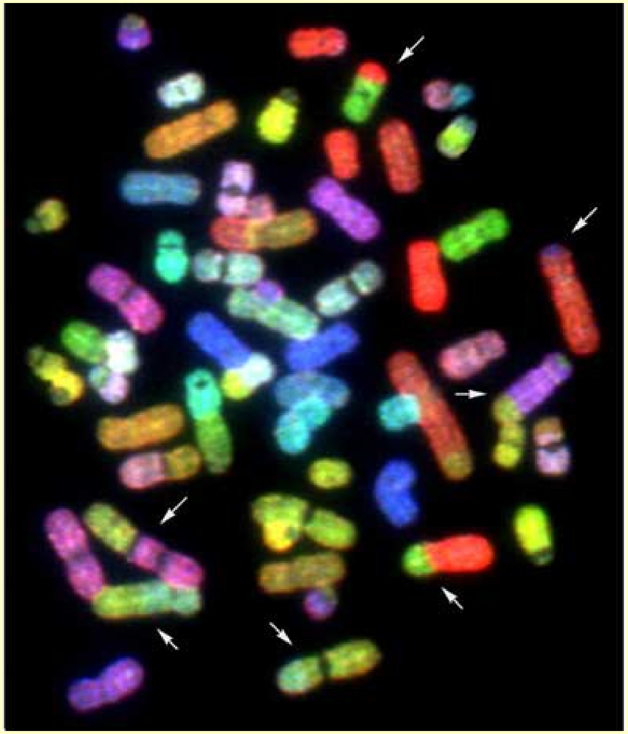
Finishing the Human Genome

http://biochem118.stanford.edu/

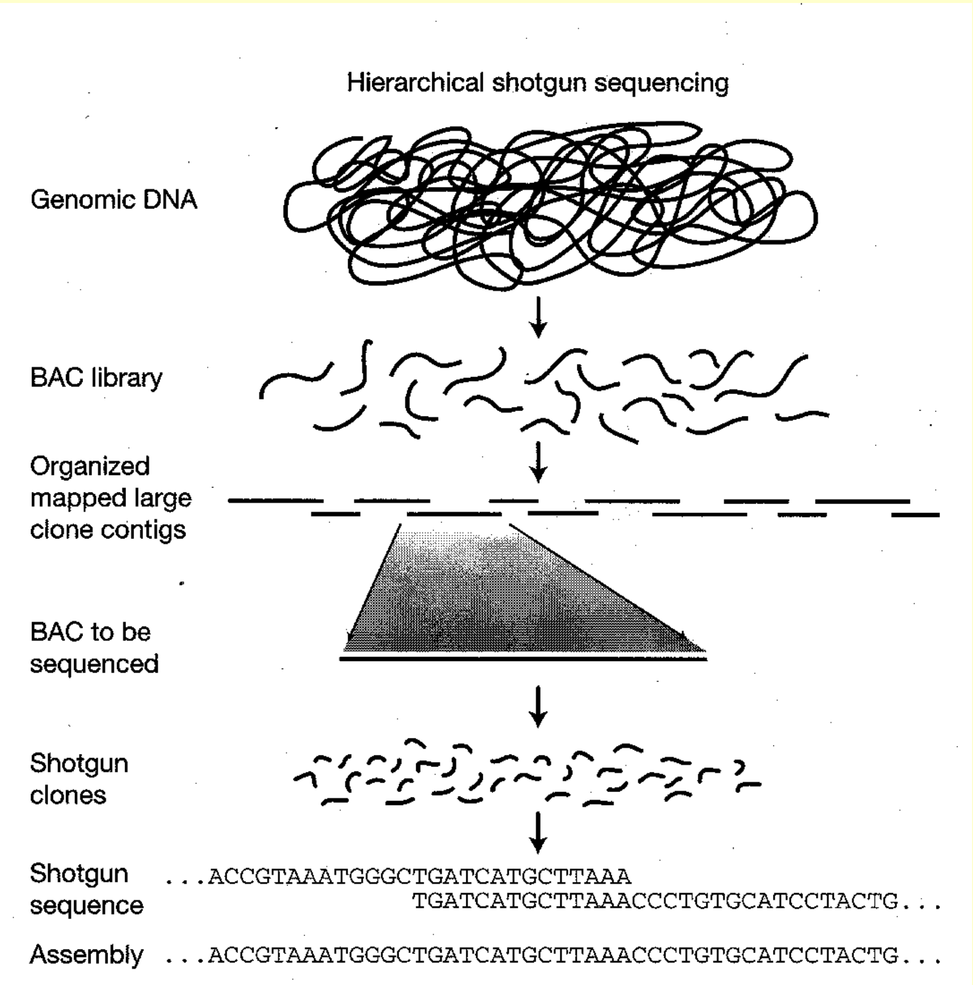
Stanford Sophomore Seminar



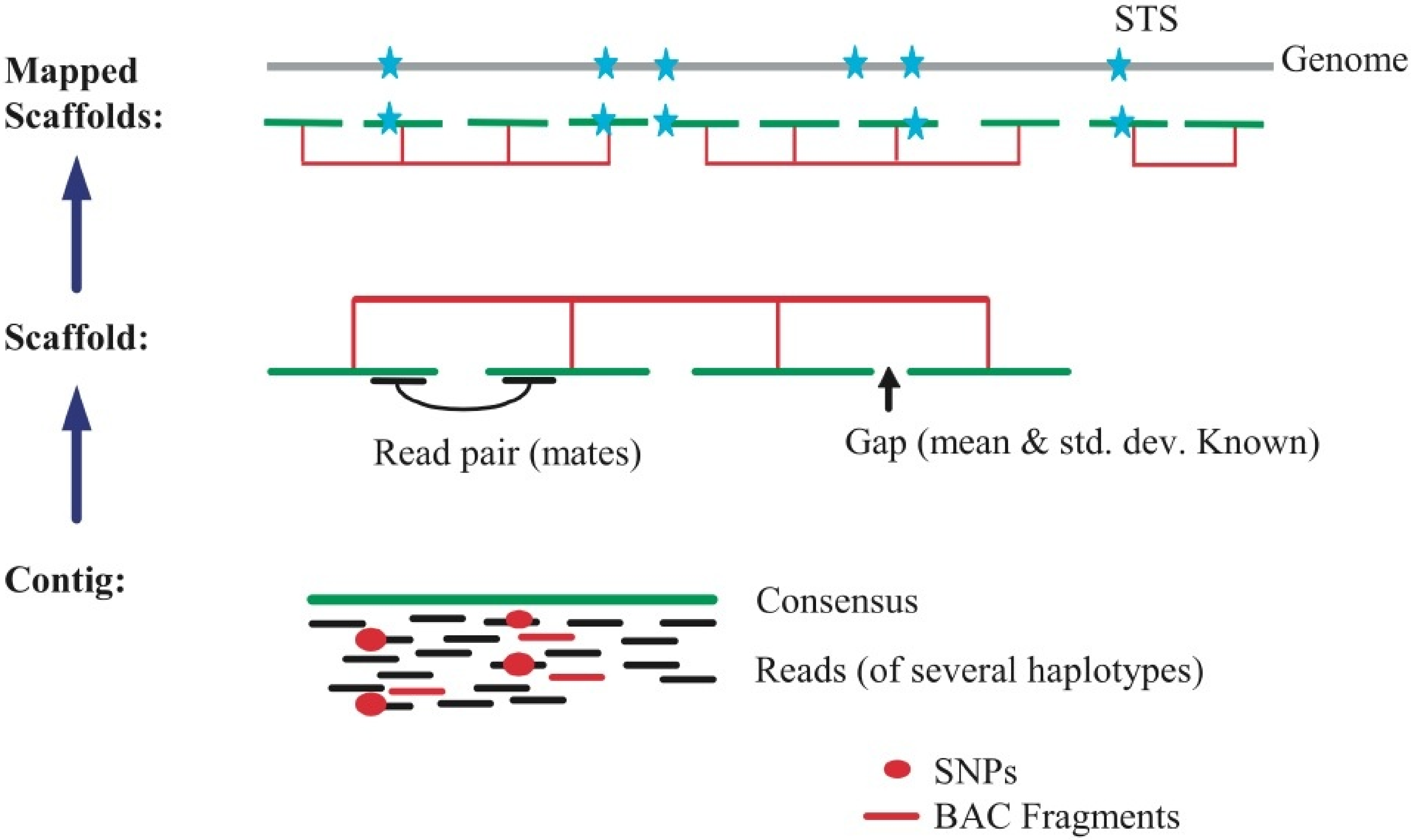
Doug Brutlag, Professor Emeritus of
Biochemistry & Medicine (by courtesy)
Stanford University School of Medicine

Public Human Genome Project Strategy

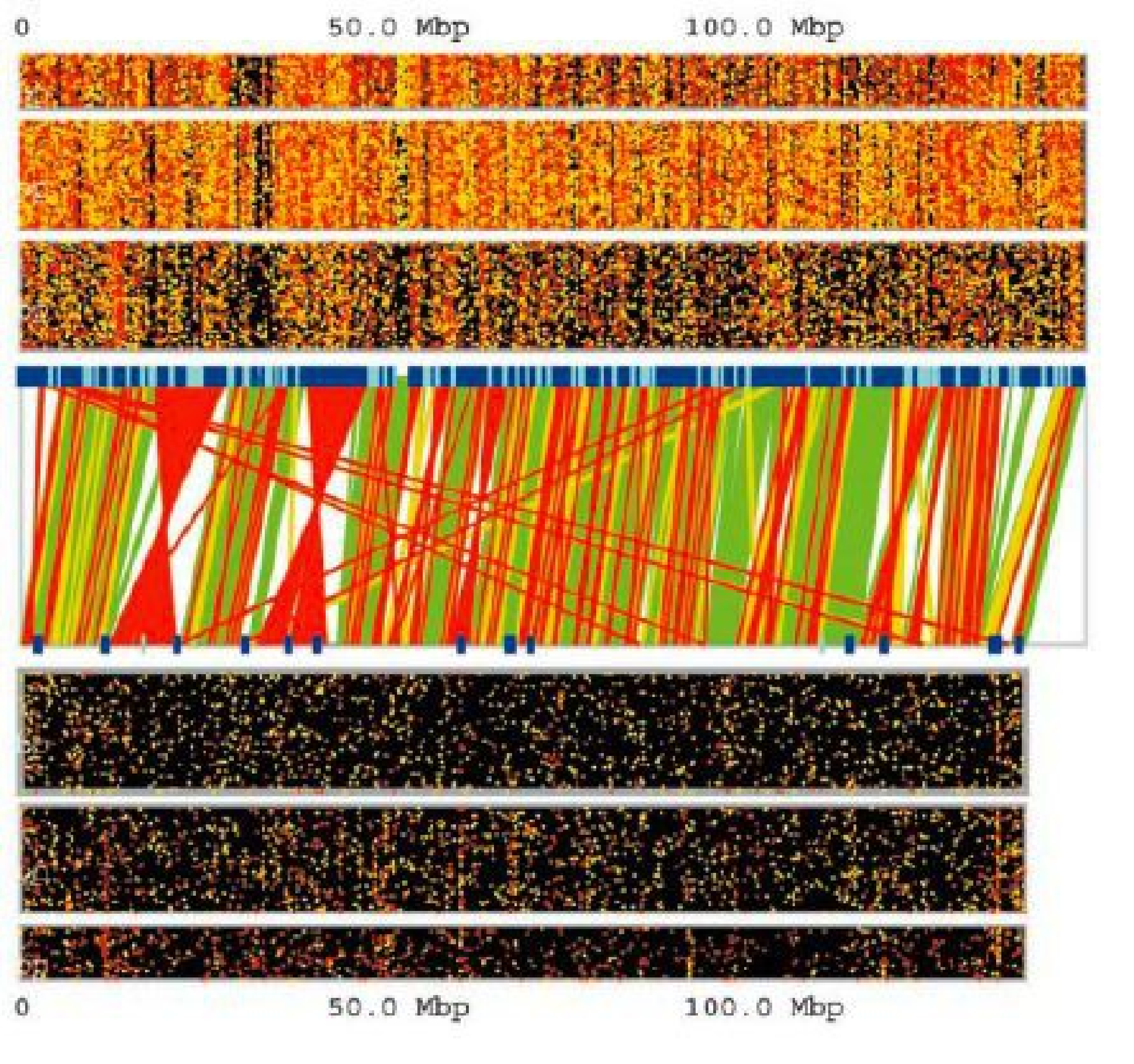
http://www.nhgri.nih.gov/


Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

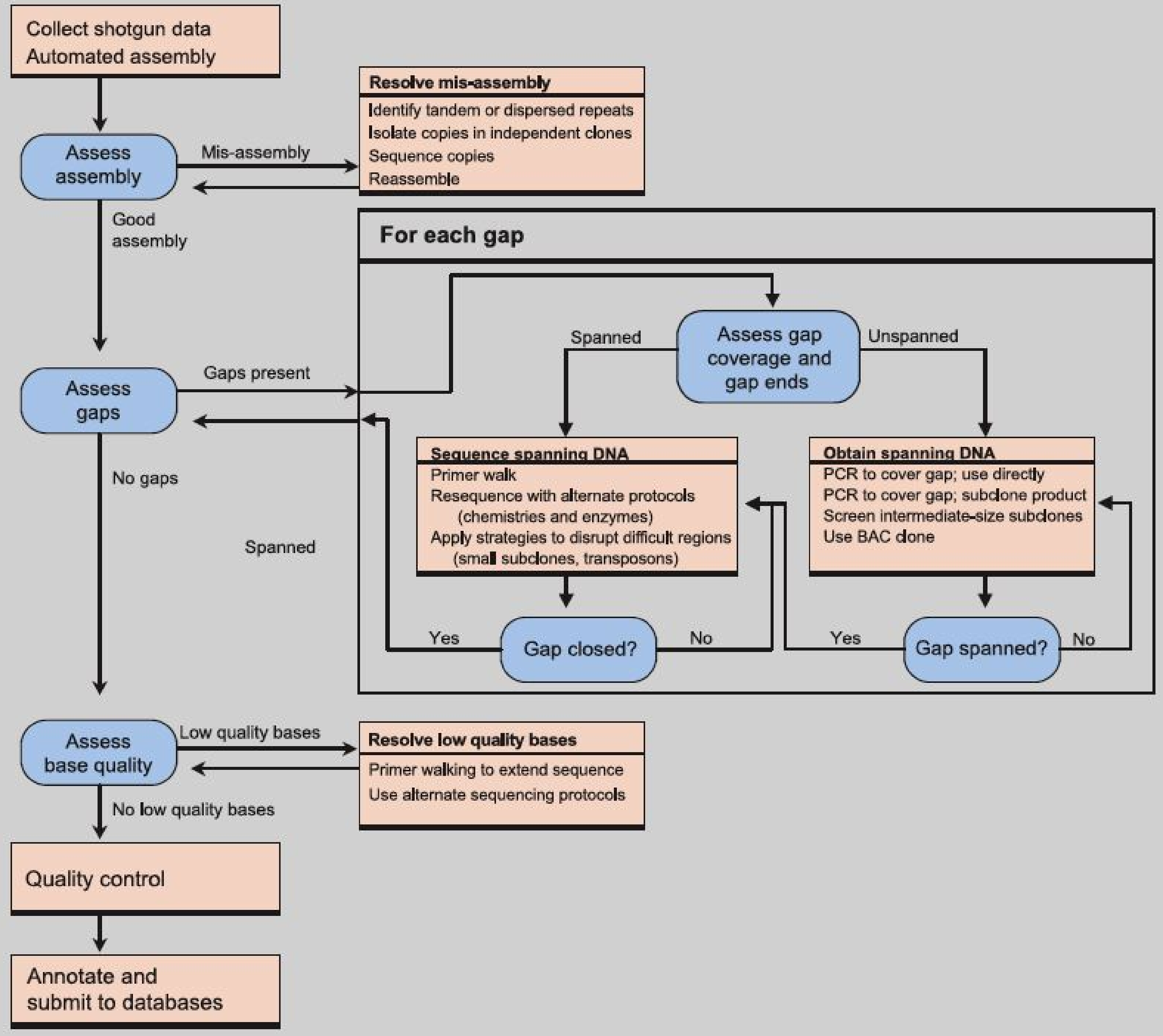
Celera Scaffolds



Chromosome 8: Public vs. Celera



Finishing Strategy for the Public Genome Project



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

Assess gaps

Gaps present

For each gap

Assess gap coverage and gap ends

Spanned

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

Finished Sequence in 2004 (Build 35)

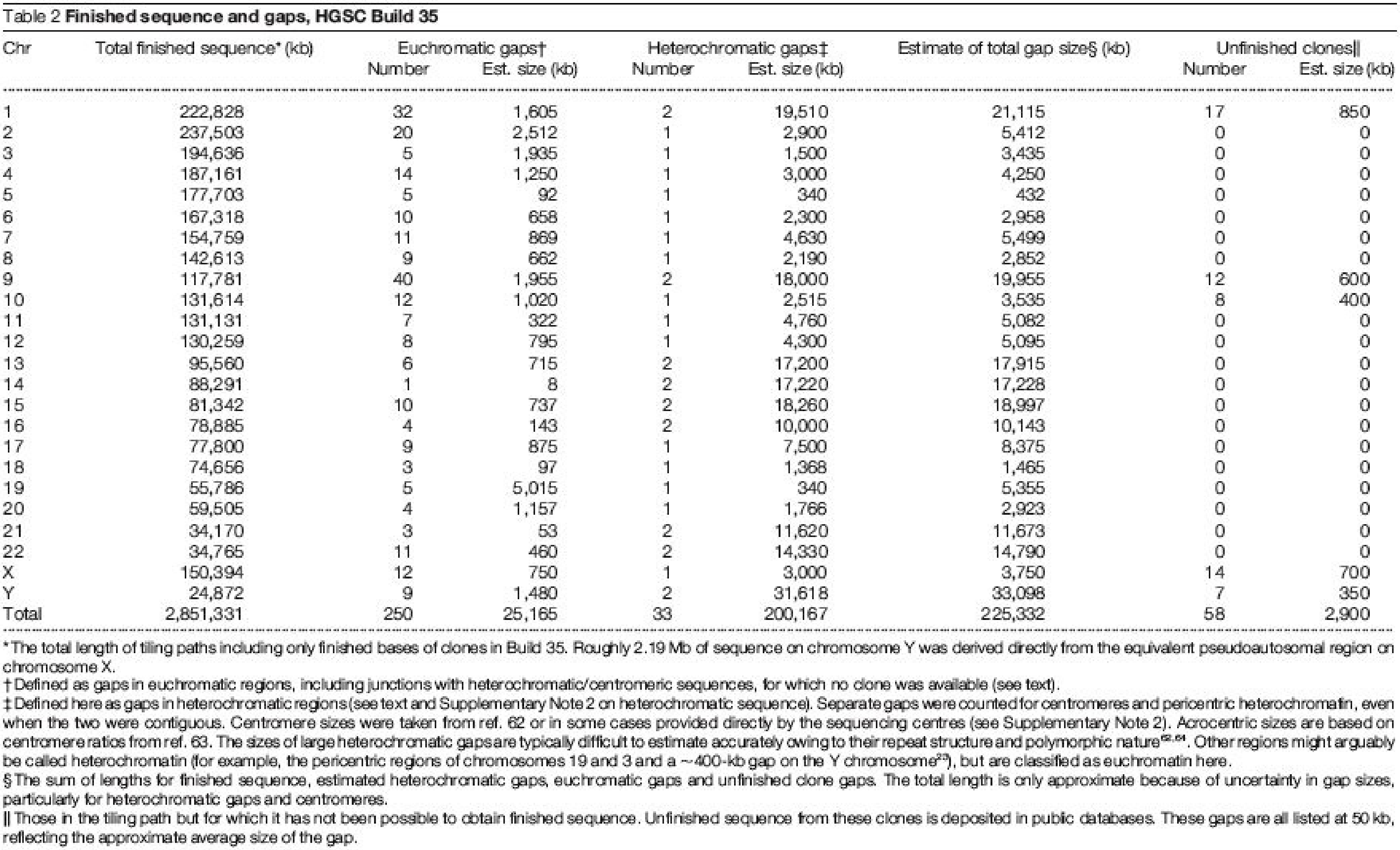


Table 2 **Finished sequence and gaps, HGSC Build 35**

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.

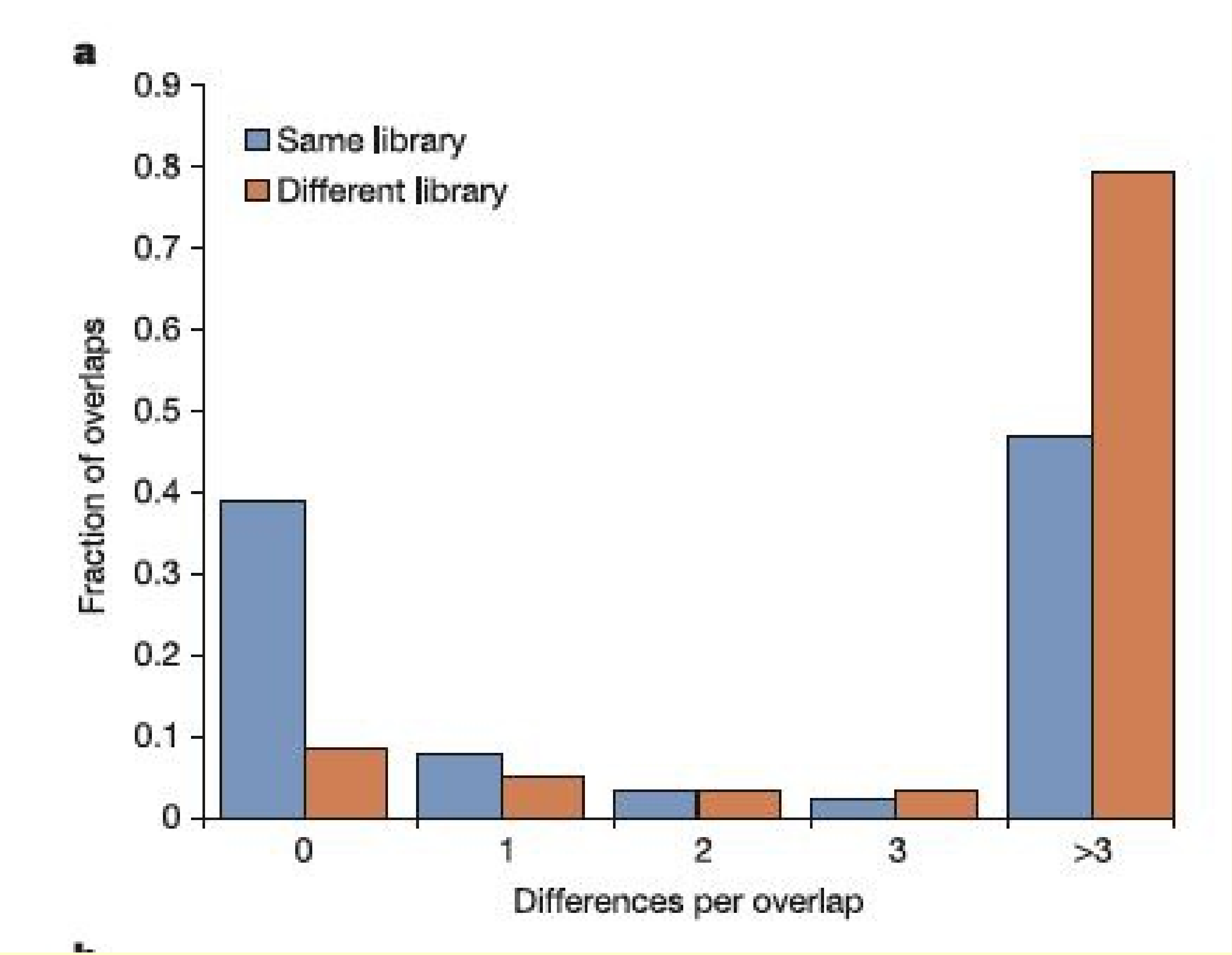
†Defined as gaps in euchromatic regions, including junctions with heterochromatic/centromeric sequences, for which no clone was available (see text).

‡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.

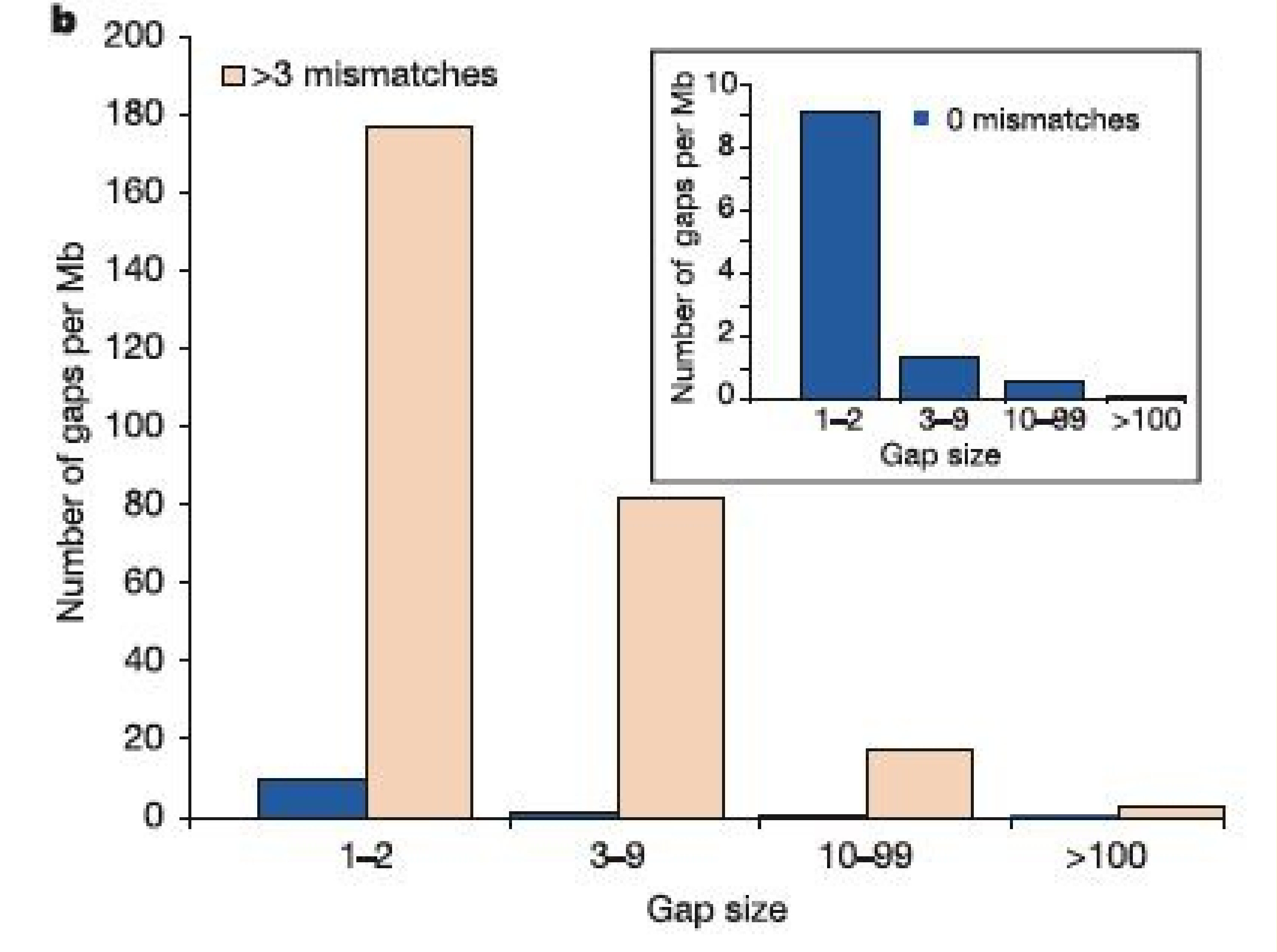
§The sum of lengths for finished sequence, estimated heterochromatic gaps, euchromatic gaps and unfinished clone gaps. The total length is only approximate because of uncertainty in gap sizes, particularly for heterochromatic gaps and centromeres.

‖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.

Substitutions in BAC Overlaps with BACs from Same or Different Libraries



Gaps in BAC Overlaps with BACs from Same or Different Libraries



Duplications and Deletions in the Human Genome

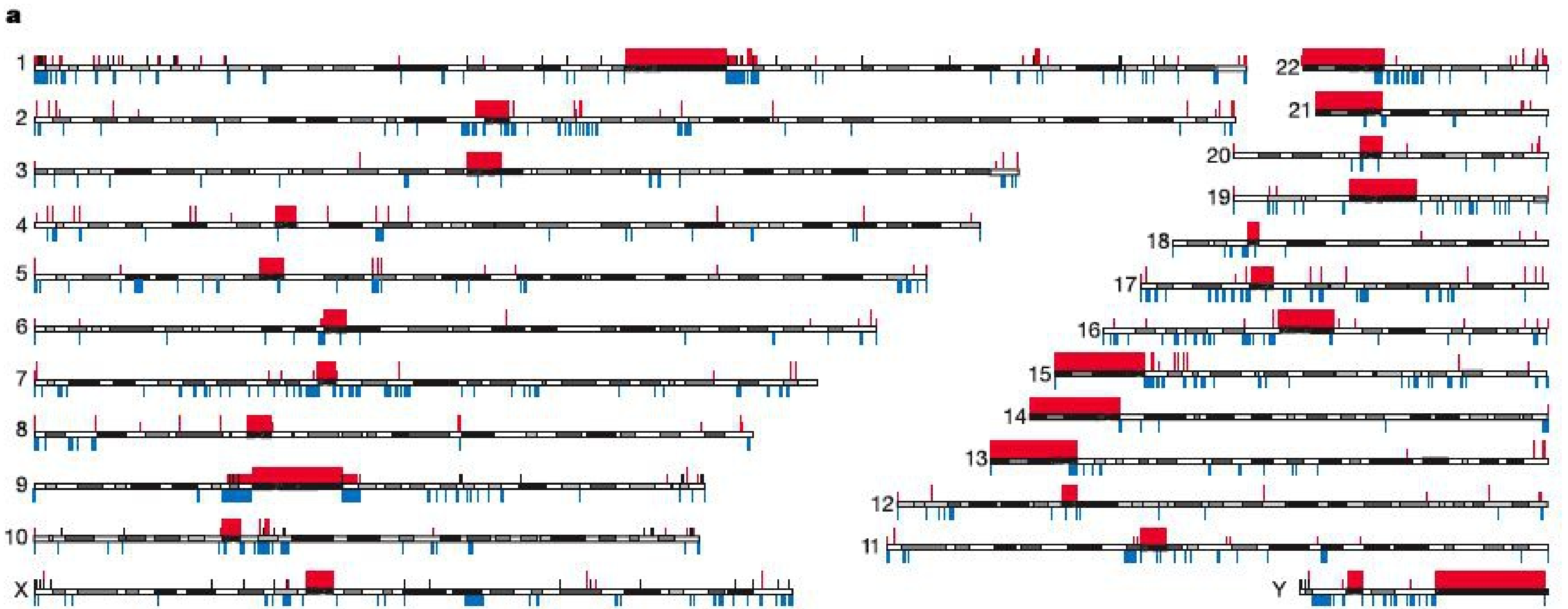
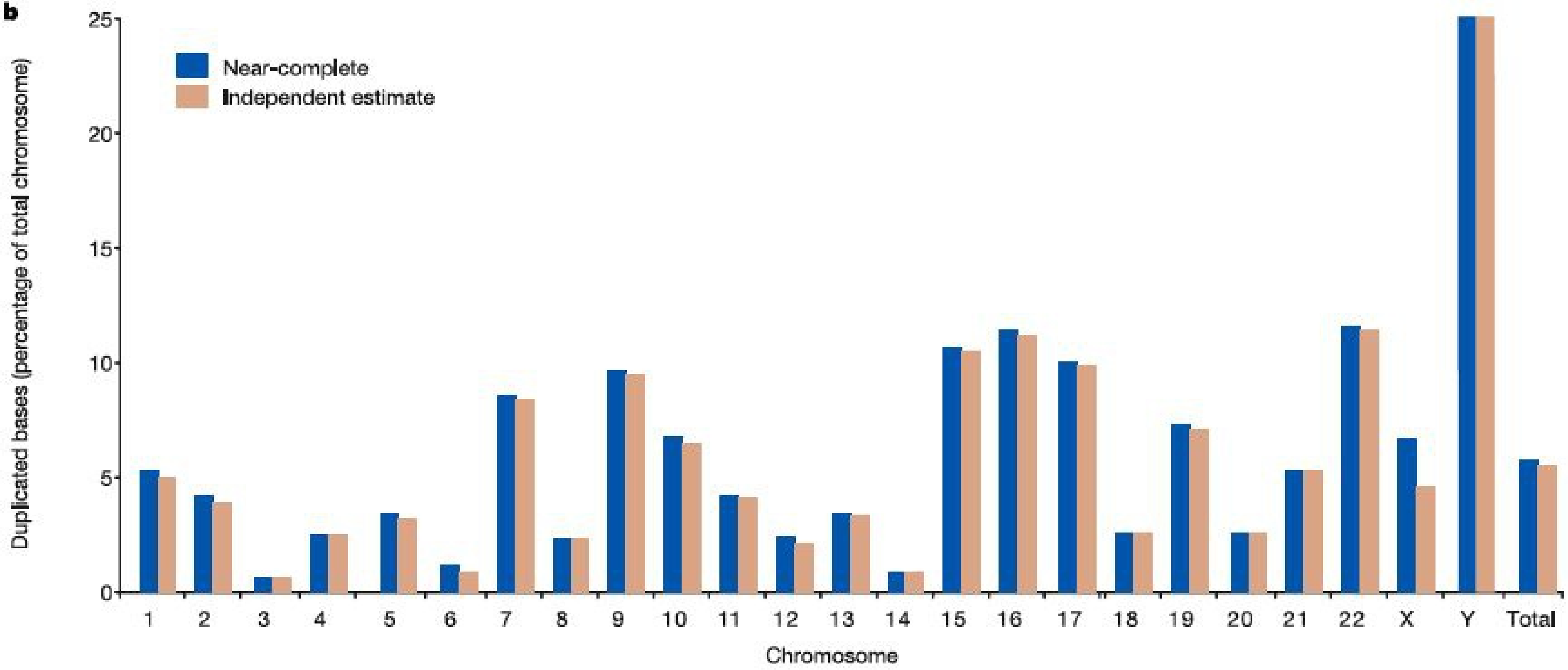


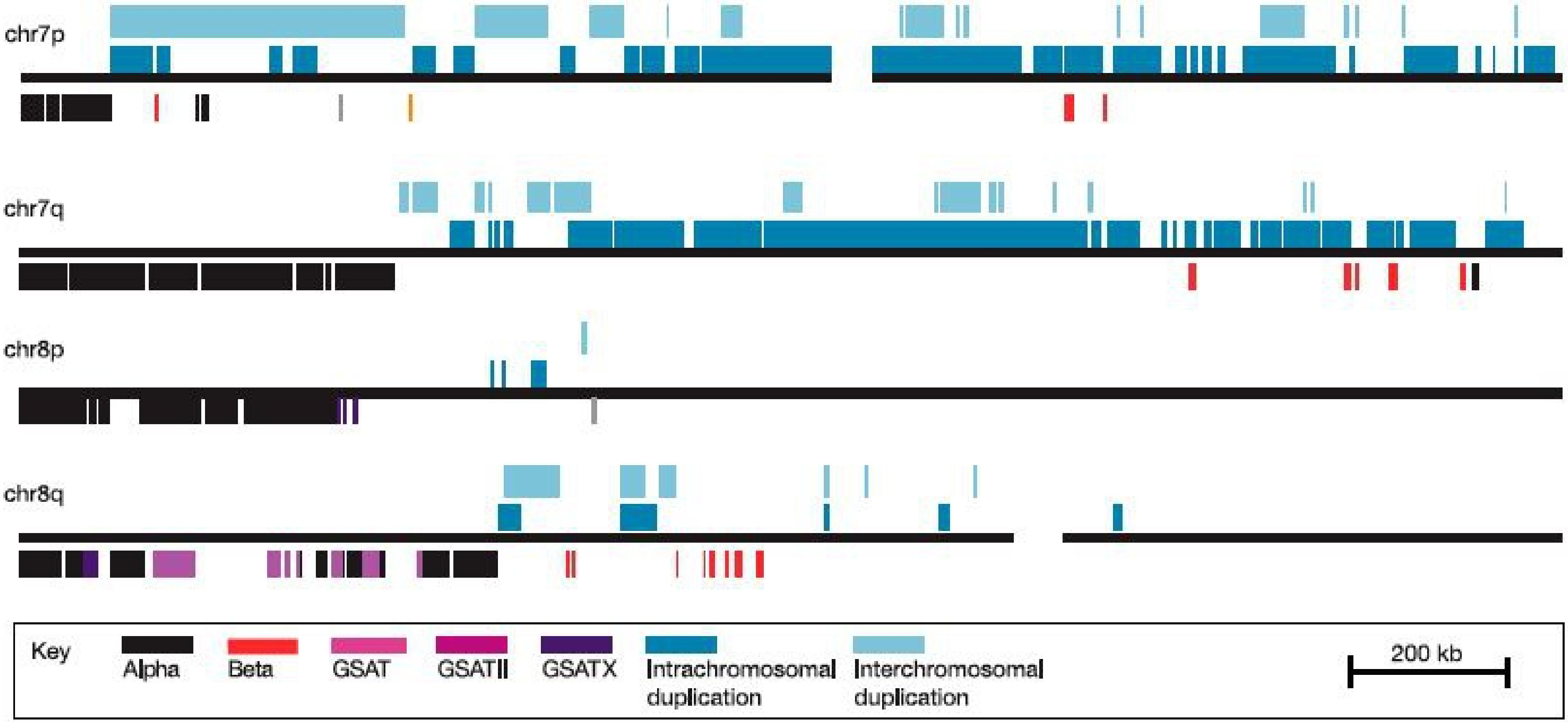
Figure 4 Segmental duplications across the genome. **a**, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length ≥10 kb and sequence identity ≥95%). Large duplications are shown to approximate scale; smaller ones are indicated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

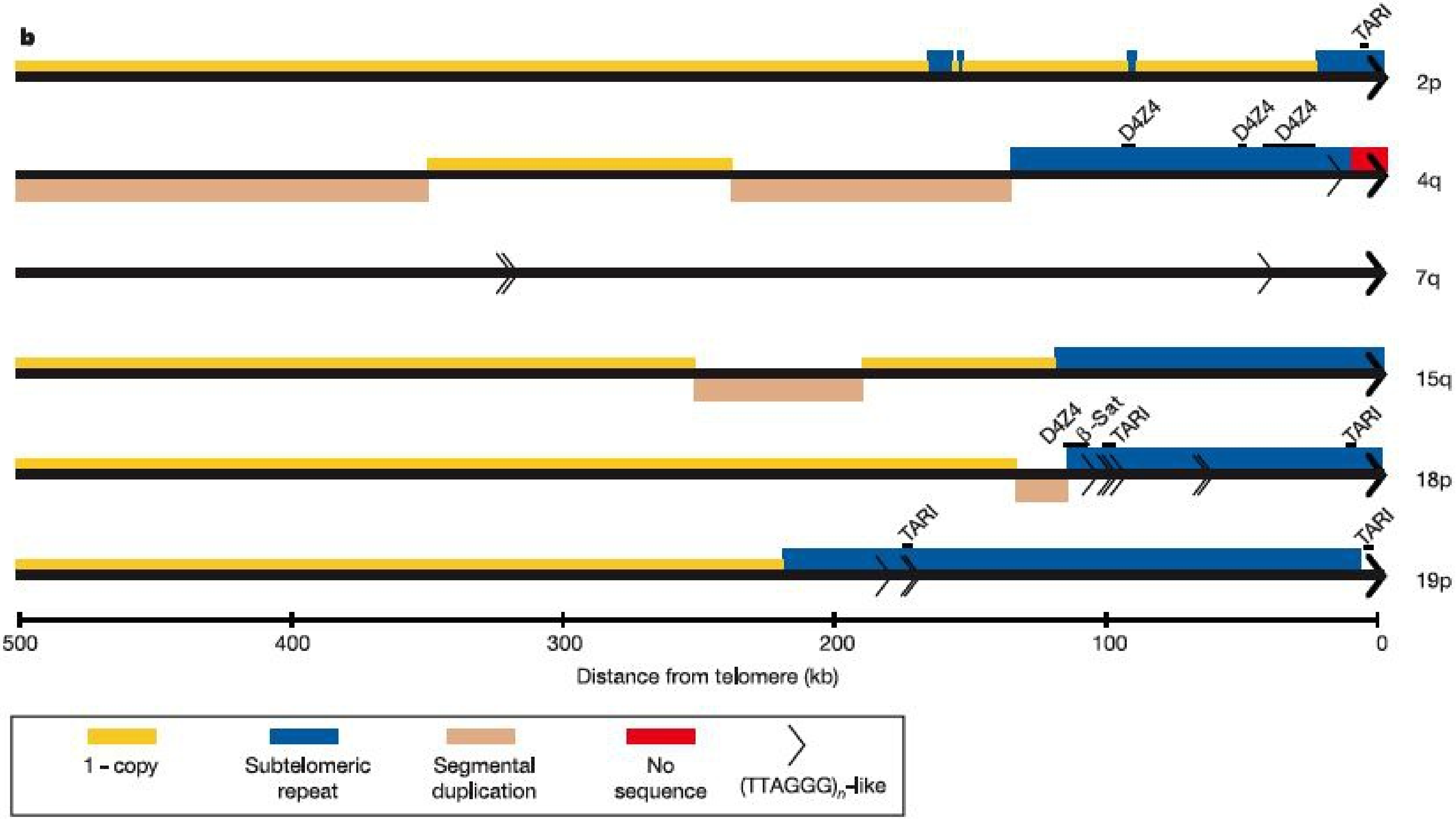


© Doug Brutlag 2015

Duplications near Centromeres

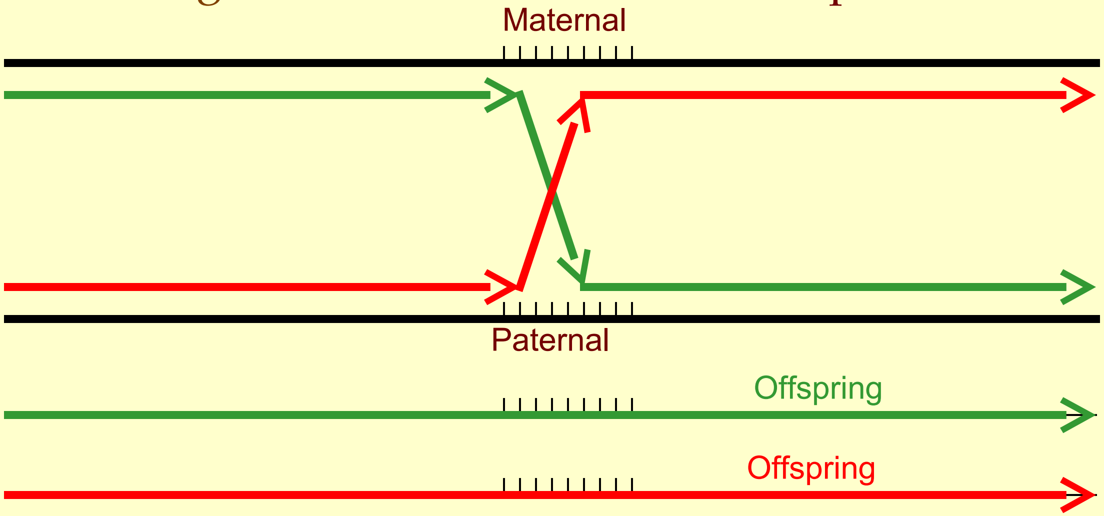


Duplications near Telomeres

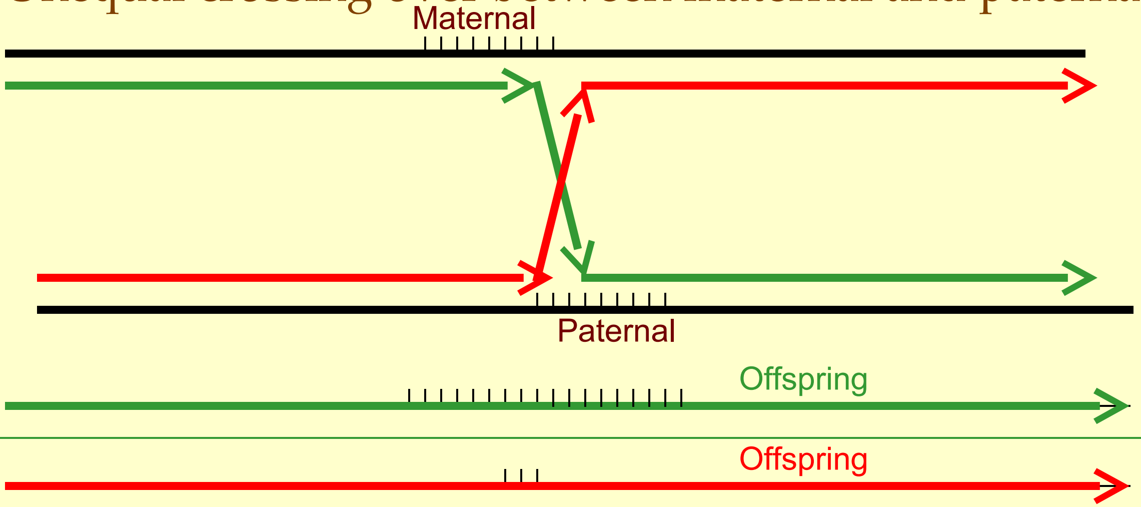


Deletions and Duplications can Arise from Unequal Crossing Over in Repeated Regions

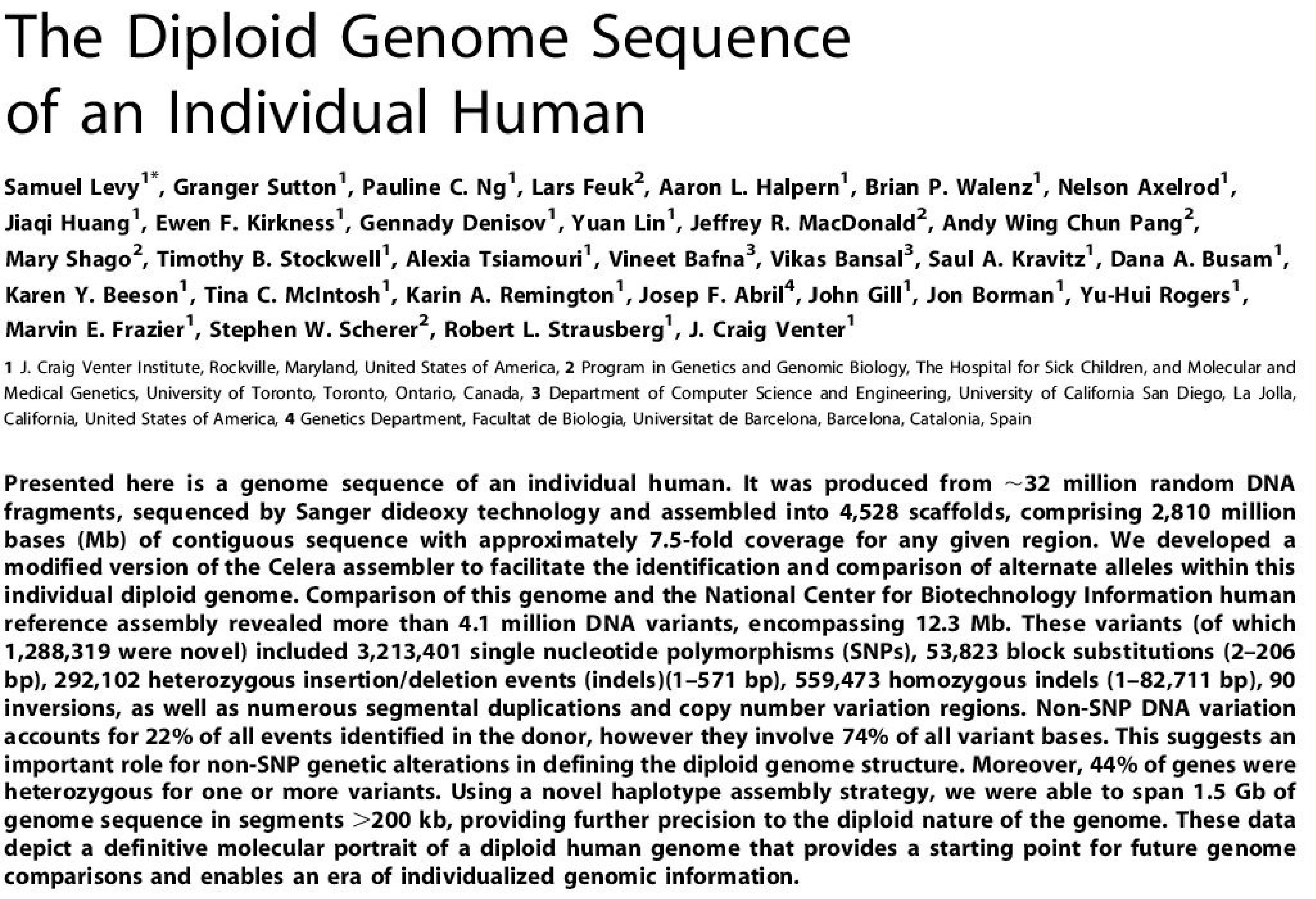
- Crossing over between **maternal** and **paternal** chromosomes



- Unequal crossing over between maternal and paternal chromosomes



The Diploid Sequence of an Individual Human (HuRef)



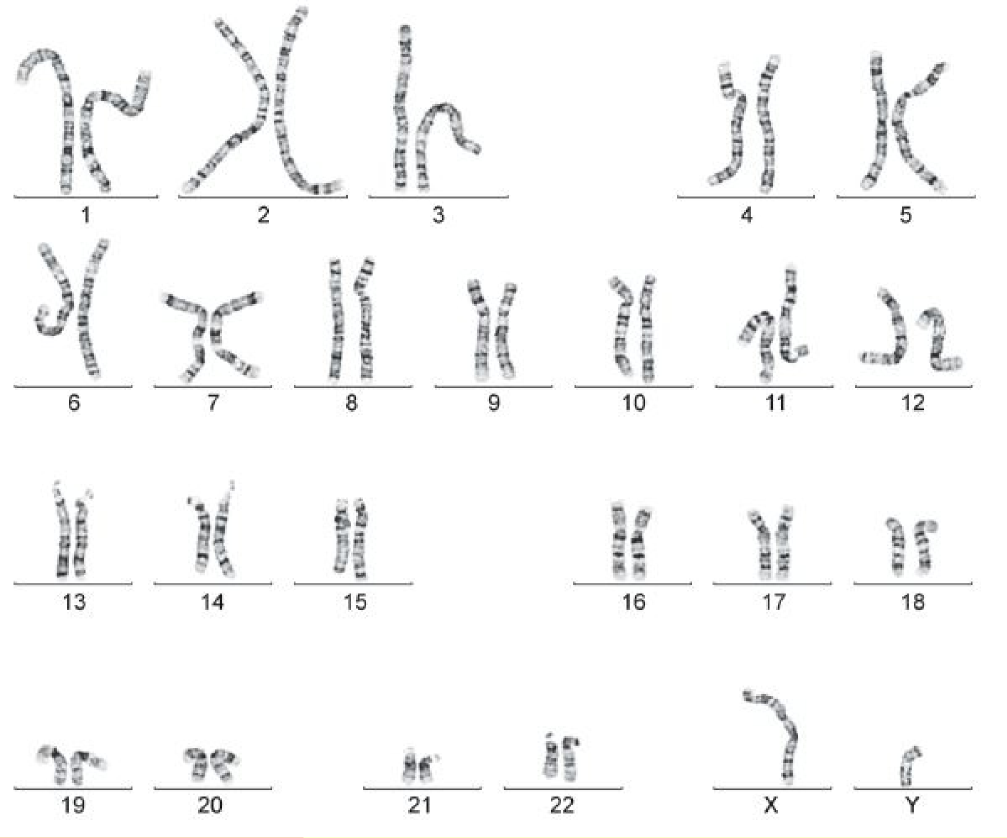
The Diploid Genome Sequence of an Individual Human

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]

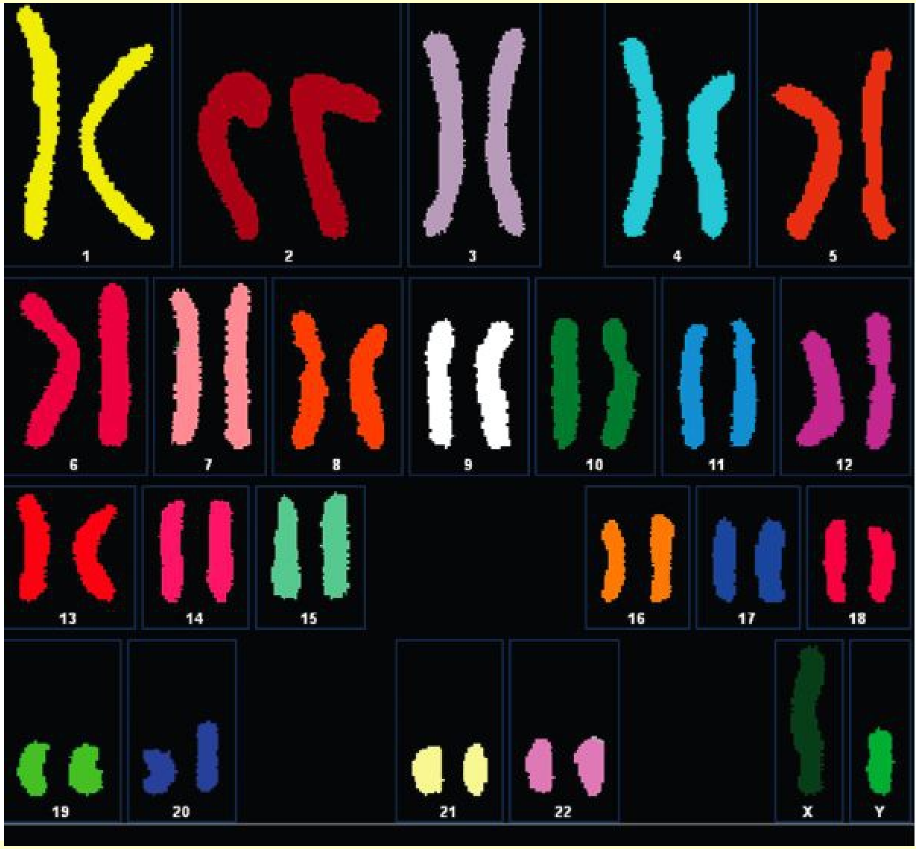
1 J. Craig Venter Institute, Rockville, Maryland, United States of America, **2** Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, **3** Department of Computer Science and Engineering, University 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.

Karyotype of J.Craig Venter



Giemsa Stain



FISH Stain

Comparing NCBI Assembly to HuRef Assembly

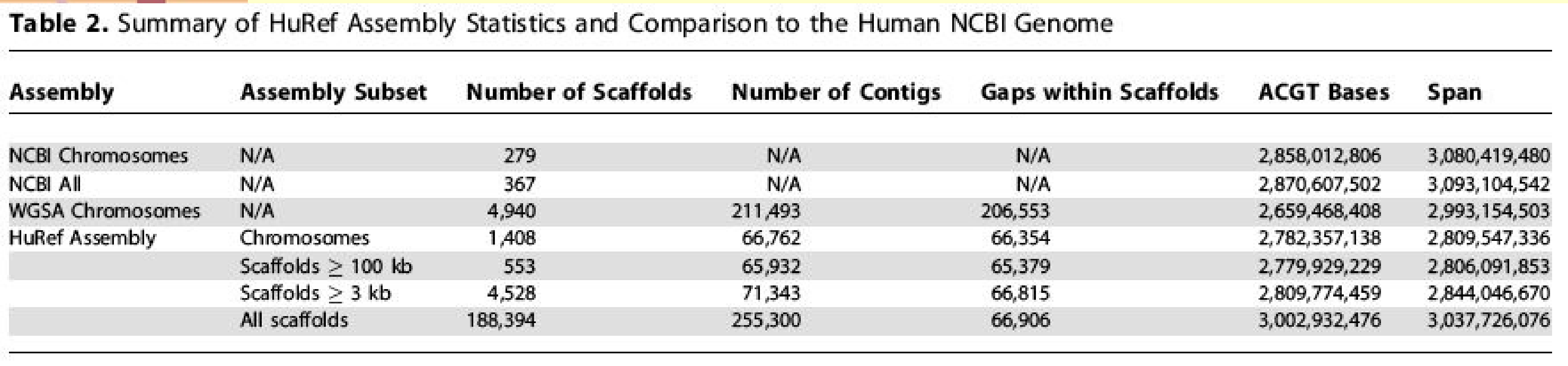
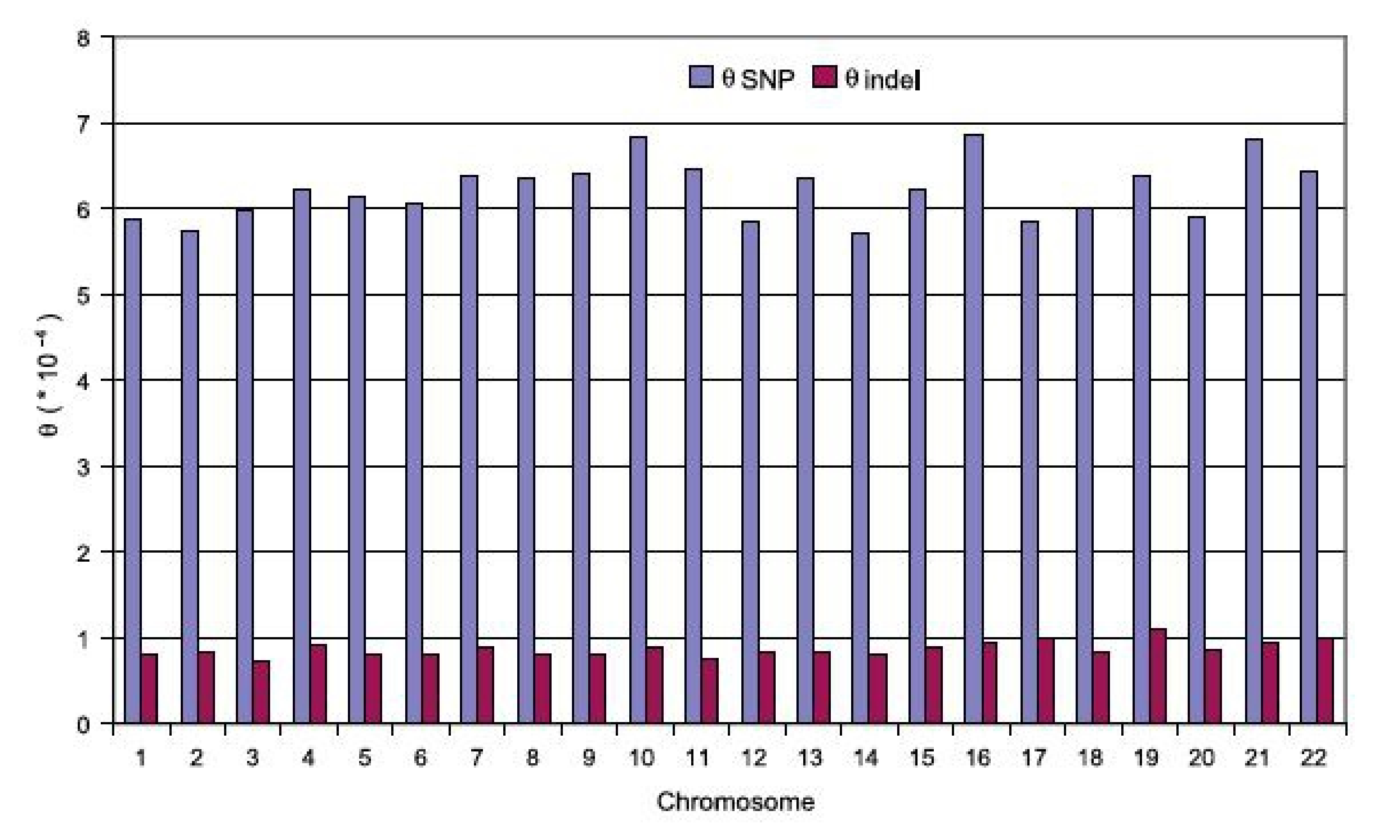


Table 2. Summary of HuRef Assembly Statistics and Comparison to the Human NCBI Genome

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

SNPs & InDels in HuRef Autosomes



© Doug Brutlag 2015