nature genetics

Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants

Adnan Derti[1–4], Frederick P Roth[1,5], George M Church[2,3] & C-ting Wu[6]

An earlier search in the human, mouse and rat genomes for sequences that are 100% conserved in orthologous segments and ≥200 bp in length identified 481 distinct sequences[1]. These human-mouse-rat sequences, which represent ultraconserved elements (UCEs), are believed to be important for functions involving DNA binding, RNA processing and the regulation of transcription and development. *In vivo* and additional computational studies of UCEs and other highly conserved sequences are consistent with these functional associations, with some observations indicating enhancer-like activity for these elements[1–9]. Here, we show that UCEs are significantly depleted among segmental duplications and copy number variants. Notably, of the UCEs that are found in segmental duplications or copy number variants, the majority overlap exons, indicating, along with other findings presented, that UCEs overlapping exons represent a distinct subset.

Our study began with the observations[1] that human-mouse-rat (HMR) UCEs are essentially single copy in the haploid genome and that the only two chromosomes from which they are absent in humans are the Y, which is normally present in only one copy in males, and chromosome 21, trisomies of which constitute the most frequent viable whole autosomal aneuploidy. These observations suggested that, in addition to their other functions, UCEs and/or the regions containing them may be dosage sensitive and that this sensitivity contributes to the integrity of the diploid genome by ensuring the presence of UCE-containing regions in exactly two copies in a diploid cell. Notably, the broad distribution of UCEs implied a genome-wide process, and their ultraconservation and individual distinctiveness suggested a mechanism involving copy counting. We tested our hypothesis by determining whether UCEs are depleted among segmental duplications and human copy number variants (CNVs). For example, because the UCEs in our study predate the segmental duplications and CNVs we analyzed, presence of a UCE in even one copy of a duplicated region would argue against dosage sensitivity for that UCE, as it would indicate that the UCE had once been included in a viable duplication.

The segmental duplications analyzed in our study are thought to have occurred within the last 40 million years and constitute up to 5% of vertebrate genomes[10,11]. In particular, we considered segmental duplications identified in the human[12], mouse[13], dog and chicken (see Methods) genomes by Scherer and colleagues, who required ≥90% identity between fragments ≥5 kb in length, as well as segmental duplications identified in the human genome by Eichler and colleagues (who required >90% identity and >1 kb length)[14]. The combined human segmental duplications encompassed 160 Mb, or 5.6% of the genome (**Fig. 1** and **Table 1**).

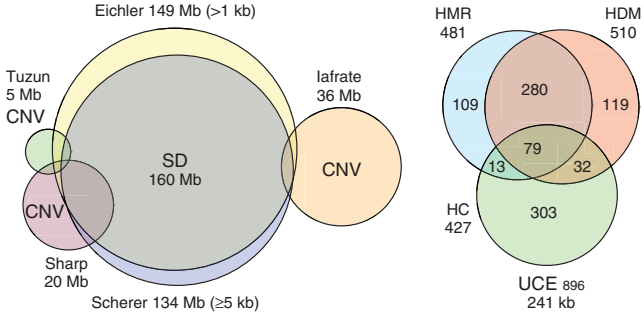

Figure 1 UCEs, segmental duplications (SD) and CNVs. Left, approximate relationships among the segmental duplications (SDs) and copy number variants (CNVs), excluding those of Sebat *et al.*[16]. See legend of **Table 1** for complete list of data sets, including the deletions (DELs, not shown here) of Hinds *et al.*[19], Conrad *et al.*[20] and McCarroll *et al.*[21] The CNVs of Iafrate *et al.*[17] overlap slightly with those of Tuzun *et al.*[15] and Sharp *et al.*[18] (not shown). Right, number and overlaps of ultraconserved elements (UCEs). Partial overlaps of <200 bp occur but are not included in the counts for this figure. Overlapping UCEs were counted as a single element, resulting in the total count (896) being less than the sum of all UCEs. HMR, human-mouse-rat; HDM, human-mouse-dog; HC, human-chicken.

[1]Department of Biological Chemistry and Molecular Pharmacology, [2]Department of Genetics and [3]Lipper Center for Computational Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA. [5]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [6]Divisions of Genetics and Molecular Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to C.-t.W. (twu@genetics.med.harvard.edu).

Received 30 March; accepted 23 August; published online 24 September 2006; doi:10.1038/ng1888

Table 1 Data sets of segmental duplications, CNVs, DELs and UCEs

Data set		Total			Length (kb)		Overlap (%)								
		N	Mb	%	Mean	s.d.	Eichler	Tuzun	Sebat	Iafrate	Sharp	Hinds	Conrad	McCarroll	UCEs
SDs	Scherer	3,750	134	4.70	36	92	92	29	16	6	49	3	15	15	0.3
	Eichler	8,033	149	5.23	19	59		36	18	7	51	3	17	18	1.3
CNVs	Tuzun	98	5	0.18	52	49			17	10	14	0	11	12	0.0
	Sebat	65	19	0.67	299	357				5	13	4	5	13	0.9
	Iafrate	236	36	1.26	152	35					4	1	3	5	0.5
	Sharp	111	20	0.70	182	84						2	5	10	0.0
DELs	Hinds	100	0.14	0.00	1	1							5	0	0.0
	Conrad	560	19	0.67	33	57								36	0.0
	McCarroll	539	9	0.32	17	51									0.0
UCEs	Combined	896	0.24	0.01	0.27	0.08									

The number of distinct elements (*N*), total length (Mb), corresponding percentage of the genome (%) and mean length (± s.d., in kb) are given for each data set. The number of elements and total bp may differ from published information because of losses due to conversion of data sets to the current human genome sequence, rejection of elements on unordered chromosomes, joining of overlapping elements and exclusion of Ns within sequences. Overlap was calculated as the percentage of bp in the smaller set contained in the larger data set. The amount of overlap between UCEs and segmental duplications (SDs), CNVs (except for those of Sebat *et al.*[16]) and DELs is less than expected. SDs, segmental duplications of Scherer and colleagues[12] and Eichler and colleagues[14]; CNVs, CNVs of Tuzun *et al.*[15], Sebat *et al.*[16], Iafrate *et al.*[17] and Sharp *et al.*[18]; DELs, deletions of Hinds *et al.*[19], Conrad *et al.*[20] and McCarroll *et al.*[21]; UCEs, combined ultraconserved elements.

Human CNVs are polymorphic among individuals and are believed to have arisen recently. We considered the deletions of Tuzun *et al.*[15] (5 Mb) and the deletions and duplications of Sebat *et al.*[16] (19 Mb), Iafrate *et al.*[17] (36 Mb) and Sharp *et al.*[18] (20 Mb). During the course of our analyses, Hinds *et al.*[19] (0.14 Mb), Conrad *et al.*[20] (19 Mb) and McCarroll *et al.*[21] (9 Mb) reported over 1,000 additional deletions, which we included in a subset of our studies, referring to them as DELs in order to distinguish studies that included them (**Fig. 1**, **Table 1** and Methods; reviewed in refs. 22, 23). Because the CNV data sets were generated with different approaches (reviewed in refs. 21, 23) and were limited in the number of individuals sampled, they represent just a fraction of CNVs that will eventually be identified. For instance, the CNVs of Sharp *et al.*[18] were sought in regions flanking duplications and represent 47 individuals, whereas those of Tuzun *et al.*[15] were found by aligning paired-end sequences from a single individual to the reference genome. Notably, the diverse nature of these data sets contributed power to our analyses, and the temporal complementarity of the relatively recent CNVs and older segmental duplications allowed us to query the impact of UCEs on genome evolution.

We assembled 481 HMR UCEs (concordant with ref. 1), 510 human-dog-mouse (HDM) UCEs and 427 human-chicken UCEs, in line with a report by the International Chicken Genome Sequencing Consortium (ICGSC)[24] (**Figs. 1** and **2** and **Supplementary Fig. 1** online) and then combined these three sets to form a fourth consisting of 896 distinct UCEs. Again, the only chromosomes lacking mammalian (HMR or HDM) UCEs are the Y and chromosome 21, although two human-chicken elements are found on chromosome 21 (**Supplementary Fig. 1**). For each set of UCEs, we then generated a million sets of sequences chosen at random from the genome, matching each random set with the UCE set being analyzed in terms of the number and length of elements. We then assessed the amount of overlap between UCEs and segmental duplications by determining whether it differed significantly from that of overlaps obtained at random.

Our data show a marked depletion of UCEs from segmental duplications in the human genome: all sets of UCEs, individually or combined, were depleted from the Scherer and Eichler segmental duplications, considered separately or together ($P \leq 1.3 \times 10^{-4}$; **Figs. 2** and **3**, **Table 2** and **Supplementary Fig. 1**). This depletion is not driven by the absence of repetitive sequences in UCEs, chromosome-specific patterns of UCE distribution, depletion of segmental duplications in conserved regions or requirement that UCEs reside in orthologous regions (**Supplementary Note** online). We also found that UCEs are depleted from segmental duplications in the mouse ($P \leq 0.0045$) and dog ($P = 0.0008$) (**Fig. 3** and **Table 2**). The increased *P* values for these depletions are due, in part, to a low genomic segmental duplication content, which may also explain the lack of the depletion of UCEs among chicken segmental duplications.

We next determined whether UCEs are depleted among CNVs (**Table 3**). Of the CNVs, the deletions of Tuzun *et al.*[15] were most appropriate because their determination through sequence comparison precluded underestimation of their boundaries. We found complete depletion among these deletions, although the depletion was not significant because of the small size of the data set ($P = 0.2$). We also queried three other sets of CNVs, including duplications as well as deletions, even though the extents of the CNVs had not been precisely determined. Although there was no depletion among the CNVs of Sebat *et al.*[16] ($P = 0.75$), we found significant depletion among the CNVs of Iafrate *et al.*[17] ($P = 0.01$) and Sharp *et al.*[18] ($P = 0.002$) (**Table 3**).

Because depletion could have resulted from underestimations of the lengths of some CNVs, we computationally extended the boundaries of the CNVs in both directions by 50, 100, 250 or 1,000 kb and reran our analyses. Although the number of overlaps increased with these parameters, the significance of depletion was maintained in all cases except when the CNVs of Iafrate *et al.*[17] were maximally extended (**Table 3**). By pooling the CNVs of Tuzun *et al.*[15], Iafrate *et al.*[17] and

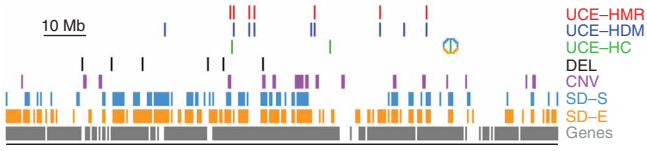

Figure 2 Genomic features on chromosome 17. Tick marks (not drawn to scale) indicate the locations of the elements identified on the right in the same color. UCEs overlapping other elements are encircled in the color of the other elements. DEL, pooled deletion variants of Hinds *et al.*[19], Conrad *et al.*[20] and McCarroll *et al.*[21]; 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]; SD-S, segmental duplications of Scherer and colleagues[12]; SD-E, segmental duplications of Eichler and colleagues[14]. See **Supplementary Figure 1** for the entire genome.

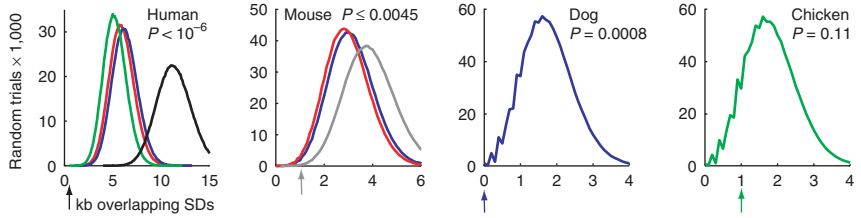

Figure 3 Mammalian UCEs are depleted among segmental duplications. Arrows indicate observed overlaps, in kb, of UCEs with genome-specific Scherer segmental duplications (SD), and curves show the overlaps of randomly chosen sequences with segmental duplications. Data from random trials with HMR UCEs are shown in red, HDM in blue, human-chicken in green, combined human UCEs in black and combined mouse UCEs in gray. For human and mouse, arrows and *P* values correspond to the worst case of any UCE set. These data are also shown in **Table 2**.

Sharp *et al.*[18] (59 Mb) and assaying those designated as duplications (30 Mb) separately from those designated as deletions (30 Mb), we further observed that depletions among duplications and deletions were of almost equal significance ($P = 0.008$ and $P = 0.006$, respectively; **Table 3**). These observations indicate that UCEs are strongly depleted among CNVs. Because CNVs represent recent events, this depletion cannot be easily explained by divergence and natural selection acting over long evolutionary time frames.

Finally, we considered the DELs discovered by Hinds *et al.*[19] using tiled oligonucleotide microarrays and those of Conrad *et al.*[20] and McCarroll *et al.*[21] from SNP analyses (**Table 1**). Because these DELs (~25 Mb) were identified through the use of oligonucleotides and SNPs, their endpoints are likely to be precise to within an average of 20 bp and 3 kb, respectively. We observed complete depletion among the pooled DELs ($P = 0.0004$) even when we extended the boundaries of the DELs by at least 3 kb and sometimes up to 9 kb in both directions (**Table 3**).

The rare overlaps we observed between UCEs and segmental duplications or CNVs present a notable pattern in that all 13 involving segmental duplications and two of the four involving CNVs (excluding those of Sebat *et al.*[16]) overlap exons. This pattern is consistent with the depletion of intergenic^{tr} and intronic but not exonic UCEs from the pooled sets of segmental duplications and CNVs when these classes of UCEs are considered separately (**Supplementary Table 1** online; a superscripted 'tr' designates words defined by transcript analysis rather than by the classical concept of the gene, which includes regulatory regions and the promoter; see **Supplementary Note** for consideration of the clustered nature of exonic, intronic and intergenic^{tr} UCEs). Additional studies indicate that exonic UCEs represent a distinct class, apparently reflecting a multiplicity of constraints, as has been proposed for UCEs in general[1,2,7,25]. First, compared with all transcribed regions, genic^{tr} UCEs exhibit a threefold enrichment in exonic sequences relative to intronic sequences ($P < 10^{-6}$; **Supplementary Table 2** online). Second, compared with all exons, exonic UCEs are enriched two- to threefold at overlaps of internal exons and 5′ UTRs, internal exons and 3′ UTRs, and 5′ UTRs and 3′ UTRs, with more pronounced enrichment if only cassette exons are considered ($P < 10^{-6}$; **Supplementary Fig. 2** and **Supplementary Table 2** online). UCEs were also slightly enriched within exons transcribed in both directions. Finally, consistent with the potential of exonic but not intronic or intergenic^{tr} UCEs to tolerate duplication, the best matches in the human genome to exonic UCEs show higher overall percentages of identity (≤98.3%) than matches to intronic (≤82.9%) and intergenic^{tr} (≤92.4%) UCEs (**Supplementary Fig. 3** online).

In summary, UCEs are depleted among eight of nine sets of segmental duplications[12,14] and CNVs[15,17–21], including DELs[19–21] (see **Supplementary Note** for demonstration of depletion among ancient duplications and a new set of CNVs). Below, we discuss three models to explain this depletion. Note that, because our data do not distinguish whether it is the UCEs or the regions containing them that are responsible for depletion, our use of 'UCE' will pertain to a UCE as well as the region in which a UCE resides.

Our first model proposes that UCEs prevent rearrangements from occurring or enhance their repair. UCEs may also be located in regions that are enriched in sequences that prevent rearrangements or that are deficient in sequences that promote rearrangements (see **Supplementary Note** for discussion of the relationship between segmental duplications and CNVs and discussion of other genomic features), although this explanation leaves open the question of how the enrichments or deficiencies could have occurred. The second model suggests that inclusion of UCEs in duplications leads to their mutation or loss. For example, if UCEs are multifunctional or nonessential, their depletion from duplicated regions could reflect subfunctionalization or loss upon duplication. This interpretation, however, leaves unanswered why UCEs are predominantly

Table 2 Overlaps of UCEs with segmental duplications

			Observed		Expected (bp)			
Genome	SD	UCEs	*N*	bp	Mean	s.d.	Min	*P*
Human	Scherer	HMR[a]	2	458	5,984	1,269	987	$<10^{-6}$
		HDM[a]	2	458	6,320	1,301	1,172	$<10^{-6}$
		HC[a]	1	232	5,277	1,184	623	$<10^{-6}$
		Combined[a]	3	690	11,374	1,778	4,082	$<10^{-6}$
	Eichler	HMR	7	1,810	6,625	1,329	1,423	6×10^{-6}
		HDM	11	2,600	6,998	1,365	1,714	1.3×10^{-4}
		HC	3	702	5,843	1,242	1,096	6×10^{-6}
		Combined	13	3,191	12,595	1,863	4,695	$<10^{-6}$
	Both[b]	Combined	13	3,191	13,548	1,928	4,143	$<2 \times 10^{-9}$
Mouse	Scherer	HMR[a]	4	1,095	3,247	944	0	0.0045
		HDM[a]	4	1,154	3,428	968	0	0.0037
		Combined[a]	4	1,154	3,930	1,045	245	0.0009
Dog	Scherer	HDM[a]	0	0	1,810	706	0	0.0008
Chicken	Scherer	HC[a]	3	1,003	1,866	715	0	0.1084

The observed overlaps as well as mean ± s.d. and minimum (Min) expected overlaps are given in bp. *P* values indicate the significance of the difference between the observed and expected overlaps. 'Combined' indicates the 896 UCEs obtained from the union of HMR, HDM and HC (human-chicken) UCEs. SD, segmental duplications of Scherer and colleagues[12] and Eichler and colleagues[14].
[a]Data are shown in **Figure 3**. [b]500 million random iterations.

Table 3 Overlaps of combined UCEs with CNVs

Data set		Ext.[a] (kb)	Observed *N*	Observed bp	Expected (bp) Mean	Expected (bp) s.d.	Expected (bp) Min	*P*
CNVs	Tuzun	0	0	0	428	353	0	0.2016
	Sebat	0	8	2,080	1,642	690	0	0.7513
	Iafrate	0	4	1,102	3,016	932	0	0.0101
	Sharp	0	0	0	1,655	693	0	0.0021
CNVs+Ext.[b]	Iafrate	50	11	2,667	4,975	1,194	662	0.0183
		100	12	3,004	6,923	1,402	1,564	0.0008
		250	33	9,118	12,586	1,867	4,354	0.0267
		1,000	125	34,345	38,400	3,073	24,812	0.0922
	Sharp	50	0	0	2,499	851	0	8.5×10^{-5}
		100	1	219	3,287	973	0	2×10^{-5}
		250	3	684	5,449	1,248	479	10^{-6}
		1,000	22	5,271	15,289	2,045	6,021	$<10^{-6}$
CNVs, pooled[b]	All	0	4	1,102	4,942	1,188	206	4.3×10^{-5}
	Duplications	0	2	414	1,855	733	0	0.0080
	Deletions	0	2	688	2,508	851	0	0.0064
DELs	Hinds	0	0	0	12	56	0	0.9498
	Conrad	0	0	0	1,579	677	0	0.0029
	McCarroll	0	0	0	777	474	0	0.0533
	Pooled	0	0	0	2,089	776	0	0.0004
DELs+Ext.[b]	Hinds	1	0	0	28	90	0	0.8921
		3	0	0	62	134	0	0.7870
		6	0	0	113	181	0	0.6517
		9	0	0	162	217	0	0.5434
		20	1	303	343	316	0	0.5594
	Conrad	1	1	278	1,674	696	0	0.0103
		3	2	590	1,861	733	0	0.0254
		6	4	1,124	2,140	786	0	0.0873
		9	6	1,696	2,420	837	0	0.1985
		20	8	2,446	3,428	993	0	0.1620
	McCarroll	1	0	0	868	502	0	0.0378
		3	0	0	1,048	551	0	0.0188
		6	0	0	1,317	617	0	0.0071
		9	0	0	1,583	677	0	0.0026
		20	2	624	2,546	856	0	0.0037

[a]Boundaries of elements were extended (Ext.) on both sides by the length specified. [b]Pooled CNVs include CNVs of Tuzun *et al.*[15], Iafrate *et al.*[17] and Sharp *et al.*[18] but not those of Sebat *et al.*[16]

altered in, or lost from, both copies of a duplicated region, and furthermore, it does not easily explain the depletion of UCEs among CNVs, whose origins may have been too recent to allow for significant divergence. It is also less attractive if UCEs are essential as well as intolerant of changes in sequence, as is suggested by their ultraconserved nature.

The third model proposes that duplications and deletions of UCEs are eliminated at the cellular or organismal level through lethality, segregation distortion or lowered fitness. For example, UCEs and/or the regions in which they reside may be dosage sensitive, consistent with the association of UCEs with specific classes of genes[1] (but see also **Supplementary Table 1**). A role for dosage sensitivity has also been hypothesized for conserved non-genic[tr] sequences, wherein the deleterious consequences of aneuploidy are proposed to arise from a *trans*-regulatory activity of these elements[26]. However, dosage sensitivity alone does not offer an immediate explanation for sequence conservation. In light of this, we suggest that UCEs may act through copy counting. Here, the maternal and paternal copies of UCEs could recognize each other directly or indirectly, perhaps by pairing, and could trigger deleterious events when irregularities of sufficient magnitude are detected in copy number or sequence. Notably, this interpretation raises the possibility that homozygosity for loss of a UCE may be less deleterious than heterozygosity would be, and it allows for ultraconservation, more than exact sequence, to be the key feature of UCEs (also see ref. 6). More notably, it simultaneously explains the single-copy nature and ultraconservation of UCEs, as significant deviations in UCE copy number or sequence would be eliminated from the population (see **Supplementary Note** for additional discussion). Note that participation of UCEs in copy counting can accommodate the involvement of UCEs in other functions as well. In fact, the enhancer-like activities of UCEs are consistent with a role of UCEs in pairing-mediated copy counting, as enhancers can participate in pairing-mediated phenomena[27,28].

Perhaps most notable is the depletion of UCEs among human CNVs. As the CNVs in our study were detected predominantly in healthy individuals, it may be that heterozygosity for duplications or deletions of UCEs is correlated with lowered fitness. Intolerance of copy number changes in UCEs may also confer a fitness advantage: if somatic loss of a UCE results in cellular lethality, then presence of that UCE may protect an individual from diseases, such as cancer, associated with loss of heterozygosity for that region.

METHODS

Databases. The genome sequences of human (hg17), mouse (mm6), rat (rn3), dog (canFam1) and chicken (galGal2) were obtained from the University of California Santa Cruz (UCSC), as were the pairwise alignments of human genomic regions with their orthologs (axtNet). The coordinates of nonrepetitive (upper-case), non-N fragments were derived from the human genome sequence. We ignored human chromosome sequences labeled 'random', which constituted a negligible fraction of the genome and did not contain UCEs, and discarded segmental duplications involving these sequences.

Human mRNA sequences in RefSeq release 15 and UniGene build 188 were obtained from the National Center for Biotechnology Information (NCBI) and aligned to the genome, and the boundaries of exonic, genic[tr] and intergenic[tr] regions were based on these alignments (**Supplementary Methods** online). Recombination rates and hotspots were obtained from the HapMap project.

Coordinates of segmental duplications, CNVs and DELs were obtained from sources cited in the text. Coordinates of human and mouse data sets using previous versions of genome sequences (hg16 and mm5, respectively), including HapMap data and deletions inferred from those[20,21], mouse segmental duplications[13], and the CNVs of Sebat *et al.*[16], Iafrate *et al.*[17] and Sharp *et al.*[18], were converted to current versions using the UCSC liftOver utility and corresponding chain files. For every set of coordinates, we joined overlapping fragments so as not to count overlaps multiple times. Additional steps included separating the individual copies of each segmental duplication, excluding variants of Tuzun *et al.*[15] spanning gaps and using the outermost coordinates for the deletions described by Hinds *et al.*[19] and Conrad *et al.*[20].

See **Supplementary Table 3** online for UCE sequences and genomic coordinates of all sequence elements and **Supplementary Methods** for genomic alignments of human mRNAs and analysis of exonic and intronic sequences.

Nomenclature of CNVs and DELs. Changes in the copy number of genomic regions are variably called CNVs, copy number polymorphisms (CNPs) and large-scale copy number polymorphisms or variants (CNPs, LCVs). Here, we use 'CNV' to refer to all the changes reported by Sebat *et al.*[16], Iafrate *et al.*[17] and Sharp *et al.*[18] as well as the deletion variants of Tuzun *et al.*[15]. We did not include the insertion variants of Tuzun *et al.*[15] because the boundaries of the inserted sequences had not been systematically determined. We also excluded the inversion variants from Tuzun *et al.*[15] because these do not change the copy number of chromosomal segments and, on their own, did not constitute a data set large enough to permit us to separately address the potential depletion of UCEs among inversions. We later expanded our study to include the deletion polymorphisms described by Hinds *et al.*[19], Conrad *et al.*[20] and McCarroll *et al.*[21] during the writing of this manuscript. We consider these variants to be CNVs, although we refer to them in this report as deletions, or DELs, in order to distinguish studies that involved them from those that did not.

Identification of UCEs. UCEs were identified by aligning orthologous blocks after removing gaps, or entire genomes where noted, using Mega BLAST with filtering disabled, a word length of 196 and a large mismatch penalty, followed by removal of matches shorter than 200 bp. HMR and HDM UCEs were obtained from the corresponding intersections of human-mouse, human-rat and human-dog UCEs. We joined overlapping UCEs so as to not count any UCE multiple times. To avoid occasional inconsistencies in coordinates of pairwise orthologous blocks, we aligned the human sequences of UCEs to the corresponding genomes to obtain their orthologous coordinates.

Determining the depletion of UCEs among segmental duplications, CNVs and DELs. To quantify expected overlaps, we kept segmental duplications, CNVs and DELs in place and randomly chose sequences from anywhere in the genome (excluding Ns), within nonrepetitive regions only or within conserved regions only (**Supplementary Note**), taking care to match the number and length distribution of the randomly chosen sequences with the number and length distribution of the set of UCEs in question and ensuring that the chosen sequences did not overlap each other. We then calculated the overlap with segmental duplications, CNVs and DELs in bp and number of fragments and then repeated the process. One million random trials were conducted except where noted in the text: namely, for combined UCEs against combined segmental duplications (5×10^8; **Table 2**) and for overlaps with imperfectly conserved sequences (10^3; **Supplementary Methods** and **Supplementary Note**).

URLs. Genome sequences, orthologous blocks, the liftOver utility and chain files are available at the UCSC Genome Browser (http://genome.ucsc.edu). The databases of segmental duplications reported by Scherer and colleagues can be obtained at http://tcag.ca, and human duplications published by Eichler and colleagues are found at http://humanparalogy.gs.washington.edu. The RefSeq and UniGene databases and the Mega BLAST and Splign programs are available from NCBI at http://www.ncbi.nih.gov. The HapMap website is http://www.hapmap.org.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank D. Haussler, G. Bejerano and M. Nobrega for valuable discussions; P. Green for introducing C.-t.W. to UCEs and suggesting they may pair; J. Aach, A. Dudley, H. Malik, S. Otto, J. Seidman, I. Yanai, members of the Church, Roth and Wu laboratories and attendees of the 2005 Epigenetics GRC for comments and ideas and D. Gurgul, Partners Research Computing at Massachusetts General Hospital, and the West Quad Computing Group and Research Information Technology Group at Harvard Medical School for computational resources. This work was supported by the Keck Foundation and by US National Institutes of Health (NIH) grants HG0017115 and HG003224 (F.P.R. and A.D.), by the NIH Centers of Excellence in Genomic Science (G.M.C. and A.D.) and by NIH grant GM61936 and HMS (C.-t.W. and A.D.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
2. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
3. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
4. Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99 (2004).
5. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
6. Poulin, F. *et al. In vivo* characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**, 774–781 (2005).
7. de la Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15**, 1061–1072 (2005).
8. Goode, D.K., Snell, P., Smith, S.F., Cooke, J.E. & Elgar, G. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86**, 172–181 (2005).
9. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
10. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
11. Bailey, J.A., Liu, G. & Eichler, E.E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).
12. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
13. Cheung, J. *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biol.* **4**, R47 (2003).
14. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
15. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
16. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
17. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
18. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
19. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
20. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
21. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
22. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
23. Eichler, E.E. Widening the spectrum of human genetic variation. *Nat. Genet.* **38**, 9–11 (2006).
24. ICGSC. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
25. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
26. Dermitzakis, E.T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
27. Duncan, I.W. Transvection effects in *Drosophila*. *Annu. Rev. Genet.* **36**, 521–556 (2002).
28. Kennison, J.A. & Southworth, J.W. Transvection in *Drosophila*. *Adv. Genet.* **46**, 399–420 (2002).