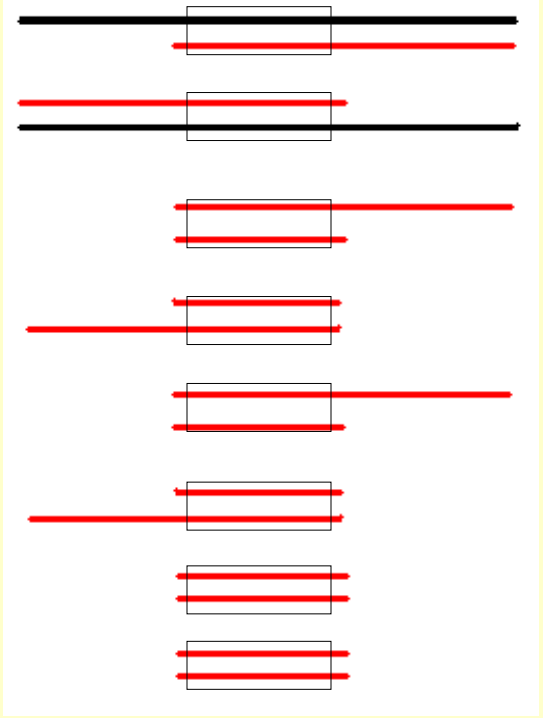
Original DNA



After Cycle 1



After Cycle 2



After Cycle 3

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

PCR Requirements

DNA

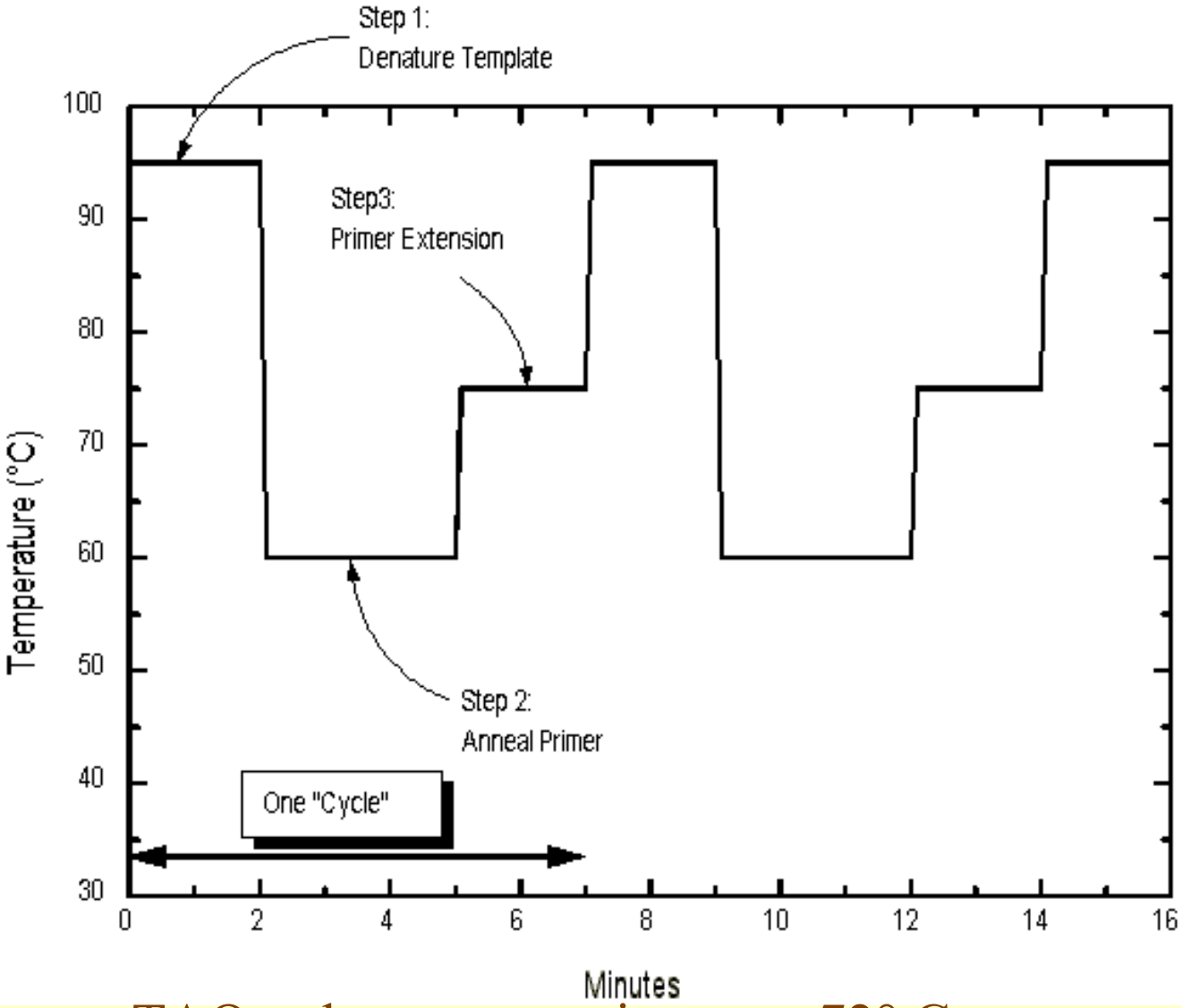
- Need to know at least the beginning and end of DNA sequence
- These flanking regions have to be unique to strand interested in amplifying
- Region of interest can be present in as little as one copy
- *Enough DNA in 0.1 microliter of human saliva to use PCR*

DNA Polymerase Enzyme

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

Thermocycler

Temperature Cycling



TAQ polymerase optimum at 72° C

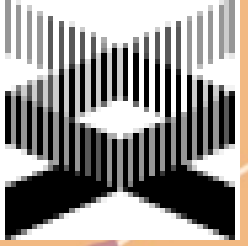
PCR Applications

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



Finished Sequence in 2004 (Build 35)

Table 2 **Finished sequence and gaps, HGSC Build 35**

Chr	Total finished sequence* (kb)	Euchromatic gaps†		Heterochromatic gaps‡		Estimate of total gap size§ (kb)	Unfinished clones‖	
		Number	Est. size (kb)	Number	Est. size (kb)		Number	Est. size (kb)
1	222,828	32	1,605	2	19,510	21,115	17	850
2	237,503	20	2,512	1	2,900	5,412	0	0
3	194,636	5	1,935	1	1,500	3,435	0	0
4	187,161	14	1,250	1	3,000	4,250	0	0
5	177,703	5	92	1	340	432	0	0
6	167,318	10	658	1	2,300	2,958	0	0
7	154,759	11	869	1	4,630	5,499	0	0
8	142,613	9	662	1	2,190	2,852	0	0
9	117,781	40	1,955	2	18,000	19,955	12	600
10	131,614	12	1,020	1	2,515	3,535	8	400
11	131,131	7	322	1	4,760	5,082	0	0
12	130,259	8	795	1	4,300	5,095	0	0
13	95,560	6	715	2	17,200	17,915	0	0
14	88,291	1	8	2	17,220	17,228	0	0
15	81,342	10	737	2	18,260	18,997	0	0
16	78,885	4	143	2	10,000	10,143	0	0
17	77,800	9	875	1	7,500	8,375	0	0
18	74,656	3	97	1	1,368	1,465	0	0
19	55,786	5	5,015	1	340	5,355	0	0
20	59,505	4	1,157	1	1,766	2,923	0	0
21	34,170	3	53	2	11,620	11,673	0	0
22	34,765	11	460	2	14,330	14,790	0	0
X	150,394	12	750	1	3,000	3,750	14	700
Y	24,872	9	1,480	2	31,618	33,098	7	350
Total	2,851,331	250	25,165	33	200,167	225,332	58	2,900

*The total length of tiling paths including only finished bases of clones in Build 35. Roughly 2.19 Mb of sequence on chromosome Y was derived directly from the equivalent pseudoautosomal region on chromosome X.

†Defined as gaps in euchromatic regions, including junctions with heterochromatic/centromeric sequences, for which no clone was available (see text).

‡Defined here as gaps in heterochromatic regions (see text and Supplementary Note 2 on heterochromatic sequence). Separate gaps were counted for centromeres and pericentric heterochromatin, even when the two were contiguous. Centromere sizes were taken from ref. 62 or in some cases provided directly by the sequencing centres (see Supplementary Note 2). Acrocentric sizes are based on centromere ratios from ref. 63. The sizes of large heterochromatic gaps are typically difficult to estimate accurately owing to their repeat structure and polymorphic nature[62,64]. Other regions might arguably be called heterochromatin (for example, the pericentric regions of chromosomes 19 and 3 and a ~400-kb gap on the Y chromosome[23]), but are classified as euchromatin here.

§The sum of lengths for finished sequence, estimated heterochromatic gaps, euchromatic gaps and unfinished clone gaps. The total length is only approximate because of uncertainty in gap sizes, particularly for heterochromatic gaps and centromeres.

‖Those in the tiling path but for which it has not been possible to obtain finished sequence. Unfinished sequence from these clones is deposited in public databases. These gaps are all listed at 50 kb, reflecting the approximate average size of the gap.

Comparison of Chromosome 7 Draft ver

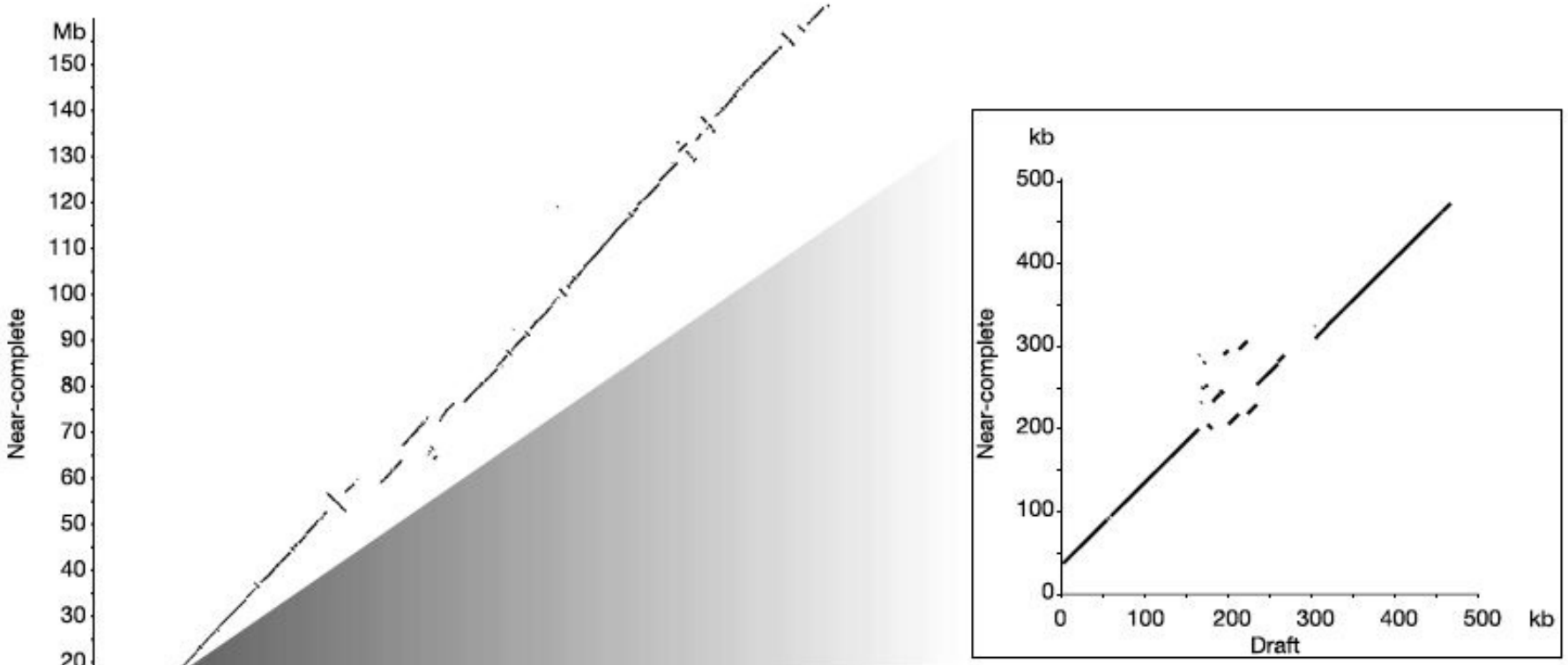
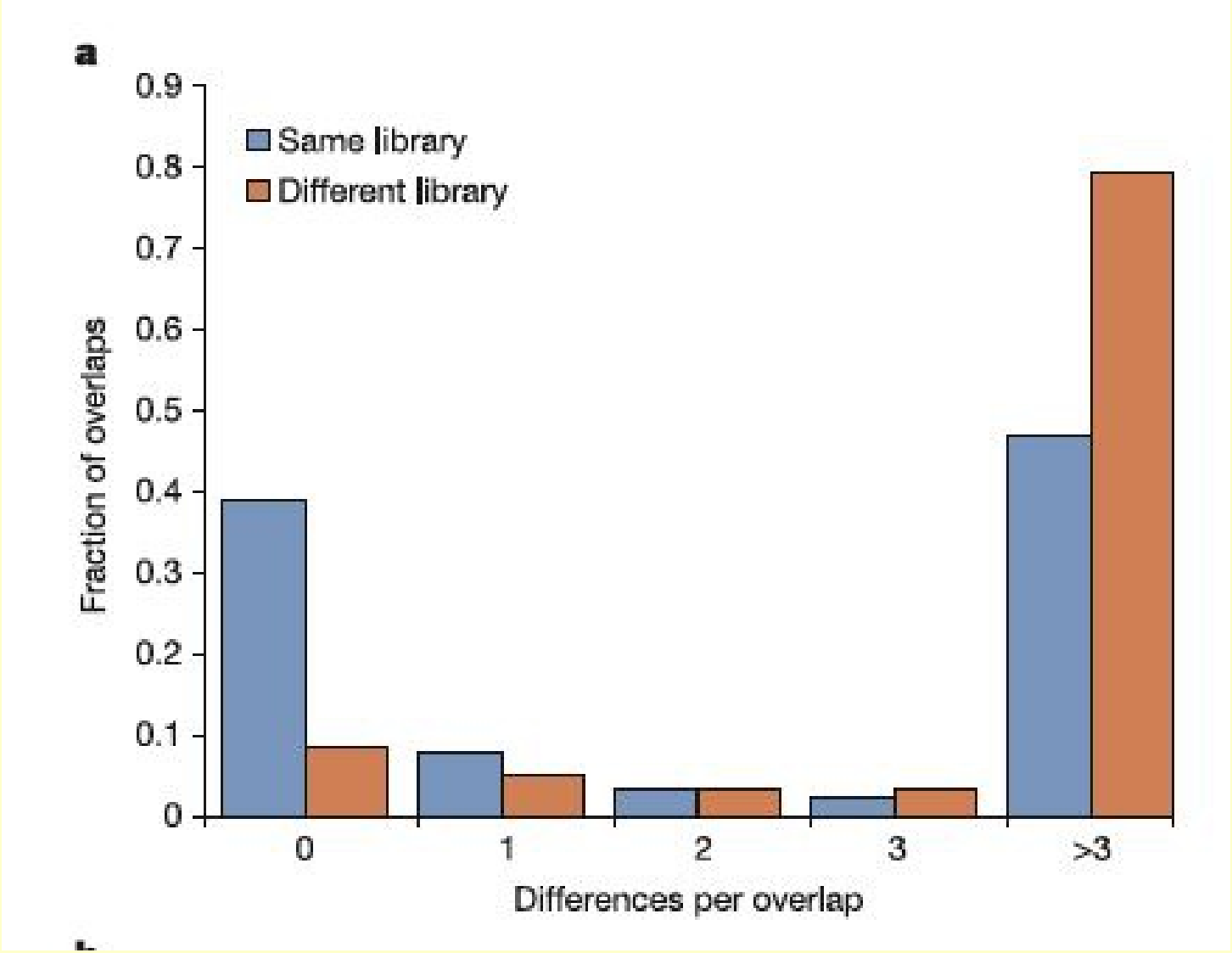
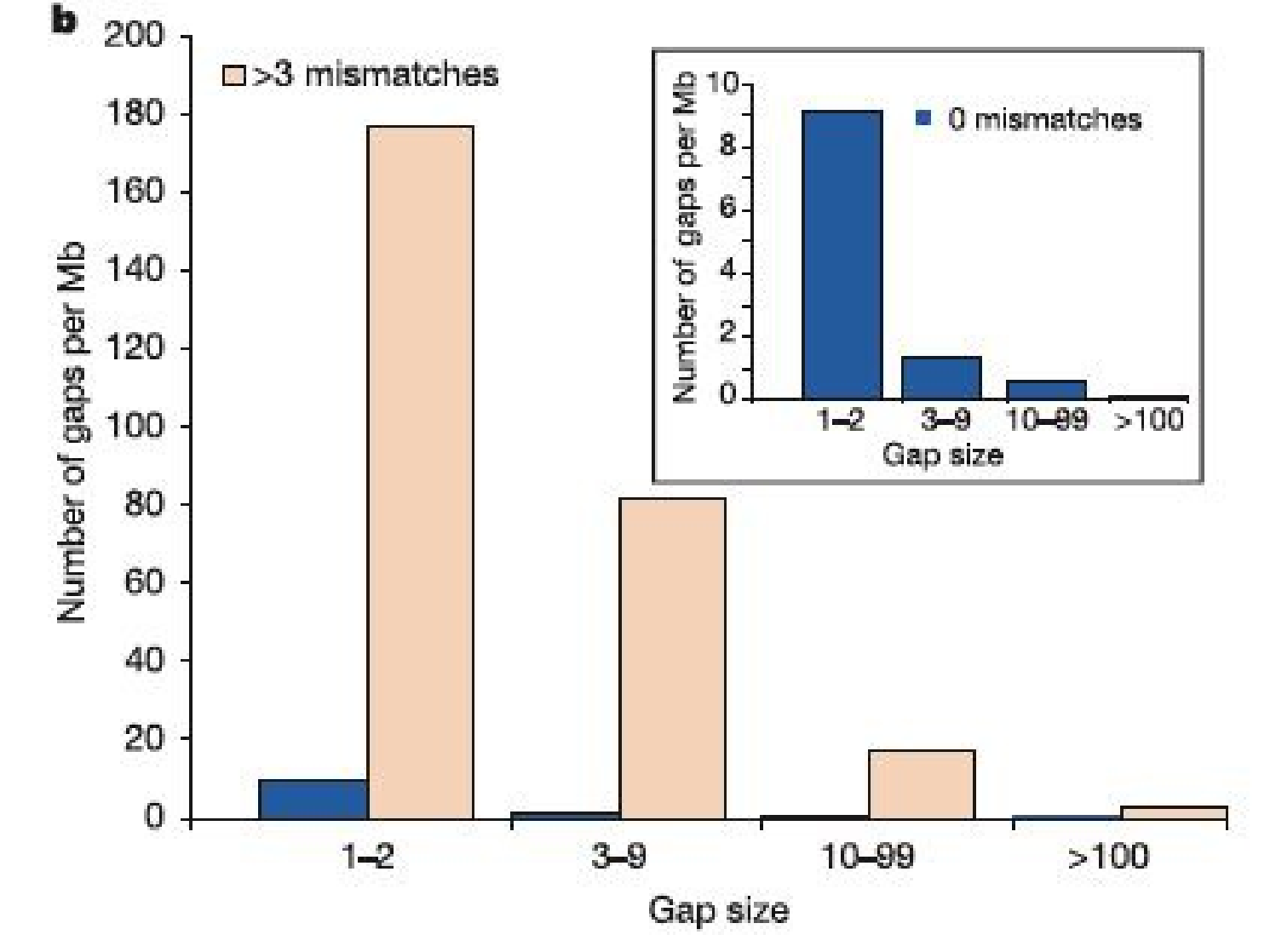


Figure 1 Comparison of previous draft sequence with current near-complete sequence of chromosome 7 (ref. 24). At large scale, there was good collinearity between draft and near-complete sequence, although some inversions were present in the draft due to lack of sufficient anchors in some regions. At finer scale, the draft sequence contained some sequence contigs for which order and orientation were not known. The inset shows a region of 500 kb with sequence derived from three overlapping BACs. BACs at each end were finished at the time of draft assembly, whereas the middle BAC was at an early stage of shotgun coverage in which contigs were not yet ordered and oriented.

Base Changes in BAC Overlaps with BACs from Same or Different Libraries



Gaps in BAC Overlaps with BACs from S



Duplications and Deletions in the Human Genome

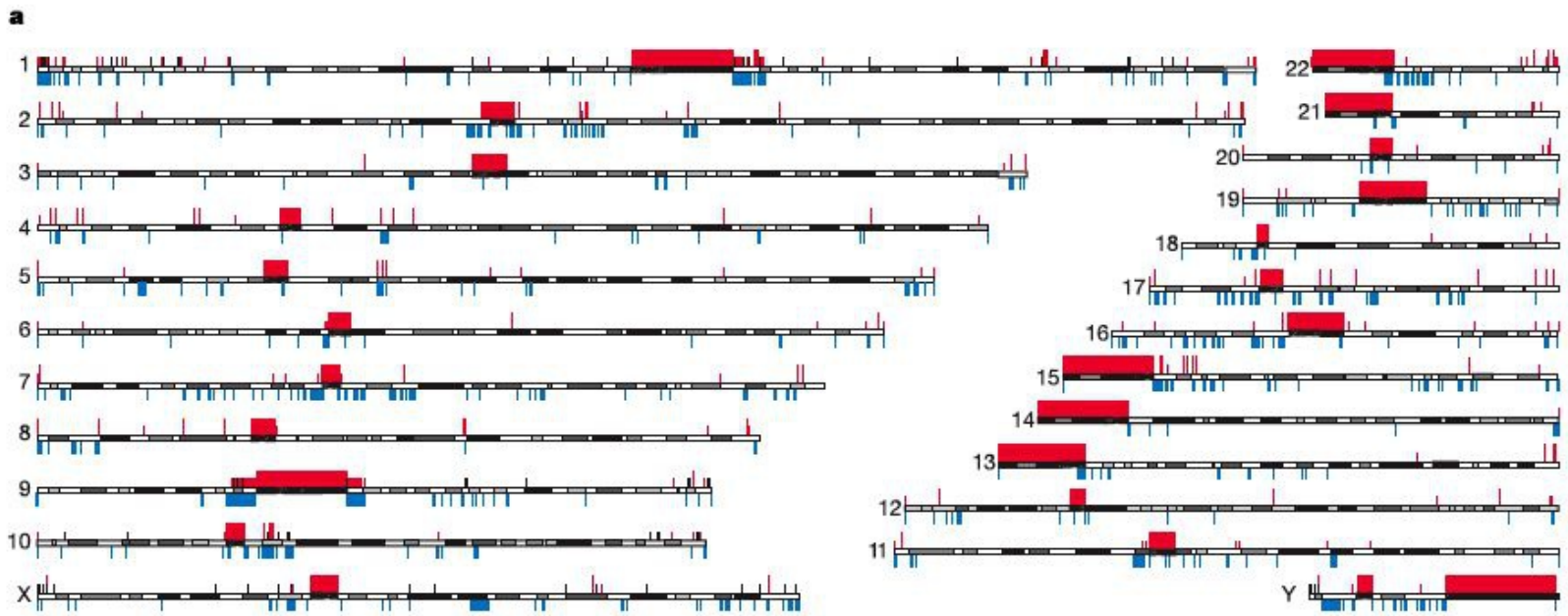
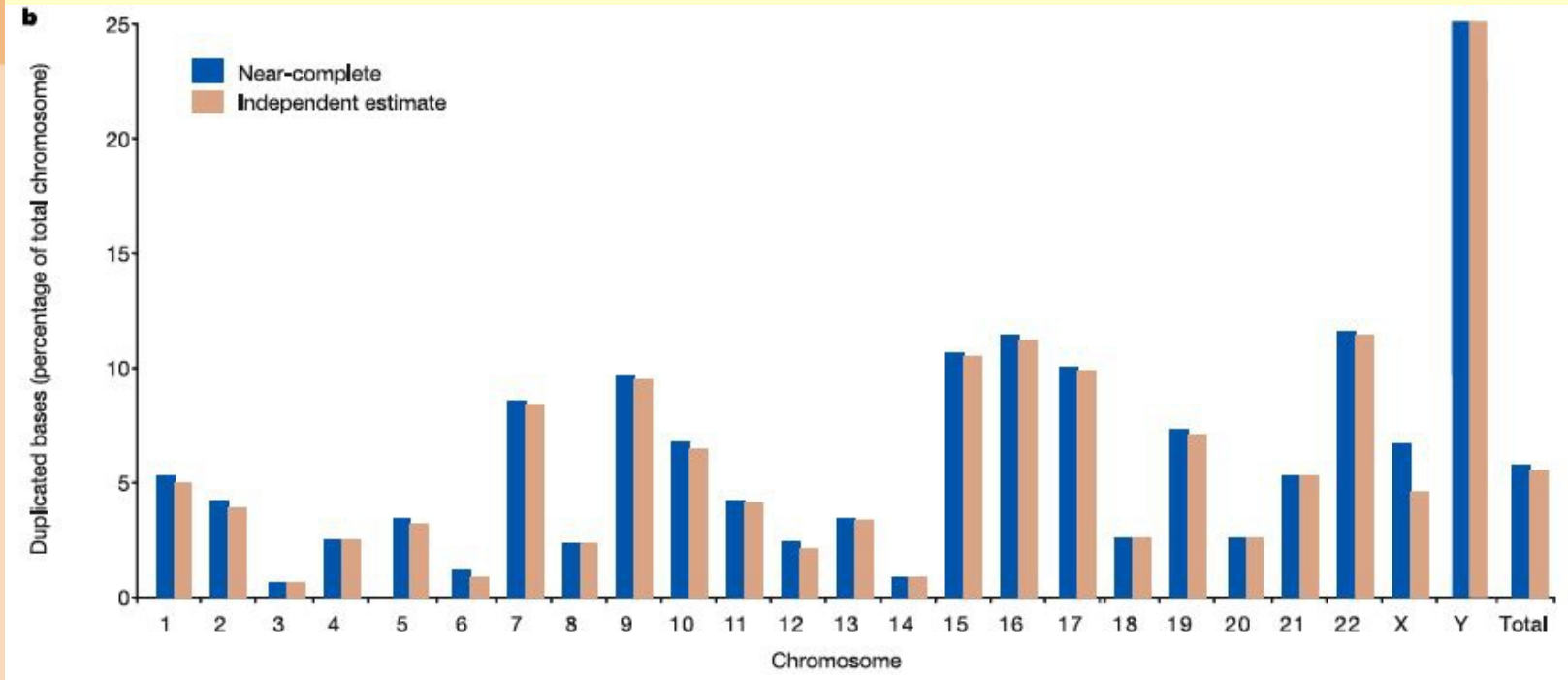
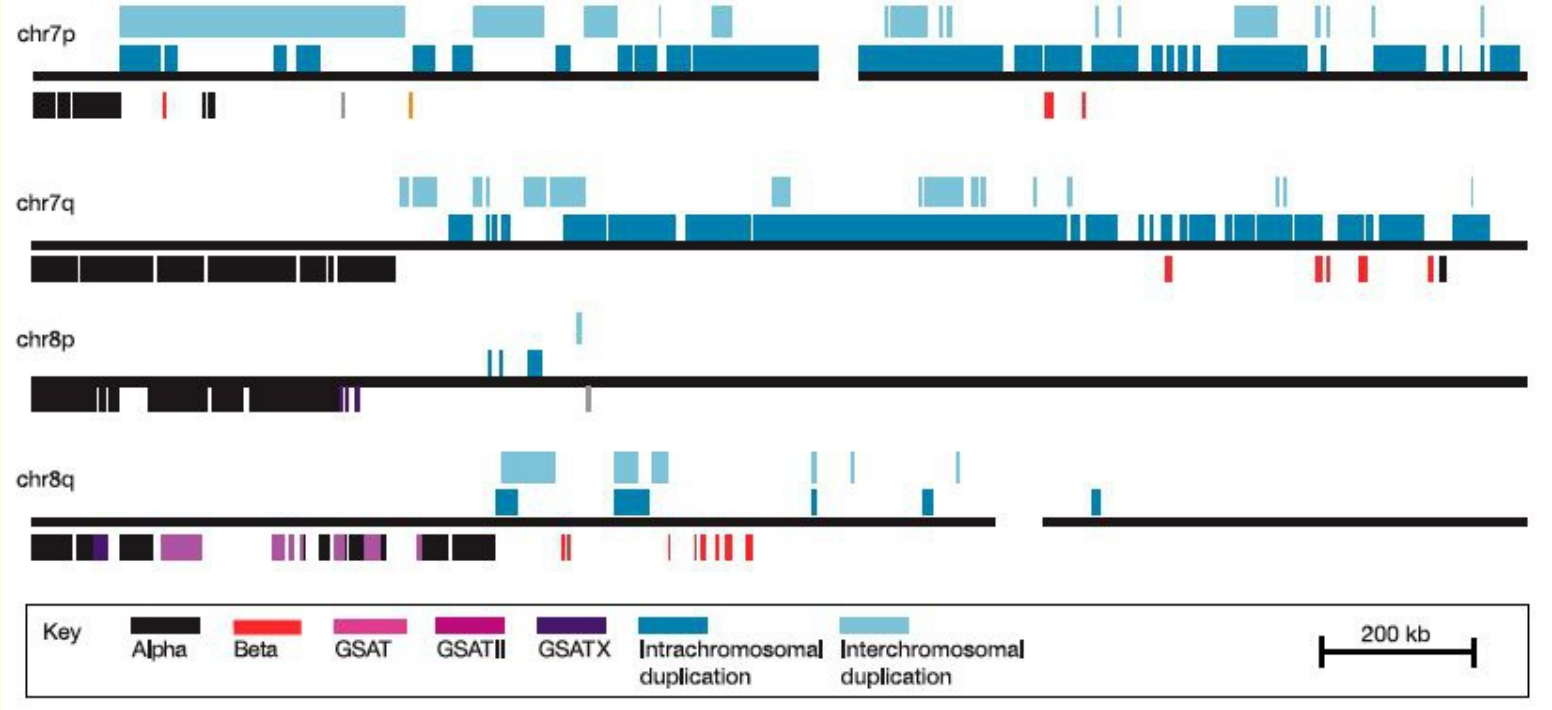


Figure 4 Segmental duplications across the genome. **a**, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length ≥10 kb and sequence identity ≥95%). Large duplications are shown to approximate scale; smaller ones are indicated as ticks. Sequence gaps are indicated above the chromosomes in red. Large gaps (>300 kb) are shown to approximate scale; smaller gaps are indicated as ticks with those that are 50 kb or smaller shown as shorter ticks. Unfinished clones are indicated as black ticks. **b**, Percentage of

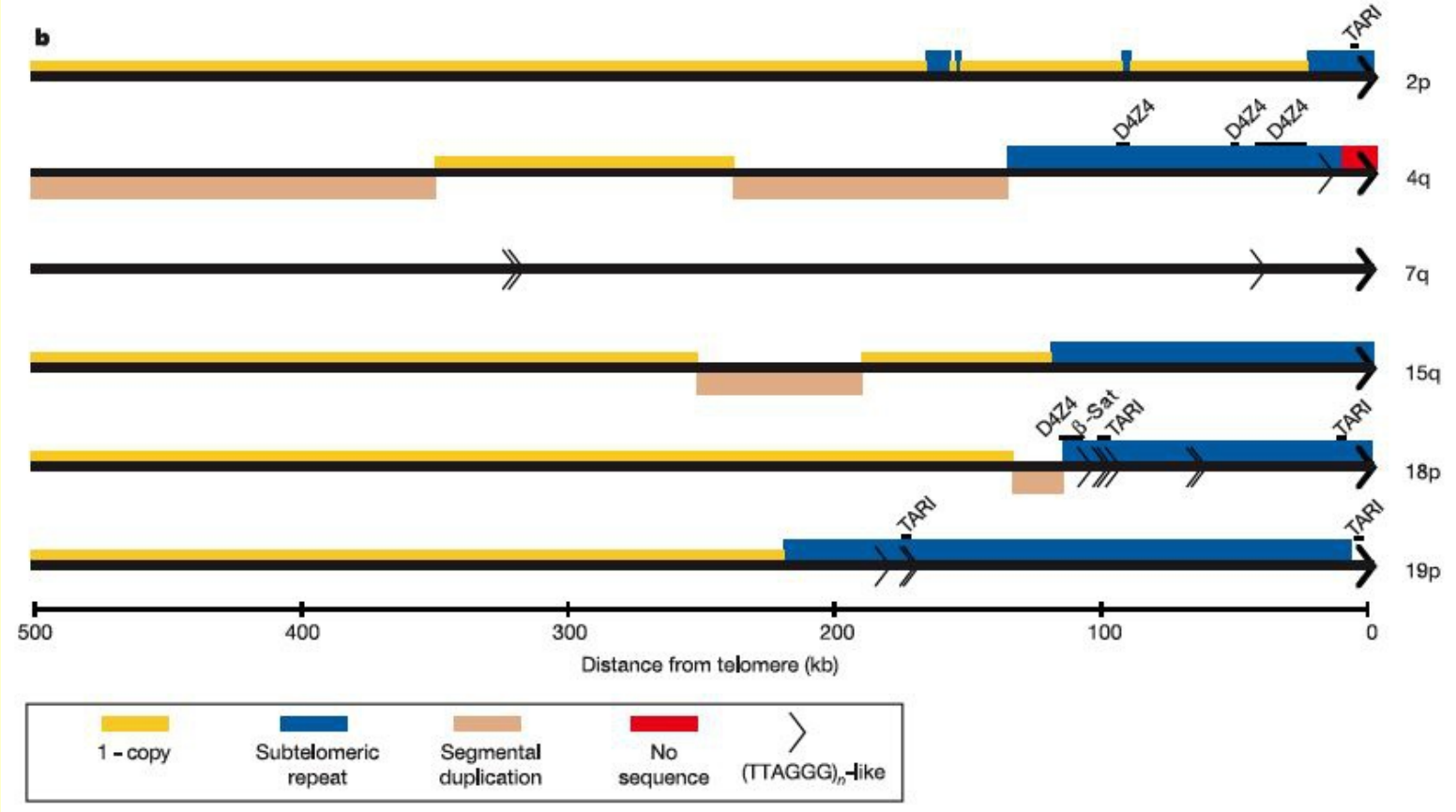
Percentage of Chromosomes Duplicated



Duplications near Centromeres

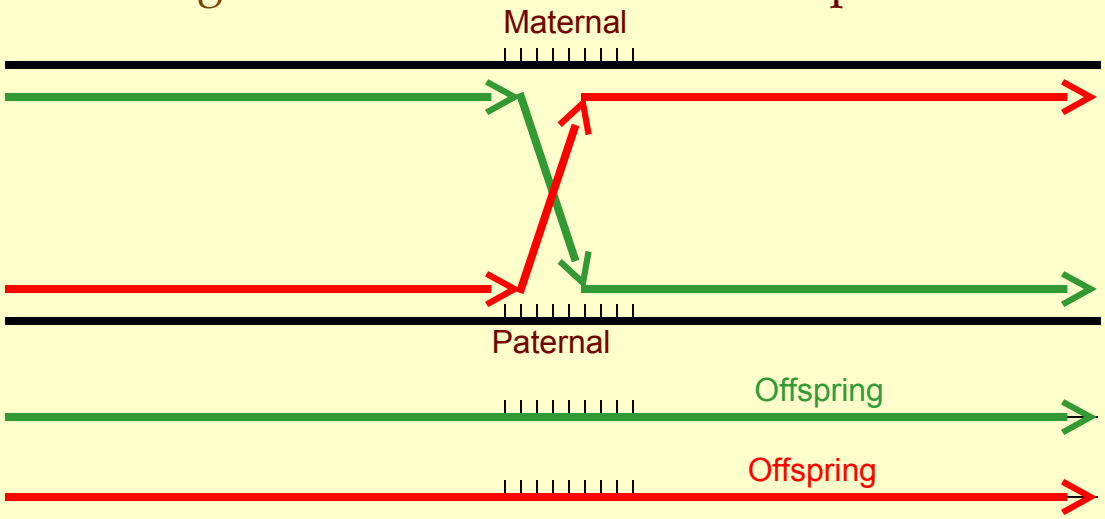


Duplications near Telomeres

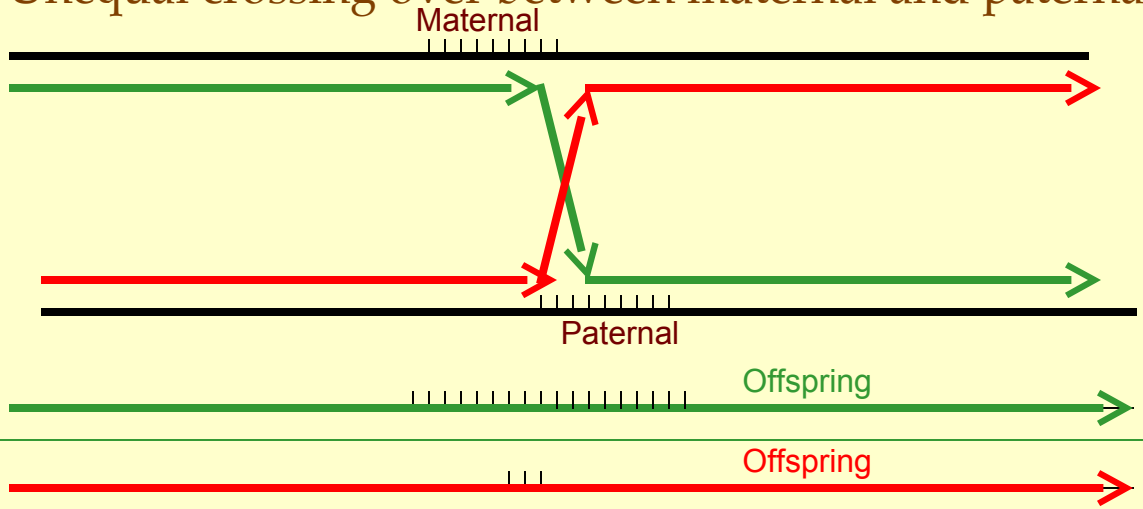


Deletions and Duplications can Arise from Unequal Crossing Over in Repeated Regions

- Crossing over between maternal and paternal chromosomes



- Unequal crossing over between maternal and paternal chromosomes



J. Craig Venter INSTITUTE

The Diploid Sequence of an Individual Human

The Diploid Genome Sequence of an Individual Human

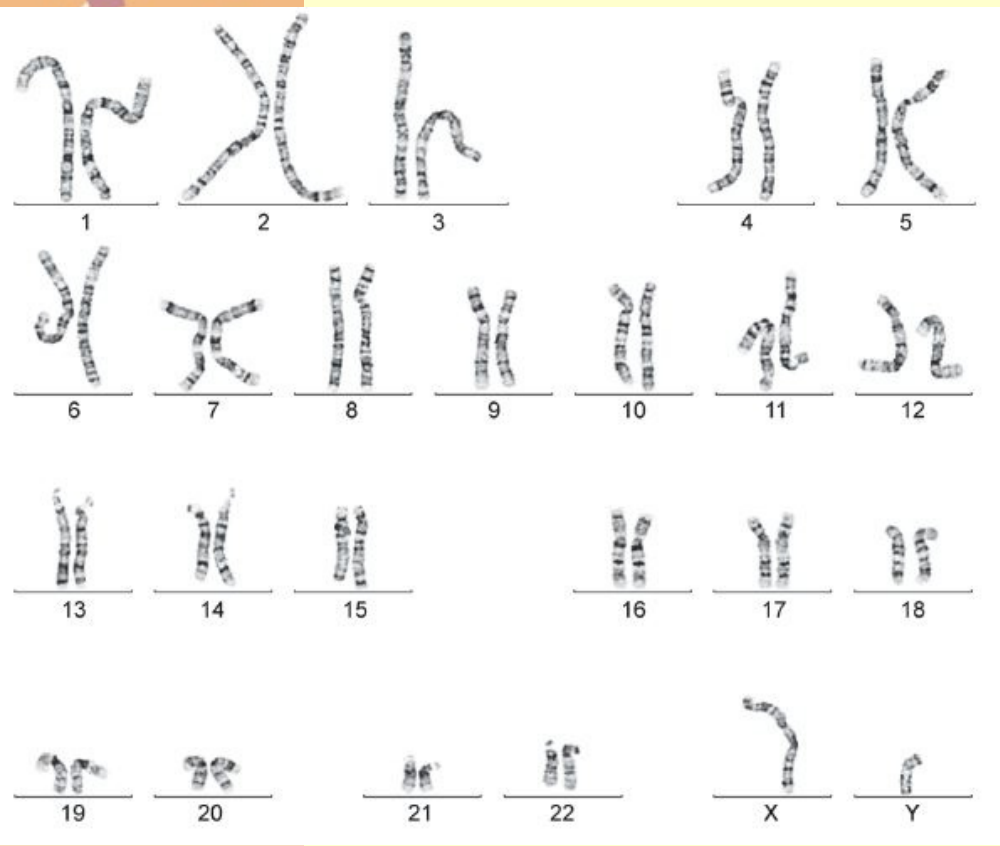
Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, 4 Genetics Department, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

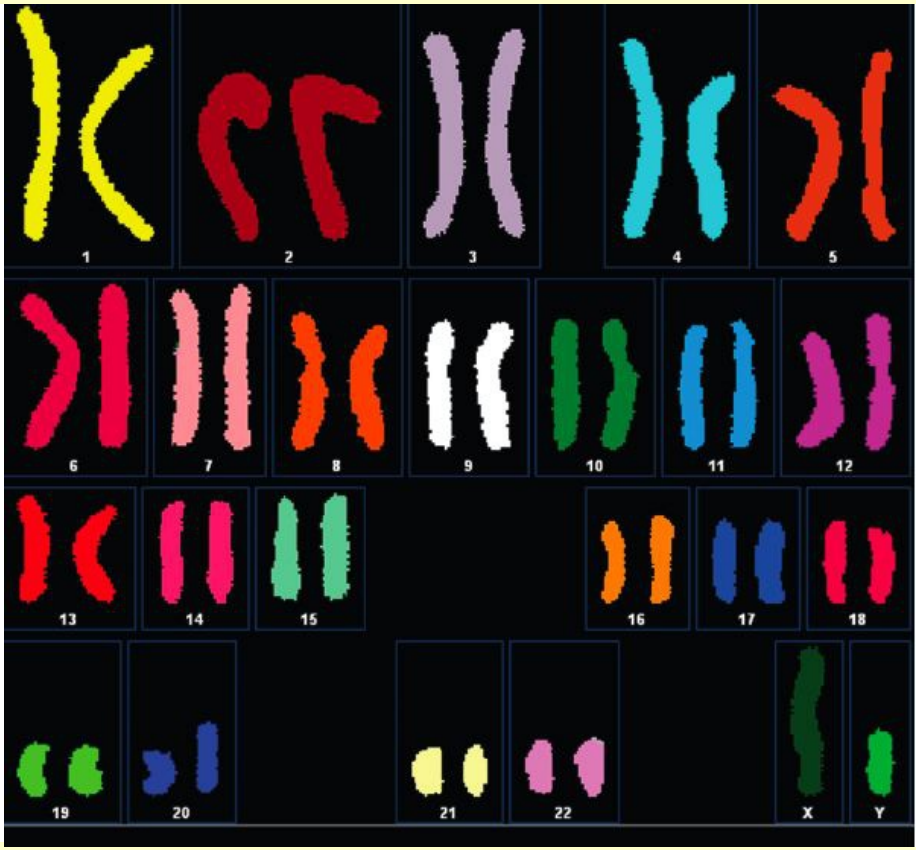
Presented here is a genome sequence of an individual human. It was produced from ~32 million random DNA fragments, sequenced by Sanger dideoxy technology and assembled into 4,528 scaffolds, comprising 2,810 million bases (Mb) of contiguous sequence with approximately 7.5-fold coverage for any given region. We developed a modified version of the Celera assembler to facilitate the identification and comparison of alternate alleles within this individual diploid genome. Comparison of this genome and the National Center for Biotechnology Information human reference assembly revealed more than 4.1 million DNA variants, encompassing 12.3 Mb. These variants (of which 1,288,319 were novel) included 3,213,401 single nucleotide polymorphisms (SNPs), 53,823 block substitutions (2–206 bp), 292,102 heterozygous insertion/deletion events (indels)(1–571 bp), 559,473 homozygous indels (1–82,711 bp), 90 inversions, as well as numerous segmental duplications and copy number variation regions. Non-SNP DNA variation accounts for 22% of all events identified in the donor, however they involve 74% of all variant bases. This suggests an important role for non-SNP genetic alterations in defining the diploid genome structure. Moreover, 44% of genes were heterozygous for one or more variants. Using a novel haplotype assembly strategy, we were able to span 1.5 Gb of genome sequence in segments >200 kb, providing further precision to the diploid nature of the genome. These data depict a definitive molecular portrait of a diploid human genome that provides a starting point for future genome comparisons and enables an era of individualized genomic information.

J. Craig Venter INSTITUTE

Karyotype of J.Craig Venter



Giemsa Stain



FISH Stain

J. Craig Venter INSTITUTE

Comparing NCBI Assembly to HuRef Assembly

Table 2. Summary of HuRef Assembly Statistics and Comparison to the Human NCBI Genome

Assembly	Assembly Subset	Number of Scaffolds	Number of Contigs	Gaps within Scaffolds	ACGT Bases	Span
NCBI Chromosomes	N/A	279	N/A	N/A	2,858,012,806	3,080,419,480
NCBI All	N/A	367	N/A	N/A	2,870,607,502	3,093,104,542
WGSA Chromosomes	N/A	4,940	211,493	206,553	2,659,468,408	2,993,154,503
HuRef Assembly	Chromosomes	1,408	66,762	66,354	2,782,357,138	2,809,547,336
	Scaffolds ≥ 100 kb	553	65,932	65,379	2,779,929,229	2,806,091,853
	Scaffolds ≥ 3 kb	4,528	71,343	66,815	2,809,774,459	2,844,046,670
	All scaffolds	188,394	255,300	66,906	3,002,932,476	3,037,726,076

J. Craig Venter INSTITUTE

SNPs & InDels in HuRef Autosomes