SUPPLEMENTARY NOTE

Consideration of repetitive sequences, chromosome-specific distribution of UCEs, orthology, and conserved genomic regions. Since UCEs are almost entirely devoid of repetitive sequences, such as Alus and simple repeats, we conducted additional studies with the Scherer SDs wherein the random sequence sets were drawn from the non-repetitive portion of the genome (Supplementary Methods); even under these circumstances, UCEs are depleted among SDs ($P < 1E-6$). We also determined whether chromosome-specific patterns of UCE distribution underlie the depletion by conducting additional analyses with the Scherer SDs and HMR UCEs in which the number and lengths of randomly chosen sequences from any particular chromosome were matched with the number and lengths of UCEs on that chromosome. Depletion was slightly enhanced ($P < 1E-6$; data not shown), arguing against chromosome-specific effects. Depletion is also not an artifact of the requirement that UCEs reside in orthologous regions, since new sets of UCEs, generated by aligning whole genomes to each other without regard to orthologous blocks (Methods), are depleted among the Scherer SDs ($P < 1E-6$).

		Obs.		Expected			
UCEs	Genome	N	bp	Avg	StDv	Min	P
HMR	Entire	2	458	5,984	1,269	987	< 1E-6
	NR			5,860	1,260	931	< 1E-6
HDM	Entire	2	458	6,320	1,301	1,172	< 1E-6
	NR			6,185	1,289	759	< 1E-6
HC	Entire	1	232	5,277	1,184	623	< 1E-6
	NR			5,166	1,176	580	< 1E-6
Combined	Entire	3	690	11,374	1,778	4,082	< 1E-6
Unconstrained UCEs							
HMR	Entire	4	984	6,285	1,321	1,312	< 1E-6
	NR			6,163	1,315	1,404	< 1E-6
HDM	Entire	2	458	6,607	1,331	1,299	< 1E-6
	NR			6,470	1,319	1,440	< 1E-6
HC	Entire	3	758	5,511	1,240	768	< 1E-6
	NR			5,403	1,233	665	4E-6
Combined	Entire	5	1,216	11,835	1,831	4,417	< 1E-6

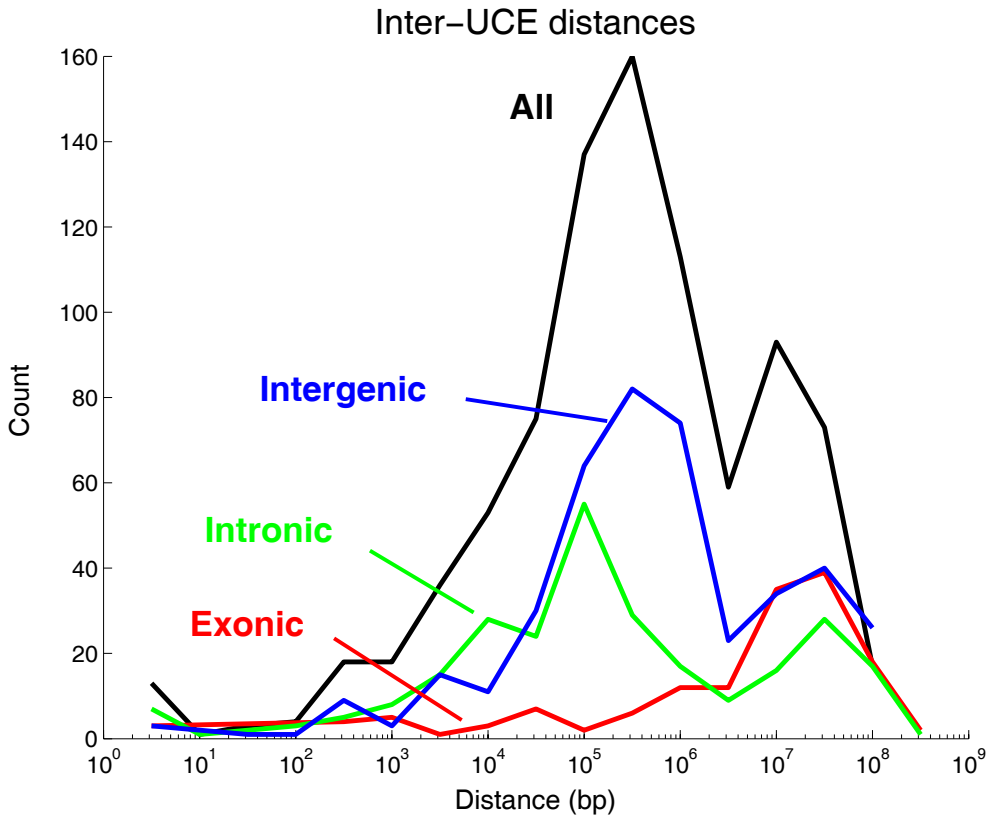
Additional analyses of overlaps of UCEs with human SDs. For original sets of UCEs (top) and new sets selected without regard for orthologous blocks (unconstrained, bottom), observed and expected overlaps with Scherer SDs. Analyses were conducted separately with sequences chosen randomly from the entire genome or from the non-repetitive (NR) portion of the genome. One million random iterations were conducted for each analysis.

To assess whether the depletion of UCEs within SDs is due to a depletion of SDs within conserved genomic regions, we repeated the random trials with random sequences drawn only from the 35% of the genome that is conserved (Supplementary Methods); depletion of UCEs among SDs remains significant under this constraint. The table below also shows that the depletion of UCEs among CNVs and DELs is not due to a depletion of CNVs and DELs within conserved genomic regions.

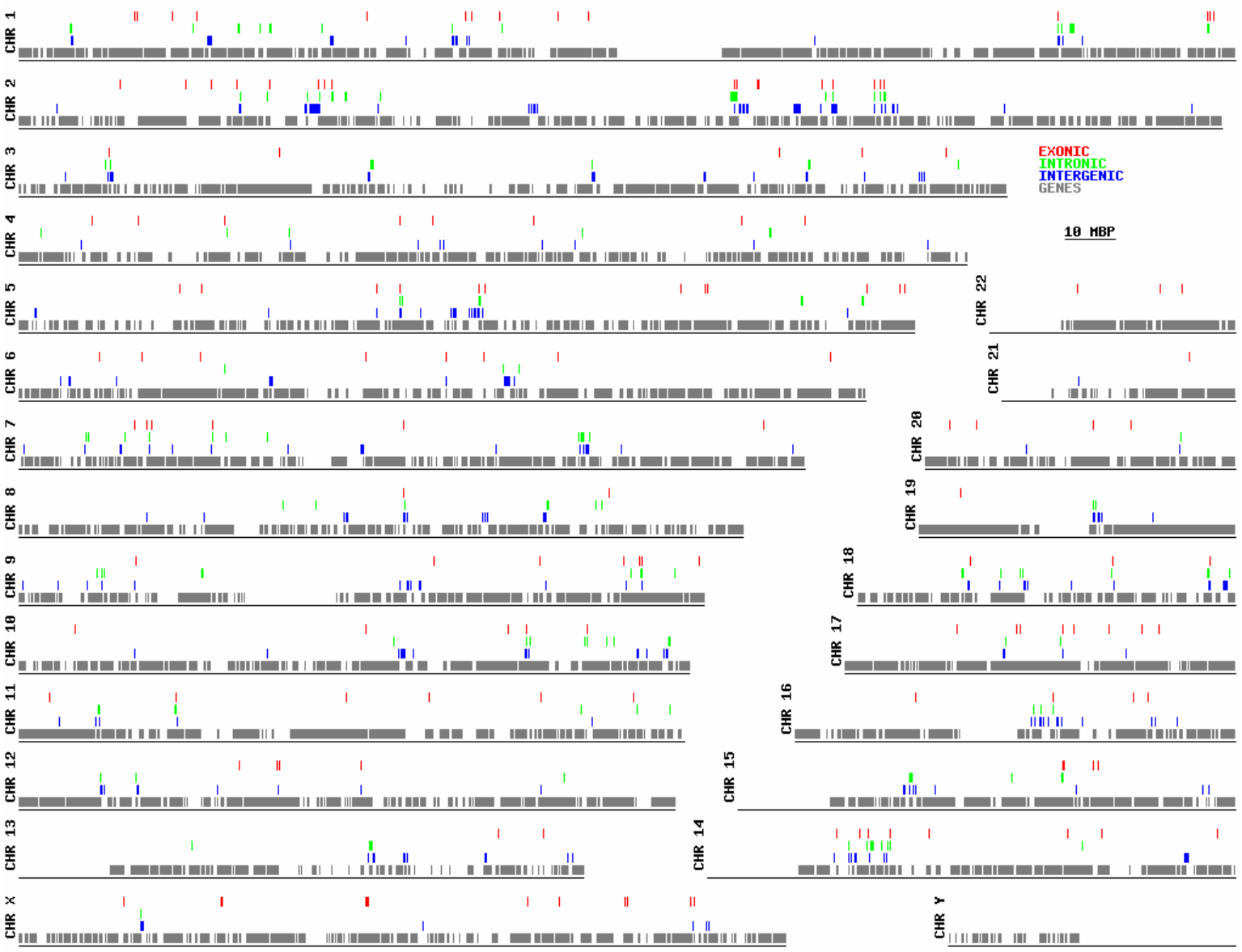
Dataset	Subset	Observed		Expected			P
		N	Bp	Avg	StDv	Min	
SDs		13	3,191	8,903	1,572	2,192	1.4E-5
CNVs	All	4	1,102	4,318	1,112	200	0.0003
	Duplications	2	414	1,635	689	0	0.0164
	Deletions	2	688	2,245	805	0	0.0133
DELs		0	0	1,520	662	0	0.0032

Overlaps in conserved regions of the genome. For pooled SDs, CNVs (excluding those of Sebat et al.[16]) and DELs, observed and expected overlaps of combined UCEs when all sequences, including random sets corresponding to UCEs, are confined to conserved genomic regions (1 Gbp; Supplementary Methods).

Clustering of UCEs does not drive depletion within SDs. Considering the different degrees of depletion observed for intergenic^{tr} and genic^{tr} UCEs, we compared their genome-wide distributions in humans. We found that intergenic^{tr} UCEs have a stronger tendency to cluster, as had been noted previously[1], raising the possibility that depletion may result from the clustered nature of UCEs.



Distribution of inter-UCE distances. Distribution of inter-UCE distances for all 896 UCEs and separately for exonic, intronic and intergenic^{tr} UCEs.



Genome map of the 172 exonic, 302 intronic and 422 intergenic[tr] UCEs.

We therefore reevaluated depletion by assuming that any two UCEs within a given distance of each other lie within a cluster and retaining the distances between them in the random trials (Supplementary Methods). These analyses were carried out using distance thresholds of 1 Kbp, 15 Kbp and 1.1 Mbp and the pooled set of SDs as well as the pooled set of CNVs. Regardless of the threshold used, depletion was maintained for the combined set of 896 UCEs as well as separately for the 422 intergenic[tr] UCEs. These data indicate that the clustering of UCEs does not explain their depletion among SDs and CNVs.

Dataset	Cluster dist.[1] (Kbp)	Obs N	Obs bp	Overlap Expected Avg	Expected StDv	Expected Min	P
SDs	0	13	3,191	13,549	1,925	5,347	< 1E-6
	1	13	3,191	13,547	2,065	5,276	< 1E-6
	15	13	3,191	13,520	2,379	4,745	< 1E-6
	1,060	13	3,191	10,276	3,139	1,661	< 1E-6
CNVs	0	4	1,102	4,942	1,188	206	4.3E-5
	1	5	1,411	5,654	1,359	748	6.7E-5
	15	5	1,411	5,654	1,593	466	3.1E-4
	1,060	5	1,411	5,041	1,990	0	0.0071
Intergenic[tr]							
SDs	0	0	0	7,488	1,397	1,778	< 1E-6
	1	0	0	8,076	1,449	1,775	< 1E-6
	15	0	0	14,245	1,867	6,559	< 1E-6
	1,060	0	0	99,777	1,705	90,296	< 1E-6
CNVs	0	2	456	2,769	870	0	0.0006
	1	2	456	2,819	879	0	0.0005
	15	2	456	3,440	967	0	6E-5
	1,060	2	456	33,081	2,555	20,556	< 1E-6

Impact of the clustered nature of combined UCEs on overlaps with pooled SDs and CNVs. Cluster dist., distance within which two UCEs were considered a cluster; [1]Actual distances used were 1,000 bp, 15,430 bp, and 1.06 Mbp. Intergenic[tr], analyses done with only intergenic[tr] UCEs and only within the intergenic[tr] portion of the genome. One million random iterations were conducted for each combination. Pooled CNVs excluded those of Sebat et al.[16].

Depletion of UCEs among SDs does not drive depletion among CNVs. Because CNVs have been reported to be enriched within SDs[15-18,20,22,23], it was possible that the depletion of UCEs among CNVs is driven by their depletion among SDs. To address this possibility, we repeated the random tests of overlap after subtracting all base pairs within pooled CNVs that were also contained in SDs (25% of all bp in CNVs). Significant depletion of UCEs within CNVs was retained even under these conditions (observed overlap: N = 4, 1102 bp; expected overlap: 3731 ± 1036 bp; P = 0.0015). The lower significance of this depletion as compared to that obtained for the depletion of UCEs among all CNVs (P = 4.3E-5; Table 3) is likely due at least in part to the smaller size of the dataset, since the average expected overlap decreased by 25% (from 4942 ± 1188 bp) while the observed overlap remained constant. These results suggest that the depletion of UCEs among CNVs is distinct from their depletion among SDs.

Consideration of ancient duplications, gene deserts, long non-repetitive fragments, recombination hotspots, and a new set of CNVs. We have found that UCEs are depleted among ancient duplications and unstable gene deserts but enriched in stable gene deserts, consistent with previous observations of ultra- as well as highly-conserved elements[1,3-5,7,25]. They are also enriched in non-repetitive fragments that are long compared to the distribution of all such fragments (unpublished). Whether these or other features can contribute to depletion and ultraconservation remains to be elucidated. Curiously, our preliminary studies suggest that UCEs are also enriched in recombination hotspots, raising the possibility that ultraconservation may result in part from higher rates of gene conversion and/or repair. Lastly, UCEs are depleted among a newly published set of CNVs[29].

		Length		Observed		Expected			
Dataset	Subset	N	Mbp	N	bp	Avg	StDv	Min (Max)	P
Ancient duplications[1]		15	31	0	0	2,623	871	0	6E-5
Gene deserts[2]	Unstable	356	493	5	1,265	41,651	3,173	26,226	< 1E-6
	Stable	169	205	212	55,449	17,320	2,167	(29,406)	< 1E-6
Hotspots[3]		21,980	233	95	24,454	19,755	2,289	(31,309)	0.0228
Locke CNVs[4]		189	35	1	219	2,908	917	0	9E-5

Overlap of combined UCEs with other genomic features. Size of datasets followed by observed and expected overlaps with combined UCEs. Max, maximum expected overlap is shown instead of minimum, and P-value is therefore that of enrichment rather than depletion. [1]Itoh et al.[30]; [2]Ovcharenko et al.[31]; [3]HapMap; Locke et al.[29]. Coordinates of all datasets were converted from a previous version of the genome as described in Methods.

Further discussion of models for depletion of UCEs among SDs and CNVs. A dosage-sensitive nature of UCEs, as proposed by the third model, is compatible with the retention of duplications or deletions that originally involved a UCE, should the change in copy number afford enough advantage to outweigh the initial deleterious aspects of gaining or losing a UCE. In this case, selective pressures could favor the degeneration of UCEs by random or targeted mutation (for example, see ref. 32). Duplicated UCEs could also evolve into new distinct UCEs. The 12 paralogous sets of UCEs reported by Bejerano et al.[1] are potentially relevant, since they originated from duplication events ~300 million years ago and then diverged even as each became ultraconserved.

There are a number of mechanisms through which copy counting and comparison could be accomplished, with those occurring via DNA:DNA, DNA:RNA or RNA:RNA[33] homology-based pairing interactions being perhaps the simplest; pairing would permit mutations, deletions and duplications of UCEs to be detected by some process as mismatched or unpaired UCEs. Furthermore, as the counting and comparison of UCEs would likely require precision, they would be anticipated to occur prior to S phase of the cell cycle, perhaps immediately after fertilization and before the first round of replication, in order to avoid complications arising from an increased genome size and the nonuniformity of replication across the genome. If rearrangement breakpoints are disruptive of UCE pairing, UCEs could also bias the genome against inversions and translocations, consistent with suggestions that conserved elements contribute to the maintenance of synteny through structural constraints imposed by their gene regulatory functions[3,5,7,8,34,35]. In fact, there are strong precedents for a role of pairing and, more

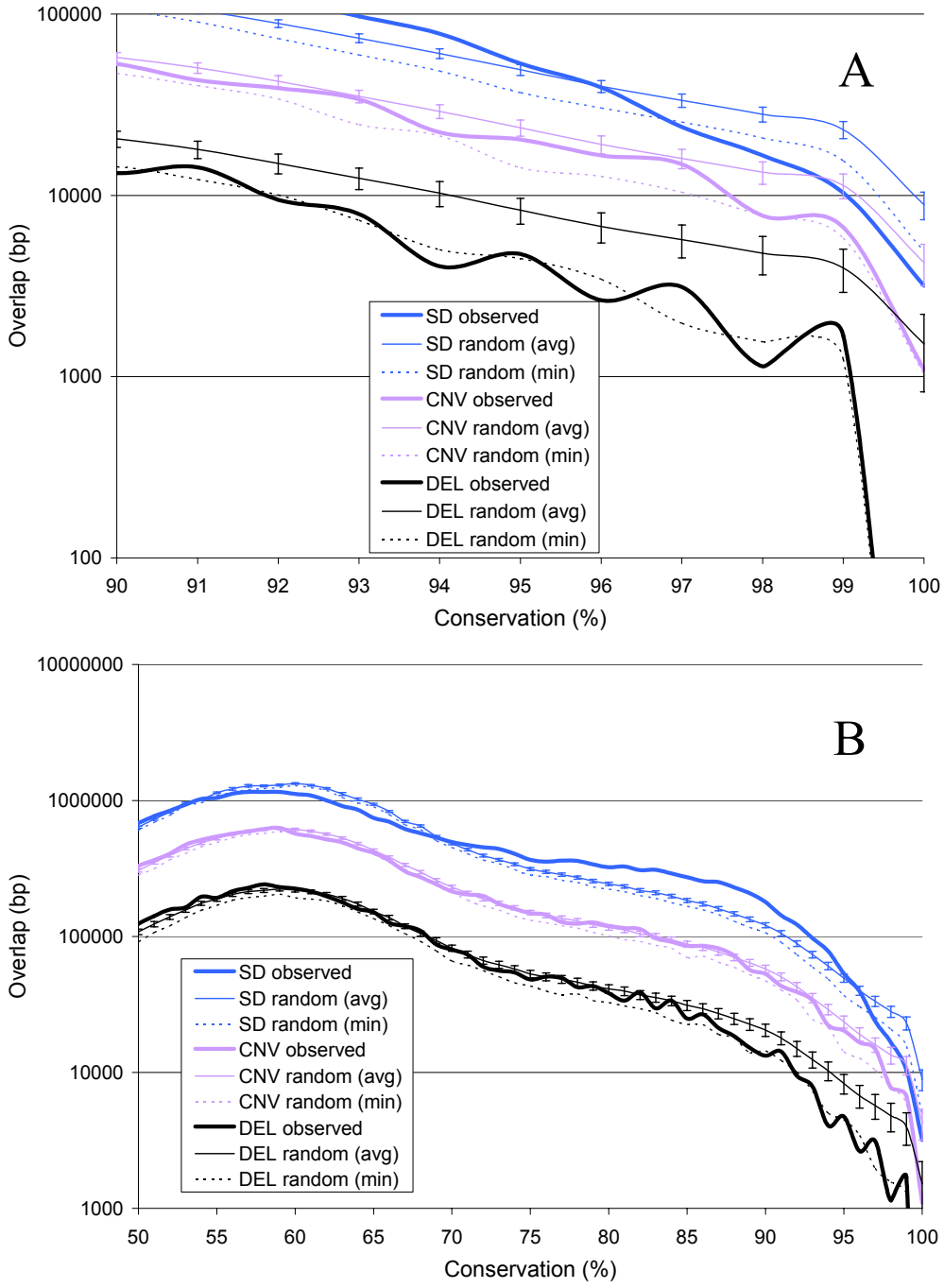
generally, of homology in gene expression and structure; pairing-modulated forms of gene regulation and host defense as well as other manifestations of homology effects have been reported in both meiotic and nonmeiotic cells of a wide variety of organisms[36]. Likewise, UCEs could act in the soma as well as germ line to modulate development, inheritance and host defense. Of special relevance is the phenomenon of meiotic silencing of unpaired DNA[37] (MSUD), wherein regions left unpaired during meiotic synapsis are silenced and can lead to deleterious phenotypes. Although the preponderance of polymorphic duplications and deletions among humans indicates that MSUD, alone, could not explain depletion, a restriction of MSUD to regions containing UCEs or an enhancement of MSUD in such regions would be consistent with our observations of depletion.

Note that since pairing requires at least a count of two, it would be difficult, should our interpretation be correct, to disentangle the mechanism of pairing from that of counting so as to determine whether the primary role of UCEs would be to pair or to count. For example, UCEs may play a role in the pairing of homologous chromosomes in meiotic or somatic cells regardless of any copy counting activity. Nonetheless, regardless of how counting might be achieved, our data suggest deleterious consequences when counts deviate from two. As for when the counting and comparison of UCEs might occur during the life cycle of an organism, we envision two points when assessment might be most effective: meiosis and fertilization. While assessment during meiosis would evaluate the genetic integrity of the germs cells of single individuals and lower the likelihood that deleterious alterations of the genome will be propagated, assessment at fertilization

would juxtapose genomes from two individuals, determining their compatibility such that investment in progeny of lowered fitness would be minimized and overall compatibility of the population at large maintained.

Evidence of depletion aside, there are observations that are not consistent with the third model in its simplest form. First, the X chromosome, which is present in only one copy in males, contains 33 of the 896 UCEs. This feature may indicate that, if UCEs are dosage-sensitive, the sensitivity does not pertain to the male X or is inhibited altogether in males. Second, the model does not address why UCEs are clustered in some regions but completely absent from others. It may be that UCEs are more effective when clustered, that regions lacking them are monitored for dosage by another mechanism, or that clustering reflects a function unrelated to dosage sensitivity. Alternatively, our selection of UCEs based on perfect conservation and a length requirement of $\geq$ 200 bp may have been too stringent, leaving functionally similar elements undetected; indeed, highly but imperfectly conserved sequences have been the focus of several recent studies[2-8,25,35,38,39]. In fact, we find that depletion of conserved elements among SDs, CNVs and DELs remains significant as conservation is lowered from 100% identity to 97%, 98% and 87% for analyses of depletion among SDs ($P < 0.001$), CNVs ($P \leq 0.002$) and DELs ($P \leq 0.03$), respectively (see Supplementary Methods for further discussion). This observation supports our model for the depletion of highly conserved elements among genomic features that alter the dosage of genomic segments, and provides an explanation for a report that imperfectly conserved elements are only rarely found as duplications[4], just as it provides an explanation for the single-copy nature of UCEs. Interestingly, preliminary

analyses of the 69 UCEs present in chimp, dog, mouse and rat but not in humans suggest that a conservation threshold of ~97-98% may be more appropriate; of these 69, the most poorly conserved, at 93.2% identity, is found in a segmental duplication, while the remaining 68 are conserved at 96.7% or higher identity and are not found in obviously duplicated regions (unpublished). Suggestively, this value of 96.7% matches our observation that depletion of imperfectly conserved sequences among SDs is lost when conservation is lowered below 97% identity. Third, we note that although UCEs are depleted among eight sets of SDs, CNVs and DELs, they are not depleted among the CNVs of Sebat et al.[16]; the basis for this difference is unclear.



Overlap of imperfectly conserved human sequences with SDs, CNVs and DELs. One thousand random trials were conducted for each dataset at each conservation threshold, therefore the minimum overlap from random trials (dashed lines) corresponds to a depletion with a significance of P = 0.001. Error bars for random overlap indicate one standard deviation. A, 90 to 100% and B, 25 to 100% conservation in decrements of 1%. SD, pooled SDs of Scherer and colleagues[12] and Eichler and colleagues[14]; CNV, pooled CNVs of Tuzun et al.[15], Iafrate et al.[17] and Sharp et al.[18] but not those of Sebat et al.[16]; DEL, pooled deletion CNVs of Hinds et al.[19], Conrad et al.[20] and McCarroll et al.[21].

Finally, while our models address ultraconservation as a consequence of negative selection and sequence constraints, it is also possible that influences other than these contribute to ultraconservation. For example, ultraconservation could reflect mechanisms that reverse the chemical reactions of mutagenesis, remembering and restoring wild-type sequences long after mutations occur[40] by, for example, marking mutated bases such that they are targeted for reversion, or distinguishing old from newly synthesized strands of DNA[41,42] such that base changes and replication errors are corrected rather than propagated. These forms of directional repair could also be enhanced by pairing, as a mutated base would manifest itself as the unambiguous odd-man-out in the set of four DNA strands constituting a paired set of UCEs.

SUPPLEMENTARY REFERENCES

29. Locke, D.P. et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* **79**, 275-90 (2006).
30. Itoh, T., Toyoda, A., Taylor, T.D., Sakaki, Y. & Hattori, M. Identification of large ancient duplications associated with human gene deserts. *Nat Genet* **37**, 1041-3 (2005).
31. Ovcharenko, I. et al. Evolution and functional classification of vertebrate gene deserts. *Genome Res* **15**, 137-45 (2005).
32. Galagan, J.E. & Selker, E.U. RIP: the evolutionary cost of genome defense. *Trends Genet* **20**, 417-23 (2004).
33. Lipman, D.J. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res* **25**, 3580-3 (1997).
34. Mackenzie, A., Miller, K.A. & Collinson, J.M. Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks? *Bioessays* **26**, 1217-24 (2004).
35. Boffelli, D., Nobrega, M.A. & Rubin, E.M. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5**, 456-65 (2004).
36. Wu, C.T. & Morris, J.R. Transvection and other homology effects. *Curr Opin Genet Dev* **9**, 237-46 (1999).
37. Shiu, P.K., Raju, N.B., Zickler, D. & Metzenberg, R.L. Meiotic silencing by unpaired DNA. *Cell* **107**, 905-16 (2001).
38. Glazov, E.A., Pheasant, M., McGraw, E.A., Bejerano, G. & Mattick, J.S. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res* **15**, 800-8 (2005).
39. Kamal, M., Xie, X. & Lander, E.S. A large family of ancient repeat elements in the human genome is under strong selection. *Proc Natl Acad Sci U S A* **103**, 2740-5 (2006).
40. Lolle, S.J., Victor, J.L., Young, J.M. & Pruitt, R.E. Genome-wide non-mendelian inheritance of extra-genomic information in Arabidopsis. *Nature* **434**, 505-9 (2005).
41. Potten, C.S., Owen, G. & Booth, D. Intestinal stem cells protect their genome by selective segregation of template DNA strands. *J Cell Sci* **115**, 2381-8 (2002).
42. Merok, J.R., Lansita, J.A., Tunstead, J.R. & Sherley, J.L. Cosegregation of chromosomes containing immortal DNA strands in cells that cycle with asymmetric stem cell kinetics. *Cancer Res* **62**, 6791-5 (2002).