Supplementary Online Material for Wall et al. “A novel DNA sequence database for analyzing human demographic history”

Resequencing pipeline

Sequence finishing and polymorphism detection made use of a customized version of the *Phred /Phrap/Consed/PolyPhred* suite (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998; Nickerson et al. 1997). In brief, all reads for a locus were aligned to a reference sequence (the hg17 [NCBI build 35] version of the human genome), and polymorphic sites were tagged by Polyphred. A handful of scripts were used to create individual alignments, on which the majority of sequence finishing occurred. Custom navigation files were used to visit all sites in an individual that vary from the reference sequence, as well as all sites tagged by Polyphred that showed evidence of heterozygosity. To catch sites that may have been missed by Polyphred, all positions that had an average Phred quality <50 were also visited. The final individual contig was given one last pass ‘by eye’ to check for gaps in coverage and heterozygous indels, and so that the edges of the sequence could be trimmed. Individual failures were marked and revisited when new reads became available.

Finished human and outgroup contigs were then multiply-aligned (either via Sequencher, Bioedit, or the Perl scripts msa or consed2Pasta, see below) and tables of polymorphism were constructed. These tables were reviewed by finishers who looked for indel mis-alignments, homozygous doubletons, and 4-gamete violations. Once polymorphisms tables were double-checked, finishing at the individual level was evaluated *en masse*. An independent finisher double-checked all edits for all reads that were assembled. At this stage, the finisher was looking for sites tagged by Polyphred that were not tagged by the original finisher, and other missed variation, especially commonly missed heterozygous indels. Furthermore, sites that were tagged as being putatively polymorphic at the individual level were viewed at the level of the locus trio, which allowed the second finisher to easily discriminate between ‘true’ polymorphism and sequencing artifacts. Each locus subset was deemed complete if it were 99% finished (i.e., human sequences had ≤ 1% bases called as ‘N’) or it had gone through three rounds of further attempts to obtain good sequence data (e.g., re-PCR, primer re-design, etc.). Gaps in the outgroup sequence were supplemented with the 2003 chimp genome. A rough schematic of our protocol is shown in **Supplementary Figure 1**.

Supplementary Table 1. List of main DNA samples used.

Population	ID	Population	ID
French Basque	HGDP01357	Melanesian	HGDP00825
French Basque	HGDP01358	Melanesian	HGDP00978
French Basque	HGDP01359	Melanesian	HGDP01027
French Basque	HGDP01360	Biaka	HGDP00451
French Basque	HGDP01361	Biaka	HGDP00452
French Basque	HGDP01362	Biaka	HGDP00454
French Basque	HGDP01364	Biaka	HGDP00455
French Basque	HGDP01370	Biaka	HGDP00457
French Basque	HGDP01371	Biaka	HGDP00458
French Basque	HGDP01372	Biaka	HGDP00459
French Basque	HGDP01374	Biaka	HGDP00460
French Basque	HGDP01375	Biaka	HGDP00464
French Basque	HGDP01376	Biaka	HGDP00470
French Basque	HGDP01377	Biaka	HGDP00479
French Basque	HGDP01378	Biaka	HGDP00981
French Basque	HGDP01379	Biaka	HGDP00985
Han	HGDP00774	Biaka	HGDP01088
Han	HGDP00775	Biaka	HGDP01091
Han	HGDP00777	Biaka	HGDP01094
Han	HGDP00778	Mandenka	HGDP00904
Han	HGDP00780	Mandenka	HGDP00905
Han	HGDP00785	Mandenka	HGDP00906
Han	HGDP00786	Mandenka	HGDP00907
Han	HGDP00815	Mandenka	HGDP00908
Han	HGDP00819	Mandenka	HGDP00911
Han	HGDP00977	Mandenka	HGDP00912
Han	HGDP01288	Mandenka	HGDP00913

Population	ID	Population	ID
Han	HGDP01290	Mandenka	HGDP00919
Han	HGDP01293	Mandenka	HGDP01199
Han	HGDP01294	Mandenka	HGDP01200
Han	HGDP01295	Mandenka	HGDP01202
Han	HGDP01296	Mandenka	HGDP01283
Melanesian	HGDP00490	Mandenka	HGDP01284
Melanesian	HGDP00491	Mandenka	HGDP01285
Melanesian	HGDP00655	Mandenka	HGDP01286
Melanesian	HGDP00658	San	GM03043
Melanesian	HGDP00661	San	JR00013
Melanesian	HGDP00662	San	JR00050
Melanesian	HGDP00663	San	JR00060
Melanesian	HGDP00664	San	JR00077
Melanesian	HGDP00787	San	JR00301
Melanesian	HGDP00788	San	JR00305
Melanesian	HGDP00789	San	JR00321
Melanesian	HGDP00823	San	JR00323
Melanesian	HGDP00824	San	JR00354

Supplementary Table 2. Basic information about the regions sequenced.

Location	# bp sequenced	# bp spanned	r[1]	Location	# bp sequenced	# bp spanned	r[1]
Autosomal				**X-linked**			
1pMB4	4242	19007	2.7	XpMB3	5515	14569	3.3
4qMB105	5867	17981	1.2	XpMB6	5529	17492	1.1
4qMB181	4625	19579	1.8	XpMB9	6134	15702	1.8
5pMB4	4939	18255	1.8	XpMB22	5091	16640	3.4
5pMB10	6433	23716	2.6	XpMB33	6494	17936	2.1
5qMB128	6773	18650	0.9	XpMB35	6484	17418	1.5
6pMB14	7126	22166	2.1	XqMB124	6322	17145	1.3
6qMB164	5066	11228	1.3	XqMB139	5679	20041	3.5
7pMB8	7136	21082	2.7	XqMB140	5516	21223	4.2
8pMB5	4749	17876	3.0	XqMB143	5667	16520	3.7
10qMB119	5491	20580	2.6	XpMB13	3918	19659	2.6
10qMB128	4654	19471	2.3	XpMB39	4024	16043	1.7
12qMB46	6560	16576	1.2	XqMB120	3622	21059	1.5
13qMB107	4455	13456	2.8	XqMB136	3737	18666	1.0
13qMB108	5190	19265	1.8	XqMB141	3881	15915	2.9
16pMB17	6773	20615	1.8	XqMB145	3984	24412	1.8
18pMB7	4624	21440	3.5	XqMB146	4074	18549	1.2
18qMB73	5127	20776	2.4	XqMB148	3770	24973	1.8
19qMB35	5788	17322	2.0	XqMB149	4107	18762	8.5
20pMB7	6781	21058	3.0	XqMB150	4180	21775	3.2

[1] Recombination rate in units of cM/Mb, from Kong et al. (2002)

Supplementary Table 3. Observed and expected number of autosomal polymorphic sites within humans in HKA test.

		Global				Africans				Non-Africans			
Locus	L	N	Obs.	Exp.	Dev.[a]	N	Obs.	Exp.	Dev.	N	Obs.	Exp.	Dev. [a]
10qMB119	5521	166	43	38.33	0.193	84	35	38.18	0.076	82	16	19.94	0.336
10qMB128	4657	156	46	41.13	0.184	74	38	39.04	0.008	82	14	19.70	0.718
12qMB46	6567	158	52	50.66	0.010	76	44	48.78	0.109	82	24	26.91	0.113
13qMB107	4455	156	47	55.34	0.325	74	42	43.09	0.007	82	21	23.06	0.073
13qMB108	5197	156	49	50.15	0.007	74	40	37.94	0.031	82	18	19.46	0.048
16pMB17	6779	166	58	56.99	0.005	84	51	53.53	0.027	82	26	28.35	0.068
18pMB7	4625	156	48	38.91	0.704	74	40	36.46	0.098	82	22	19.46	0.145
18qMB73	5128	156	72	57.95	0.850	74	65	49.72	1.064	82	29	23.78	0.446
19qMB35	5789	156	51	48.88	0.026	74	46	39.41	0.298	82	18	18.98	0.022
1pMB4	4242	156	60	54.90	0.123	74	51	47.14	0.075	82	27	24.98	0.061
20pMB7	6782	156	48	62.25	0.772	74	40	42.72	0.044	82	24	24.02	0.000
4qMB105	5868	166	40	39.62	0.001	84	37	34.81	0.042	82	14	16.82	0.224
4qMB181	4630	156	56	54.52	0.010	74	42	40.51	0.014	82	39	25.71	2.553
5pMB10	6447	158	67	75.40	0.191	76	56	56.18	0.000	82	37	31.95	0.257
5pMB4	4939	156	31	51.01	2.155	74	23	37.57	1.578	82	11	21.62	2.152
5qMB128	6778	156	65	61.39	0.051	74	51	51.19	0.000	82	36	29.79	0.437
6pMB14	7144	156	75	65.35	0.325	74	61	50.46	0.493	82	35	26.67	0.945
6qMB164	5067	156	32	39.77	0.496	74	25	34.62	0.792	82	15	20.18	0.571
7pMB8	7141	156	71	69.67	0.006	74	60	58.93	0.004	82	30	31.23	0.016
8pMB5	4755	156	83	81.76	0.004	74	64	70.71	0.109	82	36	39.40	0.082
Degrees of freedom:					19				19				19
Chi square value:					14.48				10.29				15.03
Probability from chi square distribution:					0.76				0.93				0.70

[a] Deviation= (observed - expected)^2/variance. Variance is not shown.

Supplemental Table 4. Observed and expected number of X-linked polymorphic sites within humans in HKA test.

		Global				Africans				Non-Africans			
Locus	L	N	Obs.	Exp.	Dev.[a]	N	Obs.	Exp.	Dev.	N	Obs.	Exp.	Dev.[a]
XpMB13	3918	89	21	19.42	0.058	41	14	14.33	0.003	48	11	9.58	0.118
XpMB22	5092	83	32	26.38	0.439	37	28	24.06	0.200	46	22	16.05	0.939
XpMB3	5522	83	37	26.72	1.436	37	33	22.89	1.431	46	19	13.52	1.041
XpMB33	6494	83	24	19.47	0.463	37	18	19.66	0.049	46	16	13.73	0.175
XpMB35	6505	83	37	35.59	0.017	37	28	26.11	0.040	46	13	15.63	0.191
XpMB39	4025	89	10	18.91	1.903	41	8	18.81	2.331	48	9	13.62	0.738
XpMB6	5521	85	32	26.47	0.425	39	28	22.50	0.443	46	18	13.94	0.545
XpMB9	6142	83	28	30.60	0.073	37	22	26.99	0.263	46	10	16.90	1.163
XqMB120	3627	89	19	19.97	0.021	41	16	15.22	0.017	48	9	9.36	0.008
XqMB124	6326	83	21	30.94	1.055	37	19	24.94	0.427	46	5	15.00	2.948
XqMB136	3733	89	24	24.59	0.006	41	20	17.61	0.126	48	14	11.28	0.340
XqMB139	5634	83	14	20.82	0.945	37	14	19.07	0.487	46	8	12.46	0.780
XqMB140	5525	83	36	36.14	0.000	37	33	29.34	0.123	46	20	18.37	0.057
XqMB141	3881	89	16	21.71	0.630	41	14	17.61	0.290	48	11	11.92	0.036
XqMB143	5667	83	26	30.20	0.196	37	14	23.18	1.153	46	21	18.16	0.175
XqMB145	3987	89	30	24.52	0.479	41	28	23.88	0.228	48	19	15.11	0.446
XqMB146	4075	89	29	37.82	0.603	41	26	28.96	0.085	48	14	18.09	0.371
XqMB148	3770	89	18	15.41	0.220	41	17	15.52	0.059	48	10	9.58	0.010
XqMB149	4107	89	26	21.03	0.502	41	22	18.21	0.302	48	12	10.85	0.064
XqMB150	4180	89	32	25.28	0.684	41	26	19.11	0.924	48	13	10.85	0.224
Degrees of freedom:					19				19				19
Chi square value:					19.39				16.33				15.52
Probability from chi square distribution:					0.43				0.59				0.64

[a] Deviation= (observed - expected)^2/variance. Variance is not shown.

Supplementary Table 5. Pairwise F_{ST} values.

Autosomes

	Han	Melanesians	Biaka	Mandenka	San
French Basque	0.078	0.106	0.152	0.150	0.227
Han		0.119	0.174	0.173	0.227
Melanesians			0.202	0.199	0.283
Biaka				0.039	0.080
Mandenka					0.089

X chromosome

	Han	Melanesians	Biaka	Mandenka	San
French Basque	0.087	0.244	0.295	0.169	0.318
Han		0.240	0.358	0.250	0.401
Melanesians			0.317	0.218	0.378
Biaka				0.104	0.184
Mandenka					0.170

Supplementary Table 6. Coverage of HapMap for resequenced regions

Population	Autosomes		X chromosome	
	All SNPs	MAF ≥ 0.1	All SNPs	MAF ≥ 0.1
FRE	0.47 (0.27 – 0.70)	0.57 (0.29 – 0.90)	0.42 (0.18 – 0.67)	0.50 (0.18 – 0.71)
HAN	0.43 (0.06 – 0.70)	0.59 (0.36 – 0.83)	0.41 (0.25 – 0.61)	0.46 (0.31 – 0.65)
MEL	0.49 (0.05 – 0.78)	0.54 (0.06 – 0.81)	0.43 (0.20 – 0.65)	0.46 (0.24 – 0.75)
BIA	0.30 (0.03 – 0.58)	0.43 (0.05 – 0.68)	0.33 (0.10 – 0.58)	0.42 (0.09 – 0.65)
MAN	0.32 (0.03 – 0.71)	0.45 (0.06 – 0.85)	0.36 (0.07 – 0.75)	0.43 (0.09 – 0.80)
SAN	0.28 (0.05 – 0.72)	0.36 (0.00 – 0.81)	0.30 (0.08 – 0.44)	0.30 (0.08 – 0.44)
All	0.18 (0.02 – 0.34)	0.56 (0.06 – 0.85)	0.23 (0.05 – 0.64)	0.46 (0.08 – 0.80)

FRE= French Basque; HAN = Chinese Han; MEL = Melanesians; BIA = Biaka; MAN = Mandenka; SAN = San.

The numbers in parentheses show the range of coverage across regions that contain at least 10 SNPs in the appropriate frequency class (i.e., at least 10 total SNPs or 10 SNPs with MAF ≥ 0.1 respectively.

Supplementary Table 7. F_{ST} estimates from different databases

Populations (Database)	SNPs	Autosomal F_{ST}	X chromsome F_{ST}	Reference
6 Populations (current database)	a	0.158	0.257	this paper
6 Populations (current database)	b	0.179	0.279	this paper
Basque, Han and Mandenka (current database)	a	0.139	0.179	this paper
Basque, Han and Mandenka (current database)	b	0.162	0.199	this paper
YRI, CEU, CHB/JPT (HapMap)	c	0.120	0.210	(International HapMap Consortium 2005)
African- & European-American, East Asian (TSC)	d	0.123	0.195	(Akey et al. 2002)
YRI, CEU, CHB/JPT (HapMap)	e	0.098	-	(Clark et al. 2005)
YRI, CEU, CHB/JPT (HapMap)	f	0.108	-	(Clark et al. 2005)
YRI, CEU, CHB/JPT (HapMap)	g	0.130	-	(Weir et al. 2005)
African- & European-American, Han (Perlegen)	h	0.100	-	(Weir et al. 2005)

[a] all SNPs in current database

[b] HapMap SNPs present in current database

[c] All HapMap SNPs

[d] 13,615 non-coding SNPs

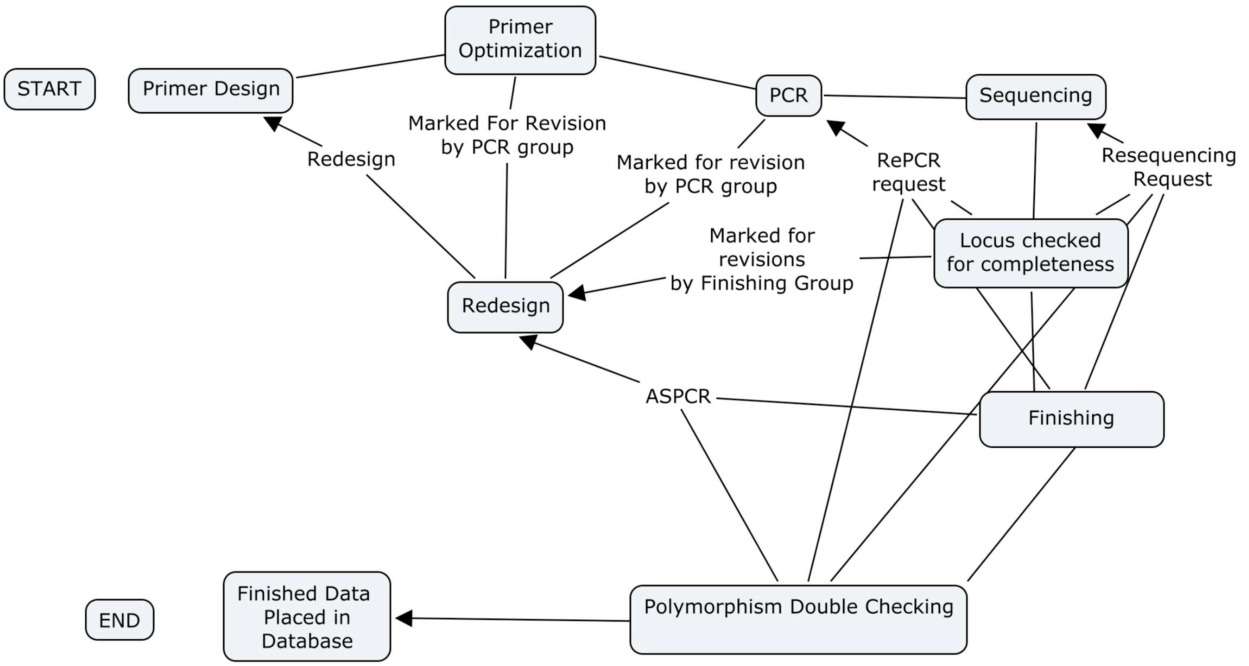
[e] ENCODE sequences

[f] HapMap SNPs in ENCODE regions

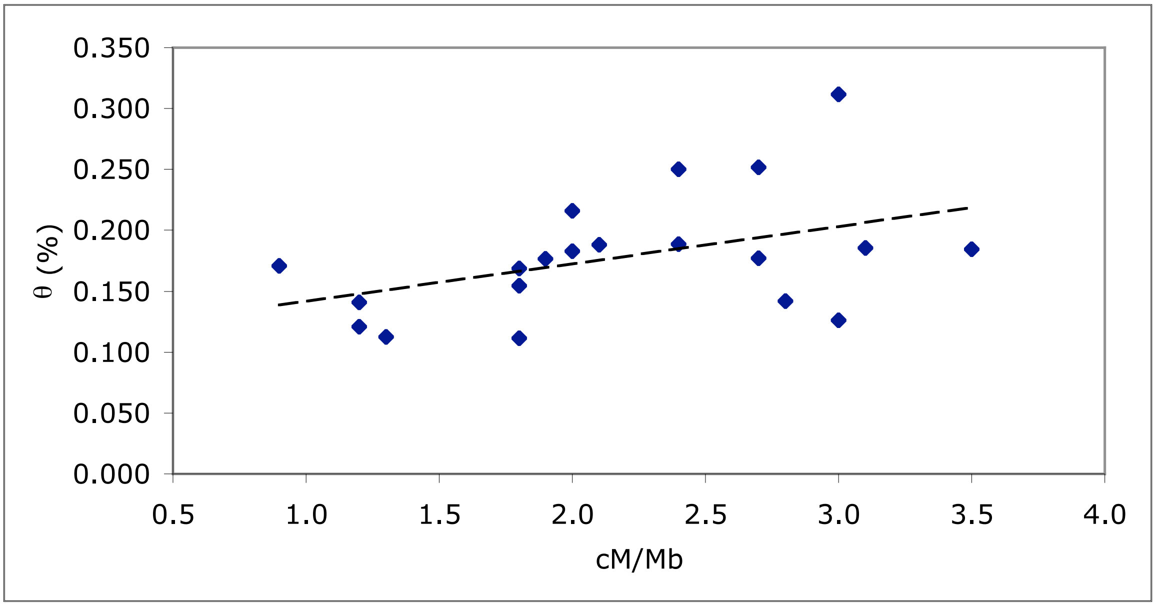
[g] HapMap SNPs segregating in all HapMap populations

[h] Perlegen SNPs segregating in all Perlegen populations

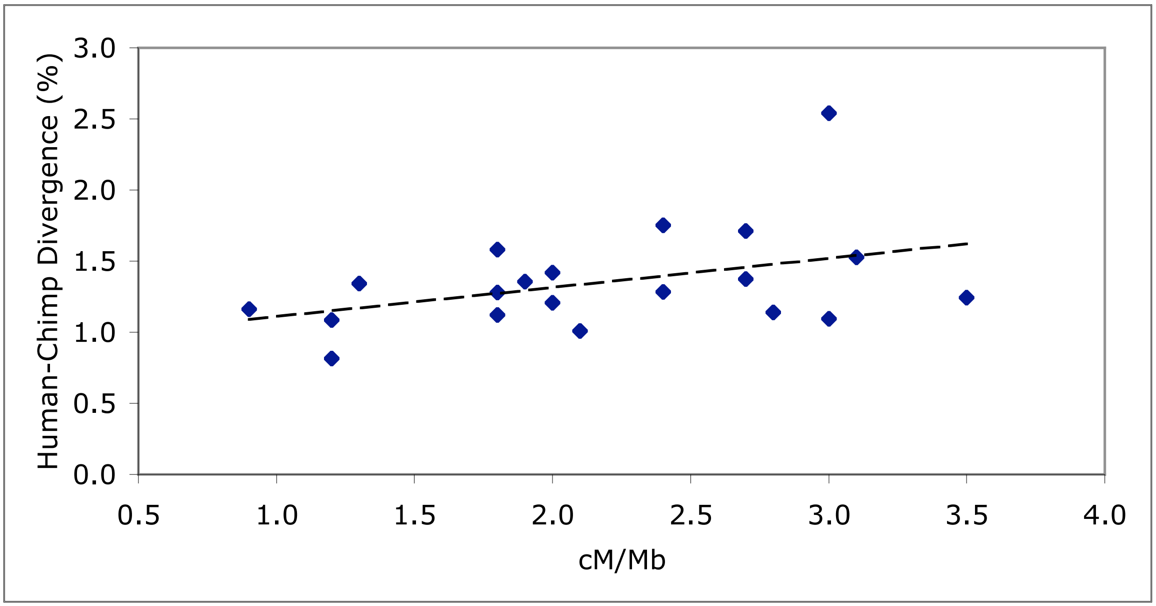
Supplementary Figure 1. Resequencing pipeline


START
Primer Design
Primer Optimization
PCR
Sequencing
Redesign
Marked For Revision by PCR group
Marked for revision by PCR group
RePCR request
Resequencing Request
Locus checked for completeness
Marked for revisions by Finishing Group
Redesign
ASPCR
Finishing
END
Finished Data Placed in Database
Polymorphism Double Checking

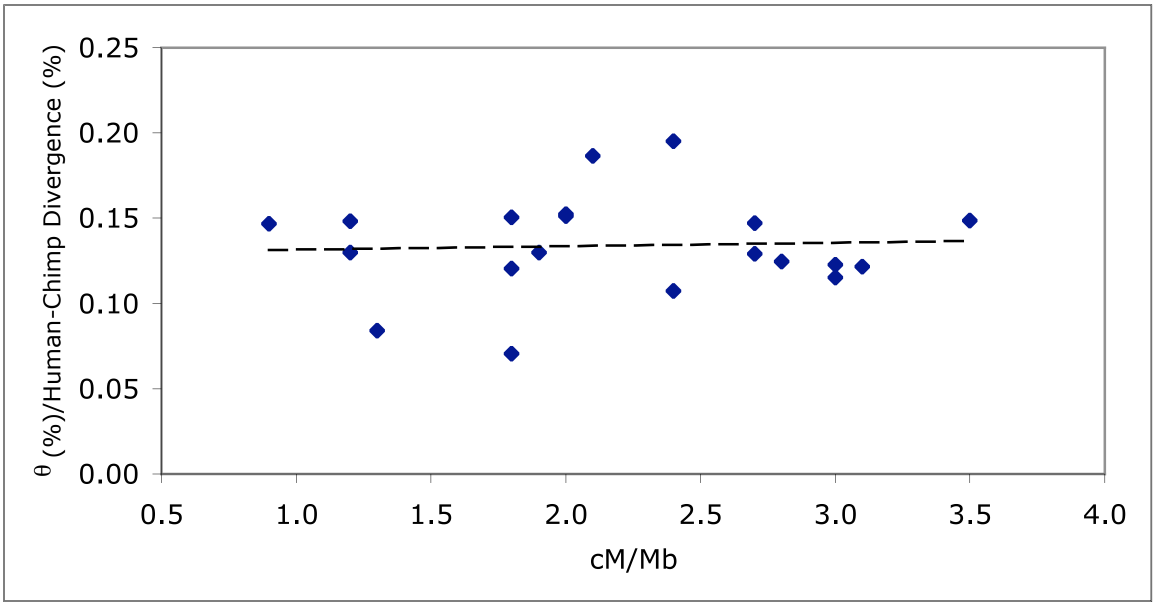
Supplementary Figure 2a. Scatterplot of human nucleotide diversity levels (θ%) *versus* recombination rate for 20 autosomal loci.



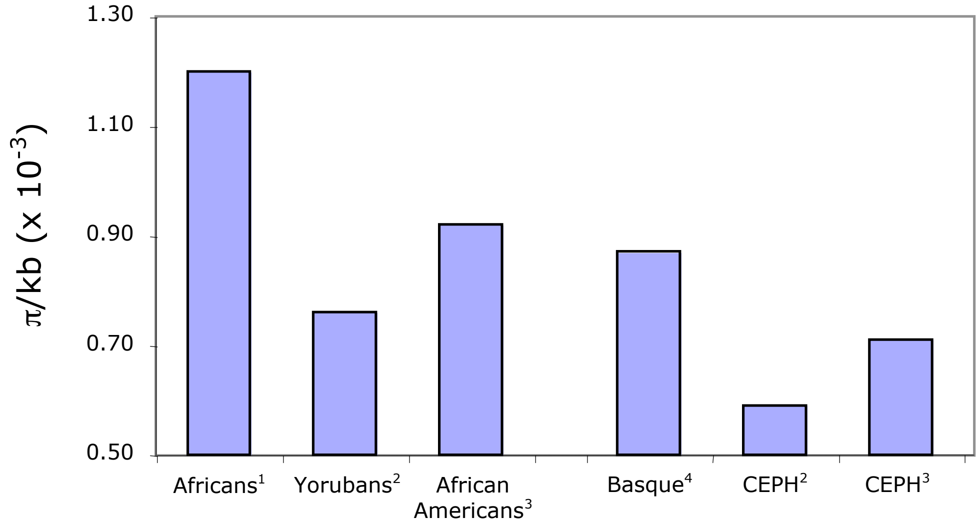
Supplementary Figure 2b. Scatterplot of human-chimpanzee divergence (%) *versus* recombination rate for 20 autosomal loci.



Supplementary Figure 2c. Scatterplot of human nucleotide diversity levels (θ)/ human-chimpanzee divergence (%) *versus* recombination rate for 20 autosomal loci.

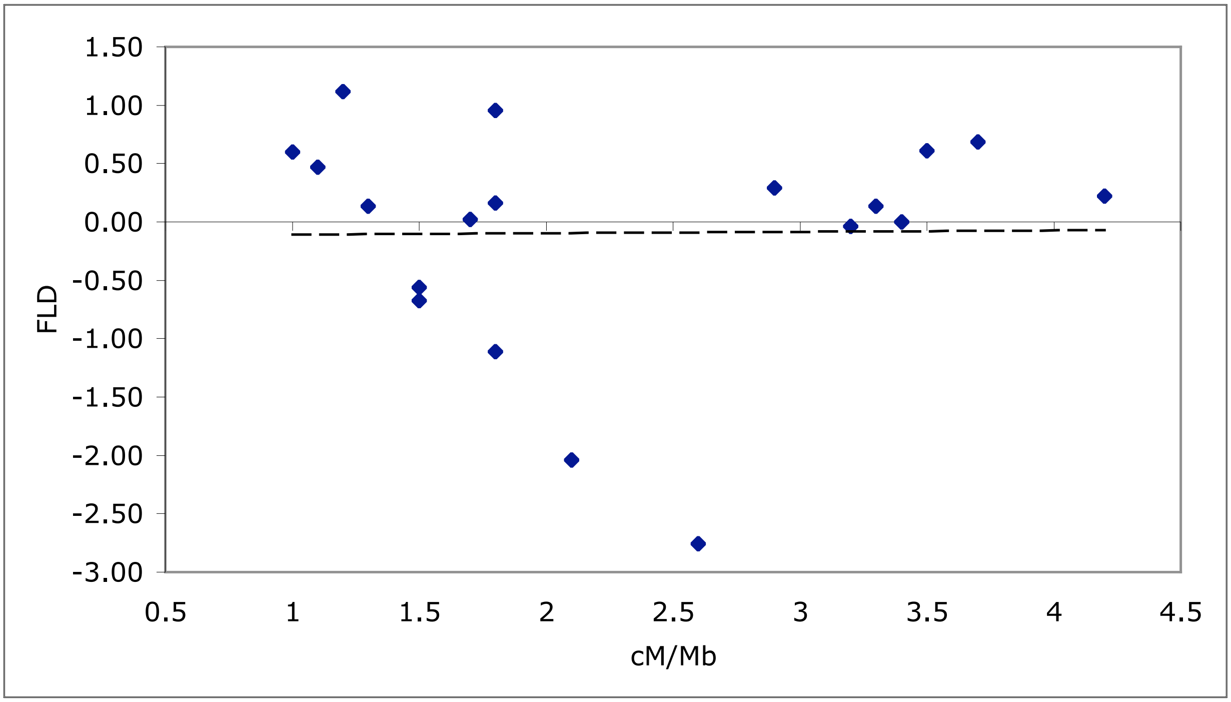


Supplementary Figure 3. Estimates of nucleotide diversity in non-genic and genic regions (see text for explanation).



[1] Three sub-Saharan African populations, this study
[2] Environmental Genome Project, 135 environmental response genes
[3] Seattle SNPs project, 300 inflammatory response genes
[4] French Basque, this study

Supplementary Figure 4. Scatterplot of Fu and L's D values (FLD) in non-Africans *versus* recombination rate for 20 X-linked loci. See **Figure 3B** in Hammer et al. (2004) for comparison of similar plot for introns associated with 15 unlinked genes on the X chromosome (see text).



References

Akey, J.M., G. Zhang, K. Zhang, L. Jin, and M.D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12:** 1805-1814.

Clark, A.G., M.J. Hubisz, C.D. Bustamante, S.H. Williamson, and R. Nielsen. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15:** 1496-1502.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

International_HapMap_Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

Nickerson, D.A., V.O. Tobe, and S.L. Taylor. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* **25:** 2745-2751.

Weir, B.S., L.R. Cardon, A.D. Anderson, D.M. Nielsen, and W.G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* **15:** 1468-1476.