MI615 Syllabus
Illustrated Topics in
Advanced Molecular Genetics
Spring 2009: CTW405 TR 9:30-10:50

DATE	TITLE	LECTURER
Thu Jan 15	**Introduction, Genomic low copy repeats**	Pierce
Tue Jan 20	**Human repetitive DNA characteristics**	Pierce
Thu Jan 22	**Gene Assembly and Library Construction**	Bradley
Tue Jan 27	Part II	Bradley
Thu Jan 29	**DNA Repair**	Gu
Tue Feb 3	Part II	Gu
Thu Feb 5	**Genomic Rescue: biochemistry**	Pierce
Tue Feb 10	Part II: **genetics**	Pierce
Thu Feb 12	**Mouse Genetics**	Spear
Tue Feb 17	Part II	Spear
Thu Feb 19	**Gene Silencing**	Lutz
Tue Feb 24	Part II	Lutz
Thu Feb 26	**Regulation of Inducible Gene Transcription**	Kaetzel
Tue Mar 3	Part II	Kaetzel
Thu Mar 5	Part III	Kaetzel
Tue Mar 10	**RNA processing**	Peterson
Thu Mar 12	Part II	Peterson
	Spring Break	
Tue Mar 24	Part III	Peterson
Thu Mar 26	Part IV	Peterson
Tue Mar 31	**Asexual organisms**	Pierce
Thu Apr 2	Part II	Pierce
Tue Apr 7	**Signaling Pathways of Viral Recognition**	Bruno
Thu Apr 9	Part II	Bruno
Tue Apr 14	**Viral oncogenesis**	Luo
Thu Apr 16	**RNAi**	Luo
Tue Apr 21	**Prions and prion diseases**	Telling
Thu Apr 23	Part II	Telling
Tue Apr 28	**Protein splicing**	Pierce
Thu Apr 30	Part II	Pierce

Course director:
Andrew Pierce
207 Combs
andrew.pierce@uky.edu
323-1455

Course calendar:
http://ical.mac.com/ajpierce/MI615

Course syllabus:
http://lectures.paralog.com/MI615.htm

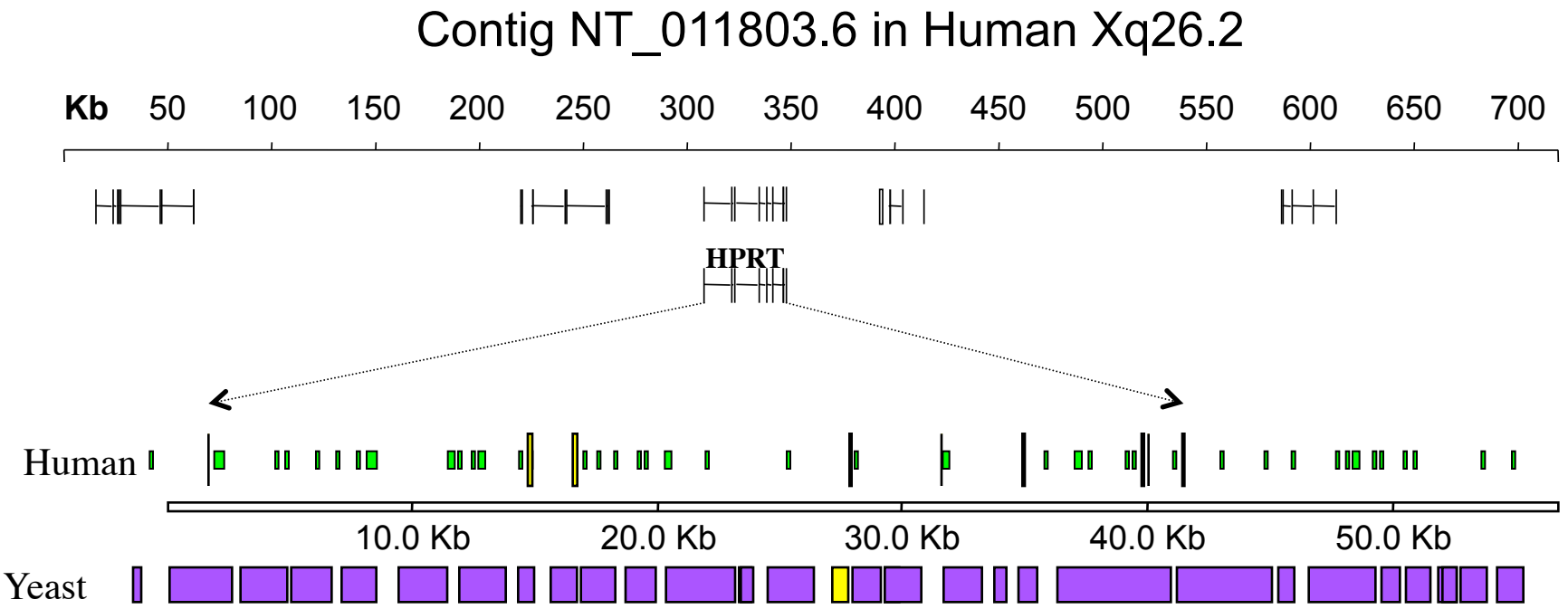
Low Copy Repeats in the Human Genome
Implications for Genomic Structure

Andrew J. Pierce

Microbiology, Immunology and Molecular Genetics
Graduate Center for Toxicology
Markey Cancer Center
University of Kentucky

MI615

Genomic Structure:
"Empty Space" and High Copy Repeats



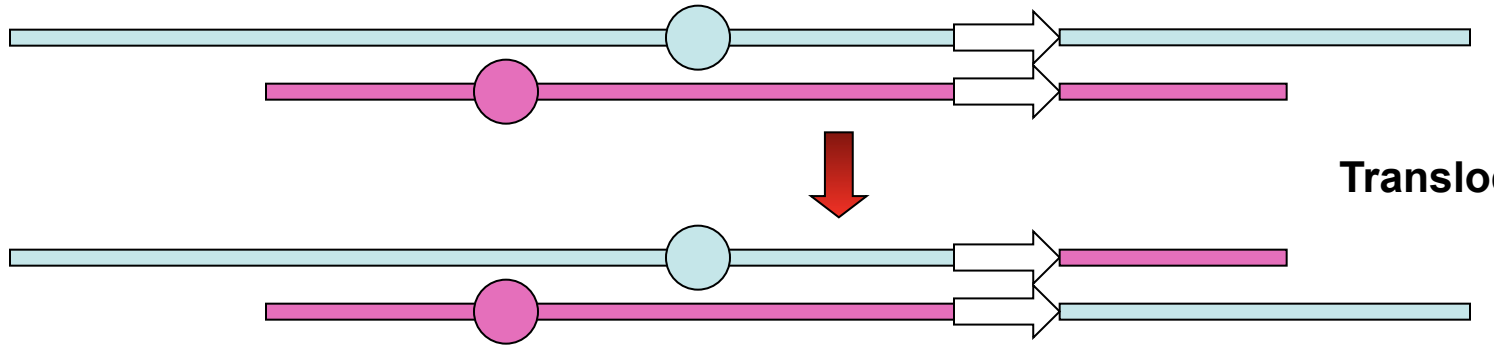
Low Copy Repeats

- 10 - 500 kb in size
- > 95% sequence identity
- usually near centromeres or telomeres
- not detectable by reassociation kinetics
  contrast with Alu-elements, LINEs, retrotransposons, satellite DNA
- problematic for sequencing purposes when longer than BAC size (150 - 200 kb)
- also called “Segmental Duplications” or “Paralogous Repeats” when locus-specific (typically > 97% sequence identity)

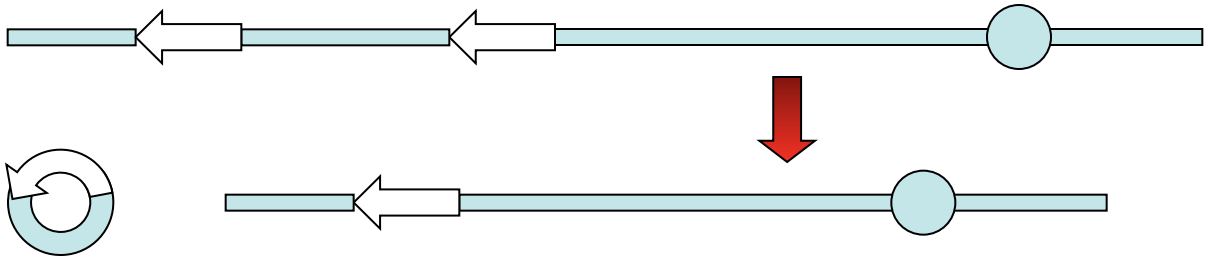
- susceptible to Non-Allelic Homologous Recombination (NAHR)
- NAHR leads to translocations, inversions and deletions

Stankiewicz P, Lupski JR. **Genome architecture, rearrangements and genomic disorders.** Trends Genet. 2002 Feb;18(2):74-82.

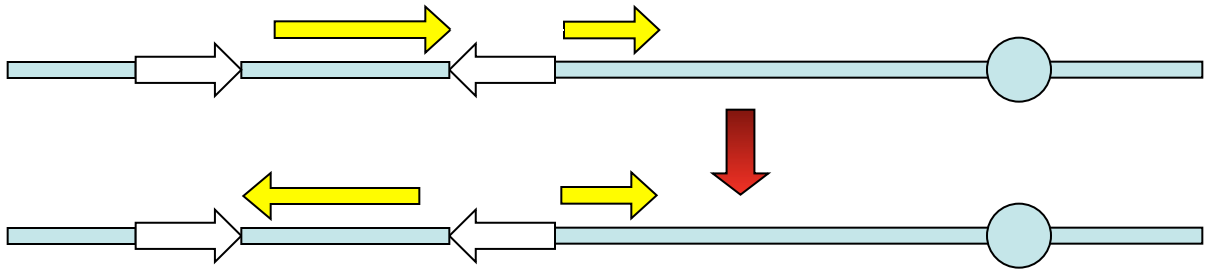
NAHR Can Cause Large-scale Genomic Rearrangements



Translocation

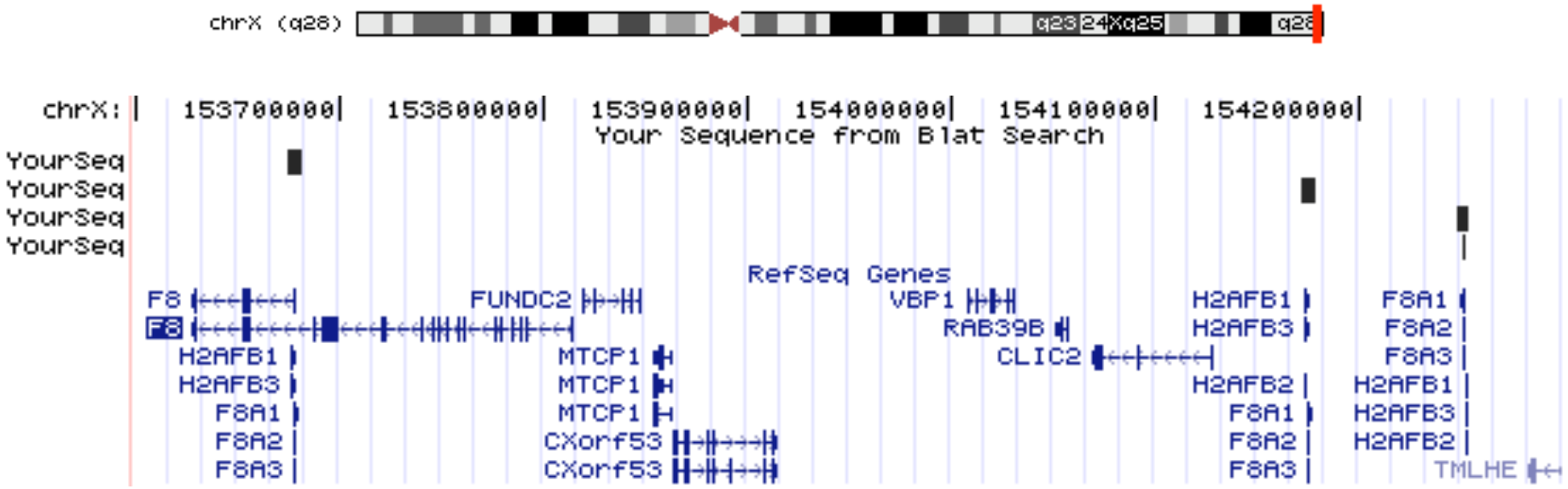


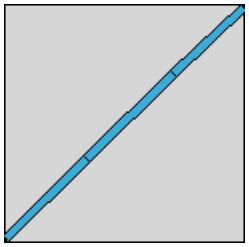
Deletion



Inversion

LCR's in Factor VIII (Xq28)



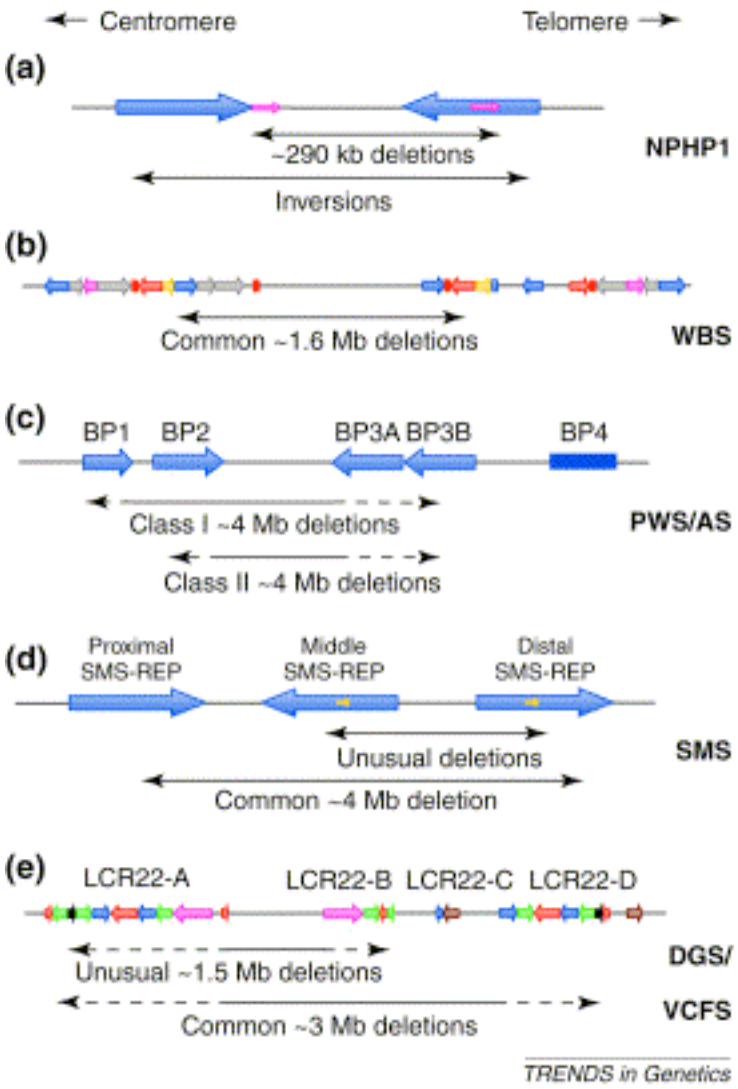


SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
6667	1	6667	6667	100.0%	X	+	153675667	153682333	6667
6603	1	6640	6667	99.8%	X	+	154172778	154179412	6635
6544	1	6640	6667	99.8%	X	-	154248016	154254654	6639

```
Query: 2281 g---gggggggggacaacagtctagaaatgcgcagacatggaattagggtcacgagttt 2337
            |   |||||||||||||||||| ||||||||||||||||||||||||||||||||||
Sbjct: 2381 gggggggggggggacaacagtccagaaatgcgcagacatggaattagggtcacgagttt 2440

Query: 4679 tatttgtaatagggatttttttttttttttttttttttttcccccgagatggagtctc 4738
            ||||||||||||||   |||||||||||||||||||  ||||||||||||||||||
Sbjct: 4781 tatttgtaataggga---ttttttttttttttttttttcccccccgagatggagtctc 4837
```

Some Genomic Disorders Mediated by LCR's



Complex structure of selected low-copy repeats (LCRs). Horizontal lines represent specific genomic regions with the centromere toward the left and telomere to the right. At the right are listed abbreviations for the disease manifested through common deletions of the regions. The colored regions refer to LCRs with the orientation given by the arrowhead. Note complex structure of LCRs consisting of both direct and inverted repeats. (a) LCRs in chromosome 2q13 responsible for rearrangements associated with familial juvenile nephronophthisis 1 (NPHP1). (b) LCRs7 flanking the Williams–Beuren syndrome (WBS) chromosome region 7q11.23. (c) LCRs15 within the Prader–Willi syndrome/Angelman syndrome (PWS/AS) chromosome region 15q11.2. (d) Smith–Magenis syndrome (SMS) repeats within 17p11.2. (e) LCRs22 within the DiGeorge syndrome (DGS) chromosome 22q11.2.

Stankiewicz P, Lupski JR. **Genome architecture, rearrangements and genomic disorders.** Trends Genet. 2002 Feb;18(2):74-82.

Low Copy Repeats: Directed Cloning/Sequencing vs Shotgun Approaches

Table 1 Comparison of segmental duplication within two human genome assemblies

Chromosome	Build34 assembly			WGSA		
	Length (bp)	Duplication (bp)	Fraction	Length (bp)	Duplication (bp)	Fraction
1	221,562,941	11,553,369	0.0521	209,662,503	2,629,537	0.0125
2	237,541,603	10,000,492	0.0421	223,960,456	2,034,342	0.0091
3	194,473,779	3,299,552	0.0170	189,481,828	2,626,848	0.0139
4	186,841,959	4,287,299	0.0229	180,981,699	2,899,296	0.0160
5	177,552,822	5,956,951	0.0336	170,281,266	1,351,471	0.0079
6	167,256,575	3,600,793	0.0215	161,428,330	1,660,626	0.0103
7	154,676,518	13,096,209	0.0847	144,247,908	5,496,523	0.0381
8	142,347,919	3,250,852	0.0228	136,878,554	1,013,574	0.0074
9	115,624,042	11,096,428	0.0960	104,630,165	1,826,794	0.0175
10	131,173,206	8,937,553	0.0681	122,948,635	3,379,280	0.0275
11	130,908,854	5,535,297	0.0423	126,253,176	4,007,704	0.0317
12	129,826,277	2,922,438	0.0225	125,900,476	2,050,906	0.0163
13	95,559,980	3,212,091	0.0336	92,484,206	1,953,930	0.0211
14	87,191,216	1,587,527	0.0182	84,198,821	951,062	0.0113
15	81,259,656	8,577,567	0.1056	74,059,970	2,599,187	0.0351
16	79,932,429	9,124,179	0.1141	66,369,068	994,790	0.0150
17	77,677,744	7,746,457	0.0997	73,627,628	3,657,946	0.0497
18	74,654,041	1,898,132	0.0254	71,253,215	370,517	0.0052
19	55,785,651	4,051,295	0.0726	51,679,110	2,835,683	0.0549
20	59,424,990	1,479,847	0.0249	57,238,069	963,199	0.0168
21	33,924,307	1,791,042	0.0528	31,584,736	410,918	0.0130
22	34,352,051	3,982,963	0.1159	31,357,605	1,590,197	0.0507
X	149,215,391	10,057,692	0.0674	121,809,144	2,105,297	0.0173
Y	24,649,555	12,745,541	0.5171	7,151,840	728,694	0.1019
Unplaced	2,592,022	980,700	0.3784	36,146,472	10,186,469	0.2818
Total	2,865,069,170	150,772,266	0.0530	2,695,614,880	60,324,790	0.0224

Segmental duplications (>90% sequence identity and >1 kb length) were calculated using the whole-genome assembly comparison method[10] for the finished human genome assembly (July 2003) and the whole-genome shotgun sequence assembly (WGSA)[8]. Due to the fragmentation of duplications within the WGSA, duplicated bases were calculated without welding across gaps in the assembly. Totals do not include gaps or centromeric/acrocentric regions of chromosomes. Both assemblies were compared using exactly the same parameters. The unplaced chromosome contains the largest proportion of WGSA duplicated sequence—28.2% (10.2 Mb based on the analysis of WGSA). Of the 21.8 Mb that could be mapped back to build34, we found that 9.2 Mb (42.3%) corresponded to duplications within our segmental duplication database.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE.
Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Low Copy Repeats: Directed Cloning/Sequencing vs Shotgun Approaches

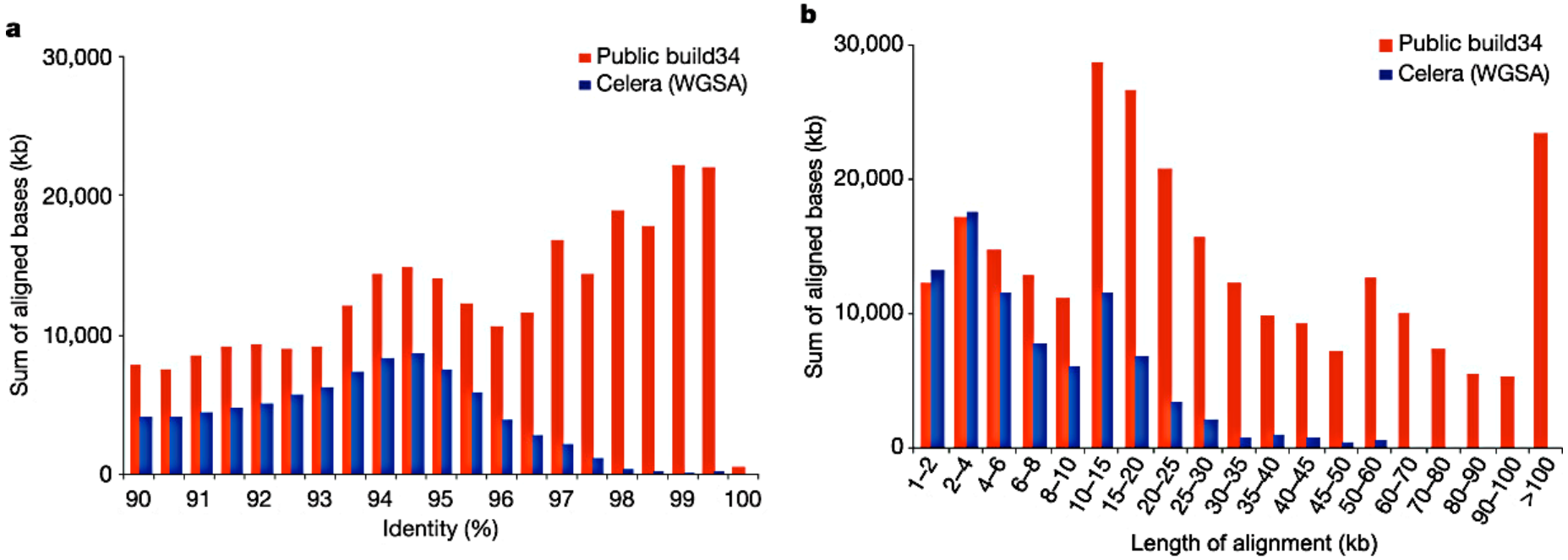
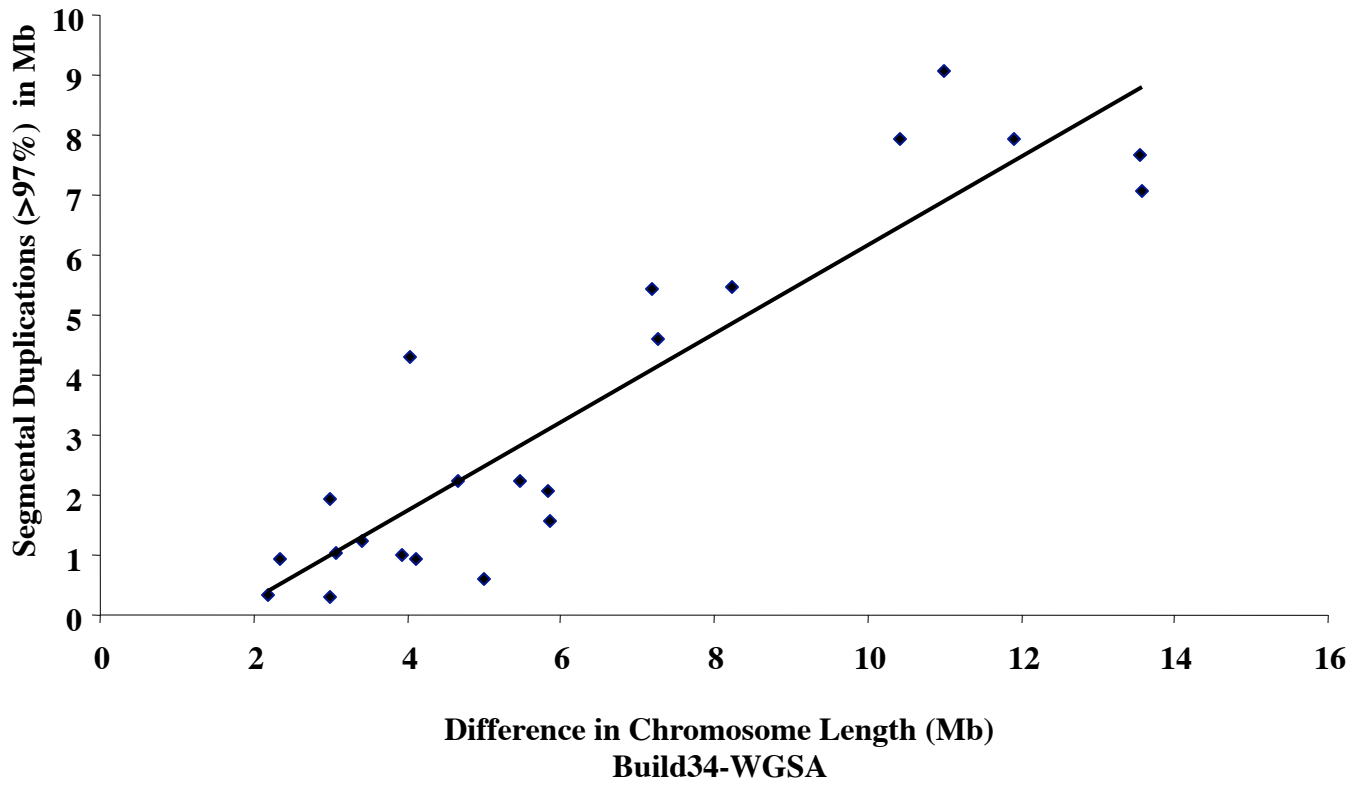


Figure 1 Sequence identity and alignment length of segmental duplications. **a**, **b**, All duplication alignments between 90–100% were categorized based on sequence identity (**a**) (0.5% bins) and the alignment length (**b**). The sum of aligned base pairs for each bin is compared between WGSA and build34 human genome sequence assemblies. The proportion of WGSA aligned base pairs begins to decline most rapidly as the sequence identity exceeds 96–97% and the length of the alignments exceeds 15 kb. Note that the reduction in WGSA alignments below 96% is probably due to the fact that divergent duplications are frequently part of larger alignments where the degree of sequence identity is higher. As highly identical alignments are lost, the embedded, more divergent pairwise alignments are also eliminated from further consideration

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE.
Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Low Copy Repeats: Directed Cloning/Sequencing vs Shotgun Approaches



Chromosome Length vs. Duplication. The difference in chromosome length (Build34-WGSA) was compared to the amount of non-redundant duplicated bases that were part of alignments >97% sequence identity. Only autosomes were considered in this analysis. A strong correlation (r^2=0.83) is observed between highly identical segmental duplications and reduced chromosome length.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE.
Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Low Copy Repeats:
Directed Cloning/Sequencing vs Shotgun Approaches

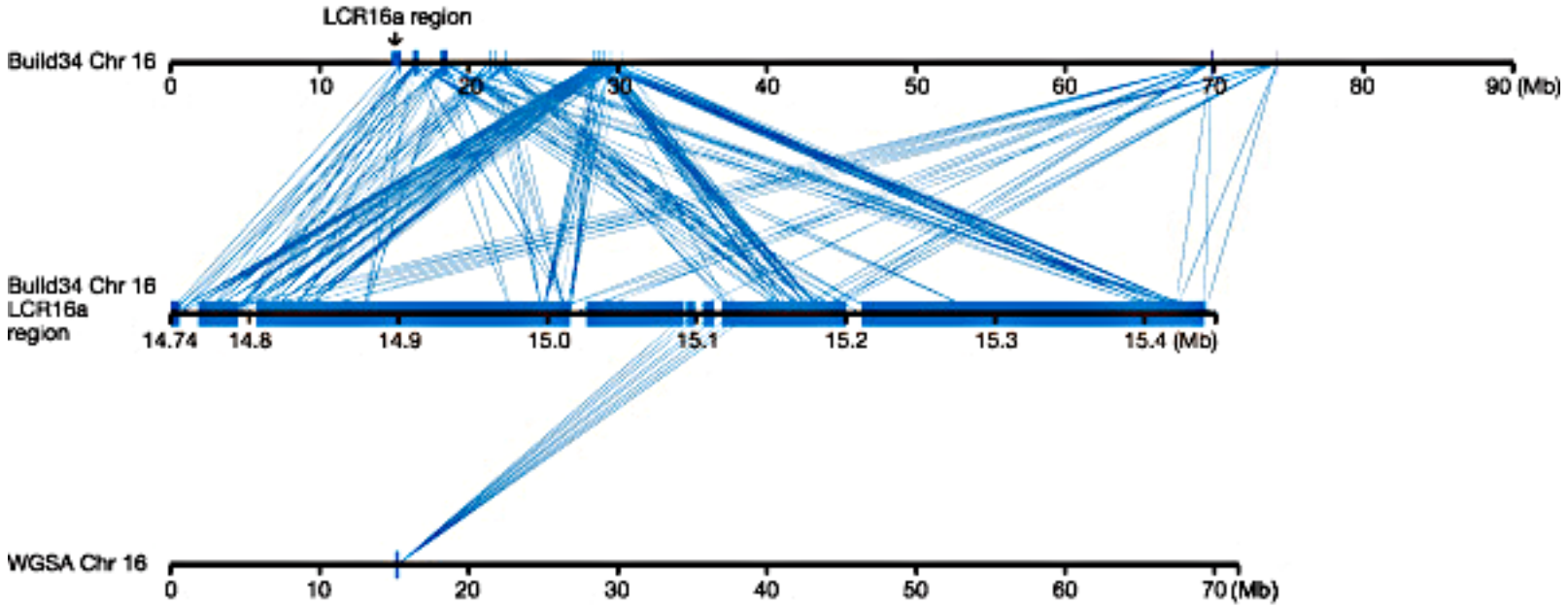


Figure 2 Distribution of LCR16a duplications in two assemblies. The pattern of duplication alignments for one 690-kb region of low-copy-repeat duplications on chromosome 16 is shown between the build34 and WGSA human genome assemblies. The entire region is duplicated to 28 distinct regions within build34 (locations have been experimentally verified) whereas only a small portion (46 kb) maps to a single location on WGSA chromosome 16

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE.
Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Low Copy Repeats by Chromosome

Human Segmental Duplication Content

Build35	Initial		Filtered		
Chrom	bp	percent (%)	bp	percent (%)	ChromSizeNoN
chr1	10,179,722	4.57	9766480	4.38	222,827,847
chr2	9,748,188	4.10	9424377	3.97	237,503,374
chr3	3,308,512	1.70	3193712	1.64	194,635,738
chr4	4,984,316	2.66	4462699	2.38	187,161,218
chr5	5,899,394	3.32	5865302	3.30	177,702,766
chr6	3,466,813	2.07	3096489	1.85	167,317,698
chr7	13,166,147	8.51	13037077	8.42	154,759,139
chr8	3,040,606	2.13	3000604	2.10	142,612,826
chr9	12,190,627	10.35	12113582	10.28	117,781,268
chr10	8,971,582	6.82	8943396	6.80	131,613,619
chr11	5,549,488	4.23	5362293	4.09	131,130,753
chr12	2,962,789	2.27	2733942	2.10	130,259,309
chr13	3,045,551	3.19	3009165	3.15	95,559,980
chr14	2,703,108	3.06	2666234	3.02	88,290,585
chr15	8,152,323	10.02	8140842	10.01	81,341,915
chr16	7,849,791	9.95	7814348	9.91	78,884,752
chr17	7,160,701	9.20	7066136	9.08	77,800,220
chr18	1,923,415	2.58	1875293	2.51	74,656,155
chr19	4,086,749	7.33	4007270	7.18	55,785,651
chr20	1,480,315	2.49	1465008	2.46	59,505,253
chr21	1,852,333	5.42	1848943	5.41	34,170,106
chr22	4,133,956	11.89	4099943	11.79	34,764,789
chrX	10,493,957	6.98	7220018	4.80	150,394,264
chrY	12,531,772	50.39	8910503	35.83	24,871,691
TOTAL	154,005,734	5.37	143,794,977	5.02	2,865,798,592

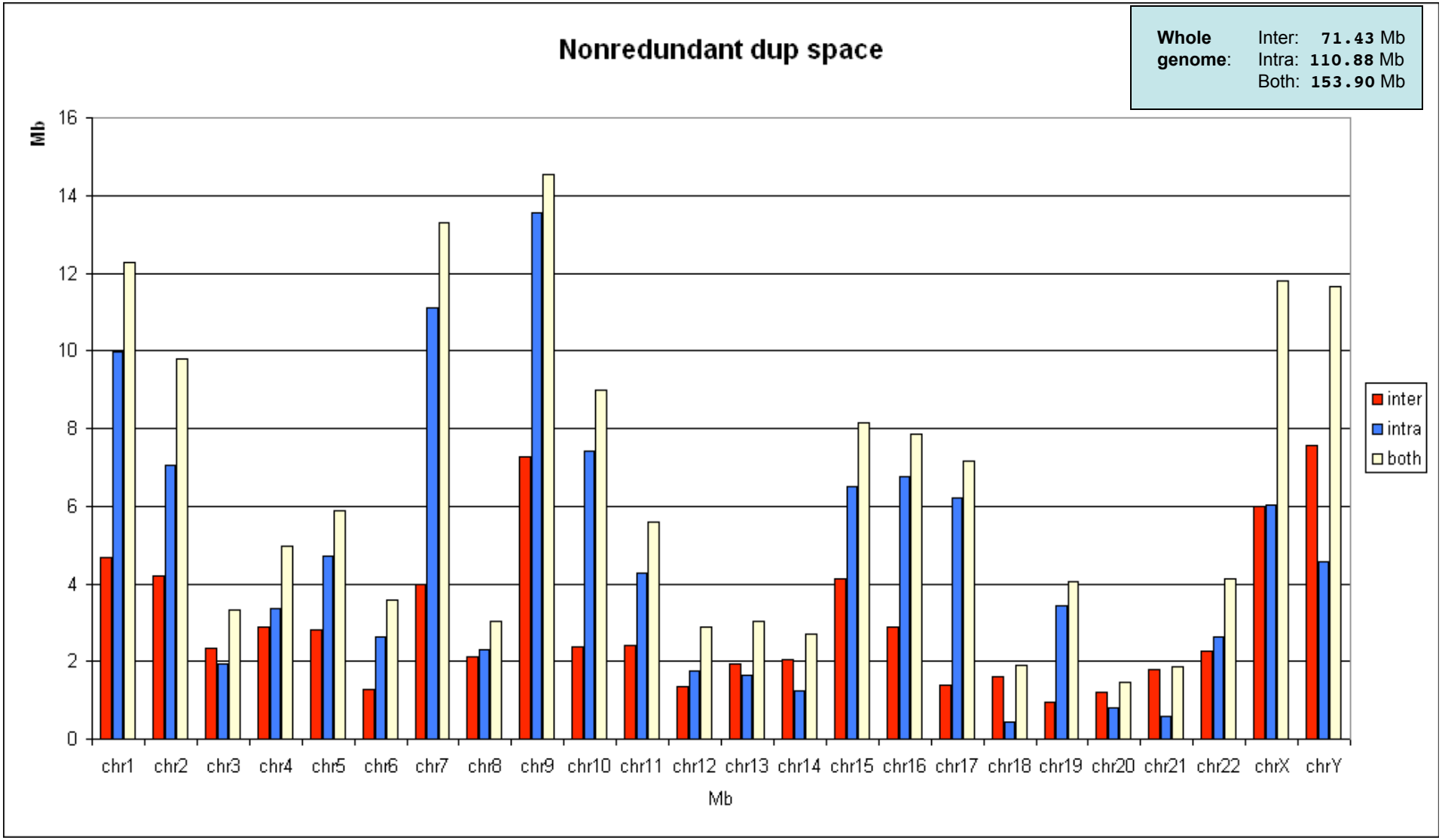
She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE.
Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Sequencing Human Disease Loci Involving Low Copy Repeats

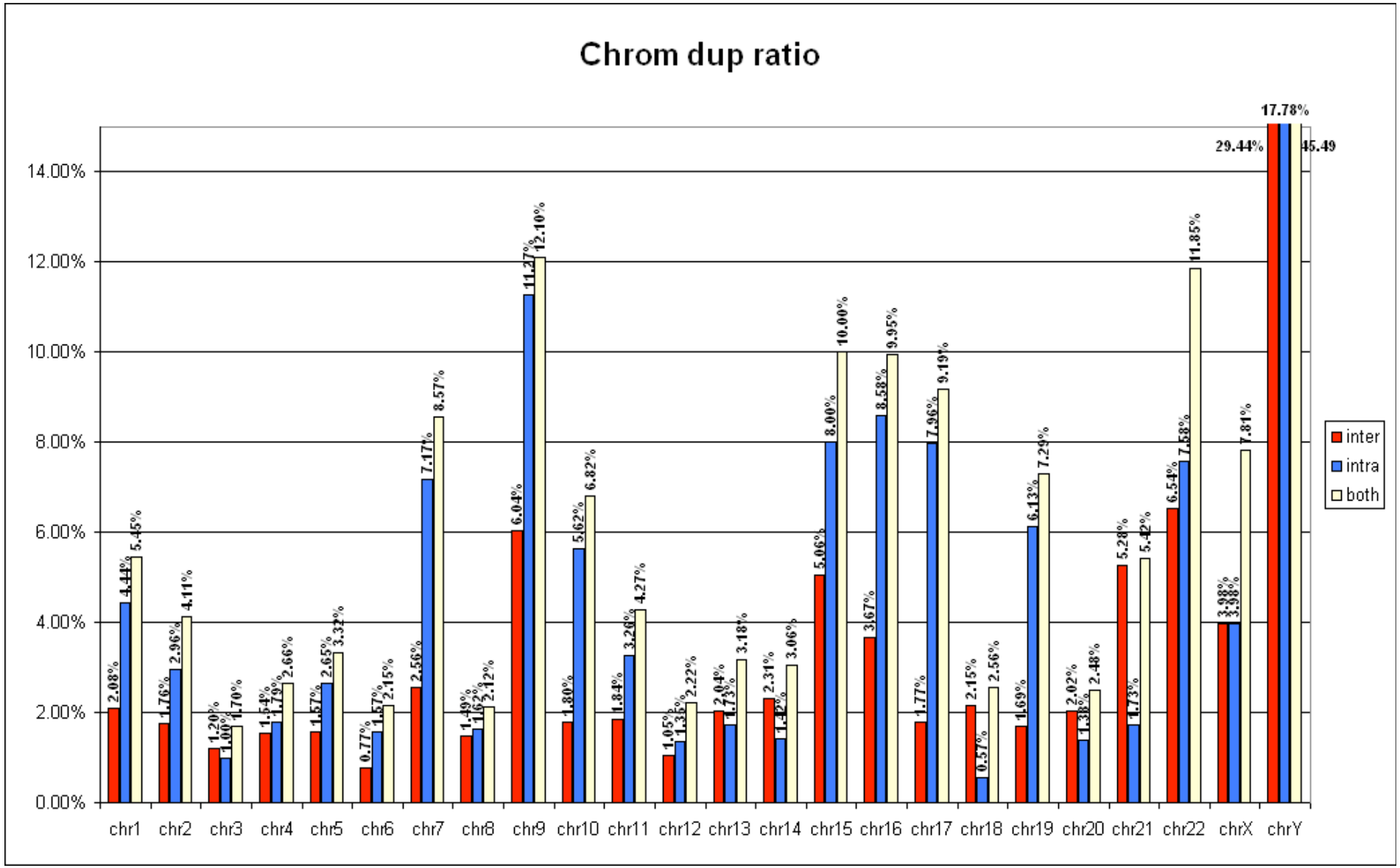
Supplementary Figure 1. Duplication in disease breakpoint regions
Five disease breakpoint regions: spinal muscular atrophy type I (SMA), Williams-Beuren syndrome (WBS), Charcot-Marie-Tooth disease (CMT1A), Prader-Willi Syndrome (PWS) and velo-cardiofacial/DiGeorge Syndrome (VCS/DG) are show in build34 genome browser view. The segmental duplication tracks show the extent of segmental duplication. Corresponding one to one mapping of WGSA on build34 is shown (blue track). 71-97% of the sequences corresponding to these large segmental duplications was absent in WGSA.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE.
Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Human Genomic Duplications >90% Sequence Identity
Total Repeated Sequence Distribution by Chromosome

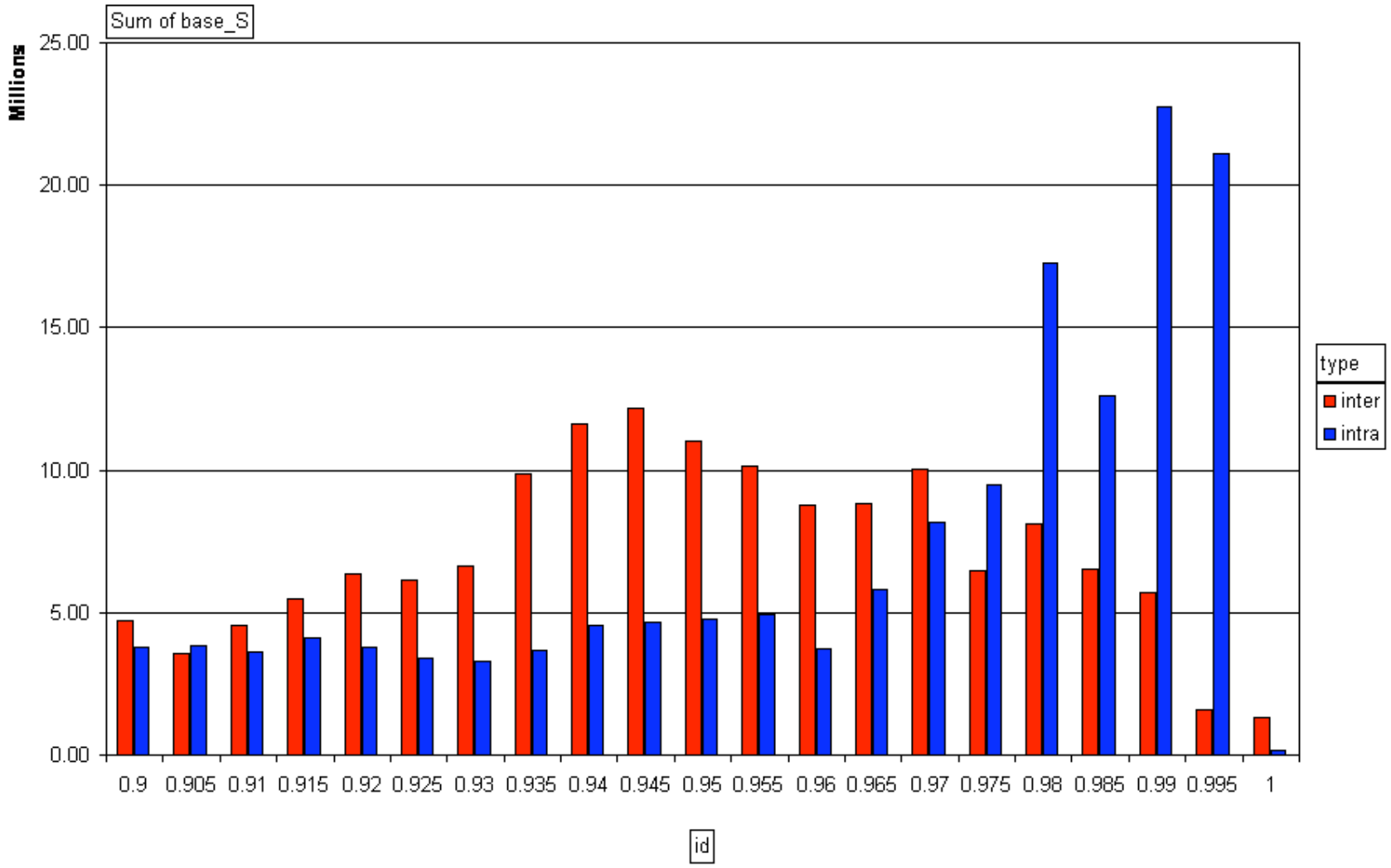


http://humanparalogy.gs.washington.edu

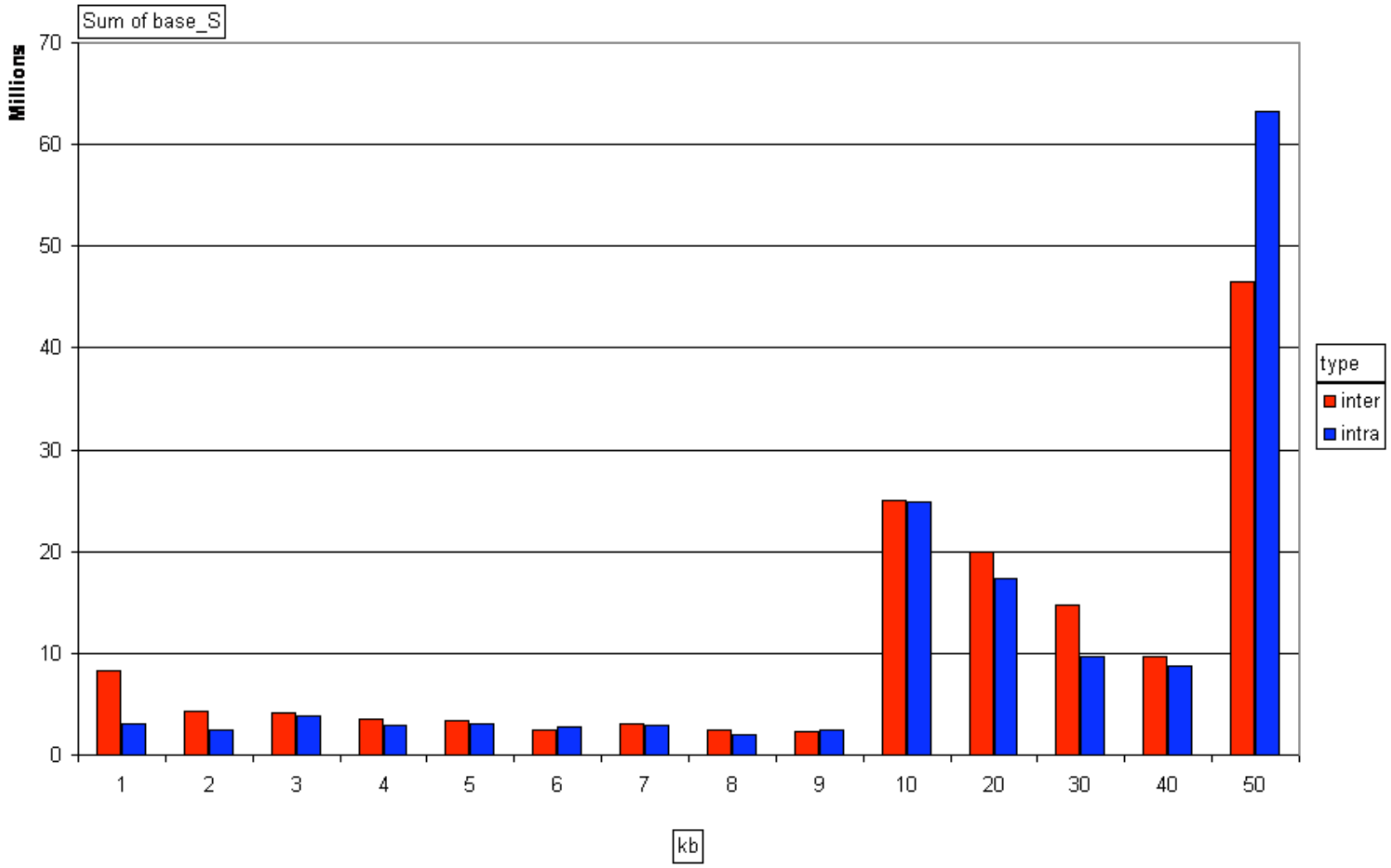


http://humanparalogy.gs.washington.edu

Human Genomic Duplications >90% Sequence Identity
Distribution by Percent Sequence Identity



Human Genomic Duplications >90% Sequence Identity Distribution by Length of Repeat



http://humanparalogy.gs.washington.edu

Human Genomic Project – Not Finished Yet