Testing for Archaic Hominin Admixture on the X-Chromosome: Model Likelihoods for the Modern Human *RRM2P4* Region from Summaries of Genealogical Topology under the Structured Coalescent

Murray P. Cox, Fernando L. Mendez, Tatiana M. Karafet, Maya Metni Pilkington, Sarah B. Kingan, Giovanni Destro-Bisol, Beverly I. Strassmann, Michael F. Hammer

SUPPLEMENTARY TABLE 1A

Polymorphism in the *RRM2P4* Central Region (Sample Panel A). SNPs, relative to the Chimpanzee outgroup, are described by UCSC March 2006 coordinates and the identifiers of GARRIGAN *et al.* (2005). Dots denote character states identical to Chimpanzee reference. African and non-African proportions are shown on the right. 'Clade A' haplotypes are highlighted.

UCSC	143219175	143219400	143219560	143219632	143219703	143220010	143220424	143220506	143220511	143220738	143220790	143220797	143220802	143220896	143220984	143221102	143221114	143221183	143221415	143221426	143221467	143221477	African	Non-African	Total
Garrigan	12	237	397	469	540	847	1261	1343	1348	1575	1627	1634	1639	1733	1821	1939	1951	2020	2252	2263	2304	2314			
Chimpanzee	T	C	G	A	T	C	A	T	G	C	A	T	A	A	G	C	T	G	T	C	T	A			
Reference	C	.	.	.	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	.	65	79	144
1	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	A	.	.	.	-	8	8
2	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	A	.	A	.	-	11	11
3	.	.	.	.	.	.	.	A	.	.	.	.	G	.	.	.	.	G	A	.	A	.	1	1	2
4	.	.	.	.	.	.	.	A	.	.	.	.	G	.	.	.	.	.	.	.	C	.	2	-	2
5	.	.	.	.	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	.	5	-	5
6	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	C	.	2	-	2
7	C	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	C	.	7	-	7
8	C	.	.	.	.	.	.	.	.	C	.	C	G	.	.	.	.	.	.	.	C	.	1	-	1
9	C	.	.	.	.	.	.	.	T	.	.	C	G	.	.	.	.	.	.	.	C	.	1	-	1
10	C	.	.	.	.	T	.	.	.	.	.	C	G	.	.	G	.	.	.	.	C	.	2	-	2
11	C	.	.	.	.	T	G	.	.	.	.	C	G	.	.	.	.	.	.	.	C	.	-	13	13
12	C	.	.	.	C	T	G	.	.	.	.	C	G	.	.	.	.	.	.	.	C	.	-	3	3
13	C	.	.	.	.	T	G	.	.	.	.	C	G	.	.	.	.	.	A	.	C	.	-	1	1
14	C	.	.	.	.	T	.	.	.	.	.	C	G	.	.	.	.	.	.	.	C	.	35	2	37
15	C	.	.	.	.	T	.	.	A	.	.	C	G	.	.	.	.	.	.	.	C	.	1	-	1
16	C	.	.	.	.	T	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	.	1	-	1
17	C	T	.	.	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	.	-	3	3
18	C	.	C	.	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	.	3	-	3
19	C	.	.	.	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	G	-	1	1
20	C	.	.	.	.	.	.	.	.	.	.	.	G	G	A	.	.	.	.	.	C	.	1	-	1
21	C	.	.	G	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	C	.	1	-	1
22	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	C	.	3	-	3
																							131	122	253

SUPPLEMENTARY TABLE 1B

Distribution of *RRM2P4* central region haplotypes among populations. Compare with the distribution maps presented by GARRIGAN *et al.* (2005). 'Clade A' haplotypes are highlighted. Population abbreviations: San, San pygmies; Bia, Biaka pygmies; Man, Mandenka; Dgn, Dogon; Dnk, Dinka; Cmn, Cameroonians; Bas, French Basque; Han, Han Chinese; Mel, Melanesians; Sel, Selkups; Fon, Forest Nentsi; Tun, Tundra Nentsi.

Haplotype	Africans						Non-Africans						Total
	San	Bia	Man	Dgn	Dnk	Cmn	Bas	Han	Mel	Sel	Fon	Tun	
Reference	7	7	11	22	11	7	9	14	8	25	21	2	144
1								5	2	1			8
2									6		4	1	11
3				1					1				2
4		1		1									2
5	2					3							5
6				2									2
7		3			1	3							7
8	1												1
9	1												1
10		1			1								2
11									8	3	2		13
12										2	1		3
13											1		1
14	8	4	5	5	6	7	2						37
15	1												1
16					1								1
17							2			1			3
18		1				2							3
19								1					1
20						1							1
21	1												1
22	3												3
	24	17	16	31	20	23	13	20	25	32	29	3	
						131						122	253

3

SUPPLEMENTARY TABLE 2A

Polymorphisms in the *RRM2P4* central region and their association with 5′ and 3′ flanking region polymorphisms (Sample Panel B). SNPs, relative to the Chimpanzee outgroup sequence, are described by UCSC March 2006 coordinates. 'Clade A' haplotypes are highlighted.

UCSC	143212181	143212381	143213316	143213404	143213713	143219175	143219400	143220010	143220424	143220506	143220738	143220797	143220802	143220896	143221102	143221114	143221183	143221426	143221467	143221477	143227149	143227156	143227378	143228252	143228301	143228333	143228414	143228526
Chimpanzee	T	G	T	C	A	T	C	C	A	T	C	T	A	A	C	T	G	C	T	A	A	G	A	A	T	A	A	A
Reference	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	.	A	.	.	.	.	.	.
1	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	A	A	.	.	.	.	.	G	.	C	.	.
2	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	A	.	.	.	.	A	.	.	.	.	.	.
3	.	.	C	.	G	C	.	T	.	.	.	C	G	.	G	.	.	.	C	.	.	A	.	.	.	.	.	.
4	.	.	C	.	.	C	.	T	.	.	.	C	G	.	.	.	.	.	C	.	.	.	.	G	.	C	.	.
5	.	T	C	.	.	C	.	T	.	.	.	C	G	.	.	.	.	.	C	.	.	.	.	G	.	C	.	.
6	.	T	C	.	.	C	.	T	.	.	.	C	G	.	.	.	.	.	C	.	.	.	.	G	G	C	.	.
7	.	T	C	.	.	C	.	T	.	.	.	C	G	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.
8	.	T	C	.	.	C	.	T	G	.	.	C	G	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.
9	C	.	.	.	.	C	.	.	.	.	.	C	G	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.
10	.	.	C	.	.	C	.	T	.	.	.	C	G	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.
11	.	.	C	.	.	C	.	.	.	.	T	C	G	.	.	.	.	.	C	.	.	.	.	G	G	C	.	.
12	.	.	.	.	.	C	.	.	.	.	.	.	G	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.
13	C	.	.	.	.	C	T	.	.	.	.	.	G	G	.	.	.	.	C	.	.	A	.	.	.	.	.	.
14	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	.	.	.	G	.	C	.	.
15	C	.	.	T	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	.	A	.	.	.	.	.	.
16	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	.	.	.	G	.	C	-	.
17	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	G	A	.	.	.	.	.	.
18	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	G	G	A	.	.	.	.	.	.
19	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	.	A	.	.	.	.	C	.
20	C	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	C	.	.	A	.	.	.	.	.	C

SUPPLEMENTARY TABLE 2B

Distribution of *RRM2P4* central and flanking region haplotypes among populations. Compare with the distribution maps presented by GARRIGAN ***et al.* (2005). 'Clade A' haplotypes are highlighted. Population abbreviations: Bia, Biaka pygmies; Man, Mandenka; San, San pygmies; Bas, French Basque; Han, Han Chinese; Mel, Melanesians.**

Haplotype	Africans			Non-Africans			Total
	Bia	Man	San	Bas	Han	Mel	
Reference	4	7	2	10	8	5	36
1				1		3	4
2					4	1	5
3	1						1
4			1				1
5	1						1
6			2				2
7	1	4		1			6
8						5	5
9	3		1				4
10	1			1			2
11			1				1
12			2				2
13				3			3
14	3						3
15		2					2
16		1					1
17					1		1
18					1		1
19					1		1
20					1		1
	14	14	9	16	16	14	
		37			46		83

SUPPLEMENTARY TABLE 3

Basic summaries of comparative X-chromosome loci.[*]

X-chromosome locus (MB)[†]	Coverage	bp	S
XpMB3	14674	5620	37
XpMB6	17512	5549	32
XpMB9	15714	6146	28
XpMB13	19663	3922	22
XpMB22	16672	5123	33
XpMB33	17936	6494	25
XpMB35	17448	6514	38
XpMB39	16054	4035	10
XqMB120	21092	3655	19
XqMB124	17150	6327	25
XqMB136	18685	3742	24
XqMB139	20056	5647	14
XqMB140	21235	5528	36
XqMB141	15920	3886	16
XqMB145	24467	4039	30
XqMB146	18573	4094	30
XqMB148	25006	3803	19
XqMB149	18763	4108	26
XqMB150	21778	4181	32
mean	18863	4864	26

[*] See supplementary methods (below) for information on how loci were chosen and sequenced.

[†] Approximate position along the X-chromosome in megabases.

SUPPLEMENTARY FIGURE LEGENDS

SUPPLEMENTARY FIGURE 1. –Recombination rates in cM/Mb across the *RRM2P4* trio as calculated from patterns of sequenced polymorphism. The horizontal dotted line indicates the human X-chromosome average.

SUPPLEMENTARY FIGURE 2. –Network representation of sequence variation across the entire *RRM2P4* locus. Circles are sized relative to haplotype frequency; mutations dividing haplotypes are indicated by solid links; mutation positions are indicated relative to the UCSC March 2006 build (minus 143,200,000). Site position numbers are listed relative to the nomenclature of GARRIGAN *et al.* (2005) in Supplementary Table 1. A cross indicates the closest point to the chimpanzee outgroup sequence. Solid and dotted cycles indicate probable recombination events; highlighted regions indicate a major sequence block likely dispersed across the network by recurrent chiasma at the 3′ recombination hotspot. African individuals are shown in black, non-African individuals in white. 'Clade A' labels the lineages putatively introgressing from ancient hominins.

SUPPLEMENTARY FIGURE 3. –Example results from joint Approximate Bayesian Inference for optimal modern effective size (N_0) and intercontinental migration rate (m) under the 2-deme RAR model. Data points represent migration and modern effective sizes variables drawn from a random uniform distribution. Approximately 10^5 prior data points (black) are plotted on a log-log scale; posterior data points (red) represent the 0.01% of replicates best fitting the observed *S*.

SUPPLEMENTARY FIGURE 4. –Effects on the minimum clade-proportion of archaic admixture into the *African* deme (A) prior to the out-of-Africa expansion and (B) following the out-of-Africa expansion. Likelihood surface generated by interpolation from mean p_{mc} values of 10^5 simulations at each point in a 10 x 10 grid (dots) covering the parameter space of admixture time and admixture proportion. Red dots indicate demographies under which the *RRM2P4* locus is a statistical outlier. (A) Moderately frequent (~2%) archaic admixture into African populations just prior to the out-of-Africa expansion (a few tens of thousand years) provide a set of valid alternative models, whereas (B) all scenarios involving admixture following the out-of-Africa expansion can be discounted.
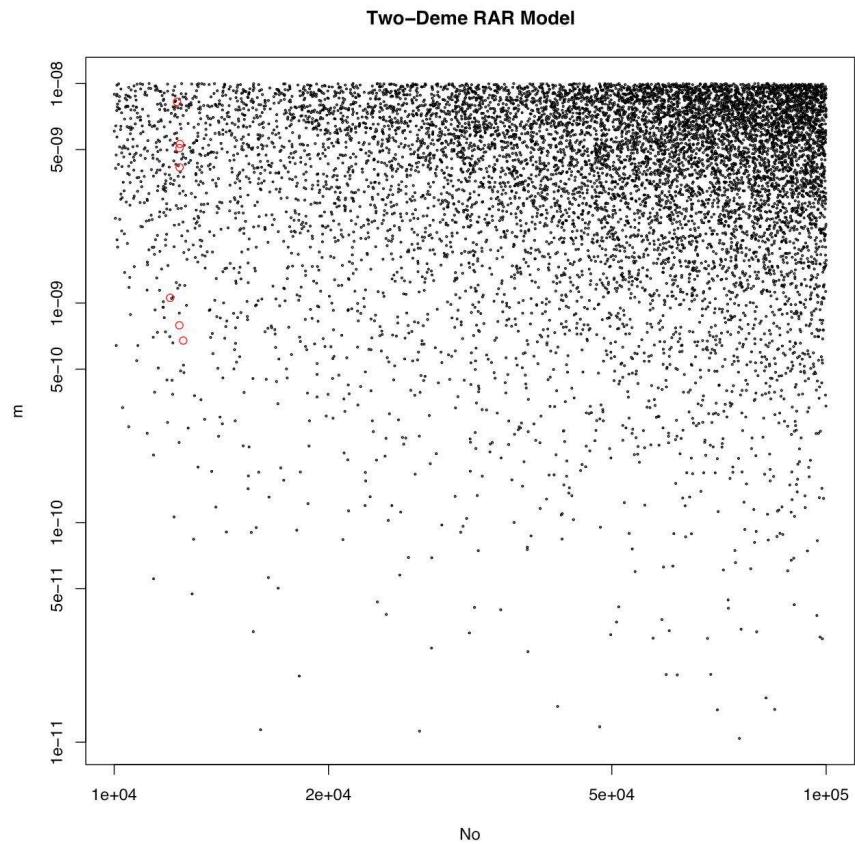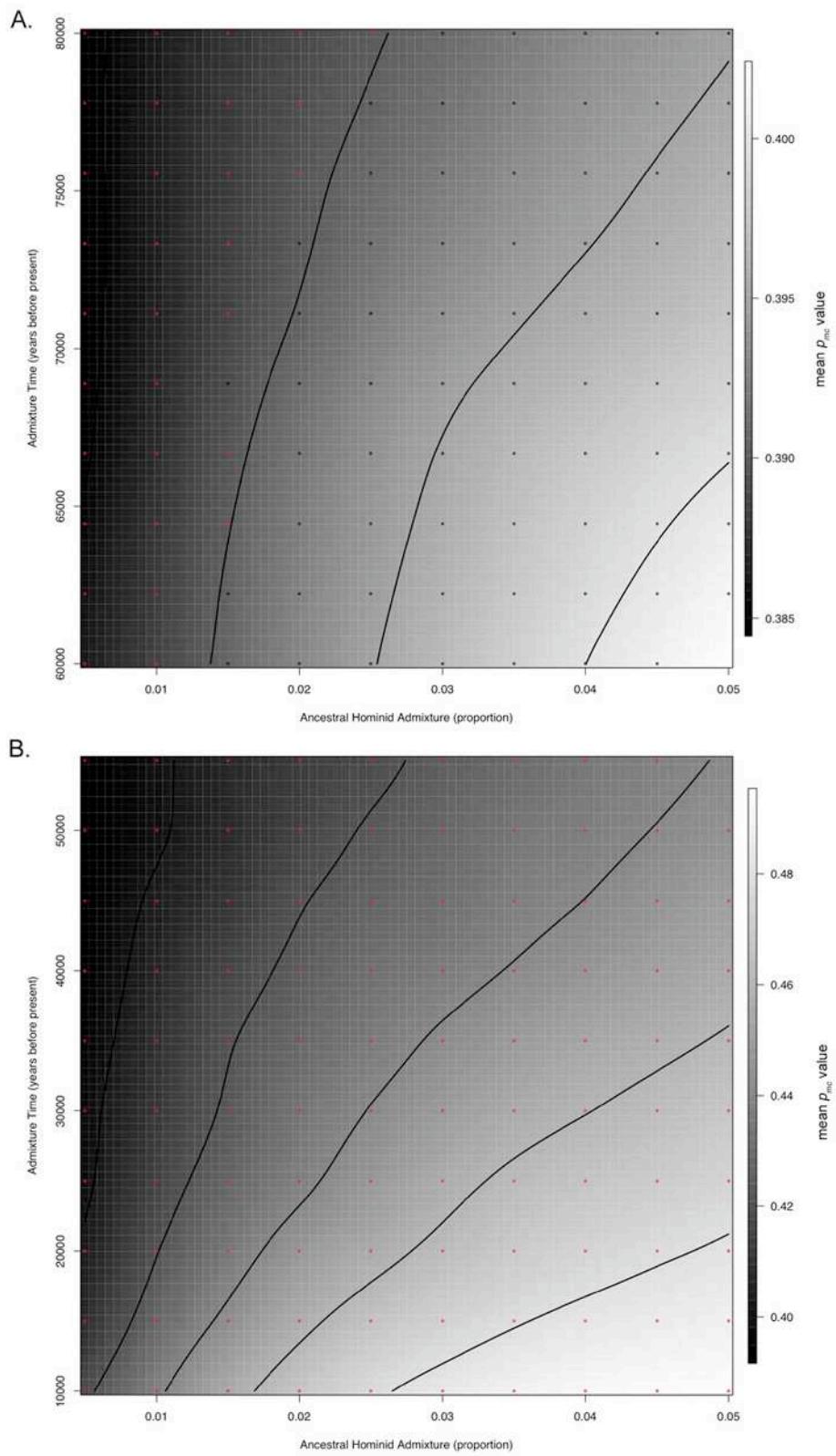
SUPPLEMENTARY FIGURE 1
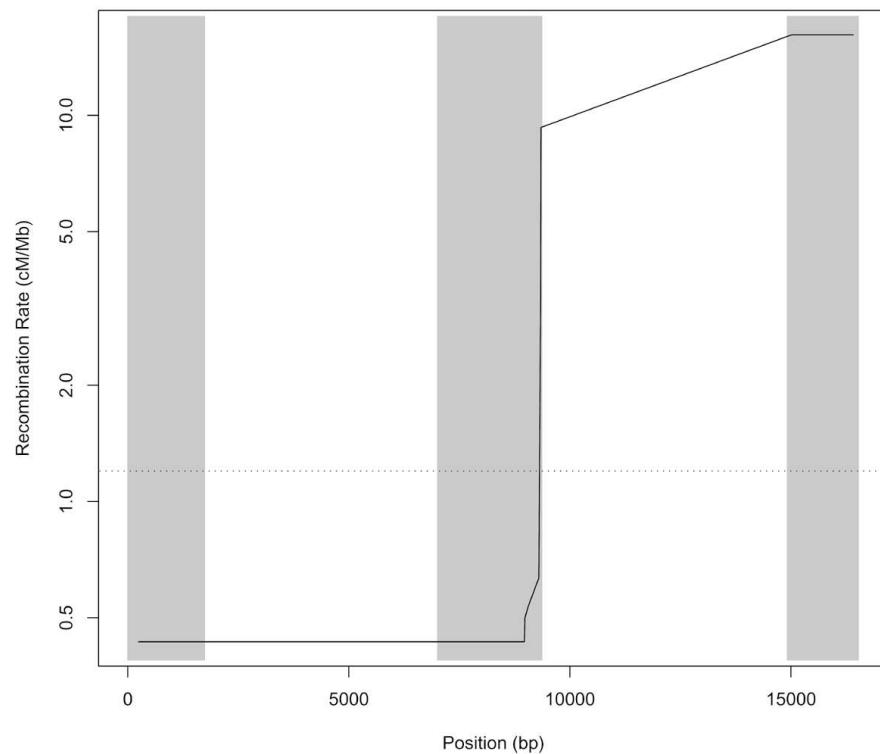
Two−Deme RAR Model
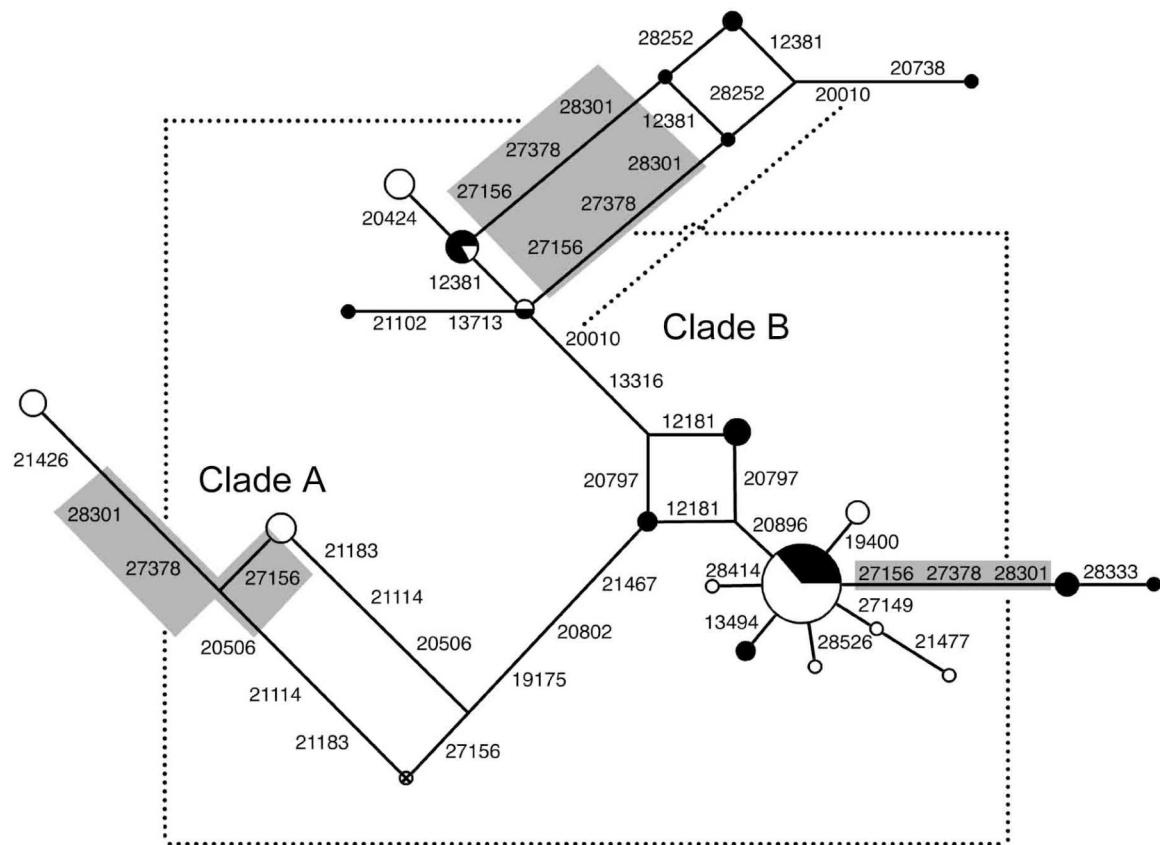
SUPPLEMENTARY METHODS

The regions used for sequencing were chosen to minimize any potential confounding effects of natural selection. As part of larger study of 20 autosomal and 20 X-linked regions, we identified 19 different ~20 kb regions of primarily single-copy non-coding (i.e., putatively non-functional) DNA in regions of medium or high recombination ($r \geq 0.9$ cM / Mb) (KONG et al. 2002). Each region was at least 50 kb away from the nearest gene, with 'gene' defined as the union of both stringent (Known Genes, cf. HSU et al. 2006) and broad (Gene Bounds, cf. BURGE and KARLIN 1997; 1998) gene-prediction definitions. Within each region, we gathered ~4–6 kb of sequence data from 3 or 4 discrete subsections that spanned most of the distance of each region (a locus trio, see **Figure 1**). Subsections were chosen to minimize the number of repetitive elements after alignment with orthologous sequences in the chimp genome. To avoid non-coding functional DNA, we rejected candidate regions that had human-chimp divergence less than the average human-chimp exonic divergence rate (~0.7%) (cf. BEJERANO et al. 2004). After scanning the whole genome for regions that met the above search criteria, we ranked these regions based on additional criteria, including (a) the quality and quantity of nearby ESTs, (b) the distance to the nearest gene beyond the minimum requirements, (c) the number of base pairs of non-repeat masked sequence included in the locus trio, (d) and the number of homo/hetero polymers included in the final target region. We then chose the top 20 X-linked regions (which included the *RRM2P4* region analyzed here) to sequence, ensuring that no two regions were within 1 Mb of each other. The locus trio design provides a cost-effective strategy to gather intermediate-range linkage disequilibrium information that would not be available if we had sequenced contiguous stretches of the same size from each region (FRISSE et al. 2001).

SUPPLEMENTARY REFERENCES

BEJERANO, G., M. PHEASANT, I. MAKUNIN, S. STEPHEN, W. J. KENT *et al.*, 2004 Ultraconserved elements in the human genome. Science **304:** 1321-1325.

BURGE, C., and S. KARLIN, 1997 Prediction of complete gene structures in human genomic DNA. J Mol Biol **268:** 78-94.

BURGE, C. B., and S. KARLIN, 1998 Finding the genes in genomic DNA. Curr Opin Struct Biol **8:** 346-354.

FRISSE, L., R. R. HUDSON, A. BARTOSZEWICZ, J. D. WALL, J. DONFACK *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet **69:** 831-843.

GARRIGAN, D., Z. MOBASHER, T. SEVERSON, J. A. WILDER and M. F. HAMMER, 2005 Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol **22:** 189-192.

HSU, F., W. J. KENT, H. CLAWSON, R. M. KUHN, M. DIEKHANS *et al.*, 2006 The UCSC Known Genes. Bioinformatics **22:** 1036-1046.

KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON *et al.*, 2002 A high-resolution recombination map of the human genome. Nat Genet **31:** 241-247.