DOI: 10.1534/genetics.108.096537

Ultraconserved Elements: Analyses of Dosage Sensitivity, Motifs and Boundaries

Charleston W. K. Chiang,*,†,‡,1 Adnan Derti,§,1 Daniel Schwartz,* Michael F. Chou,* Joel N. Hirschhorn*,†,‡ and C.-ting Wu*,2

**Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, †Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, ‡Program in Genomics and Divisions of Genetics and Endocrinology, Children's Hospital, Boston, Massachusetts 02115 and §Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115*

Manuscript received June 20, 2008
Accepted for publication October 28, 2008

ABSTRACT

Ultraconserved elements (UCEs) are sequences that are identical between reference genomes of distantly related species. As they are under negative selection and enriched near or in specific classes of genes, one explanation for their ultraconservation may be their involvement in important functions. Indeed, many UCEs can drive tissue-specific gene expression. We have demonstrated that nonexonic UCEs are depleted among segmental duplications (SDs) and copy number variants (CNVs) and proposed that their ultraconservation may reflect a mechanism of copy counting via comparison. Here, we report that nonexonic UCEs are also depleted among 10 of 11 recent genomewide data sets of human CNVs, including 3 obtained with strategies permitting greater precision in determining the extents of CNVs. We further present observations suggesting that nonexonic UCEs *per se* may contribute to this depletion and that their apparent dosage sensitivity was in effect when they became fixed in the last common ancestor of mammals, birds, and reptiles, consistent with dosage sensitivity contributing to ultraconservation. Finally, in searching for the mechanism(s) underlying the function of nonexonic UCEs, we have found that they are enriched in TAATTA, which is also the recognition sequence for the homeodomain DNA-binding module, and bounded by a change in A + T frequency.

ALIGNMENTS of reference genomes representing distantly related species have identified thousands of ultraconserved elements (UCEs) that are 100% identical (Bejerano *et al.* 2004; Derti *et al.* 2006; Stephen *et al.* 2008), of which many are ≥200 bp in length. These UCEs, which can be intergenic, intronic, or exonic, are under negative selection (Drake *et al.* 2006; Chen *et al.* 2007; Katzman *et al.* 2007) and enriched in or near specific classes of genes (Bejerano *et al.* 2004). They are therefore likely to encode important functions, including the regulation of gene expression (Pennacchio *et al.* 2006; Paparidis *et al.* 2007; Visel *et al.* 2008). A number of laboratories have indeed demonstrated the ability of nonexonic (intergenic and intronic) UCEs to direct tissue-specific expression, and thus their capacity to act as enhancers (Pennacchio *et al.* 2006; Ahituv *et al.* 2007; Paparidis *et al.* 2007; Visel *et al.* 2008). What remains puzzling is the ultraconservation of these UCEs, since enhancers appear to be quite tolerant of sequence changes (Ludwig *et al.* 2005; Fisher *et al.* 2006; Li *et al.* 2008; McGaughey *et al.* 2008; Rastegar *et al.* 2008). The ultraconservation of nonexonic UCEs may therefore reflect a multiplicity of constraints, such as dual regulatory roles at the DNA and RNA levels (Feng *et al.* 2006) or a superimposition of binding sites for multiple transcription factors [Bejerano *et al.* 2004; Boffelli *et al.* 2004; De La Calle-Mustienes *et al.* 2005; Siepel *et al.* 2005; Derti *et al.* 2006; Pennacchio *et al.* 2006; Vavouri *et al.* 2007; also see related arguments for exonic UCEs (Derti *et al.* 2006)]. The latter explanation is consistent with the putative enhancer-like activity of nonexonic UCEs as well as other highly conserved noncoding elements (CNEs) (Woolfe *et al.* 2005; McEwen *et al.* 2006; Pennacchio *et al.* 2006; Ahituv *et al.* 2007; Paparidis *et al.* 2007; Visel *et al.* 2008) and has also been hypothesized for invertebrate highly conserved elements (Siepel *et al.* 2005; Vavouri *et al.* 2007). In support of these proposals, multiple transcription factor binding sites have been observed in an intronic UCE at the *GLI3* locus (Paparidis *et al.* 2007). However, an explanation involving overlapping binding sites for transcription factors would likely require a large number of such sites to maintain ultraconservation at all or most positions within a UCE. Alternatively, in conjunction with their enhancer activities, nonexonic UCEs may embody a separate function that constrains them at the sequence level.

[1] These authors contributed equally to this work.

[2] *Corresponding author:* Department of Genetics, 77 Ave. Louis Pasteur, NRB 264, Boston, MA 02115. E-mail: twu@genetics.med.harvard.edu

Our earlier study asked whether the ultraconservation of UCEs might reflect an evolutionary constraint in addition to enhancer and/or other functions. In particular, we hypothesized a model involving copy counting via sequence comparison (Derti *et al.* 2006; also see Vavouri *et al.* 2007 for a consideration of sequence comparison). According to this model, the maternal and paternal copies of each UCE are compared at the sequence level, perhaps through pairing, such that genomes harboring mismatches of sufficient magnitude or deviations from a copy number of two suffer lowered fitness and are lost from the population over time. Such a mechanism would explain ultraconservation as well as predict that UCEs are dosage sensitive. In line with this prediction, we found that the 896 UCEs representing human–mouse–rat (HMR), human–dog–mouse (HDM), and human–chicken (HC) sequence elements ≥200 bp long are significantly depleted among human segmental duplications (SDs) consisting of sequences that are duplicated in reference genomes and ≥90% identical between fragments ≥5 kb in length (Cheung *et al.* 2003) or >1 kb in length (She *et al.* 2004). Interestingly, the statistical evidence for the depletion is due primarily to depletion of the nonexonic UCEs. The nonexonic UCEs were also significantly depleted among CNVs, which include both duplications as well as deletions of genomic material. This finding was remarkable because CNVs are relatively recent and unlikely to have been subjected to as much natural selection as have the SDs in general; although some SDs are recent and may represent CNVs, other SDs can be as old as 40 million years and are fixed in the human population (reviewed in Bailey and Eichler 2006; Cooper *et al.* 2007). These observations of depletion are corroborated by a separate analysis of HMR UCEs and an independent CNV data set (Redon *et al.* 2006). We also found that the depletions among SDs and CNVs were maintained, although gradually less so, even when the level of conservation was lowered to ~97–98% identity (see Derti *et al.* 2006 for further discussion). This finding suggests that the mechanism of copy counting via comparison, if real, may not be uncompromising, consistent with the occurrence of SNPs in UCEs (Bejerano *et al.* 2004; Drake *et al.* 2006; Chen *et al.* 2007; Katzman *et al.* 2007) and the evidence that UCEs can drift (Derti *et al.* 2006; Stephen *et al.* 2008; Visel *et al.* 2008). It also predicts that genomes may occasionally endure a change in the copy number of a UCE, should the duplication or deletion of genomic material confer a sufficient fitness advantage.

Although our depletion studies are consistent with a model for copy counting via comparison, it remains open whether nonexonic UCEs are dosage sensitive and, if so, whether the apparent sensitivity contributes directly to ultraconservation. For example, depletion of nonexonic UCEs from SDs and CNVs may simply reflect the location of the UCEs in regions or genes that are dosage sensitive, or the possibility that nonexonic UCEs are the regulatory elements of dosage-sensitive genes. With regard to the latter, ultraconservation of UCEs would stem from the importance of their regulatory function or the function of the genes that they control, but could be independent of a dosage sensitivity of the UCEs *per se*. One prediction of the proposal that nonexonic UCEs embody important gene regulatory functions is that individuals lacking such a UCE should display a mutant phenotype.

In fact, researchers have generated mice heterozygous or homozygous for a deletion of any of four tested nonexonic UCEs, and these mice appeared to be phenotypically normal, with normal viability and fertility (Ahituv *et al.* 2007). While the robust nature of the heterozygous mice can be explained by the presence of one wild-type copy of the targeted UCE, the viability and fertility of the homozygous mutant mice argue against an immediately essential function of the UCEs as the basis for ultraconservation (Ahituv *et al.* 2007). It may be that the functions of these UCEs are more long term in nature, would be apparent only in a nonlaboratory setting, or are redundant within the genome of the mouse (Ahituv *et al.* 2007). Further, it is possible that the four tested UCEs are not representative of nonexonic UCEs in general. The outcome of these studies can also be reconciled with a model for copy counting via comparison, which predicts that loss of both the maternal and the paternal copies of a UCE could mitigate the deleterious consequences of the loss of a UCE (Derti *et al.* 2006). That is, the loss of both copies of a UCE would preclude the capacity of the genome to detect that loss through sequence comparison. Again, any reduction in fitness resulting from heterozygosity for a deletion may not be detectable in the time frame of a laboratory experiment; indeed, as our studies rest on the analysis of SDs and CNVs, they speak primarily to the long-term effects of duplicating or deleting UCEs and do not address the potential of copy number changes to produce an immediate consequence. They do predict, however, that heterozygous animals have a lowered fitness that, even if not immediately measurable, in the wild would ultimately cause the deletion either to be lost or to become fixed.

In this report, we provide new observations and then consider them in light of models invoking a gene regulatory function for UCEs as well as a model in which UCEs are proposed to partake in copy counting via comparison. We begin by further assessing the apparent dosage sensitivity of nonexonic UCEs, determining whether nonexonic UCEs are depleted among the most recent CNV data sets. Importantly, these data sets include three "second-generation" maps (Korbel *et al.* 2007; Kidd *et al.* 2008; Perry *et al.* 2008), each of which achieved a level of precision greater than that provided by earlier studies and therefore afforded our analyses a breadth and rigor exceeding that of our original report (Derti *et al.* 2006). We also address whether the depletion that we observed among CNVs as well as SDs stems from the UCEs *per se* or from a dosage sensitivity of the surrounding genomic region and then consider when in evolutionary history the apparent dosage sensitivity might have arisen. Finally, we assess whether nonexonic UCEs share sequence and structural features that are not characteristic of the exonic UCEs.

MATERIALS AND METHODS

Identification of UCEs: UCEs were identified in the manner previously described (Derti *et al.* 2006; see supplemental Table 1 for coordinates). Briefly, the genome sequences of human (hg17) and chicken (galGal2) as well as the pairwise alignments of human genomic regions with their mouse (mm6), rat (rn3), dog (canFam1), and chicken orthologs (axtNet) were obtained from the University of California at Santa Cruz (UCSC) Genome Bioinformatics site (http://genome.ucsc.edu). Ungapped sequences with perfect conservation between aligned orthologous blocks were identified, and sequences <200 bp were removed. HMR and HDM UCEs were obtained by identifying the corresponding intersections of human–mouse, human–rat, and human–dog UCEs and then applying the 200-bp length threshold.

Human–horse–mouse (HHrM), human–cow–mouse (HCowM), human–opossum (HOp), and human–platypus (HPl) UCEs were constructed as described above, using pairwise alignments of human with horse (equCab1), cow (bosTau2), opossum (monDom4), and platypus (ornAna1). In these cases, alignments were available only with human genome hg18, and thus the UCEs identified were mapped back to hg17 with the liftOver utility provided by UCSC.

To separate UCEs into intergenic, intronic, and exonic subclasses, human mRNA sequences in RefSeq release 15 and UniGene build 188 were obtained from the National Center for Biotechnology Information and aligned to the genome to establish the boundaries of exonic, genic, and intergenic regions [supplemental methods of Derti *et al.* (2006)]. mRNA alignments lying outside of Refseq boundaries were discarded. UCEs contained completely within intergenic and intronic spaces were deemed intergenic and intronic, respectively, while UCEs partially or completely overlapping exonic sequences were deemed exonic.

Imperfectly conserved sequences were defined using a protocol as similar as possible to that used to define UCEs (Derti *et al.* 2006). Briefly, within human–mouse, human–rat, human–dog, and human–chicken alignments, genomic segments 200 bp long and tiled every base pair were assigned a conservation score; for each nucleotide, a score of +1 was added for a match, 0 for a mismatch, and −1 for a gap. As was the case for UCEs, overlapping human–mouse and human–rat segments formed HMR segments, and the lower of the two conservation scores was retained. The same procedure was conducted to obtain HDM segments. The maximum level of HMR, HDM, and HC conservation was used as the final conservation score. Finally, fragments conserved at 100% identity were subtracted from those conserved at ≥99% identity and, among the latter, only those of at least 200 bp in length were retained. Likewise, fragments conserved at ≥99% identity were then subtracted from fragments conserved at ≥98% identity, and so forth in steps of 1% identity. The process was iterated until all desired levels of conservation were achieved.

Compilation of CNV, insertion/deletion, and SD data sets: Our analyses considered CNV data sets published between September 2006 and May 2008. The coordinates of CNVs, insertion/deletions (indels), and SDs were obtained from sources cited in the text, with the exception of Pinto *et al.* (2007) and Jakobsson *et al.* (2008), which were obtained from the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation), and the chicken SDs, which were obtained from a database maintained by E. E. Eichler and colleagues (http://eichlerlab.gs.washington.edu/help/eray/CHICKEN/chicken.html). Following the convention of DGV, CNVs >3 Mb in length were removed from all data sets, as these tend to be very rare and, when identified, often are due to cell-line artifacts or false positives (L. Feuk, personal communication). Overlapping CNVs were merged so that any overlap with other sequences would not be counted multiple times, and CNVs reported on unordered chromosomes were excluded. CNVs reported on a different genome build were mapped to hg17 with the liftOver utility. Coordinates on human chromosome sequences that were labeled "random" or unassigned on hg17 were discarded. If a data set reported whether its CNVs were recurrent, the recurrent CNVs and the nonrecurrent CNVs were also processed separately (removal of variants >3Mb, etc.).

Of the CNVs from De Smith *et al.* (2007), only those identified on the basis of multiple probes were included. Zogopoulos *et al.* (2007) and Jakobsson *et al.* (2008) reported only recurrent CNVs (observed in more than one individual in their study population). Of the data sets of Mills *et al.* (2006), Korbel *et al.* (2007), and Kidd *et al.* (2008), only deletions were considered, since insertions cannot consistently be assigned coordinates on the reference genome. For the deletion indels of Kidd *et al.* (2008), the lengths of the actual deletions were used rather than those of the affected regions, and coordinates were inferred by centering each deletion within the respective affected sequence. The data set of Pinto *et al.* (2007), obtained from the Database of Genomic Variants, merged a CNV data set obtained from 270 HapMap individuals with that obtained from 506 unrelated individuals of northern Germany.

Depletion analysis of UCEs, imperfectly conserved elements, and exons: Depletion analyses were conducted as previously described (Derti *et al.* 2006). To quantify expected overlaps of UCEs or imperfectly conserved elements with SDs, CNVs, and/or indels, random sets of nonoverlapping sequences matched in number and length to the set of conserved elements or exons in question were selected from anywhere in the genome (excluding N's) or a fraction thereof, depending on the analysis. Overlaps with CNVs, indels, and SDs were then calculated in terms of number of fragments and base pairs. This process was iterated 1000 times, leading to an expected distribution of overlaps, which was found to follow a normal distribution (data not shown). Finally, the observed overlap was compared to the expected distribution of overlaps to determine the significance of depletion, if any. To confirm that the expected overlaps followed a normal distribution, the overlap in base pairs of each of 1000 random sets of sequences, matched to the combined set of conserved elements in length and number, with the union of all 10 CNV data sets showing a depletion of UCEs (data not shown; see Table 1 for the list of all CNV data sets and Table 2 for those showing a depletion of UCEs), was compared to 1000 values sampled from a normal distribution with a mean and standard deviation matched to those obtained in the random trials. The two distributions of overlaps were not significantly different from the two-sample Kolmogorov–Smirnov test for the equality of distribution functions ($P > 0.55$) or the Skewness/Kurtosis tests for normality ($P > 0.31$). Because the expected overlaps followed a normal distribution, we were able to estimate the significance of depletion beyond the number of empirical trials by calculating the *Z*-score [(observed overlap − mean expected overlap)/standard deviation of null distribution], which was then converted into a *P*-value by the NORMDIST function in MS Excel. Random trials were conducted 10,000 times in a few instances, resulting in minimal changes in *P*-value (data not shown). For the overlap of conserved elements and CNVs, consistent results for the *P*-value and observed/expected ratio were obtained whether these were based on the number of overlapping base pairs or on the number of overlaps (data not shown).

In some analyses, random sequences were sampled from only specific fractions of the human genome. To assess the depletion of elements from the alignable portion of the genome, random sequences were drawn only from the portion of the human genome aligned with the mouse genome. Similarly, in analyses that stratified the conserved elements into intergenic, intronic, and exonic categories, random sequences were drawn from only the intergenic, intronic, and exonic spaces of the genome, respectively, while studies addressing the potential intrinsic dosage sensitivity of nonexonic elements conserved at a specific percentage of identity drew random sequences only from within genes containing those intronic conserved elements or within a specified distance (100, 250, 500 kb and 1 and 1.5 Mb) 5′ and 3′ from those intergenic conserved elements. To assess the depletion of exons of genes containing intronic conserved elements, random sequences were drawn only from the exonic spaces of the human genome.

To assess whether the physical clustering of some UCEs influenced their depletion within CNVs, we selected a distance threshold of 15 kb on the basis of the distribution of inter-UCE distances (Derti *et al.* 2006), progressively joined UCEs if they lay within this distance of their immediate neighbor (*i.e.*, a cluster could consist of multiple UCEs), and retained the actual distances when assigning random coordinates. No single or clustered random sequences were allowed within a cluster already placed randomly, and overlaps with CNVs were calculated for the UCEs and random sequences themselves, not for the lengths of the entire clusters.

Motif identification: Motif-x was designed to detect overrepresented patterns of amino acids surrounding phosphorylation sites (Schwartz and Gygi 2005). Its algorithm employs an iterative approach of building a position weight matrix to discover motifs and to assess their statistical significance. However, the input need not be amino acid sequences. In fact, the program has been shown to extract motifs from English words and, hence, was easily adapted to search for motifs in DNA sequences. TAATTA, the most prominent motif identified by motif-x in our studies, was also the most prominent motif detected when we used two other independent motif-searching algorithms, MDScan (Liu *et al.* 2002) and Weeder (Pavesi *et al.* 2004).

To find sequence motifs overrepresented in the intergenic, intronic, and exonic UCEs, we used a slightly modified version (see details below) of the motif-x algorithm. As input, the motif-x algorithm normally takes: (i) a central character; (ii) a foreground data set from which motifs are to be extracted; (iii) a background data set against which statistical comparisons are made; (iv) a motif width; (v) an occurrence threshold, which represents the minimum number of motif observations required; and (vi) a statistical significance threshold.

Here, motif-x was run four times for each analysis, with the central character being A, T, C, or G, and these other parameters: width = 17; minimum foreground occurrences = 200 for intergenic UCEs, 150 for intronic UCEs, and 80 for exonic UCEs; significance = 0.000001 (non-Bonferroni corrected); background = "foreground unaligned" with "subtract foreground from background" option checked. The "foreground unaligned" option allows users with a FASTA formatted foreground data set to use the foreground data set as the background. Since motif-x creates a justified foreground data set from the input sequences that is centered on a single position, the "foreground unaligned" option uses all n-mers (of the specified width) from the foreground data set as the background. Hence the algorithm consists of comparing position-centered data in the foreground to nonposition-centered data in the background. The "subtract foreground from background" option simply subtracts from the background those sequences that are also in the foreground. Thus, if the foreground data set were composed of "A" centered sequences, the background data set would be composed of C-, G-, and T-centered sequences. The "subtract foreground from background" option will be available on an upcoming new version of the motif-x website (http://motif-x.med.harvard.edu).

Two additional modifications to the currently available version of motif-x were used for our analyses. First, we used a modification that allows motif-x to run in breadth-first search mode instead of the standard depth-first search mode, effectively widening the number of significant search paths that can be followed during each iteration of motif building, and thereby allowing motifs to be built along more than one path. The breadth parameter was set to 4, indicating that at each iteration of motif building, motif-x retained up to four positions that exceeded the chosen significance threshold. The normal operation of motif-x follows only one significant position at each iteration, so this modification to motif-x allowed for greater sensitivity and confidence in the extracted motifs when found in more than one way. Second, the motif logos generated by motif-x were set to show both under- and overrepresented bases at each position. (For further discussion of motif-x, see http://motif-x.med.harvard.edu.

Sequence motifs identified by motif-x were then selected for further analysis if they contained: (a) at least four contiguous letters, (b) at least five letters with one substitution, or (c) at least six letters with up to two substitutions. These criteria were occasionally loosened to allow the retention of more putative motifs. For each motif detected within the intergenic, intronic, or exonic UCEs, we determined the number of UCEs of the corresponding class containing at least one instance of the motif and/or its reverse complement, as well as the total occurrences of the motif (and/or its reverse complement). Occurrences of the motifs were also determined in the 1-kb 5′ and 1-kb 3′ flanks of the UCEs. The number of times that a motif is expected to occur within the UCEs was then calculated by assuming a uniform distribution of the motif at the frequency observed in the 1-kb 5′ and 1-kb 3′ flanks. Fold enrichment was taken as the ratio of observed to expected instances of the motif. Motifs found in intergenic and intronic UCEs were reported if their fold enrichment was >1.3, while the enrichment threshold for exonic motifs was >0.96 in the interest of including TAATTA.

To find sequence motifs in the regions immediately flanking both boundaries of the UCEs, the two immediate flanks, each equivalent to one-half of the length of the UCE in question, were joined in pairs to be used as the foreground input for motif-x. All other parameters remained as described above. For each motif detected within the immediate flanks of UCEs, fold enrichment was calculated as the ratio of observed instances of the motif within the immediate flanks to the expected instances calculated on the basis of the frequency observed either within the UCEs themselves or in sequences distal to the UCEs, the latter sequences being the 1-kb regions flanking the UCEs minus the immediate flanks of the UCEs.

Selection of A + T content- and length-matched random sequences: One hundred sets of intergenic and intronic sequences matched in length to the intergenic and intronic

TABLE 1

Data sets of CNVs and indels

	Total			Length (kb)			
Source	*N*	Mb	%	Mean	SD	Subjects	Platforms and approaches
				CNVs			
Redon *et al.* (2006)	1,439	321	11.27	223	280	268	BAC array CGH, Affymetrix 500K array
Wong *et al.* (2007)	3,632	753	26.42	207	98	105	BAC array CGH
Simon-Sanchez *et al.* (2007)	184	63	2.19	340	365	272	Illumina SNP arrays
de Smith *et al.* (2007)	1,397	89	3.13	64	230	50	Agilent 185K, 244K array CGH
Zogopoulos *et al.* (2007)	512	170	5.98	333	569	1190	Affymetrix 100K, 500K SNP arrays
Korbel *et al.* (2007)[a,b]	614	35	1.22	57	279	2	454 paired-end sequencing
Pinto *et al.* (2007)	666	202	7.07	303	432	270/506	Affymetrix 500K SNP array
Wang *et al.* (2007)	1,075	54	1.90	51	132	112	Illumina HumanHap550 array
Jakobsson *et al.* (2008)	1,283	145	5.10	113	132	405	Illumina HumanHap550 array
Perry *et al.* (2008)[a]	2,945	102	3.58	35	137	30	Agilent custom array CGH
Kidd *et al.* (2008)[a,b]	747	39	1.35	52	69	8	Fosmid end sequencing
				Indels			
Mills *et al.* (2006)[b]	196,543	7	0.25	0.04	0.42	36	Sequence reads from SNP discovery
Kidd *et al.* (2008)[a,b]	716,496	11	0.39	0.02	0.01	8	Alignment of sequence reads

The CNV and indel data sets are ordered chronologically. The number of distinct elements (*N*), total length (Mb), corresponding fraction of the genome (%), mean element length (±SD), and total number of individuals assayed are given for each data set; these numbers were calculated after overlapping elements were joined and unsequenced bases were excluded. The number of elements and total length may differ from published information because we excluded elements on unordered chromosomes and those >3 Mb, joined overlapping elements and then converted coordinates to hg17 if necessary, and excluded unsequenced bases (see materials and methods for additional information).

[a] Data sets referred to as "second-generation."

[b] Only deletions were included.

UCEs, respectively, were randomly generated in the manner described above for the depletion analyses, after which the A + T content of each such sequence was calculated. For each UCE, we then selected five sequences whose A + T content was higher than, but as close as possible to, that of the UCE. If fewer than five such sequences occurred for a given UCE, we selected the five randomly chosen sequences with the highest A + T content. In these cases, the A + T contents were not appreciably below that of the corresponding UCE.

Statistical analysis: The Kolmogorov–Smirnov goodness-of-fit test, Skewness/Kurtosis tests for normality, the two-sample Wilcoxon rank-sum (Mann–Whitney) test, and the Wilcoxon signed rank-sum test were conducted in Stata/SE8.0 for Windows (Stata, College Station, TX).

RESULTS

Our studies began by addressing whether UCEs are depleted among the 11 recent genomewide data sets of CNVs involving multiple, apparently healthy subjects, in some cases numbering in the hundreds (Redon *et al.* 2006; De Smith *et al.* 2007; Korbel *et al.* 2007; Pinto *et al.* 2007; Simon-Sanchez *et al.* 2007; Wang *et al.* 2007; Wong *et al.* 2007; Zogopoulos *et al.* 2007; Jakobsson *et al.* 2008; Kidd *et al.* 2008; Perry *et al.* 2008) (Table 1; see supplemental Table 2 for additional descriptions). Of these 11, 3 derive from second-generation technologies that employed sequencing or targeted oligonucleotide arrays to determine the extent of CNV regions with unprecedented resolution (Korbel *et al.* 2007; Kidd *et al.* 2008; Perry *et al.* 2008). We excluded variants >3 Mb long following the conventions of the Database of Genomic Variants, but included those <1 kb long (indels) (Scherer *et al.* 2007), since the latter should still be large enough to disrupt putative mechanisms involving gene regulation or sequence comparison. Furthermore, we considered only the deletions of some data sets (Mills *et al.* 2006; Korbel *et al.* 2007; Kidd *et al.* 2008; Table 1), as the sequences or positions of the duplications or insertions in these data sets could not be determined systematically (materials and methods). The overlaps between each data set and the 896 UCEs were then determined and compared to those obtained in control runs, in which we determined the number of overlaps between each data set and a thousand sets of randomly chosen sequences matched with the UCEs in terms of number and length of elements. These analyses permitted us to estimate the statistical significance of depletion in our studies (materials and methods).

We found significant depletion among all but one CNV data set ($P < 10^{-9}$ to 0.0045 for combined UCEs; Table 2), where variabilities in the significance (*P*) and strength (observed/expected) of depletion may derive, at least in part, from differences among the data sets in terms of size and the technologies used to identify CNVs. Our analyses also considered two data sets reporting indels (Table 2), demonstrating significant depletion for the larger data set (Kidd *et al.* 2008) and a notable trend toward depletion, albeit not statistically significant, in the smaller data set (Mills *et al.* 2006). These findings

TABLE 2

UCEs are depleted among human CNVs and INDELs

		Observed		Expected (bp)				
Source	UCE subset	*N*	bp	Mean	SD	Minimum	*P*	Observed/expected
				CNVs				
Redon *et al.* (2006)	Combined	78	20,205	27,123	2,651	18,998	0.0045	0.74
	Intergenic	52	13,562	13,726	1,846	7,491	0.4646	0.99
	Intronic	9	2,221	7,420	1,483	1,981	0.0002	0.30
	Exonic	17	4,422	4,975	1,162	1,216	0.3171	0.89
Wong *et al.* (2007)	Combined	247	65,822	67,922	3,700	56,731	0.2852	0.97
	Intergenic	132	34,916	31,322	2,585	23,286	0.9178	1.11
	Intronic	71	19,910	23,813	2,453	15,638	0.0558	0.84
	Exonic	44	10,996	13,435	1,593	8,552	0.0629	0.82
Simon-Sanchez *et al.* (2007)	Combined	1	208	5,334	1,251	2,052	2.1×10^{-5}	0.04
	Intergenic	0	0	2,669	851	219	0.0009	0.00
	Intronic	0	0	1,498	681	0	0.0139	0.00
	Exonic	1	208	1,058	522	0	0.0517	0.20
de Smith *et al.* (2007)	Combined	12	2,649	7,534	1,409	3,888	0.0003	0.35
	Intergenic	3	528	3,403	918	1,041	0.0009	0.16
	Intronic	1	201	2,588	893	248	0.0038	0.08
	Exonic	8	1,920	1,537	599	0	0.7387	1.25
Zogopoulos *et al.* (2007)	Combined	13	3,139	15,490	2,049	8,695	8.3×10^{-10}	0.20
	Intergenic	10	2,383	7,454	1,446	3,727	0.0002	0.32
	Intronic	2	504	4,590	1,140	1,922	0.0002	0.11
	Exonic	1	252	2,976	872	457	0.0009	0.08
Korbel *et al.* (2007)	Combined	2	537	2,944	957	466	0.0059	0.18
	Intergenic	1	326	1,344	614	0	0.0487	0.24
	Intronic	0	0	1,041	576	0	0.0354	0.00
	Exonic	1	211	692	423	0	0.1277	0.30
Pinto *et al.* (2007)	Combined	40	10,479	17,050	2,239	10,965	0.0017	0.61
	Intergenic	29	7,397	9,039	1,517	4,880	0.1395	0.82
	Intronic	7	1,991	4,150	1,078	1,492	0.0226	0.48
	Exonic	4	1,091	1,939	684	206	0.1075	0.56
Wang *et al.* (2007)	Combined	5	1,254	4,585	1,136	1,370	0.0017	0.27
	Intergenic	1	207	2,491	844	428	0.0034	0.08
	Intronic	3	772	956	546	0	0.3681	0.81
	Exonic	1	275	684	423	0	0.1668	0.40
Jakobsson *et al.* (2008)	Combined	17	3,922	12,288	1,786	7,187	1.4×10^{-6}	0.32
	Intergenic	8	1,883	5,875	1,261	2,263	0.0008	0.32
	Intronic	7	1,630	3,786	1,024	1,109	0.0176	0.43
	Exonic	2	409	2,916	845	298	0.0015	0.14
Perry *et al.* (2008)	Combined	6	1,510	8,606	1,555	3,421	2.5×10^{-6}	0.18
	Intergenic	2	564	4,566	1,158	1,669	0.0003	0.12
	Intronic	0	0	1,766	740	0	0.0085	0.00
	Exonic	4	946	1,604	653	0	0.1568	0.59
Kidd *et al.* (2008)	Combined	1	290	3,233	982	930	0.0014	0.09
	Intergenic	0	0	1,770	686	0	0.0049	0.00
	Intronic	1	290	780	504	0	0.1655	0.37
	Exonic	0	0	525	363	0	0.0740	0.00
				Indels				
Mills *et al.* (2006)	Combined	3	228	590	389	20	0.1760	0.39
	Intergenic	2	227	273	244	0	0.4252	0.83
	Intronic	1	1	213	235	0	0.1835	0.00
	Exonic	0	0	135	187	0	0.2352	0.00

(*continued*)

TABLE 2

(Continued)

		Observed		Expected (bp)				
Source	UCE subset	*N*	bp	Mean	SD	Minimum	*P*	Observed/expected
Kidd *et al.* (2008)	Combined	2	3	126	39	38	0.0008	0.02
	Intergenic	1	1	56	27	7	0.0208	0.02
	Intronic	0	0	46	24	5	0.0276	0.00
	Exonic	1	2	14	13	0	0.1780	0.14

P-values indicate the significance of the difference between the observed and expected overlaps between UCEs and CNVs, where expected overlaps were determined by assessing the overlap between CNVs and 1000 sets of random sequences matched with the UCEs in number and length. Also shown are the number (*N*) of times a UCE overlaps a CNV or indel and the total number of base pairs (bp) in the overlaps in each analysis, as well as the mean (±SD) and minimum expected overlaps.

are in accordance with previous observations (Derti *et al.* 2006) and are particularly compelling, given the strong depletion observed among the second-generation CNV data sets. Importantly, the depletion cannot be fully explained by different representations of CNVs and UCEs in genomic regions that are alignable with the mouse genome (supplemental Tables 2 and 3), nonrepetitive (supplemental Table 2), or A + T rich (supplemental Figure 1), or by the positional clustering of some UCEs (data not shown; materials and methods; also see Derti *et al.* 2006).

We repeated the analyses separately for the 422 intergenic, 302 intronic, and 172 exonic UCEs, where intergenic UCEs are defined as those that do not overlap human mRNAs, intronic UCEs as those lying entirely within intronic sequences, and exonic UCEs as those that overlap an exon (materials and methods). Consistent with our earlier observations (Derti *et al.* 2006), these studies showed that, for any particular data set showing depletion, the statistical evidence for depletion was always driven by intergenic and/or intronic UCEs, but never by exonic UCEs alone, although exonic UCEs show depletion in some data sets. Note that the pattern of depletion among the three classes of UCEs differs among the CNV data sets; these variations may reflect the different methodologies used to obtain the CNV data sets. UCEs were also depleted when deletions and duplications were analyzed separately, even though the relative strength of depletion among deletions and duplications was somewhat variable among data sets (supplemental Table 4). Interestingly, depletion of exonic UCEs was significant among the duplications of some data sets. As exonic UCEs are not significantly depleted among human SDs, this finding may suggest that, if the duplication of exonic UCEs is deleterious, the negative consequences of such duplications are mitigated on the evolutionary timescale. Finally, we observed a trend toward greater depletion among recurrent CNVs, defined by their presence in multiple individuals (supplemental Table 5; also see supplemental Table 3). This finding may reflect a higher proportion of false positives among nonrecurrent CNVs or, as some nonrecurrent CNVs may represent relatively recent variants and/or those associated with deleterious phenotypes, indicate that nonrecurrent CNVs harboring UCEs have yet to be purged from the population. In line with this interpretation, our previous study found no overlap between human SDs and the intergenic and intronic UCEs [$P < 10^{-6}$, observed/expected = 0.00 for both classes of UCEs (Derti *et al.* 2006)].

We next addressed the generality of our findings by expanding our analysis to include additional sets of mammalian UCEs (materials and methods; Figure 1). Accordingly, we extracted and characterized 499 human–horse–mouse (HHrM), 457 human–cow–mouse (HCowM), 684 human–opossum (HOp), and 399 human–platypus (HPl) elements that are ≥200 bp long. We then determined whether these elements are depleted among the union of the three second-generation human CNV data sets (supplemental Table 6; see Table 1 and materials and methods for a full description of this union CNV data set) as well as among the union of all 10 CNV data sets showing a depletion of UCEs (data not shown; see Table 2 for a list of these sets). We observed significant depletion in all cases, again being driven primarily by the nonexonic UCEs. Not surprisingly, these new sets of UCEs are also strongly depleted among human SDs.

One interpretation of these findings is that the depletion of nonexonic UCEs among SDs and CNVs does not stem from a property or function of the UCEs *per se*, but is merely the consequence of the UCEs lying within or near regions of the genome containing dosage-sensitive genes or functions. In fact, and perhaps not surprisingly, the exons of genes containing intronic UCEs are depleted among human SDs (7506 bp of observed overlap, $P = 1.92 \times 10^{-4}$, observed/expected = 0.25), among the union of all three second-generation CNV data sets (2845 bp of observed overlap, $P = 8.39 \times 10^{-5}$, observed/expected = 0.12; see Table 1 and materials and methods for a full description of this union CNV data set), and among the union of all 10 CNV data sets showing a depletion of UCEs (data not shown; see Table 2 for a list of these sets). As such, we asked whether the apparent dosage sensitivity of UCEs is distinguishable from any that might be attributable to the genes or local genomic regions in which they reside. For these analyses, we considered the human SDs, the extents of which were determined through sequence analysis and thus are well defined (Cheung *et al.* 2003; She *et al.* 2004), and the nonexonic UCEs of our original 896 UCE data set. In particular, the overlap of intronic UCEs with SDs was compared to that of sequences chosen at random from within only those genes containing intronic UCEs. Although it was clear that the power of these analyses would be limited by the small number of genes containing intronic UCEs (123 genes, ∼35 Mb or 1.2% of the genome) and the fact that these genes themselves are depleted among SDs (147,827 bp of observed overlap, $P = 0.005$, observed/expected = 0.10), we nonetheless asked whether the intronic UCEs are locally depleted among SDs as compared to other sequences within these same genes. As our earlier study had shown that elements conserved at ∼97–99% identity are also depleted within SDs, suggesting that they may share a dosage sensitivity related to that of UCEs (Derti *et al.* 2006; also see Introduction), we also examined the potential local dosage sensitivity of imperfectly conserved intronic elements, with the requirement for identity lowered from 100 to 90% in steps of 1% (Table 3). Finally, parallel analyses were conducted for intergenic UCEs.

We found that depletion was absolute for the intronic UCEs as well as elements conserved at 99% identity (observed/expected = 0.00), with a lesser but still notable depletion for elements conserved at 98% identity (observed/expected = 0.49; Table 3). As predicted by the reduced power of these analyses, statistical significance was not achieved for these individual data sets, although the combined data set of all elements conserved at ≥98% identity was found to be significantly depleted ($P = 0.013$, observed/expected = 0.20). With regard to intronic sequences that are conserved at identities of ≤97%, depletion is much less pronounced or lost completely (Table 3), consistent with our previous observation of a threshold of ∼97–98% identity for the depletion of conserved elements among SDs when random sequences were drawn from the entire genome (Derti *et al.* 2006). These data suggest that the most highly conserved intronic elements may themselves be dosage sensitive. Interestingly, we have found a striking contrast between the intronic UCEs and the exons of genes containing intronic UCEs. Even though these exons are depleted among SDs when compared to all exons of the genome (see above), they are not depleted and, in fact, are enriched among SDs when compared to other sequences within those genes (observed/expected = 3.66).

Analyses were also conducted to determine whether highly conserved intergenic elements are locally depleted among SDs relative to their neighboring sequences. In this case, we compared the depletion of intergenic UCEs and imperfectly conserved intergenic elements with that of sequences randomly selected from anywhere within 100, 250, and 500 kb and 1.0 and 1.5 Mb of the elements, choosing to run this series of analyses because, in contrast to our studies of intronic UCEs, there was no obvious logic for the selection of a unit of interest equivalent to that of a gene for intronic UCEs. As no intergenic UCE overlaps any SD in the entire genome (Derti *et al.* 2006), depletion was absolute (observed/expected = 0.00) for these elements at all distances tested although, again, the statistical significance of these depletions was limited by the reduced power of our analyses (Table 3, supplemental Figure 2). We also assessed imperfectly conserved intergenic elements for local depletion and found moderate-to-weak depletion in some cases. In contrast to the pattern of depletion observed for intronic elements, where depletion was absolute (observed/expected = 0.00) even when the requirement for conservation was reduced to 99% identity, a reduction of 1% identity for intergenic elements resulted in a pronounced reduction of depletion regardless of the distance considered (*e.g.*, observed/expected = 0.90 and 0.68 at distances of 100 and 250 kb, respectively; supplemental Figure 2).

These results argue that nonexonic UCEs as well as intronic elements conserved at 99 and 98% identity are depleted among SDs as compared to sequences chosen at random from within the genes in which they reside or, with respect to intergenic elements, the local genomic environment (defined as the region lying within a minimum of 100 kb of the element). They therefore suggest that the apparent dosage sensitivity of these elements cannot be fully explained by a potential dosage sensitivity of the surrounding sequence and therefore may be a feature intrinsic to the elements themselves. These observations are compatible with a model of copy counting via sequence comparison, which predicts a dosage sensitivity of nonexonic UCEs *per se*. They are also consistent with a regulatory, enhancer-like property of nonexonic UCEs (Pennacchio *et al.* 2006; Ahituv *et al.* 2007; Paparidis *et al.* 2007; Visel *et al.* 2008), wherein duplication of these elements in *cis* or to distant sites could lead to deleterious gene expression.

Finally, we assessed the local depletion of intronic and intergenic UCEs among second-generation CNVs. While there was variability across the individual data sets of CNVs (data not shown), we observed depletion when considering the data sets in aggregate (supplemental Table 7). Importantly, depletion of intronic and intergenic UCEs from CNVs was strong (*i.e.*, low observed/expected) although, as predicted by the reduced power of these analyses, the significance of this observation was marginal at best. These results are comparable to those observed for the depletion of these elements among SDs (Table 3). Varying degrees of depletion were also observed for some sets of imperfectly conserved ele-

TABLE 3

Assessment of the local depletion of nonexonic conserved elements (including UCEs) among human SDs

Conserved elements		Observed		Expected (bp)				Observed/ expected	Conserved elements		Genes/ intergenic	
Subset	% identity[a]	*N*	bp	Mean	SD	Minimum	*P*		*N*	Mean (bp)	*N*	Mean (kb)
Intronic	100	0	0	379	328	0	0.124	0.00	302	285	123	282
	99	0	0	665	436	0	0.064	0.00	624	264	247	255
	98	2	456	933	477	0	0.159	0.49	730	246	321	264
	97	7	1,541	1,425	570	0	0.581	1.08	789	238	393	254
	96	6	1,424	1,730	628	206	0.313	0.82	845	232	441	254
	95	22	5,057	3,095	826	891	0.991	1.63	1,136	229	604	228
	94	35	7,977	2,525	764	637	1.000	3.16	1,276	226	710	230
	93	45	10,830	2,901	797	616	1.000	3.73	1,381	226	773	233
	92	59	13,065	3,434	867	1,064	1.000	3.80	1,650	224	927	218
	91	87	20,148	7,166	1,267	2,967	1.000	2.81	2,028	224	1,113	204
	90	108	24,716	6,582	1,197	2,923	1.000	3.76	2,233	223	1,207	206
Intergenic	100	0	0	718	432	0	0.048	0.00	422	264	239	249
	99	15	3,590	3,979	1,035	1,087	0.354	0.90	1,196	267	510	281
	98	32	7,524	7,560	1,313	3,403	0.489	1.00	1,606	244	721	275
	97	44	10,233	7,706	1,377	3,708	0.967	1.33	2,011	237	893	284
	96	78	18,112	12,065	1,648	7,265	1.000	1.50	2,393	231	1,097	277
	95	100	22,406	14,816	1,806	9,307	1.000	1.51	2,747	229	1,299	275
	94	160	35,756	20,291	2,203	13,803	1.000	1.76	3,417	227	1,575	278
	93	216	49,441	29,889	2,591	22,228	1.000	1.65	4,109	226	1,806	281
	92	285	64,626	37,751	2,879	28,483	1.000	1.71	4,750	225	2,089	284
	91	315	71,554	38,372	2,855	29,266	1.000	1.86	5,521	224	2,326	289
	90	423	95,786	53,817	3,460	43,936	1.000	1.78	6,372	223	2,547	298

For analyses of intronic conserved elements (including intronic UCEs), the 1000 sets of random sequences matched with the conserved elements in number and length were chosen only from within the genes containing the intronic elements conserved at the indicated percentage of identity. For analyses of intergenic conserved elements (including intergenic UCEs), the random sequences were chosen from anywhere within the 100 kb lying 5′ and 3′ of these elements (see supplemental Figure 2 for results of analyses using other distances). The number and mean length of the conserved elements are indicated, as are those of the genes or regions flanking intergenic elements. The SDs were taken from Scherer and colleagues (Cheung *et al.* 2003) and Eichler and colleagues (She *et al.* 2004).

[a] These data sets are not overlapping (materials and methods).

ments (see legend of supplemental Table 7 for discussion of the potential implications of this finding with regard to dosage-sensitive functions and the influence of natural selection).

We next turned our attention to when in evolutionary history the apparent dosage sensitivity of UCEs might have come into play. In particular, if dosage sensitivity is an intrinsic feature of UCEs, then we might expect the apparent dosage sensitivity to be as ancient as are the UCEs themselves. Specifically, as HMR UCEs became fixed ∼300–400 million years ago (Bejerano *et al.* 2004; Stephen *et al.* 2008), at about the time when the sauropsidian (bird and reptile) and mammalian lineages diverged (International Chicken Genome Sequencing Consortium 2004; Figure 1), might the dosage sensitivity of nonexonic UCEs have been in effect at that time? We had previously demonstrated a depletion of UCEs among the SDs of mouse and dog, whose common ancestor with humans dates to ∼90–95 million years ago (reviewed in Murphy *et al.* 2004), but the results were inconclusive with respect to the SDs of chicken (Derti *et al.* 2006). Taking advantage of a recent map of chicken SDs, we have now observed a significant depletion of UCEs (2090 bp of observed overlap, $P = 0.0010$, observed/expected = 0.35). Assuming a single origin of the apparent dosage sensitivity of UCEs, this finding suggests that such a dosage sensitivity of UCEs is at least as ancient as the divergence of sauropsids and mammals (Figure 1). This result is also consistent with that of an alternative approach, which examined the depletion of UCEs among human SDs that were defined by length and identity criteria lower than those generally used (Cheung *et al.* 2003; She *et al.* 2004) and therefore represented duplication events older than those previously analyzed (Derti *et al.* 2006; C. W. K. Chiang, unpublished results). These findings are remarkable given the distinct and strong forces that have shaped the evolution of the lineages (International Chicken Genome Sequencing Consortium 2004; Warren *et al.* 2008).

In brief, our analyses have raised the possibility that dosage sensitivity is an integral feature of the nonexonic UCEs *per se*. To better understand these elements, we

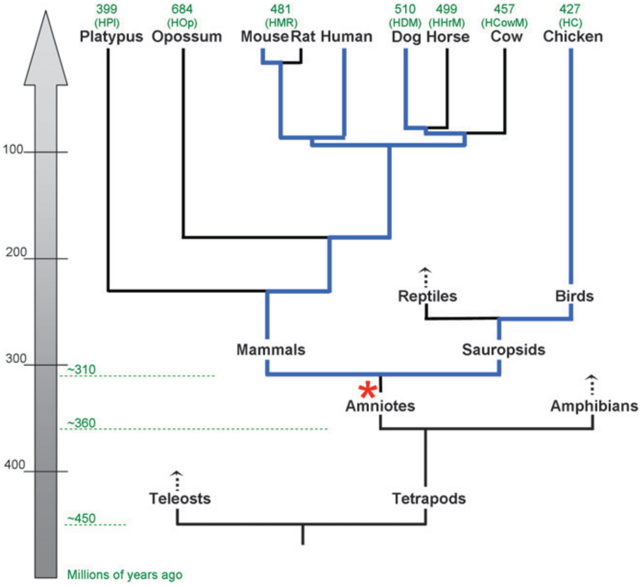

FIGURE 1.—Evolution of vertebrates. Major radiation events during vertebrate evolution are shown, with the number of millions of years since the divergence early in vertebrate lineages indicated (INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM 2004; MCEWEN *et al.* 2006; STEPHEN *et al.* 2008). The number of UCEs shared between the human, mouse, and a third species are shown in green above the designation for the third species except in the cases of the opossum, platypus, and chicken, whose genomes were aligned only with the human genome. HMR, HC, and HDM UCEs were reported previously (BEJERANO *et al.* 2004; INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM 2004; DERTI *et al.* 2006), and the remaining UCEs are reported in this study. On the basis of the evidence that nonexonic UCEs are depleted from human (DERTI *et al.* 2006 and this article), mouse (DERTI *et al.* 2006), dog (DERTI *et al.* 2006), as well as chicken SDs (this article), and assuming a single origin for the apparent dosage sensitivity of nonexonic UCEs, it is likely that a dosage sensitivity of nonexonic UCEs was in place before the mammalian and the avian lineages diverged (*) and persisted at least along the lineages highlighted in blue.

asked whether there may be structural features that distinguish them as discrete entities, distinct from flanking chromosomal regions (*e.g.*, see GARDINER *et al.* 2006). For example, because any chromosomal region in an individual is, barring structural changes, on average ~99.9% identical to its homolog (INTERNATIONAL HAPMAP CONSORTIUM 2005), a mechanism involving copy counting via comparison would likely have to distinguish a UCE from its flanking regions so that the comparison of the maternal and paternal copies is confined to the UCE. Furthermore, if the comparison of hundreds of nonexonic UCEs were to be accomplished by a single mechanism, the targeted UCEs would necessarily share features so that they could all be recognized by that single mechanism. Similarly, it may be that regulatory regions that are also ultraconserved share structural features that distinguish them from other regulatory elements. With this in mind, we examined the UCEs for sequence motifs and distinct boundaries. Note that our structural analyses were conducted with populations of UCEs in aggregate rather than with individual elements and are based entirely on primary sequence information. As such, they do not reflect the heterogeneity among members of any population or the numerous epigenetic mechanisms by which genetic elements can be modulated, considerations that lie beyond the scope of this report. Nevertheless, the analyses of UCE populations revealed interesting trends and commonalities.

We began by applying the motif-x program (SCHWARTZ and GYGI 2005), originally designed for the analysis of protein sequences, to search for nucleic acid sequence motifs in our initial set of 896 UCEs (MATERIALS AND METHODS). All motifs identified by motif-x as overrepresented in the UCEs were then validated by assessing their enrichment within UCEs as compared to their occurrence in the 1-kb regions flanking the UCEs. Several motifs were found to be enriched twofold or higher in the intergenic and/or intronic UCEs, while no motif achieved that level of enrichment in the exonic UCEs (Table 4). As the boundaries of UCEs could as easily be defined by motifs just outside the UCEs, we also identified motifs enriched in the regions flanking UCEs (Table 4 legend). Here, we focus on motifs within the UCEs.

The motifs most overrepresented in intergenic and intronic UCEs contain the sequence TAAT, the core recognition sequence for homeodomain-containing proteins (FRAENKEL *et al.* 1998). Of these, the most prominent among both the intergenic and intronic UCEs is TAATTA (Table 4; Figure 2A), which is also the recognition sequence for the engrailed homeodomain protein. Intriguingly, this motif is one of the 10 A + T-rich motifs enriched at the boundaries of CNEs identified through comparison of the human and Fugu genomes (ABNIZOVA *et al.* 2007) and also is embedded in several other longer motifs found enriched among mammalian CNEs (XIE *et al.* 2007). We also identified a number of other motifs, 3 of which are similar to motifs found in a separate study of human–Fugu conserved elements, including one UCE (PENNACCHIO *et al.* 2006).

The TAATTA motif is 3.5- and 2.3-fold enriched in intergenic and intronic UCEs, respectively. This level of enrichment for TAATTA is greater than would be expected from the A + T content of intergenic and intronic UCEs (data not shown) or for any other combination of three A's and three T's (supplemental Table 8). To address the possibility that the genome is generally enriched in TAATTA *vs.* any other hexameric combinations of three T's and three A's, we compared the distributions of TAATTA among intergenic and intronic UCEs to those in five randomly chosen sets of intergenic and intronic sequences matched in A + T content and length to the respective sets of UCEs (MATERIALS AND METHODS). These analyses showed a clear enrichment of TAATTA in the UCEs ($Z = 9.034$,

TABLE 4

Enrichment of motifs within UCEs and their flanking regions

	UCEs with motif		Instances in UCEs				UCEs with motif		Instances in UCEs		
Motif	*N*	%	Observed	Expected	Observed/expected	Motif	*N*	%	Observed	Expected	Observed/expected
			Intergenic						Intronic		
(C/G)*TAATTA*	127	30	229	55	4.14	*TAATTA*GG	14	5	14	5	2.89
*TAATTA*GA	31	7	33	9	3.79	*TAATTA*	*116*	*38*	*161*	*68*	*2.35*
TAATTA	*168*	*40*	*268*	*76*	*3.52*	C(A/T)GCCG	32	11	37	19	1.95
AATTA(A/G)*	301	71	749	287	2.61	GTAAT	219	73	407	216	1.89
CCATTA*	124	29	151	62	2.44	CTGCAG	50	17	53	31	1.68
TAAATGA*	67	16	80	39	2.06	TAATG	234	77	493	298	1.65
GNTAA(A/G/T)T	262	62	436	212	2.05	GC(A/C)GC	163	54	269	163	1.65
AAATNGC	121	29	157	85	1.85	GNAAATG	133	44	171	105	1.62
TTTATG	160	38	222	122	1.82	CNGCAG	133	44	232	144	1.61
T(A/C/G)TGACA	99	23	115	64	1.79	GTAAATG	62	21	69	44	1.55
TAAANTG	182	43	244	141	1.73	CTGAC	148	49	205	134	1.53
TTTGCA	160	38	228	135	1.69	TCATTT	174	58	257	169	1.52
T(G/T)ACAG	200	47	276	166	1.66	CTGCNA	173	57	252	167	1.51
TGTCA	274	65	472	287	1.64	CATTTC	107	35	148	102	1.44
TGCANA	278	66	634	395	1.60	CAGC	279	92	917	662	1.39
T(A/C)AT(G/T)T	362	86	912	577	1.58	CTGC	275	91	877	645	1.36
TAAAT	346	82	861	557	1.55	TTGAT	198	66	341	252	1.36
(C/T)TNAT	421	100	4325	2840	1.52	TTTGAT	112	37	144	107	1.35
CA(A/T)TT	399	95	1429	943	1.52	TNNCAG	294	97	1217	919	1.32
AAATNG	311	74	572	378	1.51	TAAATG	124	41	173	132	1.31
TTGAT	257	61	413	284	1.45						
TGAT	399	95	1380	997	1.38						
AAATG	348	82	827	620	1.33						
TTTGA(A/T)	242	57	360	273	1.32						
TNGCAG	226	54	320	245	1.30						
			Exonic								
GCAG(C/G)	124	72	246	189	1.30						
GAAGA	89	52	150	131	1.15						
ACAG	161	94	509	450	1.13						
ATGA	159	92	476	426	1.12						
CAC	172	100	1300	1165	1.12						
AGC	170	99	1355	1223	1.11						
GAGC	128	74	284	257	1.10						
ATG(A/C)	170	99	810	738	1.10						
ACA(C/G)	169	98	856	781	1.10						
CAG	172	100	1771	1630	1.09						
ACAA	159	92	564	521	1.08						
(C/G)ANC	160	93	649	603	1.08						
CNCCA	137	80	358	342	1.05						

(*continued*)

TABLE 4

(Continued)

	UCEs with motif		Instances in UCEs				UCEs with motif		Instances in UCEs		
Motif	*N*	%	Observed	Expected	Observed/expected	Motif	*N*	%	Observed	Expected	Observed/expected
(C/G)AANA	172	100	1178	1159	1.02						
ACA(A/G)A	148	86	369	370	1.00						
ACAAA	123	72	227	228	1.00						
AGG	171	99	1350	1386	0.97						
TAATTA	*22*	*13*	*27*	*28*	*0.96*						

The number (*N*) of UCEs that have at least one copy of the motif(s) identified by motif-x as well as the percentage (%) of the UCE subset that that number represents are shown. Also shown are the total observed and expected occurrences of the motif(s) in the UCEs, where the expected occurrences were determined by the frequency of the motif(s) in the 1-kb 5′ and 3′ flanks of the UCEs. Fold enrichment of motif(s) in the UCEs was calculated by dividing the observed by the expected occurrence. The intergenic, intronic, and exonic UCEs contain 422, 302, and 172 elements, respectively. TAATTA was not identified as a motif in the exonic UCEs, but is included for comparison. Motifs containing TAATTA are noted in italics. TAATTA is among 10 motifs enriched at the boundaries of CNEs (Abnizova *et al.* 2007) and is found in longer motifs enriched among mammalian CNEs (Xie *et al.* 2007). Three other motifs (*) were found in a separate study of human–Fugu conserved elements, some of which are HMR UCEs (Pennacchio *et al.* 2006). Motif discovery was also conducted for sequences flanking both sides of the UCEs, each flank of a UCE being one-half of the length of the UCE. Here, enrichment for a motif was calculated in two ways: relative to its occurrence in sequences distal to the flanks and relative to its occurrence within the UCEs (materials and methods). Relative to occurrences distal to the flanks, the motifs most enriched in the flanks of intergenic, intronic, and exonic UCEs were CGCCG (1.64-fold enriched), CGCAG (1.89-fold enriched), and GCTCC (1.21-fold enriched), respectively. Relative to occurrences within UCEs, the motifs most enriched in the flanks of intergenic UCEs were CGCCG (4-fold enriched), AA(A/G)AAAA (2.89-fold enriched), CCCG (2.34 fold enriched), and CCACC (2.14-fold enriched); in the flanks of intronic UCEs was CCCC (1.78-fold enriched); and in the flanks of exonic UCEs, was GCTCC (1.23-fold enriched).

$P \ll 10^{-15}$ for intergenic UCEs, Figure 2B; $Z = 8.133$, $P \ll 10^{-15}$ for intronic UCEs, data not shown). This enrichment was maintained even if one degeneracy in either of the final two noncore positions of the motif was allowed ($Z = 15.050$, $P \ll 10^{-15}$ for intergenic UCEs, Figure 2C; $Z = 12.929$, $P \ll 10^{-15}$ for intergenic UCEs, data not shown).

The TAATTA motif is found in 39.8% of intergenic and 38.4% of intronic UCEs as compared to 14.5 and 13.0% of length- and A + T-content-matched random intergenic and intronic sequences, respectively. Allowing degeneracy in one of the two noncore positions, the frequencies reach 94.7 and 93.0% as compared to 77.9 and 81.8% of length- and A + T-content-matched random sequences, with enrichments of 2.1- and 1.7-fold over flanking sequences, respectively. Finally, we have found that the frequency of TAATTA increases at the transition from flanking sequences into the nonexonic UCEs (supplemental Figure 2), suggesting that the boundaries of nonexonic UCEs may be biologically significant.

To better characterize the boundaries of UCEs, we assessed the base composition of the UCEs and their flanking regions. As A + T-rich motifs and overall A + T richness can influence DNA topology, nucleosome positioning, and higher-order chromatin structure (Minsky 2004; Segal *et al.* 2006), this analysis also addressed the potential for one or more of these chromosomal features to distinguish UCEs from their surrounding genomic regions. First, we found that intergenic and intronic UCEs are similar in A + T richness (∼0.63) and are both significantly more A + T rich relative to their flanks, the flanks being computationally designated as one-half of the length of the UCEs (0.63 *vs.* 0.59, $P \ll 10^{-15}$ and 0.63 *vs.* 0.62, $P = 0.0126$ for intergenic and intronic UCEs, respectively, by paired Wilcoxon signed-rank test; also see Figure 3). In contrast, the exonic UCEs are less A + T rich than both the intergenic and intronic UCEs and, furthermore, are significantly less A + T rich than are their flanks (0.58 *vs.* 0.60, $P = 0.0236$ by paired Wilcoxon signed-rank test; also see Figure 3).

We also paralleled a study that had revealed a sharp drop in the A + T frequency at and just beyond the boundaries of CNEs (Walter *et al.* 2005; Vavouri *et al.* 2007). As detailed below, we found a similar drop for the nonexonic UCEs but, strikingly, not for the exonic UCEs. Here, we considered the intergenic, intronic, and exonic UCEs separately. We calculated the A + T frequency for each position of the 50 and 100 bp at the ends and center, respectively, of the UCEs in addition to that of the 1 kb lying 5′ and 3′ to them (Figure 3). We found that the boundaries of intergenic and intronic UCEs, when assayed as a population, are marked by a sharp drop in A + T frequency just beyond the UCEs. The drop is most dramatic in the case of intergenic UCEs, where the A + T frequency of the

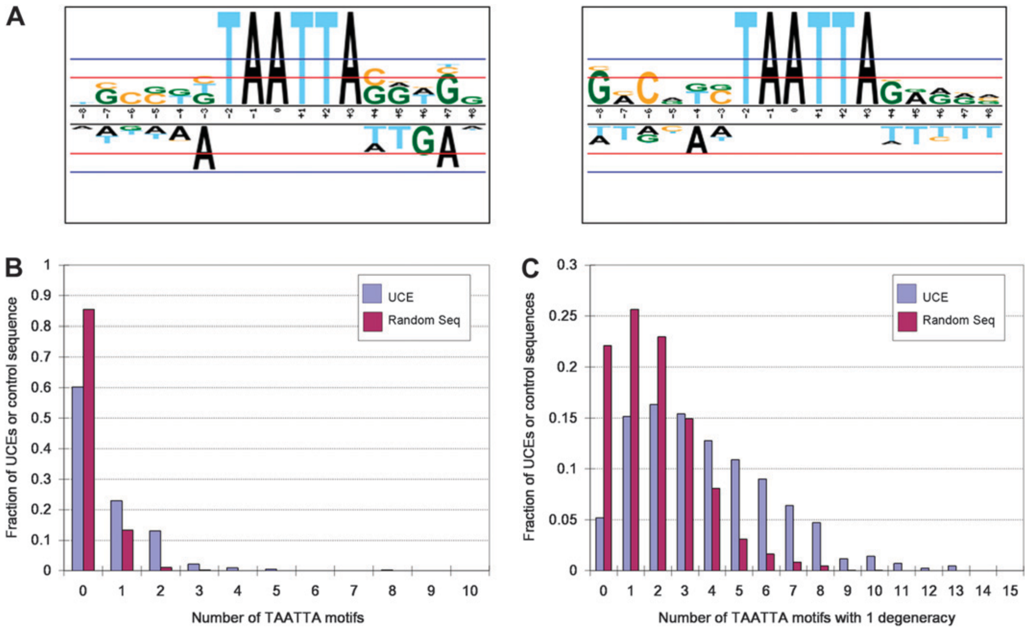

FIGURE 2.—TAATTA is enriched in intergenic and intronic UCEs. (A) Examples of motif-x sequence logo outputs of the TAATTA motif detected in intergenic (left) and intronic (right) UCEs. The vertical axes are in units of −log(binomial probability), which is a measure of the statistical significance of nucleotides at each position in the motif. Thus, the heights of nucleotides above and below the two black midlines are proportional to their binomial probability of over- and underrepresentation in the UCE data set, respectively, with more significant nucleotides positioned closer to the midline. Positions fixed by motif-x (*e.g.*, positions −2 through +3 in the examples shown) do not have additional nucleotide options. To become fixed, a nucleotide must both exceed the top blue line in height ($P < 0.0001$ after Bonferroni correction) and meet the user-defined occurrence threshold (MATERIALS AND METHODS). The red line indicates the threshold necessary for a nucleotide to achieve statistical significance ($P < 0.05$ after Bonferroni correction). (B and C) TAATTA is enriched among intergenic UCEs compared to randomly chosen length- and A + T-content-matched control sequences (MATERIALS AND METHODS). By a two-sample Mann–Whitney test, the distributions of the numbers of TAATTA (B) or TAATTA with one degeneracy in either of the last two positions (C) are clearly shifted to the right for the UCEs (blue bars) relative to the matched random sequences (red bars) ($Z = 9.034$, $P \ll 10^{-15}$ for TAATTA; $Z = 15.050$, $P \ll 10^{-15}$ for TAATTA with one degeneracy). These data attest to an enrichment of TAATTA in intergenic UCEs. Similar results were obtained from our analyses of intronic UCEs (data not shown).

flanking sequences is notably lower than that of the UCEs. It is also evident for intronic UCEs, although the A + T frequency gradually rises over a region of ~200 bp extending beyond the UCEs to approximately that of the UCEs (Figure 3). The exonic UCEs stand in stark contrast to the nonexonic UCEs, as they are not bounded by a sharp drop in A + T frequency. Instead, they exhibit a gradual decrease in A + T frequency within the UCEs. This pattern was generally maintained even when we separately analyzed the boundaries of only those exonic UCEs lying entirely within exons or only the exonic boundaries of UCEs spanning exon–intron junctions (data not shown).

Interestingly, we continue to detect changes in A + T frequency at the boundaries of elements as we lower the requirement for conservation to 90% identity and less (supplemental Figures 3, 4, and 5). This finding is consistent with observations of CNEs (WALTER *et al.* 2005; VAVOURI *et al.* 2007) and suggests that a change in A + T frequency may be a general marker for the boundaries of conserved elements (also see WALTER *et al.* 2005; VAVOURI *et al.* 2007), as it has been proposed for the punctuation of transcription units and perhaps other discrete genetic elements as well (ZHANG *et al.* 2004). Significantly, the drop in A + T frequency immediately flanking the nonexonic UCEs is seen with nonexonic elements conserved at identities as low as 80% but then is no longer obvious at ~75% identity, suggesting that this pattern may be a characteristic feature of highly conserved nonexonic elements. Importantly, however, both the intergenic and intronic UCEs have higher A + T content than all the corresponding elements conserved at lower identity, suggesting that a high A + T content in conjunction with a drop in A + T frequency at the boundaries may be a distinguishing signature of the nonexonic UCEs.

DISCUSSION

We have found that UCEs are depleted among 10 of the 11 most recently reported data sets of CNVs and that the significance of depletion is driven by the nonexonic UCEs, strongly confirming previous observations (DERTI *et al.* 2006). In addition, highly conserved nonexonic elements appear to be dosage sensitive themselves relative to their surrounding regions, the strength of depletion being strong and/or absolute for UCEs and elements conserved at 99 and 98% identity when considering intronic elements, while, perhaps interestingly, being strong only for the UCEs when considering intergenic elements. By demonstrating that nonexonic UCEs are depleted from SDs of chicken, we also provide evidence that a dosage sensitivity of nonexonic UCEs

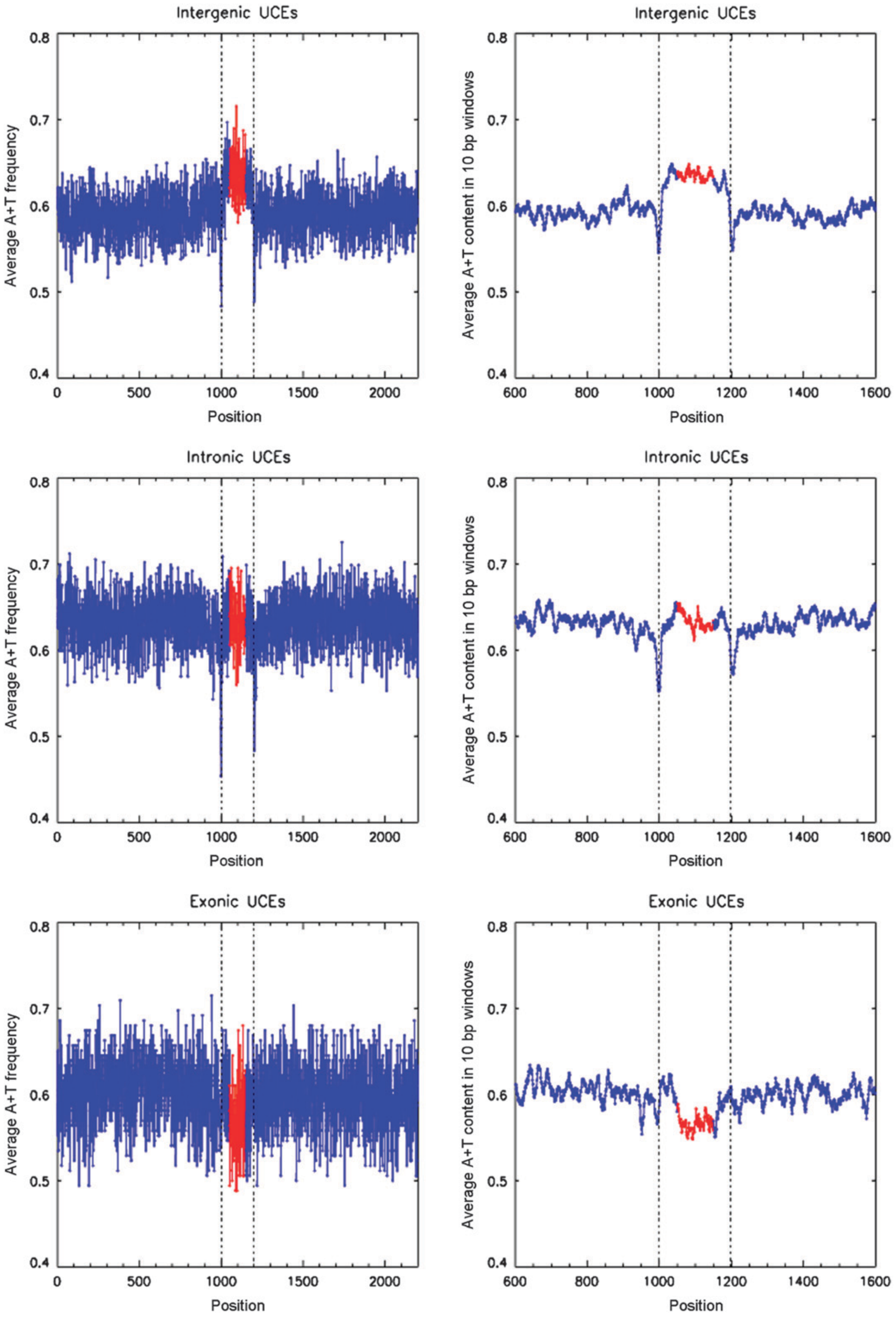

FIGURE 3.—A drop in A + T content flanks the intergenic and intronic UCEs. For each UCE, we extracted 1 kb of 5′ and 3′ flanking sequence together with 50 bp of sequence from either end of the UCE (blue). The middle 100 bp of each UCE was also included (red). The frequency of A + T nucleotides at each base-pair position was then computed over all 422 intergenic, 302 intronic, and 172 exonic UCEs (left). Additionally, the running average of A + T frequency was computed over an 11-bp window centered around each nucleotide position and is shown from positions 600 to 1600 (right). Vertical dotted lines mark the first base pair inside the UCE at both ends. Note that the most extreme single-base-pair drop in A + T frequency at or near both boundaries of the nonexonic UCEs (left) may be due to an ascertainment bias in computationally defining conserved elements. Specifically, since nonexonic UCEs appear to be more A + T rich than their flanks and obvious biases in favor of A-to-G and T-to-C substitutions in mouse conserved noncoding elements have been noted (DERMITZAKIS *et al.* 2004), it is possible that A/T-to-G/C substitutions are the likely cause of mismatches between the human genome and another genome. This single-base-pair drop, however, cannot fully explain the overall drop in A + T frequency flanking the intergenic and intronic UCEs.

may be as ancient as are the elements themselves, suggesting that such a sensitivity may have played a role in the ultraconservation of these elements during vertebrate evolution. Finally, intergenic and intronic UCEs, but not their exonic counterparts, are overall twofold or more enriched for a handful of motifs relative to their flanking sequences and bounded by a sharp change in A + T frequency. Note that while our studies have focused on the properties of nonexonic UCEs, they do not rule out the possibility that exonic UCEs share some structural and/or functional features with the nonexonic UCEs. Exonic UCEs also appear to harbor features that are specific to themselves, identifying them as a distinct class of genetic elements with functions beyond their protein-coding capacity (see BEJERANO *et al.* 2004; DERTI *et al.* 2006; LAREAU *et al.*

2007; Ni *et al.* 2007). Indeed, signatures for the intergenic, intronic, and exonic UCEs may well involve one or more of the many epigenetic marks that can augment the information content of a sequence of DNA; changes in A + T content may be correlated with changes in DNA methylation, the positioning of nucleosomes, or shifts in the strength of pairing between the two strands of DNA, to mention just a few possibilities.

This dichotomy between nonexonic and exonic UCEs is also noteworthy considering the evolutionary distinction between the two (Stephen *et al.* 2008): nonexonic, more than exonic, UCEs experienced an increase in number after the divergence of tetrapods (sauropsids, mammals, and amphibians) and teleosts (bony fish), accompanied by a remarkable deceleration of the molecular clock (Stephen *et al.* 2008) (Figure 1). These and related observations demonstrating that CNEs experienced a deceleration of the molecular clock prior to the divergence of tetrapods and teleosts (McEwen *et al.* 2006) have suggested that nonexonic UCEs and such CNEs played and continue to play key roles in vertebrate evolution and development. Aligned with this proposal is evidence supporting the hypothesis that these UCEs and CNEs function as highly important gene regulatory elements, such as enhancers, thereby providing a reason for their perfect or remarkably high conservation. This interpretation is widely supported by the capacity of these elements to direct transcription (Woolfe *et al.* 2005; McEwen *et al.* 2006; Pennacchio *et al.* 2006; Ahituv *et al.* 2007; Paparidis *et al.* 2007; Visel *et al.* 2008) and is also consistent with the enrichment in nonexonic UCEs for the recognition sequence of the homeodomain protein DNA-binding module (also see Abnizova *et al.* 2007), which is found in many transcription factors (Fraenkel *et al.* 1998).

Our findings are also consistent with an explanation for ultraconservation in which nonexonic UCEs are maintained through copy counting via sequence comparison and wherein significant departures from normal sequence or copy number lead to a decrease in fitness. Here, a pairing mechanism could align the two copies of each element and then use the motifs and boundaries of the elements to constrain comparisons to within the elements. In fact, UCEs appear to be enriched in recombination hotspots, raising the possibility that they may be near their homologs more frequently than are other loci (Derti *et al.* 2006 and C. W. K. Chiang, unpublished results; also see Derti *et al.* 2006 for further discussion). In addition to explaining how the conservation of these elements can be sustained through evolution, such a mechanism would also liken UCEs to sentinels of genomic integrity, "genomic canaries" that act in conjunction with other mechanisms (reviewed in Birchler and Veitia 2007) to monitor the dosage of chromosomal segments. Given the plethora of homology-driven phenomena found throughout the living kingdom and the involvement of pairing in several of these phenomena (reviewed in Wu and Morris 1999; Duncan 2002; Grant-Downton and Dickinson 2004; McKee 2004; Zickler 2006), a homology-based mechanism leading to sequence conservation may not be surprising. Indeed, CNEs have been identified in invertebrates (Kent and Zahler 2000; Siepel *et al.* 2005; Vavouri *et al.* 2007), raising the possibility that invertebrates may also have employed a form of sequence comparison, as has been considered elsewhere (Vavouri *et al.* 2007).

In light of this discussion, we wonder whether the evolutionary constraints brought on by the complexities of development could have generated a multitude of sequences, such as enhancers, at the optimal copy number of two and invariant enough to qualify as the initial targets for primitive forms of copy counting via comparison. These putative original targets could have reflected the sequence specificities of the proposed machineries for copy counting; for example, enrichment in nonexonic UCEs for TAATTA could indicate an early role of homeodomain-containing transcription factors, consistent with hypotheses proposing that transcription and enhancers participate in pairing and other homology-based phenomena (reviewed in McKee 2004; Lee and Wu 2006 and references within). While some target elements could have continued to act also as key players in development, others could have relinquished their original roles over time and become nonessential with respect to developmental programs, persisting only as sites of copy counting regardless of whether their sequence composition retained the capacity to direct transcription. Maintenance of conserved sequences through copy counting via comparison, in turn, could have played a role in preserving powerful regulatory mechanisms, thereby influencing vertebrate evolution even as it left footprints of ultraconservation throughout the genome.

We thank M. L. Bulyk, G. M. Church, B. A. Cohen, M. B. Eisen, L. Feuk, C. Lee, S. A. McCarroll, N. Patterson, J. Pritchard, D. Reich, P. Sklar, and S. Sunyaev for valuable discussions and sharing data; Vivian U for assistance in the graphical representation in Figure 3 and suggestions for programming; members of the Hirschhorn laboratory as well as Jack Bateman, Richard Emmons, Amber Hohl, Matt Jakubik, Lillian Merriam, and Ben Williams for comments and ideas; and the Research Information Technology Group and the West Quad Computing Group at Harvard Medical School for computational resources. This work was supported by a graduate research fellowship from the National Science Foundation (C.W.K.C.), a grant from the Keck Foundation (A.D.), a grant from the National Institutes (NIH) of Health (A.D.; HG003224), grants from the Department of Energy and NIH/National Institute of General Medical Sciences (D.S. and M.F.C.; DE-FG02-03ER63445 and GM068763 awarded to G.M. Church), a training grant from the National Eye Institute (M.F.C.; T32 EY07110), and Harvard Medical School (C.-t.W.).

Note added at proof: Two additional CNV data sets (Cooper *et al.* 2008; McCarroll *et al.* 2008), constructed through the use of independent commercialized second-generation SNP genotyping arrays, became available while this article was under review. We found absolute depletion (observed/expected = 0.00) of UCEs among both data sets ($P = 0.0067$ and $P = 0.0010$ for Cooper *et al.* 2008 and McCarroll *et al.* 2008, respectively), where the statistical evidence for depletion was driven by depletion of the intergenic UCEs ($P = 0.0276$ and $P = 0.0098$, respectively); absolute depletion of intronic ($P = 0.1388$ and $P = 0.1043$, respectively) and exonic ($P = 0.1807$ and $P = 0.1070$) UCEs, however, was not significant. These findings, which are consistent with our analyses of other second-generation data sets, indicate that the observed depletions are not contingent on the platforms or methodologies used to detect CNVs and therefore support our proposal that the depletion of UCEs among CNVs is a general feature of the human genome.

LITERATURE CITED

Abnizova, I., K. Walter, R. Te Boekhorst, G. Elgar and W. R. Gilks, 2007 Statistical information characterization of conserved non-coding elements in vertebrates. J. Bioinform. Comput. Biol. **5:** 533–547.

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal *et al.*, 2007 Deletion of ultraconserved elements yields viable mice. PLoS Biol. **5:** e234.

Bailey, J. A., and E. E. Eichler, 2006 Primate segmental duplications: crucibles of evolution, diversity and disease. Nat. Rev. Genet. **7:** 552–564.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent *et al.*, 2004 Ultraconserved elements in the human genome. Science **304:** 1321–1325.

Birchler, J. A., and R. A. Veitia, 2007 The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell **19:** 395–402.

Boffelli, D., M. A. Nobrega and E. M. Rubin, 2004 Comparative genomics at the vertebrate extremes. Nat. Rev. Genet. **5:** 456–465.

Chen, C. T., J. C. Wang and B. A. Cohen, 2007 The strength of selection on ultraconserved elements in the human genome. Am. J. Hum. Genet. **80:** 692–704.

Cheung, J., X. Estivill, R. Khaja, J. R. MacDonald, K. Lau *et al.*, 2003 Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. **4:** R25.

Cooper, G. M., D. A. Nickerson and E. E. Eichler, 2007 Mutational and selective effects on copy-number variants in the human genome. Nat. Genet. **39:** S22–S29.

Cooper, G. M., T. Zerr, J. M. Kidd, E. E. Eichler and D. A. Nickerson, 2008 Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat. Genet. **40:** 1199–1203.

de la Calle-Mustienes, E., C. G. Feijoo, M. Manzanares, J. J. Tena, E. Rodriguez-Seguel *et al.*, 2005 A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res. **15:** 1061–1072.

Dermitzakis, E. T., E. Kirkness, S. Schwarz, E. Birney, A. Reymond *et al.*, 2004 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.

Derti, A., F. P. Roth, G. M. Church and C. T. Wu, 2006 Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat. Genet. **38:** 1216–1220.

de Smith, A. J., A. Tsalenko, N. Sampas, A. Scheffer, N. A. Yamada *et al.*, 2007 Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum. Mol. Genet. **16:** 2783–2794.

Drake, J. A., C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh *et al.*, 2006 Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. **38:** 223–227.

Duncan, I. W., 2002 Transvection effects in Drosophila. Annu. Rev. Genet. **36:** 521–556.

Feng, J., C. Bi, B. S. Clark, R. Mady, P. Shah *et al.*, 2006 The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev. **20:** 1470–1484.

Fisher, S., E. A. Grice, R. M. Vinton, S. L. Bessling and A. S. McCallion, 2006 Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science **312:** 276–279.

Fraenkel, E., M. A. Rould, K. A. Chambers and C. O. Pabo, 1998 Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures. J. Mol. Biol. **284:** 351–361.

Gardiner, E. J., L. Hirons, C. A. Hunter and P. Willett, 2006 Genomic data analysis using DNA structure: an analysis of conserved nongenic sequences and ultraconserved elements. J. Chem. Inf. Model. **46:** 753–761.

Grant-Downton, R. T., and H. G. Dickinson, 2004 Plants, pairing and phenotypes: Two's company? Trends Genet. **20:** 188–195.

International Chicken Genome Sequencing Consortium, 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432:** 695–716.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature **437:** 1299–1320.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451:** 998–1003.

Katzman, S., A. D. Kern, G. Bejerano, G. Fewell, L. Fulton *et al.*, 2007 Human genome ultraconserved elements are ultraselected. Science **317:** 915.

Kent, W. J., and A. M. Zahler, 2000 Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res. **10:** 1115–1125.

Kidd, J. M., G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas *et al.*, 2008 Mapping and sequencing of structural variation from eight human genomes. Nature **453:** 56–64.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert *et al.*, 2007 Paired-end mapping reveals extensive structural variation in the human genome. Science **318:** 420–426.

Lareau, L. F., M. Inada, R. E. Green, J. C. Wengrod and S. E. Brenner, 2007 Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature **446:** 926–929.

Lee, A. M., and C. T. Wu, 2006 Enhancer-promoter communication at the *yellow* gene of *Drosophila melanogaster*: diverse promoters participate in and regulate *trans* interactions. Genetics **174:** 1867–1880.

Li, X. Y., S. Macarthur, R. Bourgon, D. Nix, D. A. Pollard *et al.*, 2008 Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. **6:** e27.

Liu, X. S., D. L. Brutlag and J. S. Liu, 2002 An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. **20:** 835–839.

Ludwig, M. Z., A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan *et al.*, 2005 Functional evolution of a cis-regulatory module. PLoS Biol. **3:** e93.

McCarroll, S. A., F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh *et al.*, 2008 Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. **40:** 1166–1174.

McEwen, G. K., A. Woolfe, D. Goode, T. Vavouri, H. Callaway *et al.*, 2006 Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res. **16:** 451–465.

McGaughey, D. M., R. M. Vinton, J. Huynh, A. Al-Saif, M. A. Beer *et al.*, 2008 Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. **18:** 252–260.

McKee, B. D., 2004 Homologous pairing and chromosome dynamics in meiosis and mitosis. Biochim. Biophys. Acta **1677:** 165–180.

Mills, R. E., C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui *et al.*, 2006 An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. **16:** 1182–1190.

Minsky, A., 2004 Information content and complexity in the high-order organization of DNA. Annu. Rev. Biophys. Biomol. Struct. **33:** 317–342.

Murphy, W. J., P. A. Pevzner and S. J. O'Brien, 2004 Mammalian phylogenomics comes of age. Trends Genet. **20:** 631–639.

Ni, J. Z., L. Grate, J. P. Donohue, C. Preston, N. Nobida *et al.*, 2007 Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. **21:** 708–718.

Paparidis, Z., A. A. Abbasi, S. Malik, D. K. Goode, H. Callaway *et al.*, 2007 Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. Dev. Growth Differ. **49:** 543–553.

Pavesi, G., P. Mereghetti, G. Mauri and G. Pesole, 2004 Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. **32:** W199–W203.

Pennacchio, L. A., N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega *et al.*, 2006 In vivo enhancer analysis of human conserved non-coding sequences. Nature **444:** 499–502.

Perry, G. H., A. Ben-Dor, A. Tsalenko, N. Sampas, L. Rodriguez-Revenga *et al.*, 2008 The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet. **82:** 685–695.

Pinto, D., C. Marshall, L. Feuk and S. W. Scherer, 2007 Copy-number variation in control population cohorts. Hum. Mol. Genet. **16** (Spec. no. 2): R168–R173.

Rastegar, S., I. Hess, T. Dickmeis, J. C. Nicod, R. Ertzer *et al.*, 2008 The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. Dev. Biol. **318:** 366–377.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry *et al.*, 2006 Global variation in copy number in the human genome. Nature **444:** 444–454.

Scherer, S. W., C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler *et al.*, 2007 Challenges and standards in integrating surveys of structural variation. Nat. Genet. **39:** S7–S15.

Schwartz, D., and S. P. Gygi, 2005 An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. **23:** 1391–1398.

Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field *et al.*, 2006 A genomic code for nucleosome positioning. Nature **442:** 772–778.

She, X., Z. Jiang, R. A. Clark, G. Liu, Z. Cheng *et al.*, 2004 Shotgun sequence assembly and recent segmental duplications within the human genome. Nature **431:** 927–930.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. **15:** 1034–1050.

Simon-Sanchez, J., S. Scholz, H. C. Fung, M. Matarin, D. Hernandez *et al.*, 2007 Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum. Mol. Genet. **16:** 1–14.

Stephen, S., M. Pheasant, I. V. Makunin and J. S. Mattick, 2008 Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. **25:** 402–408.

Vavouri, T., K. Walter, W. R. Gilks, B. Lehner and G. Elgar, 2007 Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. **8:** R15.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis *et al.*, 2008 Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat. Genet. **40:** 158–160.

Walter, K., I. Abnizova, G. Elgar and W. R. Gilks, 2005 Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. Trends Genet. **21:** 436–440.

Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner *et al.*, 2007 PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. **17:** 1665–1674.

Warren, W. C., L. W. Hillier, J. A. Marshall Graves, E. Birney, C. P. Ponting *et al.*, 2008 Genome analysis of the platypus reveals unique signatures of evolution. Nature **453:** 175–183.

Wong, K. K., R. J. deLeeuw, N. S. Dosanjh, L. R. Kimm, Z. Cheng *et al.*, 2007 A comprehensive analysis of common copy-number variations in the human genome. Am. J. Hum. Genet. **80:** 91–104.

Woolfe, A., M. Goodson, D. K. Goode, P. Snell, G. K. McEwen *et al.*, 2005 Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. **3:** e7.

Wu, C. T., and J. R. Morris, 1999 Transvection and other homology effects. Curr. Opin. Genet. Dev. **9:** 237–246.

Xie, X., T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis *et al.*, 2007 Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc. Natl. Acad. Sci. USA **104:** 7145–7150.

Zhang, L., S. Kasif, C. R. Cantor and N. E. Broude, 2004 GC/AT-content spikes as genomic punctuation marks. Proc. Natl. Acad. Sci. USA **101:** 16855–16860.

Zickler, D., 2006 From early homologue recognition to synaptonemal complex formation. Chromosoma **115:** 158–174.

Zogopoulos, G., K. C. Ha, F. Naqib, S. Moore, H. Kim *et al.*, 2007 Germ-line DNA copy number variation frequencies in a large North American population. Hum. Genet. **122:** 345–353.

Communicating editor: G. Stormo