Supplementary Information for 'The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes'

Michael F. Hammer, August E. Woerner, Fernando L. Mendez, Joseph C. Watkins, Murray P. Cox, Jeffrey D. Wall

Nature Genetics: doi: 10.1038/ng.651

Supplementary Table 1. Summaries of nucleotide diversity[a] and divergence in expanded resequencing dataset.

Population	Sample Size[b]	Segregating Sites	π (%)	D(%)[c]	π /D(%)
Autosomes					
Mandenka	30.8	1,771	0.115	3.490	0.034
Biaka	29.3	1,886	0.121	3.493	0.036
San	18.5	1,563	0.122	3.490	0.036
Han	32.0	1,101	0.082	3.490	0.024
Basque	32.0	1,076	0.082	3.490	0.024
Oceanians	27.4	903	0.073	3.492	0.022
X chromosome					
Mandenka	16.7	451	0.092	2.719	0.034
Biaka	14.3	431	0.089	2.721	0.032
San	9.0	366	0.086	2.720	0.032
Han	16.0	260	0.055	2.725	0.020
Basque	16.0	290	0.065	2.722	0.024
Oceanians	15.3	258	0.057	2.720	0.020

[a] Mean diversity for 61 autosomal and 30 X-linked loci.

[b] Mean number of alleles sequenced per locus per population.

[c] D = human-orangutan sequence divergence (Orangutan Genome Project).

Note- Aligning human sequences with the recently available orangutan genome yielded slightly higher levels of divergence than that found in the study of Hammer et al.[1], especially for the X chromosome (4.8%). This difference may result from the fact that Hammer et al.[1] used primers designed from human sequences to PCR-amplify orthologous regions from orangutan.

1. Hammer, M.F., Mendez, F.L., Cox, M.P., Woerner, A.E. & Wall, J.D. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* **4**, e1000202 (2008).

Nature Genetics: doi: 10.1038/ng.651

Supplementary Table 2. Coverage[a] (Mb) for four bins at increasing distances from a human gene[b].

distance from	Marmoset		Rhesus		Orangutan		Gorilla		Chimpanzee	
nearest gene (cM)	A	X	A	X	A	X	A	X	A	X
0.0-0.1	509.56	29.60	585.85	37.16	628.95	42.88	393.36	23.02	665.99	38.00
0.1-0.2	75.87	5.29	85.68	6.52	90.46	7.07	55.83	3.64	51.88	6.49
0.2-0.3	33.84	2.12	38.45	2.69	40.09	2.93	24.64	1.43	42.33	2.55
0.3-0.4	15.60	1.04	17.70	1.29	18.64	1.48	11.03	0.71	19.68	1.33
total length	634.87	38.05	727.68	47.66	778.15	54.36	484.86	28.80	779.88	48.38

[a] Coverage is defined as a base that is present in the outgroup and aligned with a homologous base in at least one ingroup.

[b] See supplementary material for gene definitions

Note- Repeated and ultraconserved regions were discard from alignments.

Nature Genetics: doi: 10.1038/ng.651

Supplementary Table 3. Summary of X chromosome/autosomal data for European population samples[a] in three studies.

	Hammer et al.[1]		This study		Public Data (all bins)		Public Data (Bin 4)	
	A	X	A	X	A	X	A	X
n samples	32	16	32	16	6	6	6	6
n loci	20	20	61	30	19,568	831	439	28
n bases (Mb)	3.60	1.56	11.44	2.34	778.15	54.36	18.64	1.48
S	338	200	1,076	290	1,251,083	49,246	33,068	2,169
π	0.086	0.071	0.082	0.065	0.090	0.052	0.103	0.090
D (H-O)	3.43	2.59	3.49	2.72	3.74	2.95	3.78	3.01
π/D	0.025	0.028	0.024	0.024	0.025	0.018	0.027	0.030
N_X/N_{aut}	1.12		1.00		0.71		1.11	

[a] Hammer et al.[1] and this study: French Basque; Public data: See Supplementary Materials

1. Hammer, M.F., Mendez, F.L., Cox, M.P., Woerner, A.E. & Wall, J.D. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* **4**, e1000202 (2008)

Nature Genetics: doi: 10.1038/ng.651

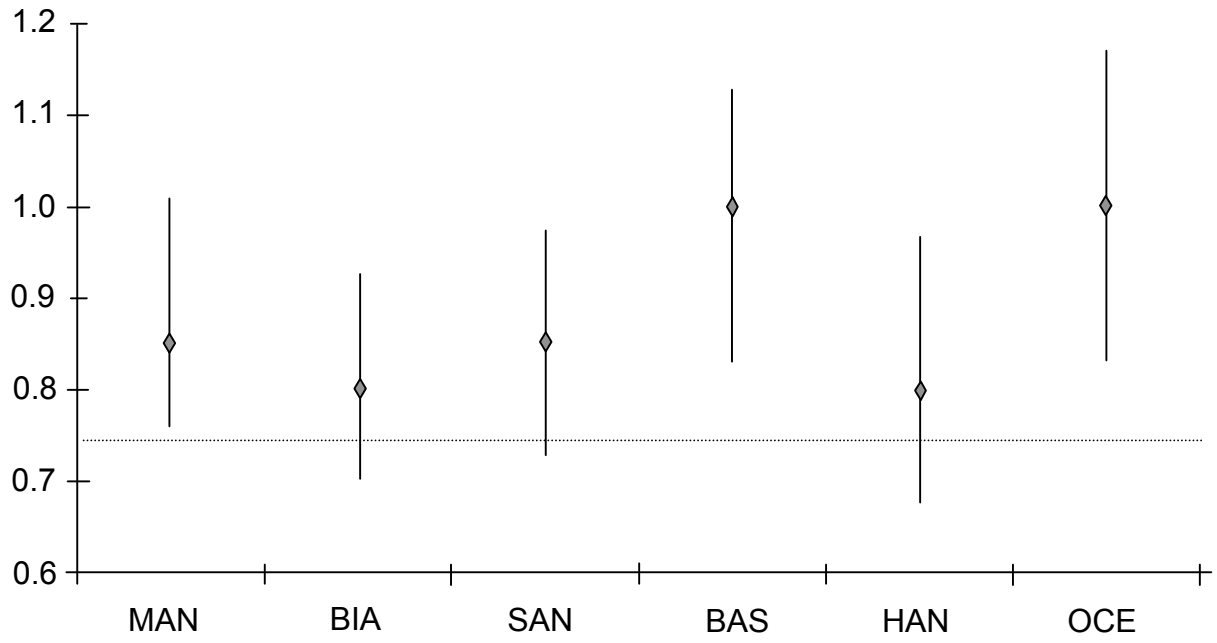
Supplementary Table 4. X chromosomal/autosomal ratios of π/D.

bin[a]	Marmoset	Rhesus	Orangutan	Gorilla	Chimp
0.0-0.1	0.60	0.63	0.67	0.67	0.70
0.1-0.2	0.65	0.68	0.74	0.72	0.77
0.2-0.3	0.77	0.79	0.91	0.78	0.92
0.3-0.4	0.95	0.95	1.11	1.05	1.06

[a] distance from nearest gene (cM)

Nature Genetics: doi: 10.1038/ng.651

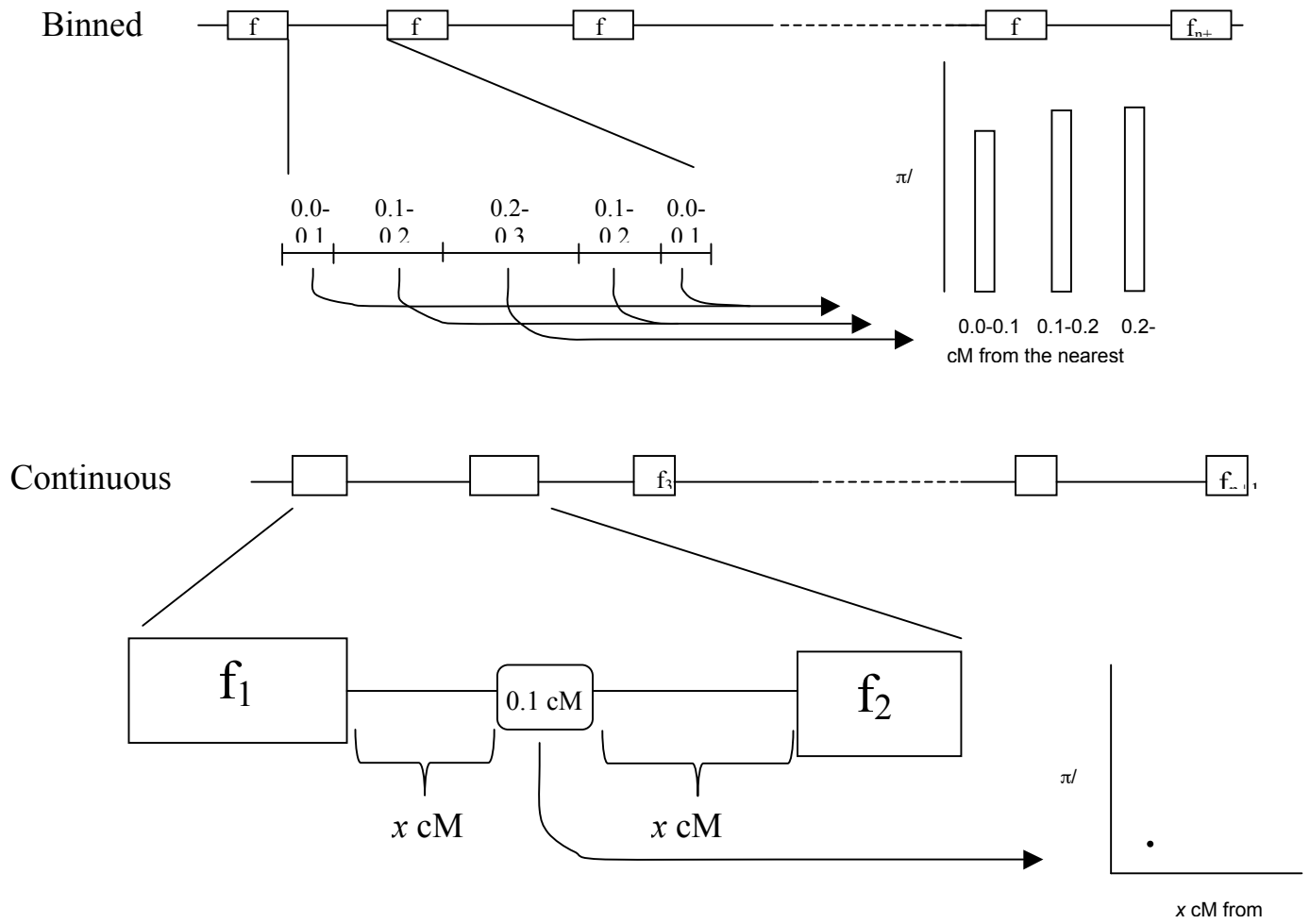
Supplementary Figure 1. Point estimates and 95% CIs for N_X/N_{aut} ratios for 91-locus dataset. The tick represents the point estimate, while the vertical bar shows the estimated 95% confidence interval (see Hammer et al.[1] for methods). The dotted line represents the expected ratio (0.75) under a neutral model with breeding sex ratio of 1. The mean value of ~0.90 is consistent with a 2-3-fold excess of breeding females. Three letter population codes are as follows: Mandenka (Man), Biaka (Bia), San (San), French Basque (Bas), Han Chinese (Han), Oceanians (Oce).



1. Hammer, M.F., Mendez, F.L., Cox, M.P., Woerner, A.E. & Wall, J.D. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* **4**, e1000202 (2008)

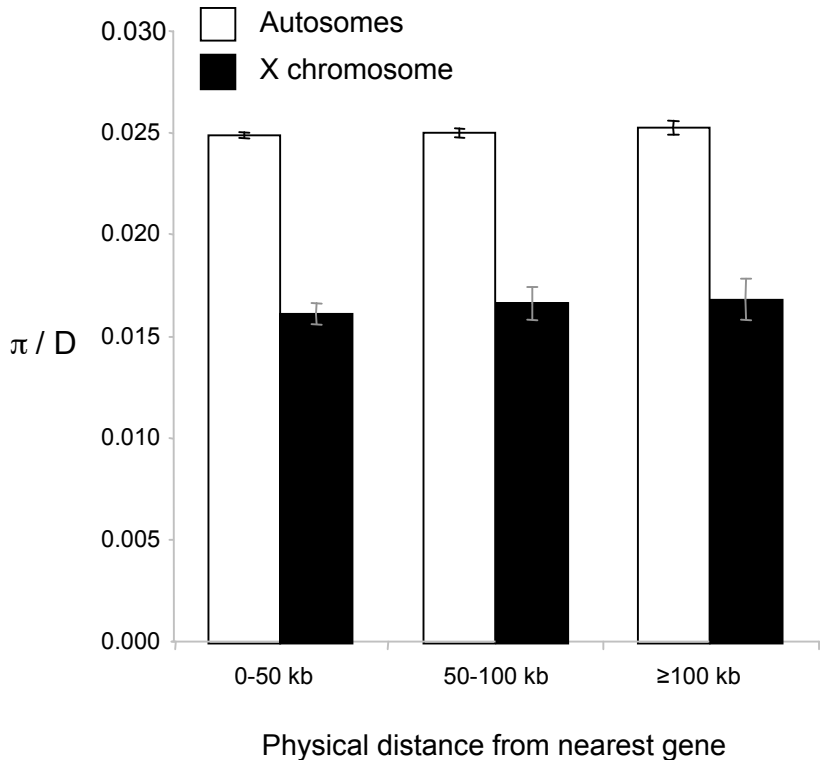
Nature Genetics: doi: 10.1038/ng.651

Supplementary Figure 2. Strategies for sampling genomes. The binned approach (top panel) places each neutral region into several bins (0.1 cM) that are defined by their distance to the nearest gene (f_1, f_2, f_3, etc.). The continuous approach (bottom panel) takes each non-genic interval in the genome, finds the medial 0.1 cM subsection (as defined by the genetic map), and defines that as the neutral region.



Nature Genetics: doi: 10.1038/ng.651

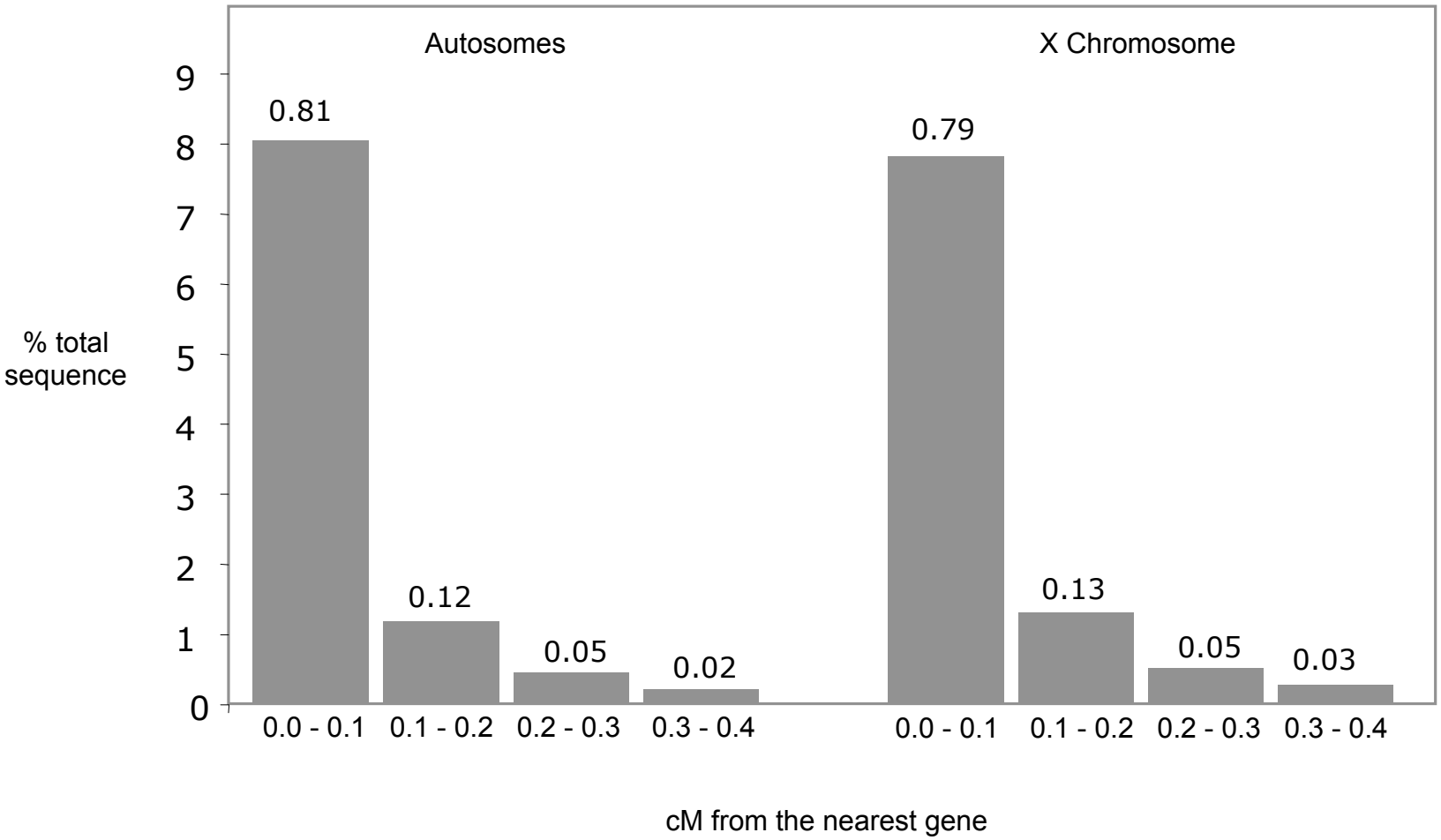
Supplementary Figure 3. Diversity (p/D) on the X chromosome and the autosomes as a function of physical distance from genes (as per [2]). The values shown are means ± standard errors of the mean. Note the different scales shown here and in Fig 1 (e.g., if 1 cM = 1 Mb, then the bar labeled > 100 kb here, corresponds to the three right bars in Fig 1).



2. Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet* **41**, 66-70 (2009).

Nature Genetics: doi: 10.1038/ng.651

Supplementary Figure 4. Percentage of genome sequence in each bin (binned approach) (see **Table S2**).



Nature Genetics: doi: 10.1038/ng.651

Supplementary Methods

Genomes Sampled. Six human genomes (Venter[1], Watson[2], three CEU samples NA12878, NA12156[3] and NA07022, and one personal genome project (NA20431)[4], broadly characterizable as being of European descent, were sampled for these analyses. We reconstructed diploid chromosome files by using the hg18 (2006) human genome as a template, overlaying SNPs on top of this template, and then masking out the regions where the coverage is missing. Watson and Venter's genomes were constructed using SNP files from the Genome Variants track from the UCSC genome database and coverage estimates from the 1000 genomes browser (http://browser.1000genomes.org). The first two CEU genomes were constructed using the SNPs ascertained in Kidd et al.[3], remapped to the hg18 human genome (using liftover, which is available at: http://www.genome.ucsc.edu/cgi-bin/hgLiftOver). Coverage for these CEU samples was estimated from the qualityaligns files[3], where a base was considered covered if it and its immediate neighbors had a Phred quality of at least 30. Coverage and SNP information for the genomes found in Drmanac et al.[4] were constructed from variation files provided at http://www.completegenomics.com/dataRelease/variations.aspx, which use assembly versions 1.2.0.11 and 1.2.0.14 for NA07022 and NA20431, respectively. To avoid missed heterozygous sites, which would reduce the apparent diversity of the autosomes more than the X chromosome, each reconstructed diploid chromosome was then subsampled to a single haploid chromosome by randomly picking one of the two alleles from the SNP files. Multiple outgroup sequences, including an orangutan, were obtained from the 44-way vertebrate alignments[5,6]. Unless otherwise stated, the orangutan was used as the outgroup in all analyses.

Regions Sampled. We took the union of both relatively conservative (UCSC Genes) and lax

Nature Genetics: doi: 10.1038/ng.651

(Gene Bounds and Spliced Ests) gene predictions[7,8] from the UCSC genome browser database[9] for the hg18 (2006) human genome to produce an inclusive definition of putatively functional genomic regions. Recombination rate estimates were taken from HapMap Phase II[10] and the recombination rates for the X chromosome was further scaled by 2/3 as per Payseur and Nachman[11]. By coupling the complement of the putatively functional regions with the fine-scale recombination rate estimates from HapMap, we were able to associate each non-genic region to its genetic distance to the nearest gene. Each non-genic interval of the genome is represented two-fold in this analysis. The "binned" approach places each neutral region into several bins that are defined by their distance to the nearest gene (see Figure S2). The bin size used was 0.1 cM. The second "continuous" approach takes each non-genic interval in the genome, finds the medial 0.1 cM subsection of said interval (medial as defined by the genetic map), and defines that as the neutral region (see Figure S2). To be conservative, we use several inclusive definitions of genes and attempt to control for the effects that conserved non-genic sequence, simple repeats, and duplications may have on patterns of nucleotide variability. Thus, each non-genic region was further filtered by removing simple repeats[12], duplicated regions, as defined by the segmental duplications and the self chain tracks[13,14], and conserved non-genic sequence, defined by the 28-way vertebrate alignment "most conserved" track[6]. After filtering, regions with less than 1kb of coverage (where a base is considered covered if it is defined both in the outgroup and in at least one of the ingroups) were excluded from the analysis. From each non-genic region we then computed π, using the *libsequence* C++ library[15], and divergence, using an in-house script.

Calculations. We ran two-tailed Mann-Whitey U tests to determine whether samples of values were drawn from the same parental distribution. For the local regression analysis, we first

Nature Genetics: doi: 10.1038/ng.651

determined a line of best fit of π/D *versus* genetic distance from genes for both autosomal data and for X chromosome data. Because the residual standard deviation depends on genetics distance, we use an iterated weighted least squares regression approach. To begin the iteration, a line is determined using ordinary least squares. To prepare for the weighted regression line in iteration k+1, the weights are determined from local regression[16] of the squares of the residuals from iteration k. This is continued until the sum of the squares of the difference between the weights from two consecutive iterations divided by the squares of the weights from the latest iteration is less that 10^{-6}. This procedure is the basis for a test of the slopes $ß_{aut}$ and $ß_X$ of these two lines:

$$H_0\ ß_X \leq 3/4\ ß_{aut}\ \textit{versus}\ H_1\ ß_X > 3/4\ ß_{aut}$$

Code to perform the iteration was written in R, which gives an output value for the F-statistic. Given 1 degree of freedom in the numerator, we take the square root of the F-statistic to get the t-statistic.

Nature Genetics: doi: 10.1038/ng.651

References for Supplementary Methods

1. Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).
2. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-6 (2008).
3. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
4. Drmanac, R. et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* (2009).
5. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-15 (2004).
6. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
7. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank: update. *Nucleic Acids Res* **32**, D23-6 (2004).
8. Hsu, F. et al. The UCSC Known Genes. *Bioinformatics* **22**, 1036-46 (2006).
9. Karolchik, D. et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**, D773-9 (2008).
10. International_HapMap_Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
11. Payseur, B.A. & Nachman, M.W. Gene density and human nucleotide polymorphism. *Mol Biol Evol* **19**, 336-40 (2002).
12. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
13. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17 (2001).
14. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**, 11484-9 (2003).
15. Thornton, K. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325-7 (2003).
16. Cleveland, W., Grosse, E. & Shyu, W. Local regression models. in *Statistical Models* (eds. Chambers, J. & Hastie, T.) (CRC Press, Boca Raton, 1992).

Nature Genetics: doi: 10.1038/ng.651