

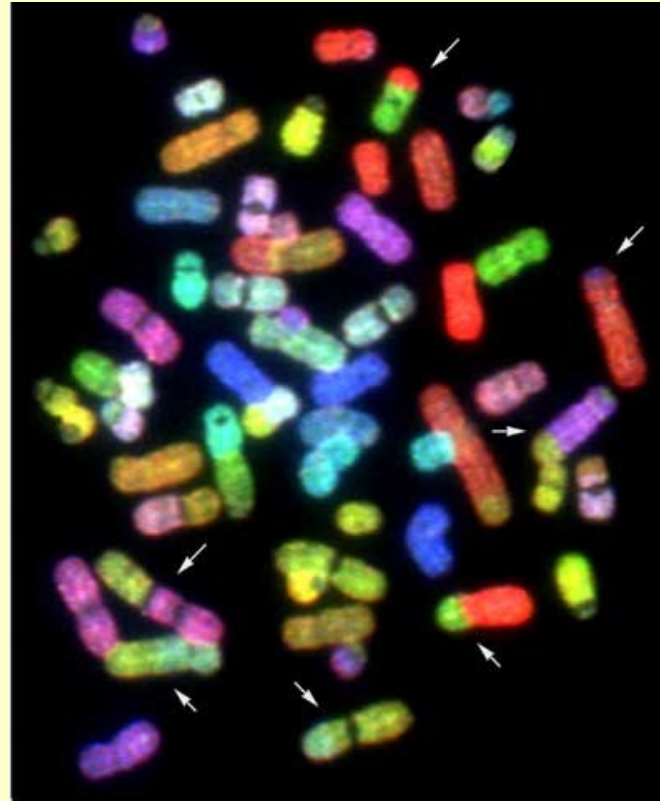
# Computational Molecular Biology

## Biochem 218 – BioMedical Informatics 231

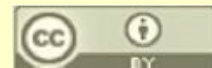
<http://biochem218.stanford.edu/>

---

### The Human Genome Project



Doug Brutlag  
Professor Emeritus  
Biochemistry & Medicine (by courtesy)



# Maeve's Office Hours

---

- Monday and Friday 3:30 to 5:00 PM
- Beckman Center B403A
- Please Email Maeve to set up an appointment if you want to meet with her.
  - [maeveo@stanford.edu](mailto:maeveo@stanford.edu)
- Take the elevator to the fourth floor of Beckman and turn left. The hallway leads directly to the lab B403 and my office is inside the lab at B403A.

# Current Topics in Genome Analysis 2010

<http://www.genome.gov/12514288>

## Current Topics in Genome Analysis 2010


*A lecture series covering contemporary areas in genomics and bioinformatics*

January 12 - March 23, 2010

[Current Topics HOME](#) : [Course Syllabus and Handouts](#) : [Course Mailing List](#) : [Course CME Credits](#) :  
[Course Teleconference Sites](#) : [Course Lectures on the Web](#) : [Course Lectures on DVD](#)

### Course Syllabus and Handouts

All lectures are on Tuesday mornings from 10:00 am to 11:30 am. Lectures are held in the Lipsett Amphitheatre, NIH Clinical Center (Building 10).

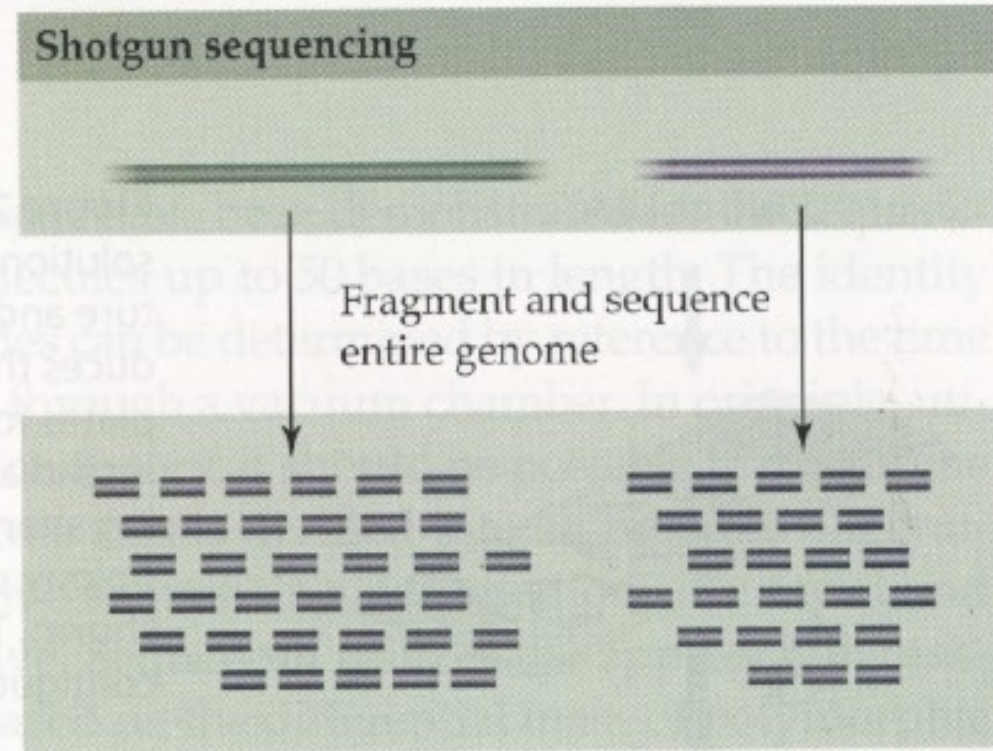
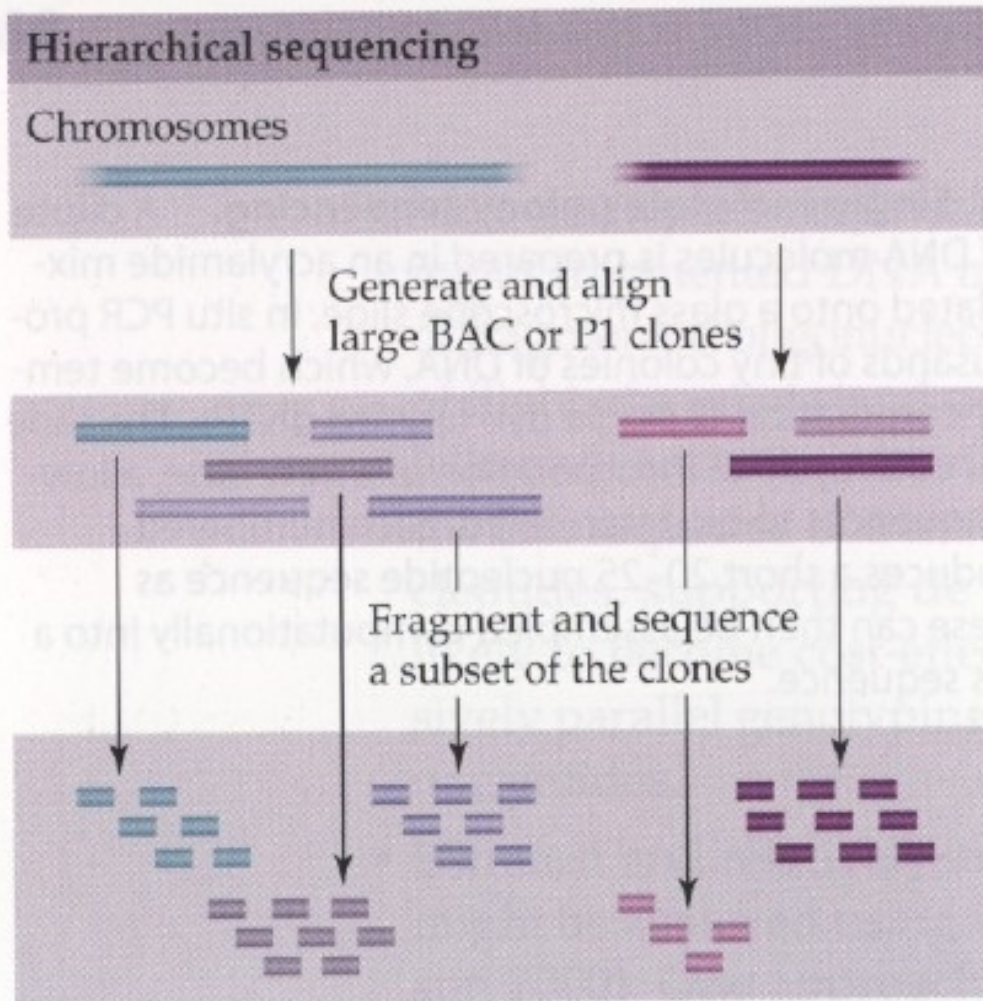
All handouts are in  format. To view, download the free Adobe Acrobat Reader.



- |                    |   |
|--------------------|---|
| <b>January 12</b>  | <b>The Genomic Landscape circa 2010</b><br><i>Eric Green, NHGRI</i><br><a href="#">January 12 Handout</a> (Color)<br><a href="#">January 12 Handout</a> (Grayscale) |
| <b>January 19</b>  | <b>Biological Sequence Analysis I</b><br><i>Andy Baxevanis, NHGRI</i>   |
| <b>January 26</b>  | <b>Biological Sequence Analysis II</b><br><i>Andy Baxevanis, NHGRI</i>  |
| <b>February 2</b>  | <b>Mining Data from Genome Browsers</b><br><i>Tyra Wolfsberg, NHGRI</i>   |
| <b>February 9</b>  | <b>Next-Generation Sequencing Technologies</b><br><i>Elliott Margulies, NHGRI</i>   |
| <b>February 16</b> | <b>Large-Scale Expression Analysis</b><br><i>Paul Meltzer, NCI</i>  |
| <b>February 23</b> | <b>Regulatory and Epigenetic Landscapes of Mammalian Genomes</b><br><i>Laura Elnitski, NHGRI</i>  |
| <b>March 2</b>     | <b>Introduction to Population Genetics</b><br><i>Lynn Jorde, University of Utah</i>   |
| <b>March 9</b>     | <b>Genome-Wide Association Studies</b><br><i>Karen Mohlke, University of North Carolina</i>   |
| <b>March 16</b>    | <b>Genomics of Microbes and Microbiomes</b><br><i>Julie Segre, NHGRI</i>  |
| <b>March 23</b>    | <b>Pharmacogenomics</b><br><i>Howard McLeod, University of North Carolina</i>   |



# Hierarchical Sequencing vs. Whole Genome Shotgun Sequencing



from Gibson & Muse, A Primer of Genome Science  
<http://www.sinauer.com/genomics/>

# 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.
  - <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Cit>
  - Use multiple length clones 2 kb, 10 kb and 50 kb
  - Use clone end sequencing generating mate-pairs
  - 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
  - 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.
  - <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt>
  - 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

# History of Whole Genome Assembly

---

1997



Let's sequence  
the human  
genome with the  
shotgun strategy



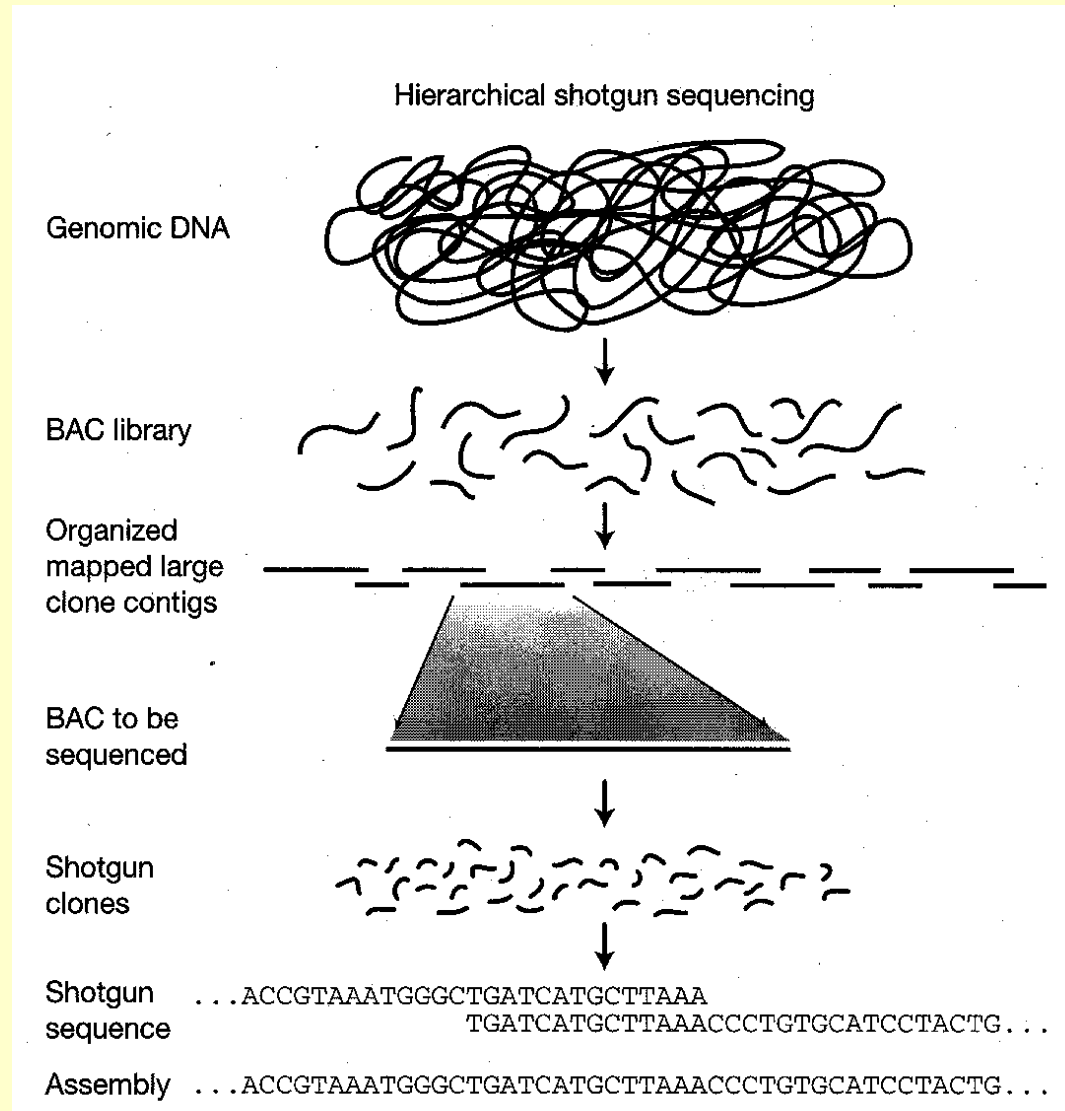
That is  
impossible, and a  
bad idea anyway

Phil Green

Gene Myers

# Public Human Genome Project Strategy

<http://www.genome.gov/>



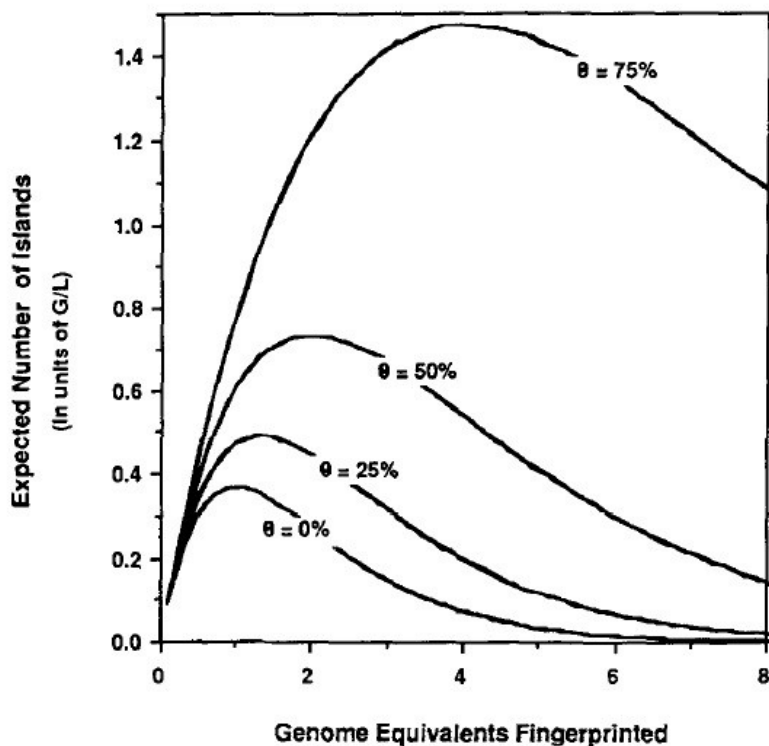
# Contig Formation as Mapping Progresses

## Lander & Waterman 1988

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=3294162](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3294162)

### MATHEMATICAL ANALYSIS OF RANDOM CLONE FINGERPRINTING

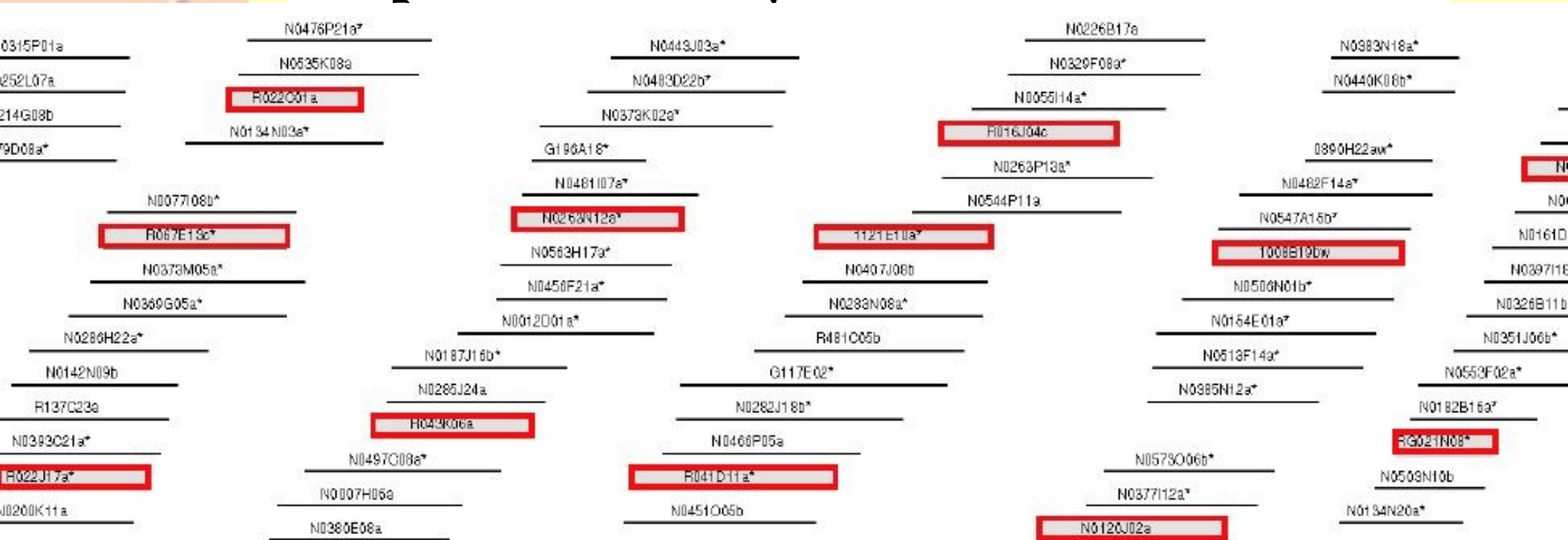
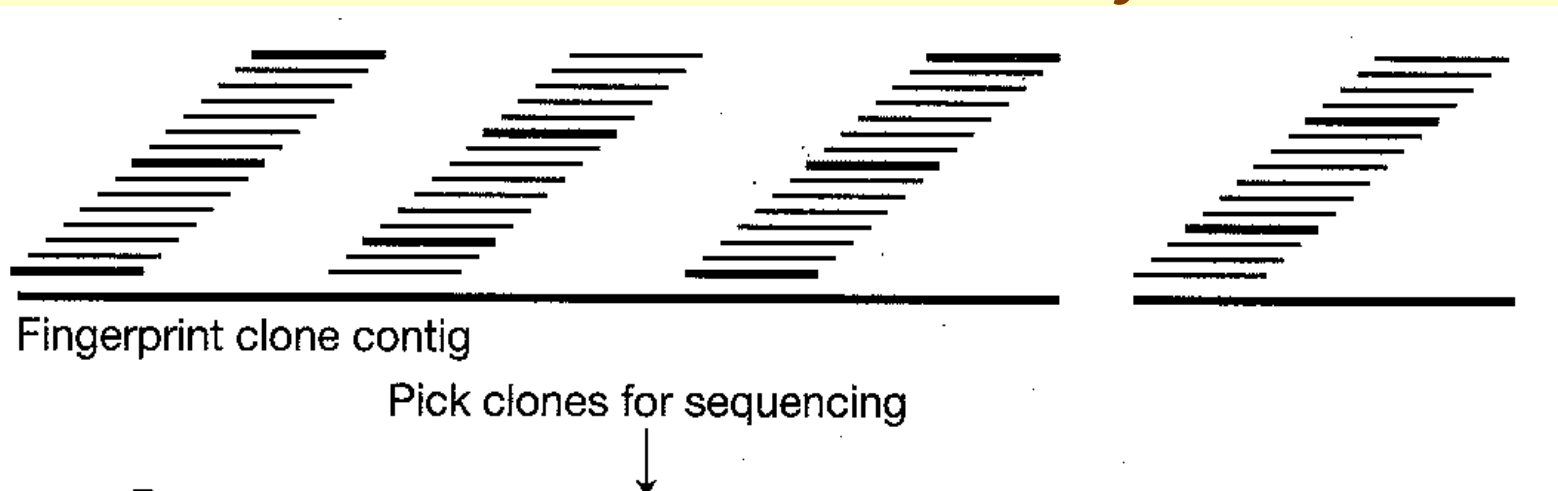
233



$G$  = 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 Genome Assembly Process



# BAC and PAC Libraries in Public Human Genome Project

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Cit>

**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	<i>HindIII</i>	120	74,496
Caltech C	CTC	BAC	Human sperm	All	<i>HindIII</i>	125	263,040
Caltech D1 (CITB-H1)	CTD	BAC	Human sperm	All	<i>HindIII</i>	129	162,432
Caltech D2 (CITB-E1)		BAC	Human sperm	All			
				2,501–2,565	<i>EcoRI</i>	202	24,960
				2,566–2,671	<i>EcoRI</i>	182	46,326
				3,000–3,253	<i>EcoRI</i>	142	97,536
RPCI-1	RP1	PAC	Male, blood	All	<i>Mbol</i>	110	115,200
RPCI-3	RP3	PAC	Male, blood	All	<i>Mbol</i>	115	75,513
RPCI-4	RP4	PAC	Male, blood	All	<i>Mbol</i>	116	105,251
RPCI-5	RP5	PAC	Male, blood	All	<i>Mbol</i>	115	142,773
RPCI-11	RP11	BAC	Male, blood	All		178	543,797
				1	<i>EcoRI</i>	164	108,499
				2	<i>EcoRI</i>	168	109,496
				3	<i>EcoRI</i>	181	109,657
				4	<i>EcoRI</i>	183	109,382
				5	<i>Mbol</i>	196	106,763
Total of top							1,482,502



# Total Genome Sequence Information 2001

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=11237011](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11237011)

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

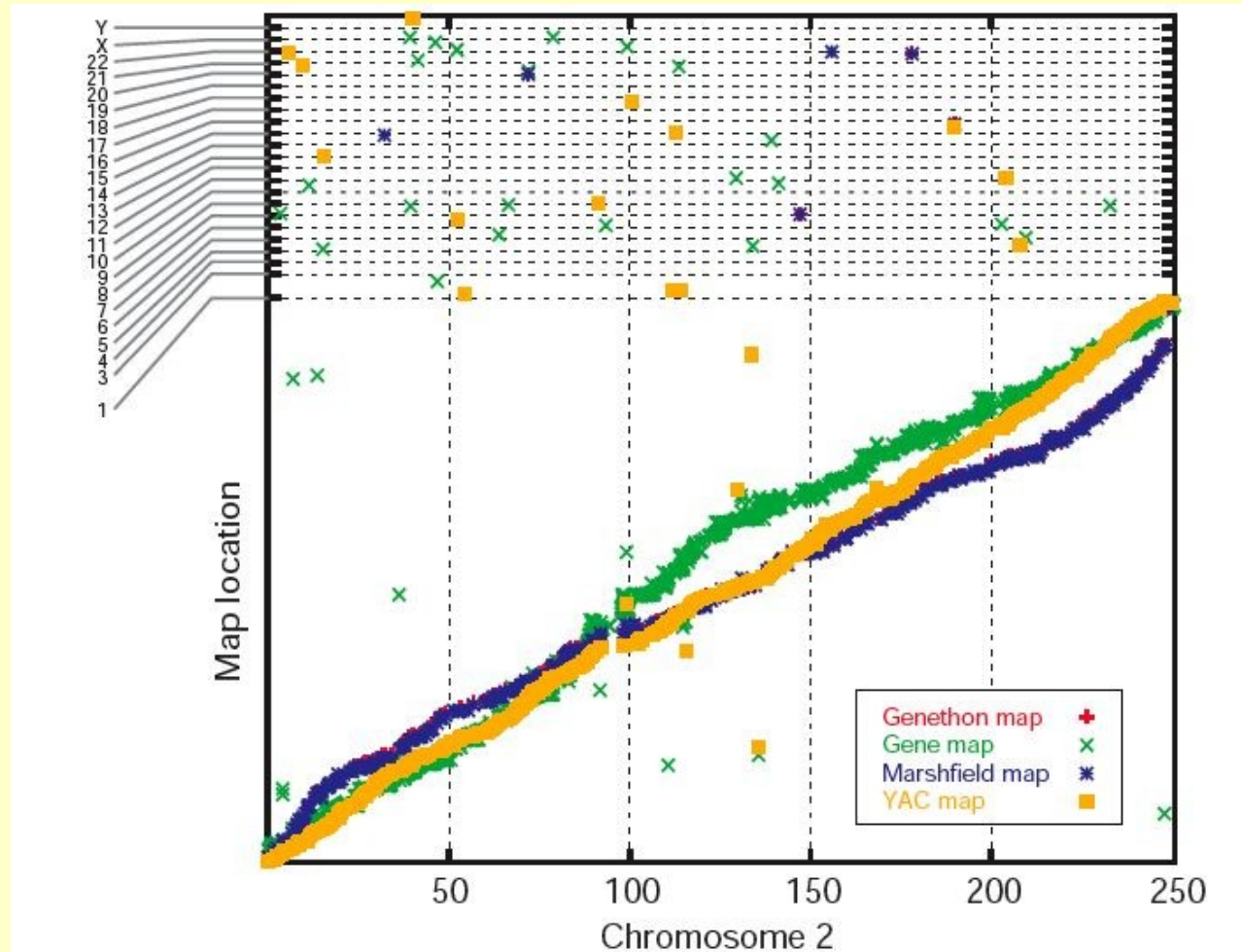
Sequence status	Number of clones	Total clone length (Mb)	Average number of sequence reads per kb*	Average sequence depth†	Total amount of raw sequence (Mb)
Finished	8,277	897	20–25	8–12	9,085
Draft	18,969	3,097	12	4.5	13,395
Predraft	2,052	267	6	2.5	667
Total					23,147

\* 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.

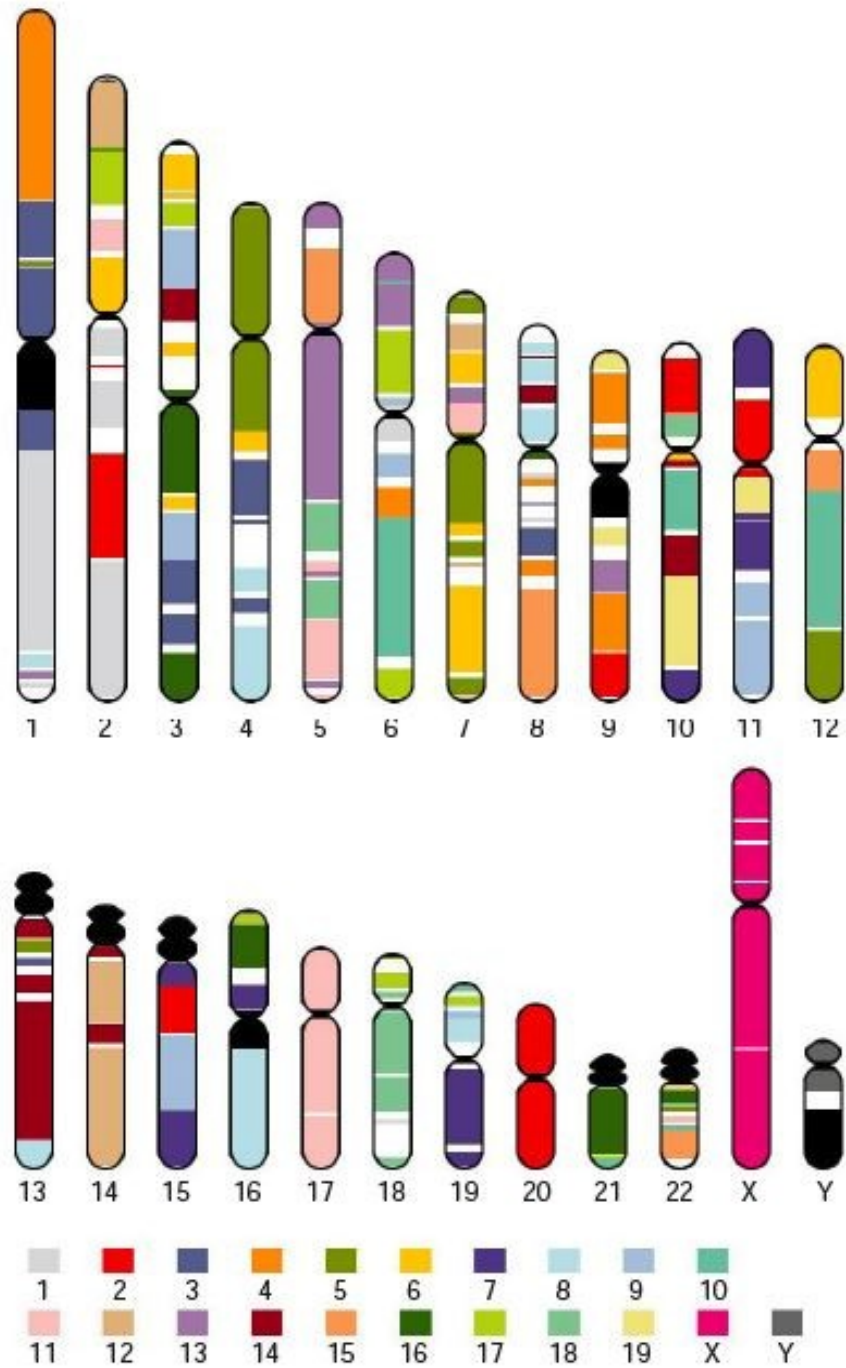
† The average depth in high quality bases ( $\geq 99\%$  accuracy) was estimated from information provided by each sequencing centre. The average varies among the centres, and the number may vary considerably for clones with the same sequencing status. For draft clones in the public databases (keyword: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

# Comparing Chromosome 2 Sequence Versus Genetic Maps

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Pubmed&dopt=Citation&list\\_uids=11237011](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11237011)



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



**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.

# Synteny Between Human and Mouse

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retr>



# 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		

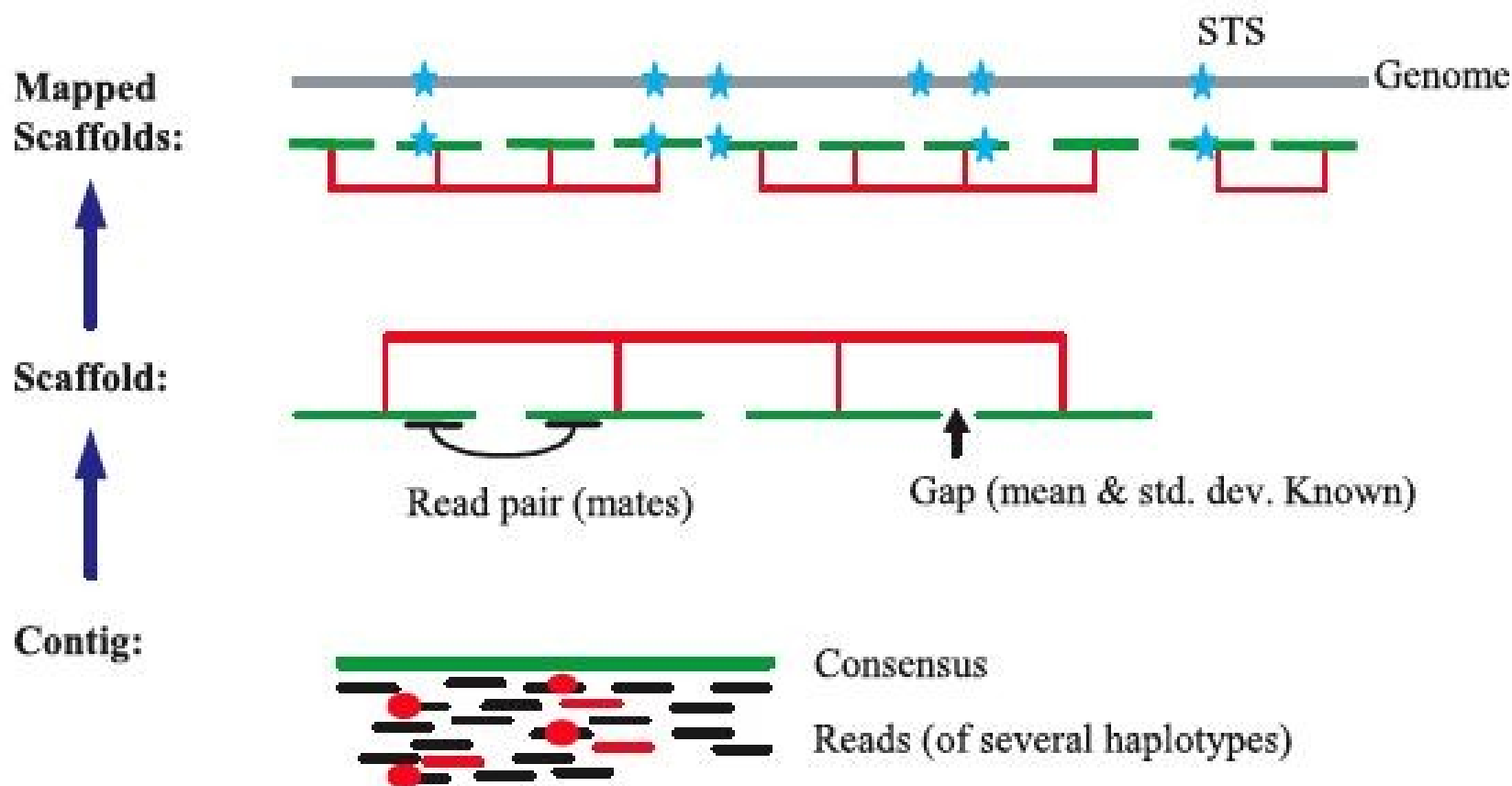
\*Insert size and SD are calculated from assembly of mates on contigs.

†% Mates is based on laboratory tracking of sequencing runs.



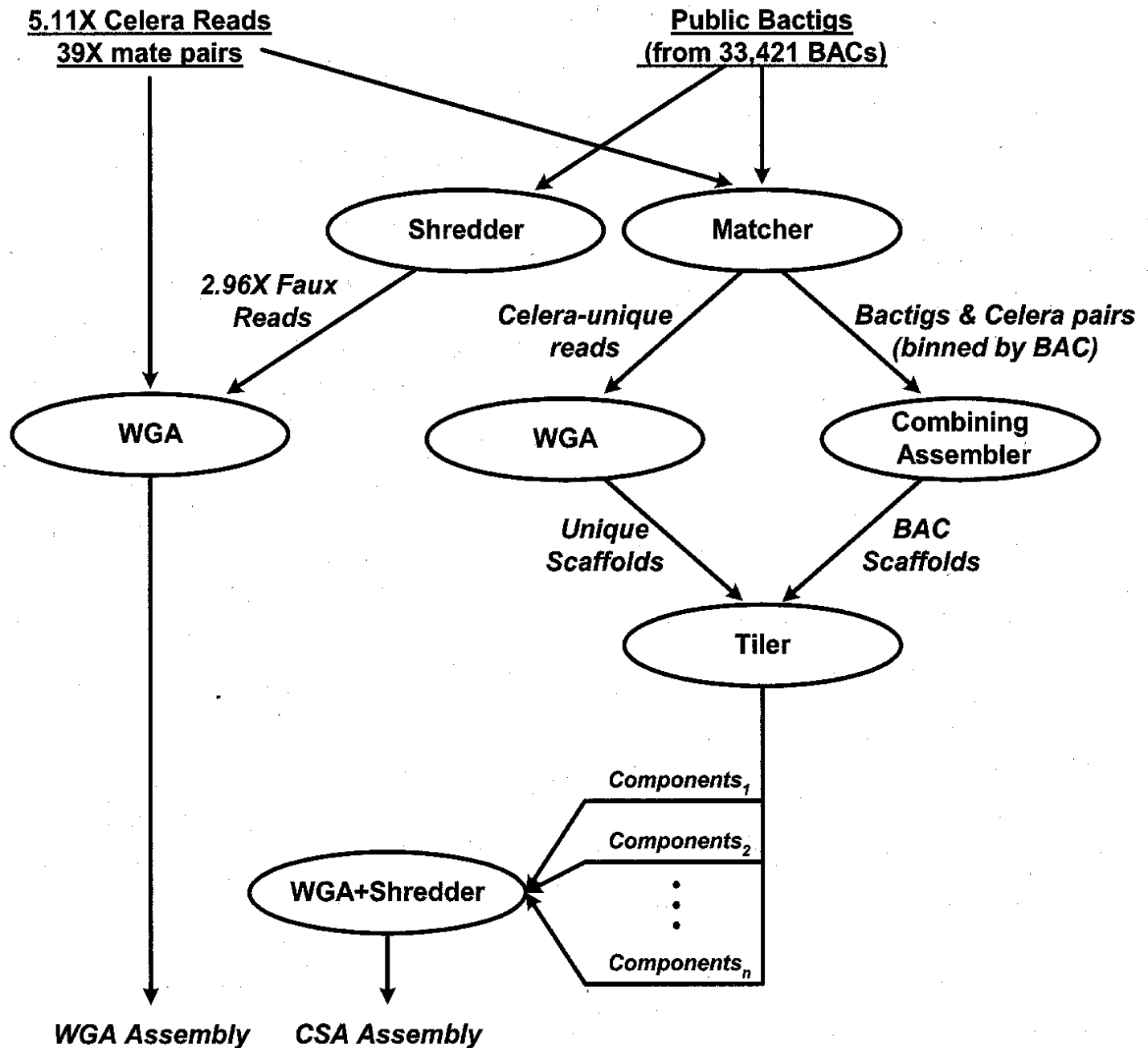
# Celera Scaffolds

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



**Fig. 3.** Anatomy of whole-genome assembly. Overlapping shredded contig 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.

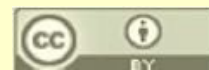
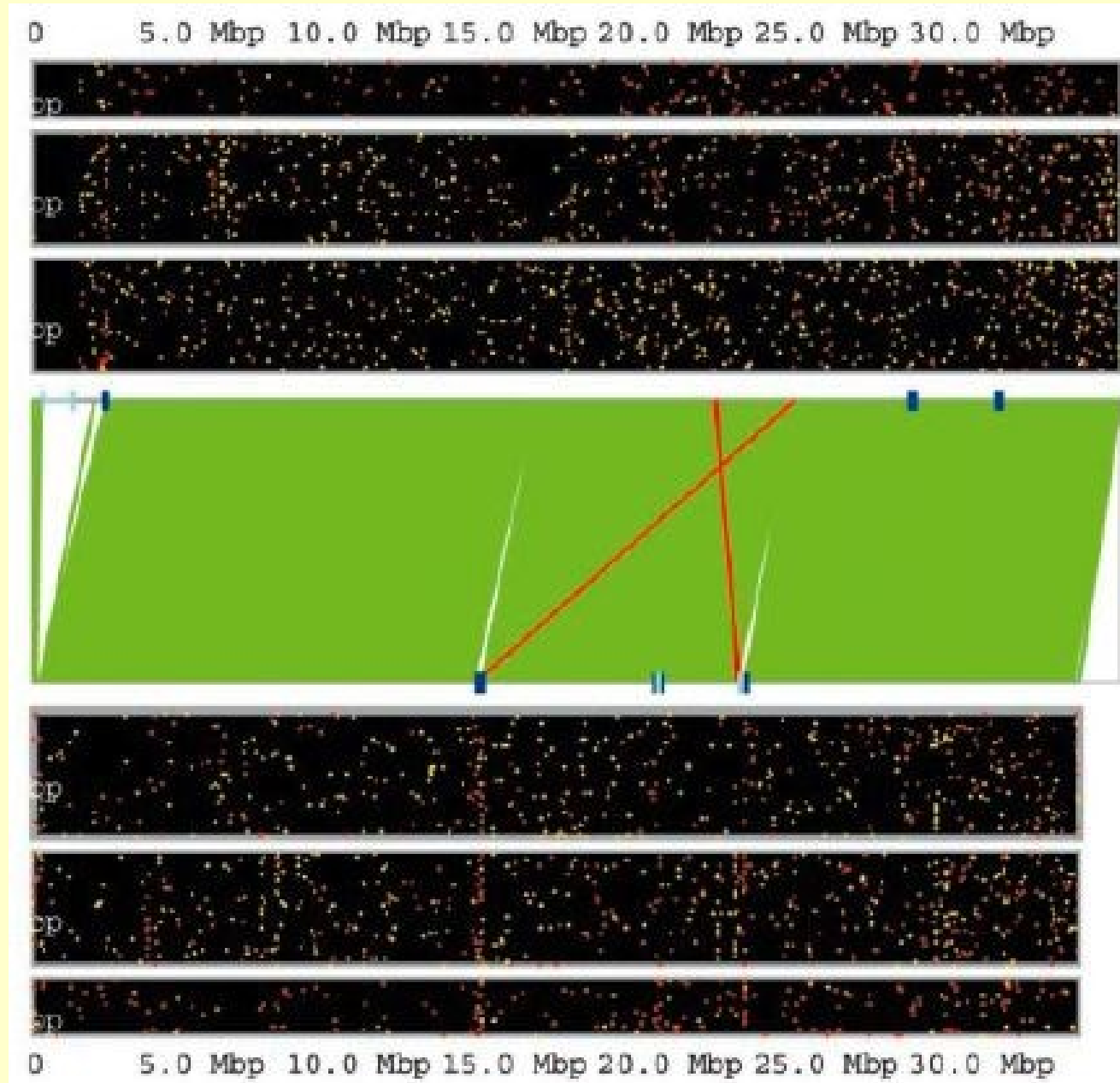
# Celera Assembler





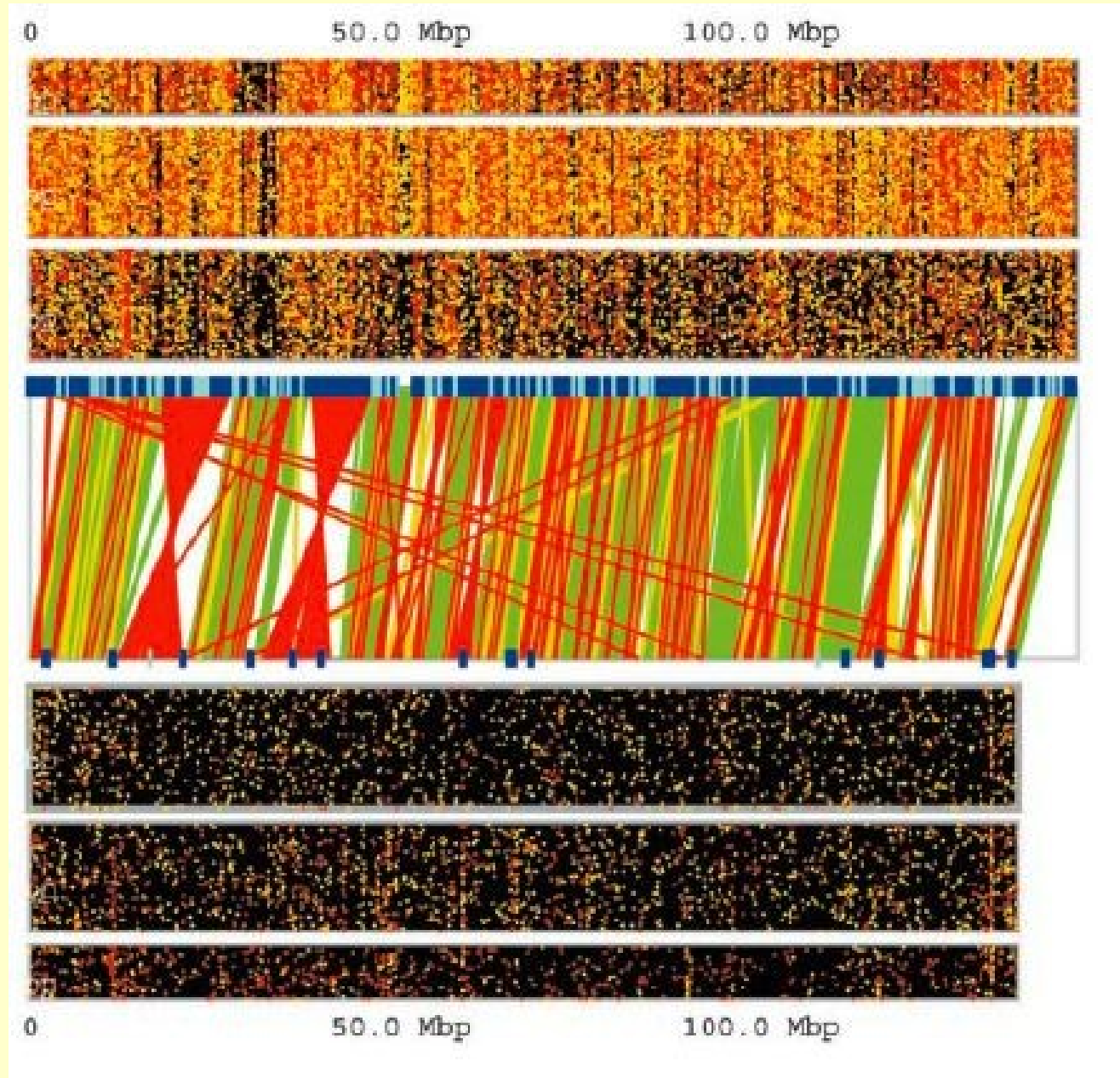
# Chromosome 21: Public vs Celera Assemblies

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



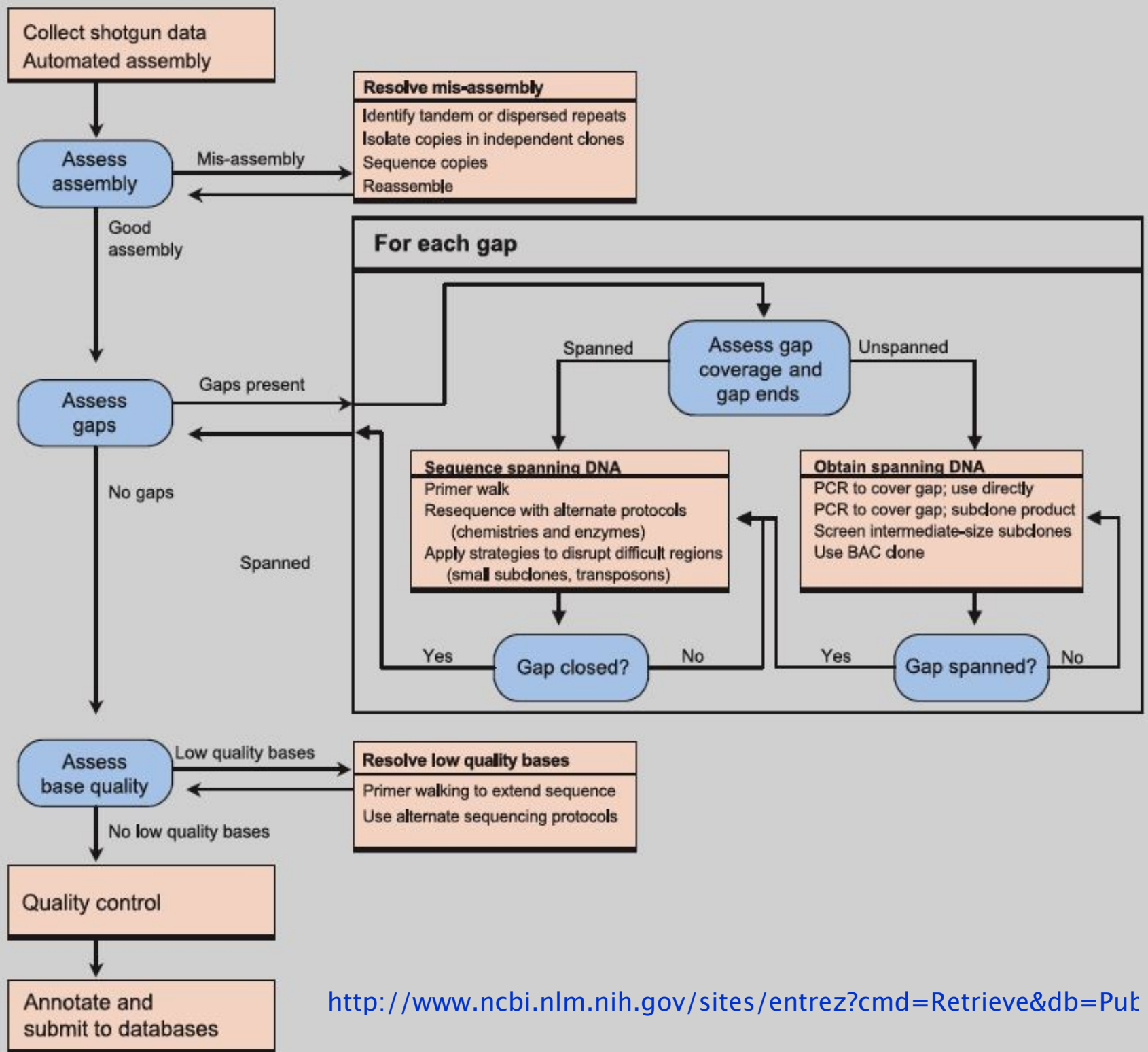
# Chromosome 8: Public vs. Celera

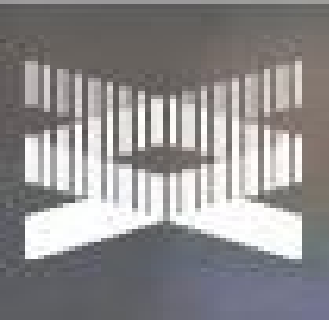
<http://www.sciencemag.org/cgi/content/full/291/5507/1304>





# Finishing Strategy for the Public Genome Project





# Finished Sequence in 2004 (Build 35)

[http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids)

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

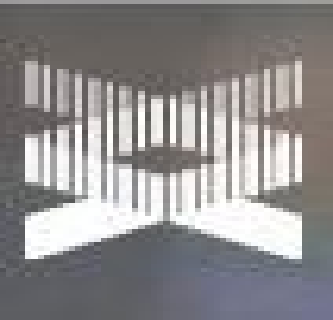
\* The total length of tiling paths including only finished bases of clones in Build 35. Roughly 2.19 Mb of sequence on chromosome Y was derived directly from the equivalent pseudoautosomal region on chromosome X.

† 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<sup>62,64</sup>. 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<sup>63</sup>), but are classified as euchromatin here.

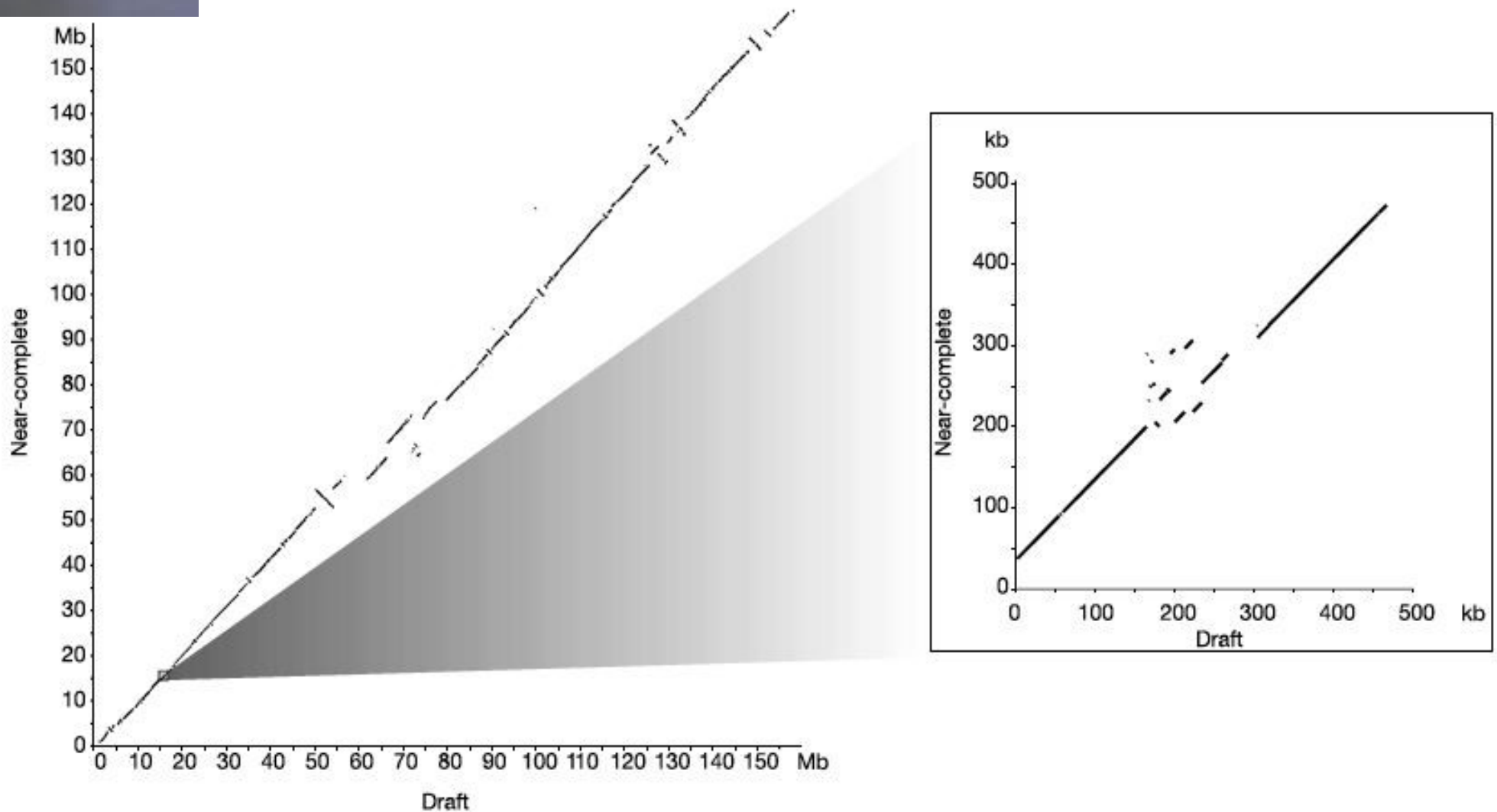
§ The sum of lengths for finished sequence, estimated heterochromatic gaps, euchromatic gaps and unfinished clone gaps. The total length is only approximate because of uncertainty in gap sizes, particularly for heterochromatic gaps and centromeres.

|| 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.



# Chromosome 7: Draft versus Finished Sequence

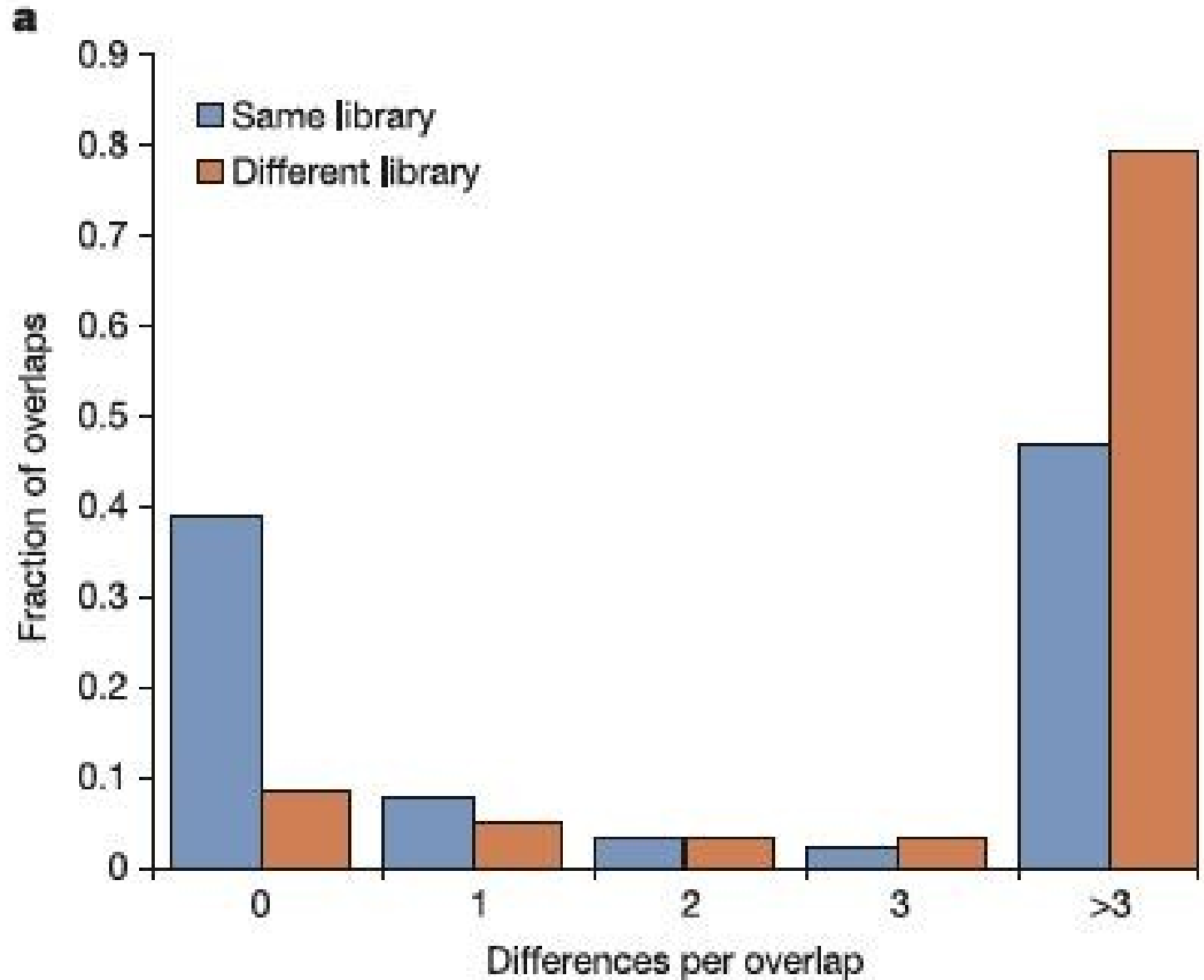
<http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&dopt=Citati>



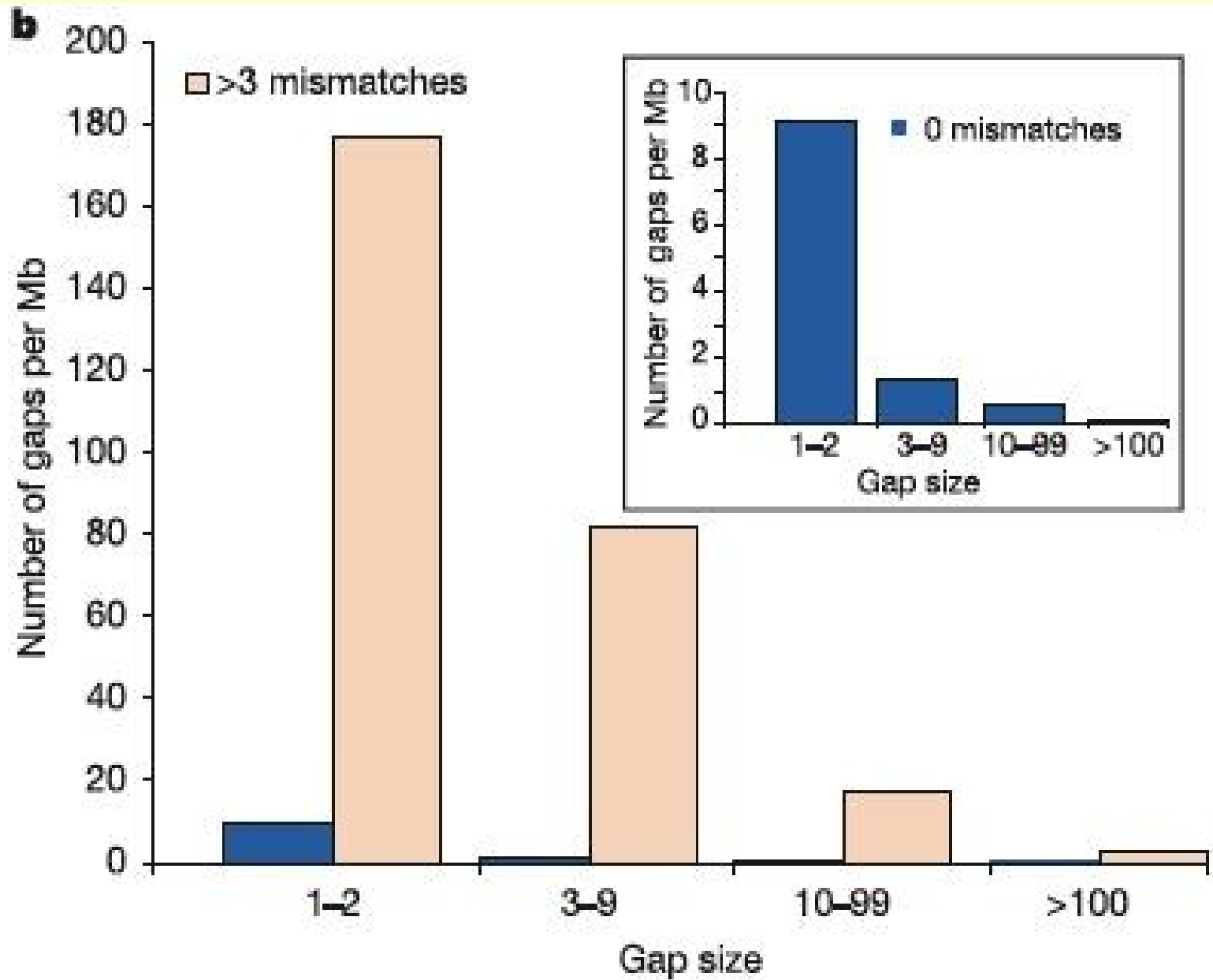
**Figure 1** Comparison of previous draft sequence with current near-complete sequence of chromosome 7 (ref. 24). At large scale, there was good collinearity between draft and near-complete sequence, although some inversions were present in the draft due to lack of sufficient anchors in some regions. 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.

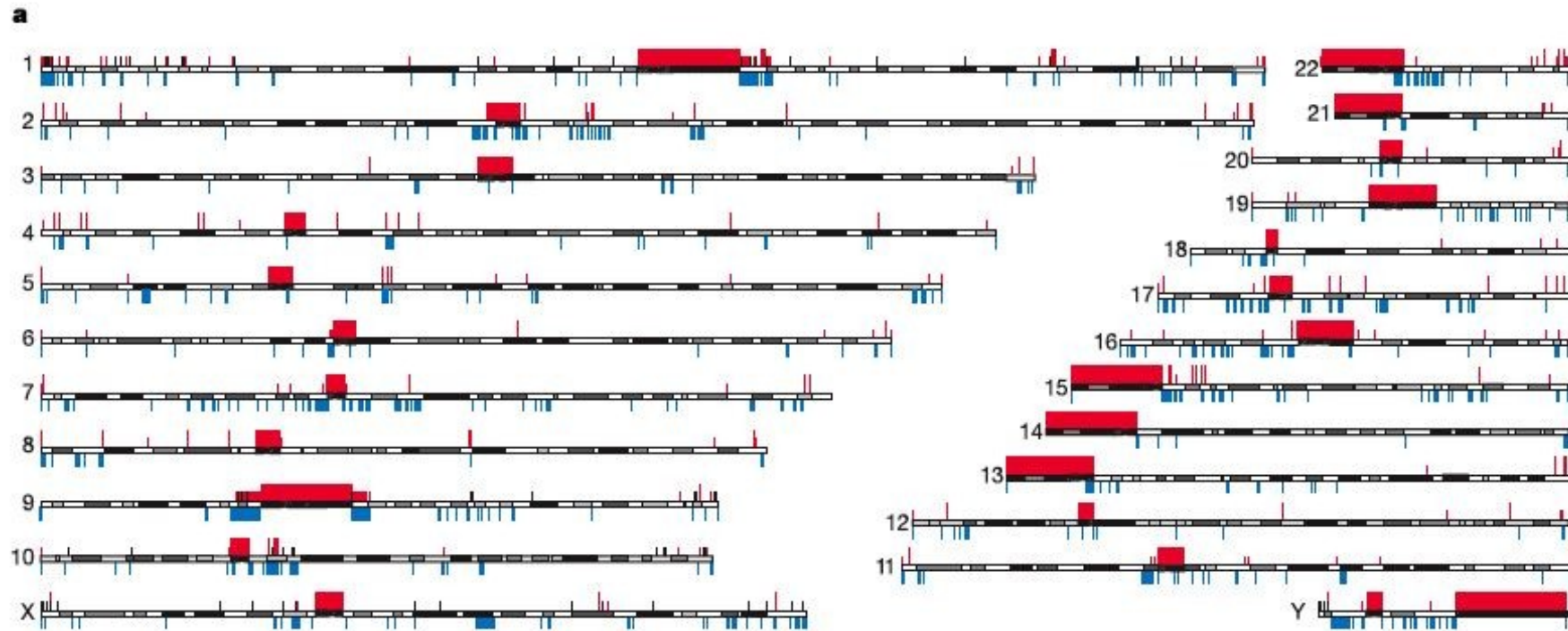
# Substitutions in BAC Overlaps with BACs



# Gaps in BAC Overlaps with BACs from



# Duplications and Deletions in the Human Genome

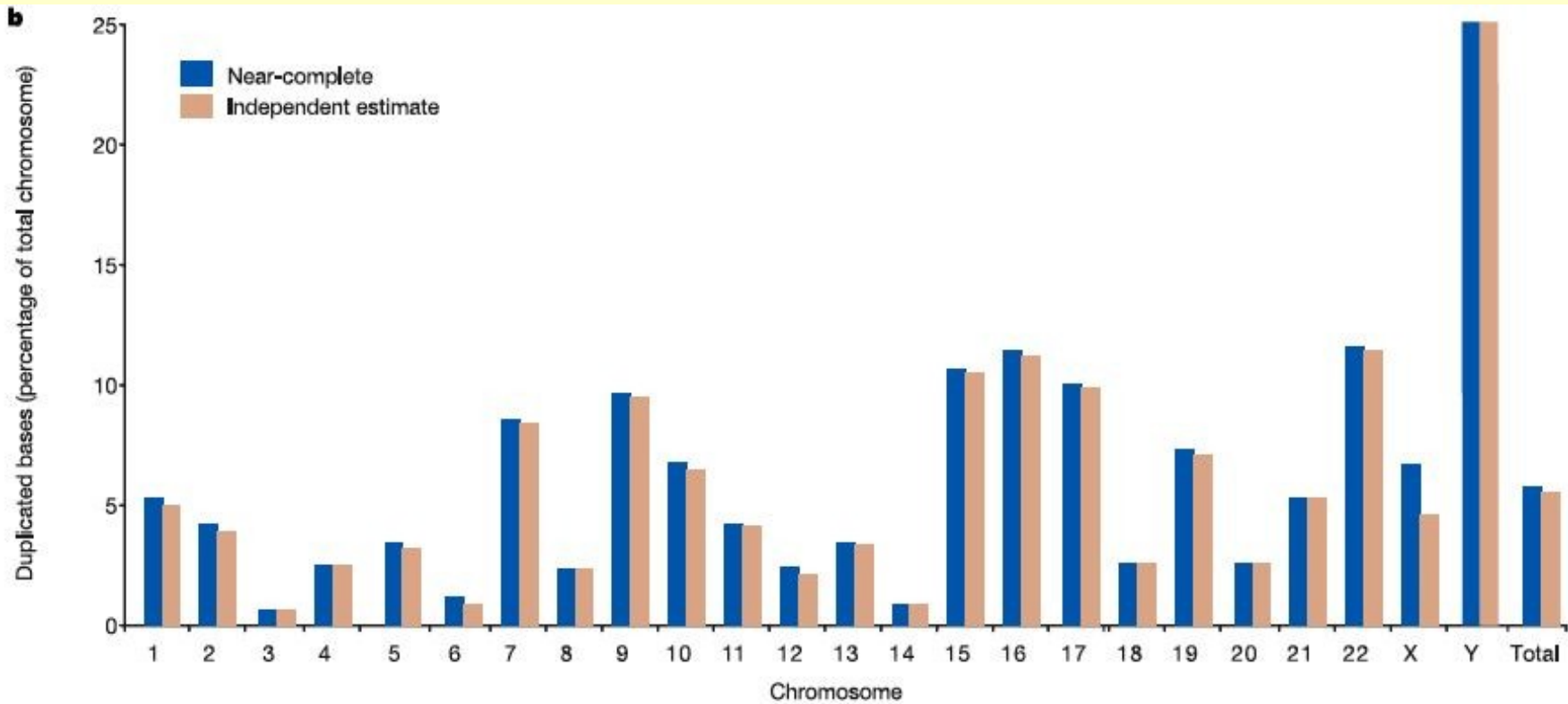


**Figure 4** Segmental duplications across the genome. **a**, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length  $\geq 10$  kb and sequence identity  $\geq 95\%$ ). Large duplications are shown to approximate scale; 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

<http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&do>



# Available Next Generation Sequencing Technologies

---

- **Illumina – Solexa**
  - <http://www.illumina.com/>
- **Illumina Technology**
  - [http://www.illumina.com/technology/sequencing\\_technology.ilmn](http://www.illumina.com/technology/sequencing_technology.ilmn)
- **454 Life Sciences**
  - <http://www.454.com/>
- **454 Life Sciences Technology**
  - <http://www.454.com/products-solutions/how-it-works/index.asp>
- **Applied Biosystems Inc. (ABI) SOLID Sequencing**
  - <http://solid.appliedbiosystems.com/>
- **ABI SOLID Sequencing Technology**
  - [http://www3.appliedbiosystems.com/AB\\_Home/applicationstechno](http://www3.appliedbiosystems.com/AB_Home/applicationstechno)

# Next Generation Sequencing Technologies in Development

---

- **Pacific BioSciences**

- <http://www.pacificbiosciences.com/>

- **Pacific BioSciences Technology**

- <http://www.pacificbiosciences.com/index.php?q=technology-introduction>

- **Helicos**

- <http://www.helicosbio.com/>

- **Helicos Technology**

- <http://www.helicosbio.com/Technology/TrueSingleMoleculeSequen>

- **Complete Genomics**

- <http://www.completegenomics.com/>

- **Complete Genomics Technology**

- <http://www.completegenomics.com/technology/technicalDetails.aspx>