22

Ultra-Conserved Elements in Vertebrate and Fly Genomes

Mathias Drton

Nicholas Eriksson

Garmay Leung

Ultra-conserved elements in an alignment of multiple genomes are consecutive nucleotides that are in perfect agreement across all the genomes. For aligned vertebrate and fly genomes, we give descriptive statistics of ultra-conserved elements, explain their biological relevance, and show that the existence of ultra-conserved elements is highly improbable in neutrally evolving regions.

22.1 The data

Our analyses of ultra-conserved elements are based on multiple sequence alignments produced by `MAVID` [Bray and Pachter, 2004]. Prior to the alignment of multiple genomes, homology mappings (from Mercator [Dewey, 2005]) group into bins genomic regions that are anchored together by neighboring homologous exons. A multiple sequence alignment is then produced for each of these alignment bins. `MAVID` is a global multiple alignment program, and therefore homologous regions with more than one homologous hit to another genome may not be found aligned together. Table 22.1 shows an example of Mercator's output for a single region along with the beginning of the resulting `MAVID` multiple sequence alignment.

Species	Chrom.	Start	End		Alignment
Dog	chrX	752057	864487	+	`A----AACCAAA---------`
Chicken	chr1	122119382	122708162	−	`TGCTGAGCTAAAGATCAGGCT`
Zebra fish	chr9	19018916	19198136	+	`------ATGCAACATGCTTCT`
Puffer fish	chr2	7428614	7525502	+	`---TAGATGGCAGACGATGCT`
Fugu fish	asm1287	21187	82482	+	`---TCAAGGG-----------`

Table 22.1. *Mercator output for a single bin, giving the position and orientation on the chromosome. Notice that the Fugu fish genome has not been fully assembled into chromosomes (cf. Section 4.2).*

The vertebrate dataset consists of 10,279 bins over 9 genomes (Table 22.2). A total of 4,368 bins (42.5%) contain alignments across all 9 species. The evolutionary relationships among these species (which first diverged about 450 million years ago) are shown in Figure 22.1. For a discussion of the problem of placing the rodents in the phylogenetic tree, see Section 21.4 and Figure 21.4.

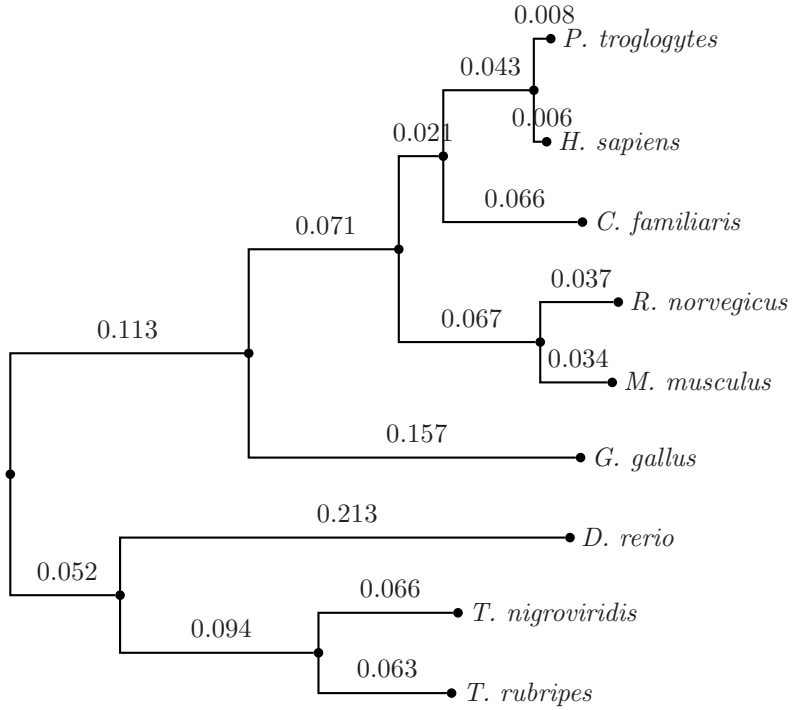


Fig. 22.1. Phylogenetic tree for whole genome alignment of 9 vertebrates.

With the exception of the probability calculations in phylogenetic tree models, our subsequent findings on ultra-conserved elements do not depend on the form of this tree.

Species	Genome Size	Genome Release Date
Zebra fish (*Danio rerio*)	1.47 Gbp	11/27/2003
Fugu fish (*Takifugu rubripes*)	0.26 Gbp	04/02/2003
Puffer fish (*Tetraodon nigroviridis*)	0.39 Gbp	02/01/2004
Dog (*Canis familiaris*)	2.38 Gbp	07/14/2004
Human (*Homo sapiens*)	2.98 Gbp	07/01/2003
Chimp (*Pan troglogytes*)	4.21 Gbp	11/13/2003
Mouse (*Mus musculus*)	2.85 Gbp	05/01/2004
Rat (*Rattus norvegicus*)	2.79 Gbp	06/19/2003
Chicken (*Gallus gallus*)	1.12 Gbp	02/24/2004

Table 22.2. *Genomes in the nine-vertebrate alignment with size given in billion base pairs.*

The fruit fly dataset consists of 8 *Drosophila* genomes (Table 22.3). Of the 3,731 alignment bins, 2,985 (80.0%) contain all 8 species, which reflects the smaller degree of evolutionary divergence. A phylogenetic tree for these 8 species, which diverged at least 45 million years ago, is illustrated in Figure 22.2.

The pilot phase of the ENCODE project (cf. Sections 4.3 and 21.2) provides an additional dataset of vertebrate sequences homologous to 44 regions of the human genome. There are 14 manually selected regions of particular biological

Species	Genome Size	Genome Release Date
D. melanogaster	118 Mbp	04/21/2004
D. simulans	119 Mbp	08/29/2004
D. yakuba	177 Mbp	04/07/2004
D. erecta	114 Mbp	10/28/2004
D. ananassae	136 Mbp	12/06/2004
D. pseudoobscura	125 Mbp	08/28/2003
D. virilis	152 Mbp	10/29/2004
D. mojavensis	177 Mbp	12/06/2004

Table 22.3. *Genomes in the eight-Drosophila alignment with size given in million base pairs.*

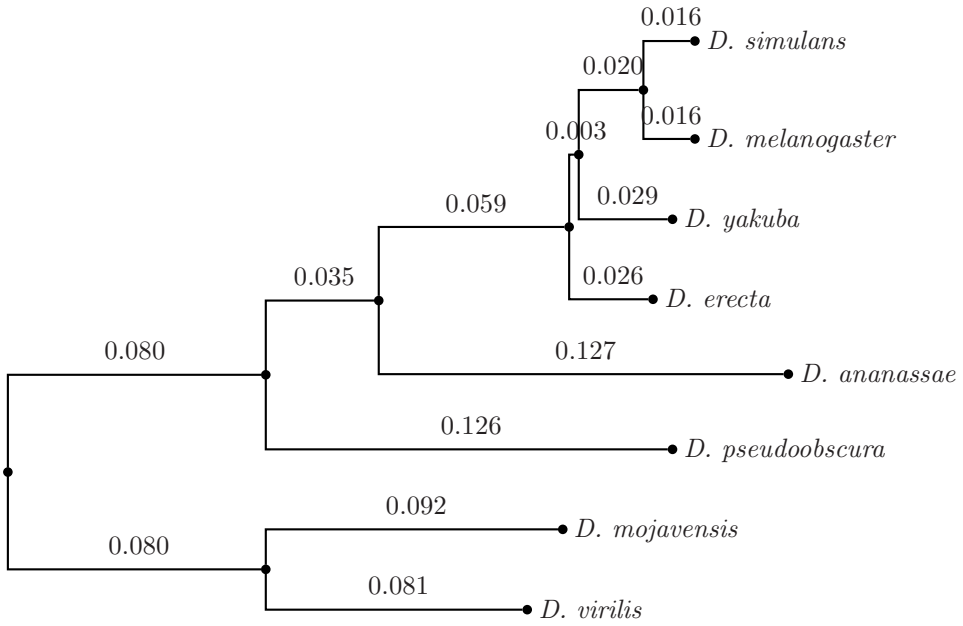


Fig. 22.2. Phylogenetic tree for whole genome alignment of 8 *Drosophila* species.

interest and 30 randomly selected regions with varying degrees of non-exonic conservation and gene density. Each manually selected region consists of 0.5–1.9 Mbp, while each randomly selected region is 0.5 Mbp in length. This gives a total of about 30 Mbp, approximately 1% of the human genome.

Varying with the region under consideration, a subset of the following 11 species is aligned along with the human genome in the preliminary October 2004 freeze: chimp, baboon (*Papiocynocephalus anubis*), marmoset (*Callithrix jacchus*), galago (*Otolemur garnettii*), mouse, rat, dog, armadillo (*Dasypus novemcintus*), platypus (*Ornithorhynchus anatinus*), and chicken. This collection of species lacks the three fish of the nine-vertebrate alignment. Armadillo and platypus sequences are only available for the first manually picked ENCODE region, and sequences for every region are only available for human, mouse, rat, dog and chicken. The number of species available for each region varies between 6 and 11 for manually selected regions, and between

8 and 10 for randomly selected regions. For each region, `Shuffle-LAGAN` [Brudno *et al.*, 2003b] was applied between the human sequence and each of the other available sequences to account for rearrangements. Based on these re-shuffled sequences, a multiple sequence alignment for each region was produced with `MAVID`.

The three sets of multiple alignments are available for download at `http://bio.math.berkeley.edu/ascb/chapter22/`.

22.2 Ultra-conserved elements

A position in a multiple alignment is *ultra-conserved* if for all species the same nucleotide appears in the position. An *ultra-conserved element* of length ℓ is a sequence of consecutive ultra-conserved positions $(n, n+1, \ldots, n+\ell-1)$ such that positions $n-1$ and $n+\ell$ are not ultra-conserved.

Example 22.1 Consider a subset of length 24 of a three-genome alignment:

```
G--ACCCAATAGCACCTGTTGCGG
CGCTCTCCA---CACCTGTTCCGG
CATTCT---------CTGTTTTGG
    *          *****  **
```

where ultra-conserved positions are marked by a star *. This alignment contains three ultra-conserved elements, one of length 1 in position 5, one of length 5 covering positions 16–20, and one of length 2 in positions 23–24. □

22.2.1 Nine-vertebrate alignment

We scanned the entire nine-vertebrate alignment described in Section 22.1 and extracted 1,513,176 ultra-conserved elements, whose lengths are illustrated in Figure 22.3. The median and the mean length of an ultra-conserved element is equal to 2 and 1.918, respectively.

We will focus on the 237 ultra-conserved elements of length at least 20, covering 6,569 bp in sum. These 237 elements are clustered together; they are only found in 113 of the 4,368 bins containing all 9 species. The length distribution is heavily skewed toward shorter sequences as seen in Figure 22.3, with 75.5% of these regions shorter than 30 bp and only 10 regions longer than 50 bp.

The longest ultra-conserved element in the alignment is 125 bp long:

```
CTCAGCTTGT CTGATCATTT ATCCATAATT AGAAAATTAA TATTTTAGAT GGCGCTATGA
TGAACCCATT ATGGTGATGG GCCCCGATAT CAATTATAAC TTCAATTTCA ATTTCACTTA
CAGCC.
```

The next-longest ultra-conserved elements are two elements of length 85, followed by one element for each one of the lengths 81, 66, 62, 60, 59, 58, and 56. In particular, there is exactly one ultra-conserved element of length 42, which is the *"meaning of life"* element discussed in [Pachter and Sturmfels, 2005].

A number of the ultra-conserved elements are separated only by a few (less

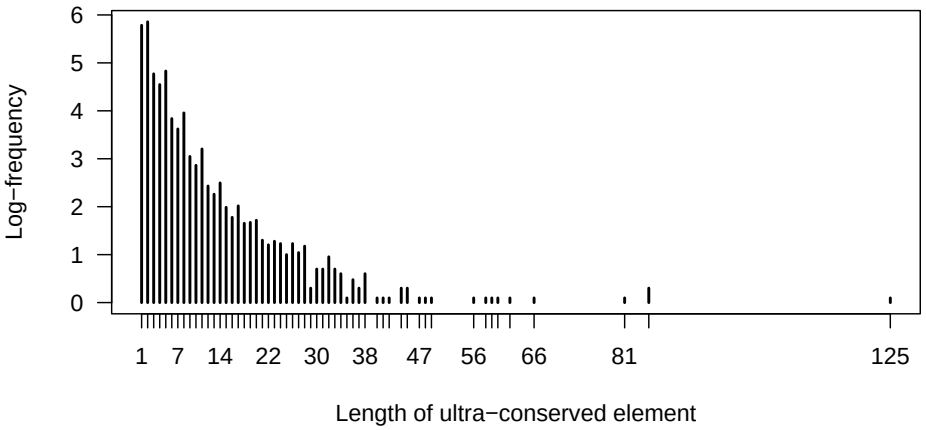


Fig. 22.3. Frequencies of vertebrate ultra-conserved elements ($\log_{10}$-scale).

than 10), ungapped, intervening positions. In 18 cases, there is a single intervening position. Typically, these positions are nearly ultra-conserved, and display differences only between the fish and the other species. Collapsing the ultra-conserved elements separated by fewer than 10 bases reduces the number of ultra-conserved elements to 209, increases the base coverage to 6,636 bp, and brings the total number of regions greater than 50 bp in length to 26.

In the human genome, the GC-ratio (proportion of G and C among all nucleotides) is 41.0%. The ultra-conserved elements are slightly more AT-rich; for the 237 elements of length 20 or longer, the GC-ratio is 35.8%. However, GC-content and local sequence characteristics were not enough to identify ultra-conserved regions using data from only one genome.

22.2.2 ENCODE alignment

The 44 ENCODE regions contain 139,043 ultra-conserved elements, 524 of which are longer than 20 bp. These long elements cover 17,823 bp. By base coverage, 73.5% of the long elements are found in the manually chosen regions. The longest one is in region ENm012, of length 169 and consists of the DNA sequence:

```
AAGTGCTTTG TGAGTTTGTC ACCAATGATA ATTTAGATAG AGGCTCATTA CTGAACATCA
CAACACTTTA AAAACCTTTC GCCTTCATAC AGGAGAATAA AGGACTATTT TAATGGCAAG
GTTCTTTTGT GTTCCACTGA AAAATTCAAT CAAGACAAAA CCTCATTGA.
```

This sequence does not contain a subsequence of length 20 or longer that is ultra-conserved in the nine-vertebrate alignment, but the 169 bp are also ultra-conserved in the nine-vertebrate alignment if the three fish are excluded from consideration. The only overlap between the nine-vertebrate and ENCODE ultra-conserved elements occurs in the regions ENm012 and ENm005, where there are 3 elements that are extensions of ultra-conserved elements in the nine-vertebrate alignment.

Table 22.4 shows the number of species aligned in the 44 ENCODE alignments and the respective five longest ultra-conserved elements that are of length 20 or larger. Omitted randomly selected regions do not contain any ultra-conserved elements of length at least 20.

Manually selected			Randomly selected		
Region	Spec.	Ultra-lengths	Region	Spec.	Ultra-lengths
ENm001	11	$28, 27, 23, 20_2$	ENr122	9	22
ENm002	8	$39, 28, 27, 26_4$	ENr213	9	$30, 27, 26, 24, 23_2$
ENm003	9	$38, 28_2, 26, 25_2$	ENr221	10	$36_2, 32_2, 29$
ENm004	8	$35, 26_2, 25, 20$	ENr222	10	$29, 22$
ENm005	10	$114, 62, 38, 34, 32$	ENr231	8	$26, 23, 20$
ENm006	8	–	ENr232	8	$26, 25, 20$
ENm007	6	–	ENr233	9	$25, 24, 20$
ENm008	9	$23, 22$	ENr311	10	$42, 31, 25, 21$
ENm009	10	–	ENr312	9	$60, 31, 22, 20_4$
ENm010	8	$86, 68, 63, 61, 60_2$	ENr313	9	27
ENm011	7	–	ENr321	10	$68, 44, 38, 37, 35$
ENm012	9	$169, 159, 125_2, 123$	ENr322	9	$126, 80, 79, 61, 55$
ENm013	10	$30, 26, 23, 22$	ENr323	8	$53, 50, 45, 42, 29$
ENm014	10	$41_2, 39, 26_2$	ENr331	9	26
			ENr332	10	26
			ENr334	8	$79, 50, 44, 37, 32$

Table 22.4. *Number of species and lengths of ultra-conserved elements in ENCODE alignments. Subindices indicate multiple occurrences.*

22.2.3 Eight-Drosophila alignment

There are 5,591,547 ultra-conserved elements in the *Drosophila* dataset with 1,705 elements at least 50 bp long and the longest of length 209 bp. We focused on the 255 *Drosophila* ultra-conserved elements of length at least 75 bp, covering 23,567 bp total. These regions are also found clustered together, occurring over 163 bins out of the 2,985 bins with all 8 species aligned together. The shortest distance between consecutive ultra-conserved elements is 130 bp, and therefore regions were not collapsed for this dataset. The mean and median length of ultra-conserved elements are 2.605 and 2, respectively. The length distribution of all ultra-conserved elements is shown in Figure 22.4. This set of ultra-conserved elements is also somewhat more AT-rich, with a GC-ratio of 38.8% (for those elements of length at least 75 bp) compared with a GC-ratio of 42.4% across the entire *D. melanogaster* genome.

22.3 Biology of ultra-conserved elements

22.3.1 Nine-vertebrate alignment

Using the UCSC genome browser annotations of known genes for the July 2003 (hg16) release of the human genome, we investigated which ultra-conserved elements overlap known functional regions. Intragenic regions cover 62.6% of the

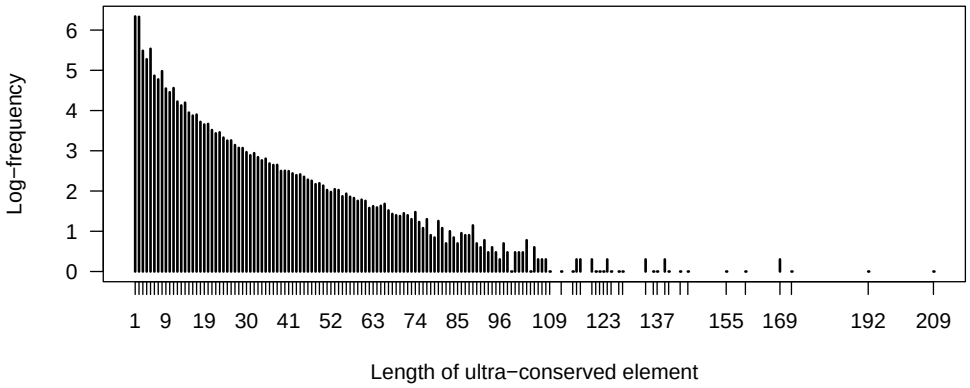
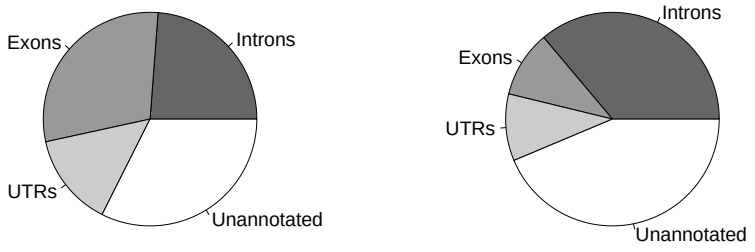


Fig. 22.4. Frequencies of *Drosophila* ultra-conserved elements ($\log_{10}$-scale).

bases of the 209 collapsed ultra-conserved elements described in Section 22.2.1. However, intragenic coverage increases to 67.6% for short elements (less than 30 bp) and drops to 56.3% for longer elements (at least 30 bp), as shown in Figures 22.5(a) and 22.5(b). While shorter ultra-conserved elements tend to correspond to exons, longer ones are generally associated with introns and unannotated regions. Nine ultra-conserved elements cover a total of 306 bp in the intronic regions of *POLA*, the alpha catalytic subunit of DNA polymerase. Six other genes are associated with more than 100 bp of ultra-conserved elements. Four of these genes are transcription factors involved in development (*SOX6*, *FOXP2*, *DACH1*, *TCF7L2*). In fact, elements near *DACH* that were highly conserved between human and mouse and also present in fish species have been shown to be *DACH* enhancers; see [Nobrega *et al.*, 2003].



(a) 150 elements ≥ 20 and < 30 bp

(b) 59 elements ≥ 30 bp

Fig. 22.5. Functional base coverage of collapsed vertebrate ultra-conserved elements based on annotations of known human genes.

Among the 237 uncollapsed ultra-conserved elements of length at least 20, 151 are in intragenic regions of 96 genes. The remaining 86 elements did not overlap any annotated gene. However, by grouping together elements that have the same upstream and downstream flanking genes, there are only 27 super-

regions to consider, with 51 unique flanking genes. There are 6 super-regions with at least 99 bp overlapping with ultra-conserved elements. At least one of the flanking genes for each of these 6 super-regions is a transcription factor located 1–314 kb away (*IRX3*, *IRX5*, *IRX6*, *HOXD13*, *DMRT1*, *DMRT3*, *FOXD3*, *TFEC*). The overall average distance to the closest flanking gene on either side is 138 kb and ranges from 312 bp to 1.2 Mbp.

It is a natural question whether the genes near or overlapping ultra-conserved elements tend to code for similar proteins. We divided the set of 96 genes with ultra-conserved overlap into 3 groups based on where in the gene the overlap occurred: exon, intron or untranslated region (UTR). If ultra-conserved elements overlap more than one type of genic region, then the gene is assigned to each of the appropriate groups. The 51 genes flanking ultra-conserved elements in unannotated regions form a fourth group of genes.

The Gene Ontology (GO) Consortium (`http://www.geneontology.org`) provides annotations for genes with respect to the molecular function of their gene products, the associated biological processes, and their cellular localization [Ashburner *et al.*, 2000]. For example, the human gene *SOX6* is annotated for biological process as being involved in cardioblast differentiation and DNA-dependent regulation of transcription. Mathematically, each of the three ontologies can be considered as a partially ordered set (*poset*) in which the catagories are ordered from most to least specific. For example, cardioblast differentiation is more specific than cardiac cell differentiation, which in turn is more specific than both cell differentiation and embryonic heart tube development. If a gene possesses a certain annotation, it must also possess all more general annotations; therefore GO consists of a map from the set of genes to order ideals in the three posets. We propose that this mathematical structure is important for analyzing the GO project.

In this study, we only considered molecular function and biological process annotations. These annotations are available for 46 of the 54 genes with exonic overlap, for all of the 28 with intronic overlap, for 14 of the 20 with UTR overlap, and for 30 of the 51 genes flanking unannotated elements. Considering one GO annotation and one of the 4 gene groups at a time, we counted how many of the genes in the group are associated with the considered annotation. Using counts of how often this annotation occurs among all proteins found in Release 4.1 of the Uniprot database (`http://www.uniprot.org`), we computed a p-value from Fisher's exact test for independence of association with the annotation and affiliation with the considered gene group. Annotations associated with at least 3 genes in a group and with an unadjusted p-value smaller than $3.0 \cdot 10^{-2}$ are reported in Table 22.5. DNA-dependent regulation of transcription and transcription factor activity are found to be enriched in non-exonic ultra-conserved elements, corresponding to previously reported findings [Bejerano *et al.*, 2004, Boffelli *et al.*, 2004, Sandelin *et al.*, 2004, Woolfe *et al.*, 2005]. Conserved exonic elements tend to be involved in protein modification.

We scanned the human genome for repeated instances of these ultra-

GO Annotation	p-value
Exons (14)	
protein serine/threonine kinase activity	$4.545 \cdot 10^{-3}$
transferase activity	$1.494 \cdot 10^{-2}$
neurogenesis	$1.654 \cdot 10^{-2}$
protein amino acid phosphorylation	$2.210 \cdot 10^{-2}$
Introns (10)	
regulation of transcription, DNA-dependent	$8.755 \cdot 10^{-4}$
transcription factor activity	$2.110 \cdot 10^{-3}$
protein tyrosine kinase activity	$4.785 \cdot 10^{-3}$
protein amino acid phosphorylation	$1.584 \cdot 10^{-2}$
protein serine/threonine kinase activity	$2.806 \cdot 10^{-2}$
UTRs (3)	
regulation of transcription, DNA-dependent	$1.403 \cdot 10^{-4}$
transcription factor activity	$3.971 \cdot 10^{-3}$
Flanking within 1.2 Mbp (4)	
transcription factor activity	$3.255 \cdot 10^{-11}$
regulation of transcription, DNA-dependent	$2.021 \cdot 10^{-8}$
development	$5.566 \cdot 10^{-3}$

Table 22.5. *GO annotations of genes associated with vertebrate ultra-conserved elements. The number of GO annotations tested for each group are in parentheses. For each group, only GO annotations associated with at least 3 genes in the group were considered.*

conserved elements and found that 14 of the original 237 elements have at least one other instance within the human genome. Generally, the repeats are not ultra-conserved except for some of the seven repeats that are found both between *IRX6* and *IRX5* and between *IRX5* and *IRX3* on chromosome 16. These genes belong to a cluster of Iroquois homeobox genes involved in embryonic pattern formation [Peters *et al.*, 2000]. These repeated elements include two 32 bp sequences that are perfect reverse complements of each other and two (of lengths 23 bp and 28 bp) that are truncated reverse complements of each other. Overall, there are 5 distinct sequences within 226 bp regions on either side of *IRX5* that are perfect reverse complements of each other. The reverse complements are found in the same relative order (Figure 22.6). Furthermore, exact copies of the two outermost sequences are found both between *IRX4* and *IRX2* and between *IRX2* and *IRX1* on chromosome 5. Both of these regions are exactly 226 bp long. The repetition of these short regions and the conservation of their relative ordering and size suggests a highly specific coordinated regulatory signal with respect to these Iroquois homeobox genes and strengthens similar findings reported by [Sandelin *et al.*, 2004].

The longest ultra-conserved element that is repeated in the human genome is of length 35 and is found 18 additional times. None of these 18 instances are ultra-conserved, but this sequence is also found multiple times in other vertebrate genomes: 13 times in chimp, 10 times in mouse, 5 times in both rat and dog, 4 times in tetraodon, 3 times in zebra fish, and twice in both fugu and chicken. Of the 19 instances found in the human genome, two are found in

```
54102348 TGTAATTACAATCTTACAGAAACCGGGCCGATCTGTATATAAATCTCACCATCCAATTAC
54102408 AAGATGTAATAATTTTGCACTCAAGCTGGTAATGAGGTCTAATACTCGTGCATGCGATAA
54102468 TCCCCTCTGGATGCTGGCTTGATCAGATGTTGGCTTTGTAATTAGACGGGCAGAAAATCA
54102528 TTATTTCATGTTCAAATAGAAAATGAGGTTGGTGGGAAGTTAATTT

55002049 AAATTAACTTCCCACCAACCTAATTTTTTCCTGAACATGAAATAATGATTTTCTGCCCGT
55002109 CTAATTACAAAGCCAACATCTGATCAAGCCAGCATCCAGAGGGGATTATCGCATGCACGA
55002169 GTATTAGACCTCATTACCAGCTTGAGTGCAAAATTATTACATCTTGTAATTGGATGGTGA
55002229 GATTTATATACAGATCGGCCCGGTTTCTGTAAGATTGTAATTACA
```

Fig. 22.6. Sequences found on either side of *IRX5*. Positions underlined with a thick line are ultra-conserved with respect to the nine-vertebrate alignment. Sequences underlined with a thin line are not ultra-conserved but their reverse complement is. Indices are with respect to human chromosome 16.

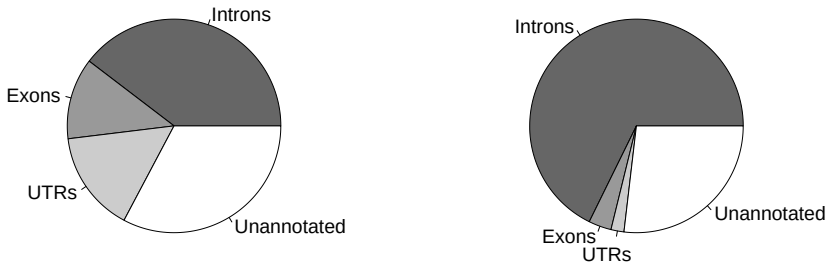
well-studied actin genes, *ACTC* and *ACTG*, and the remainder are found in predicted retroposed pseudogenes with actin parent genes. These predictions are based on the retroGene track of the UCSC genome browser. Retroposed pseudogenes are the result of the reverse transcription and integration of the mRNA of the original functional gene. Actins are known to be highly conserved proteins, and β- and γ-actins have been shown to have a number of non-functional pseudogenes [Ng *et al.*, 1985, Pollard, 2001]. The precise conservation of this 35 bp sequence across a number of human actin pseudogenes may suggest that these integration events may be relatively recent changes in the human genome.

22.3.2 ENCODE alignment

Based on the annotations of known human genes provided by the UCSC Genome Browser, 69.2% of the bases of the ultra-conserved elements of length at least 20 in the ENCODE alignment overlap intragenic regions. Shorter sequences (less than 50 bp) have far more overlap with exons and UTRs than longer sequences (at least 50 bp), as illustrated in Figures 22.7(a) and 22.7(b). These longer sequences are heavily biased towards intronic overlap, accounting for 67.7% of these sequences by base coverage.

Values for the gene density and non-exonic conservation level (human–mouse) are available for the randomly selected ENCODE regions (see Section 21.2). For these regions, the base coverage by ultra-conserved elements is not correlated with gene density (Pearson correlation = -0.0589) and is moderately correlated with non-exonic conservation (Pearson correlation = 0.4350).

While we do not repeat the gene ontology analysis from the previous section, we note that the regions with the greatest number of ultra-conserved elements by base coverage are regions with well-known genes involved in DNA-dependent transcriptional regulation (Table 22.6). The elements in these 5 regions account for 80.3% of the bases of the ultra-conserved elements found in this dataset.



(a) 445 elements ≥ 20 and < 50 bp (b) 79 elements ≥ 50 bp

Fig. 22.7. Functional base coverage of ultra-conserved elements found in ENCODE regions based on annotations of known human genes.

The 35 longest ultra-conserved elements, of length at least 69 bp, are also all found in these 5 regions.

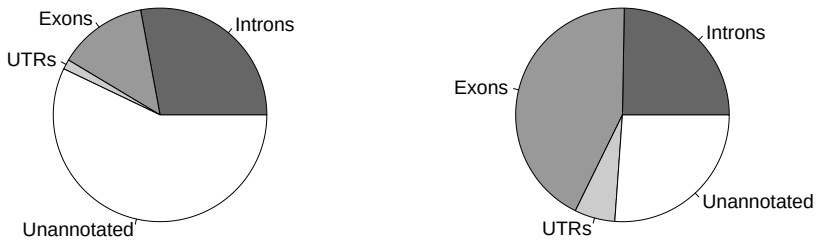
	Ultra Coverage (bp)	Transcription Factor Genes	# Aligned Species
ENm012	9,086	*FOXP2*	9
ENr322	2,072	*BC11B*	9
ENm010	1,895	*HOXA1-7,9-11,13*; *EVX1*	8
ENm005	718	*GCFC*; *SON*	10
ENr334	549	*FOXP4*; *TFEB*	8

Table 22.6. *ENCODE regions with the greatest number of ultra-conserved elements by base coverage and their associated transcription factor genes.*

22.3.3 Eight-Drosophila alignment

We analyzed the 255 ultra-conserved elements of length at least 75 bp using the Release 4.0 annotations of *D. melanogaster*. These elements overlap 95 unique genes. Although the intragenic overlap for shorter elements (less than 100 bp) is only 42.9%, this proportion increases to 68.2% for the elements that are at least 100 bp in length (Figures 22.8(a) and 22.8(b)). Unlike the vertebrate dataset, longer regions are associated with exons, while shorter regions tend to correspond to unannotated elements.

The three genes with the greatest amount of overlap with ultra-conserved elements are *para* (765 bp), *nAcRα-34E* (426 bp) and *nAcRα-30D* (409 bp). All three of these genes are involved in cation channel activity, and the ultra-conserved elements correspond mostly with their exons. As with the nine-vertebrate dataset, the full set of 95 *D. melanogaster* genes is assessed for GO annotation enrichment, using all Release 4.0 *D. melanogaster* genes as the background set (Table 22.7). GO annotations exist for 78 of these 95 genes, which we did not differentiate further according to where in the gene overlap with an ultra-conserved element occurred. Genes involved in synaptic trans-



(a) 196 elements ≥ 75 and < 100 bp (b) 59 elements ≥ 100 bp

Fig. 22.8. Functional base coverage of ultra-conserved elements found in the *Drosophila* alignment based on annotations of known *D. melanogaster* genes.

mission are strongly over-represented in genes that have an ultra-conserved element overlap with their exons, introns and UTRs. These genes include those involved with ion channel activity, signal transduction and receptor activity, playing roles in intracellular signaling cascades, muscle contraction, development, and behavior. RNA binding proteins are also found to be over-represented. Another group of over-represented genes are those involved in RNA polymerase II transcription factor activity. These genes are strongly associated with development and morphogenesis.

The 130 ultra-conserved elements found in unannotated regions are grouped together into 109 regions by common flanking genes. These regions are flanked by 208 unique genes, 134 of which have available GO annotations. The distance from these ultra-conserved elements to their respective nearest gene ranges from 0.2–104 kb and is 16 kb on average. A number of transcription factors involved with development and morphogenesis are found within this set of genes. Five of the 10 flanking genes with ultra-conserved sequences both upstream and downstream are transcription factors (*SoxN*, *salr*, *toe*, *H15*, *sob*). In total, 44 unique transcription factors are found across the intragenic and flanking gene hits.

Ten of the original 255 ultra-conserved elements are repeated elsewhere in the *D. melanogaster* genome. However, all of these repeats correspond to annotated tRNA or snRNA, but not to homologous exons or regulatory regions. There are 10 ultra-conserved elements that overlap with tRNA (757 bp in sum), two that overlap with snRNA (191 bp in sum), and one that overlaps with ncRNA (81 bp). None of the ultra-conserved elements correspond to annotated rRNA, regulatory regions, transposable elements, or pseudogenes.

22.3.4 Discussion

We studied ultra-conserved elements in three very different datasets: an alignment of nine distant vertebrates, an alignment of the ENCODE regions in

GO Annotation	p-value
Exons, Introns, and UTRs (41)	
synaptic transmission	$3.290 \cdot 10^{-9}$
specification of organ identity	$1.044 \cdot 10^{-6}$
ventral cord development	$3.674 \cdot 10^{-6}$
RNA polymerase II transcription factor activity	$4.720 \cdot 10^{-6}$
muscle contraction	$8.714 \cdot 10^{-6}$
voltage-gated calcium channel activity	$3.548 \cdot 10^{-5}$
RNA binding	$7.650 \cdot 10^{-5}$
synaptic vesicle exocytosis	$3.503 \cdot 10^{-4}$
leg morphogenesis	$3.503 \cdot 10^{-4}$
calcium ion transport	$6.401 \cdot 10^{-4}$
Flanking within 104 kb (58)	
regulation of transcription	$8.844 \cdot 10^{-7}$
neurogenesis	$5.339 \cdot 10^{-6}$
ectoderm formation	$8.285 \cdot 10^{-6}$
endoderm formation	$2.125 \cdot 10^{-5}$
salivary gland morphogenesis	$5.870 \cdot 10^{-5}$
Notch signaling pathway	$1.591 \cdot 10^{-4}$
leg joint morphogenesis	$1.788 \cdot 10^{-4}$
RNA polymerase II transcription factor activity	$2.381 \cdot 10^{-4}$
salivary gland development	$4.403 \cdot 10^{-4}$
signal transducer activity	$5.308 \cdot 10^{-4}$
foregut morphogenesis	$8.004 \cdot 10^{-4}$

Table 22.7. *GO annotations of genes associated with Drosophila ultra-conserved elements. The number of GO annotations tested for each group are in parentheses. For each group, each tested GO annotation is associated with at least 3 genes in the group.*

mammals, and an alignment of eight fruit flies. As Figures 22.5, 22.7, and 22.8 show, ultra-conserved elements overlap with genes very differently in the three datasets. In particular, in the *Drosophila* dataset, exonic conservation is much more substantial. This conservation at the DNA level is very surprising, as the functional constraint on coding regions is expected to be at the amino acid level. Therefore, the degeneracy of the genetic code should allow synonymous mutations (see Section 21.3) to occur without any selective constraint.

The GO analysis showed that non-coding regions near or in genes associated with transcriptional regulation tended to contain ultra-conserved elements in all datasets. In *Drosophila*, ultra-conserved elements overlapped primarily with genes associated with synaptic transmission. While the exonic conservation in *Drosophila* is due in part to a much shorter period of evolution, the exact conservation of exons whose gene products are involved in synaptic transmission may be fly-specific.

Non-coding regions that are perfectly conserved across all 9 species may be precise regulatory signals for highly specific DNA-binding proteins. In particular, repeated ultra-conserved elements such as those found near the Iroquois homeobox genes on chromosome 16 are excellent candidates for such regulatory elements. Of course, it is interesting to note that the degree of conservation in

our ultra-conserved elements exceeds what is observed for other known functional elements, such as splice sites. We discuss the statistical significance of ultra-conservation in Section 22.4.

Many of our results mirror those of previous studies. However, these studies have considered long stretches of perfectly conserved regions across shorter evolutionary distances [Bejerano *et al.*, 2004], or aligned regions above some relatively high threshold level of conservation [Boffelli *et al.*, 2004, Sandelin *et al.*, 2004, Woolfe *et al.*, 2005]. We have focused on ultra-conserved elements across larger evolutionary distances. As a result, we have not captured all regions containing high levels of conservation, but have identified only those regions that appear to be under the most stringent evolutionary constraints.

22.4 Statistical significance of ultra-conservation

Which ultra-conserved elements are of a length that is statistically significant? In order to address this question, we choose a model and compute the probability of observing an ultra-conserved element of a given length for the nine-vertebrate and *Drosophila*-alignments. First we consider phylogenetic tree models. These models allow for dependence of the occurrence of nucleotides in the genomes of different species at any given position in the aligned genomes, but make the assumption that evolutionary changes to DNA at one position in the alignment occur independently from changes at all other, and in particular, neighboring positions. Later we also consider a Markov chain, which does not model evolutionary changes explicitly but incorporates a simple pattern of dependence among different genome positions.

Before being able to compute a probability in a phylogenetic tree model, we must build a tree and estimate the parameters of the associated model. The tree for the nine-vertebrate alignment is shown in Figure 22.1. The topology of this tree is well-known, so we assume it fixed and use `PAML` [Yang, 1997] to estimate model parameters by maximum likelihood. As input to `PAML`, we choose the entire alignments with all columns containing a gap removed. The resulting alignment was 6,300,344 positions long for the vertebrates and 26,216,615 positions long for the *Drosophila*. Other authors (see Chapter 21 or [Pachter and Sturmfels, 2005]) have chosen to focus only on synonymous substitutions in coding regions, since they are likely not selected for or against and thus give good estimates for neutral substitution rates. However, our independence model does not depend on the functional structure of the genome; that is, it sees the columns as i.i.d. samples. Thus, we believe that it is more appropriate to use all the data available to estimate parameters.

There are many phylogenetic tree models (Section 4.5) and we concentrate here on the Jukes–Cantor and HKY85 models. With the parameter estimates from `PAML`, we can compute the probability p_{cons} of observing an ultra-conserved position in the alignment. Recall that the probability $p_{i_1 \ldots i_s}$ of seeing the nucleotide vector $(i_1, \ldots, i_s) \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^s$ in a column of the

alignment of s species is given by a polynomial in the entries of the transition matrices $P_e(t)$, which are obtained as $P_e(t) = \exp(Qt_e)$ where t_e is the length of the edge e in the phylogenetic tree and Q is a rate matrix that depends on the model selected.

Under the Jukes–Cantor model for the nine-vertebrate alignment, the maximum likelihood (ML) branch lengths are shown in Figure 22.1 and give the probabilities

$$p_{\texttt{AAAAAAAAA}} = \cdots = p_{\texttt{TTTTTTTTT}} = 0.01139...$$

Thus, the probability of a conserved column under this model is $p_{\text{cons}} = 0.0456$. If we require that the nucleotides are identical not only across present-day species but also across ancestors, then the probability drops slightly to 0.0434.

Under the HKY85 model for the nine-vertebrate alignment, the ML branch lengths are very similar to those in Figure 22.1 and the additional parameter is estimated as $\kappa = 2.4066$ (in the notation of Figure 4.7, $\kappa = \alpha/\beta$). The root distribution is estimated to be almost uniform. These parameters give the probabilities

$$p_{\texttt{AAAAAAAAA}} = \cdots = p_{\texttt{TTTTTTTTT}} = 0.00367,$$

which are much smaller than their counterpart in the Jukes–Cantor model. The HKY85 probability of a conserved column is $p_{\text{cons}} = 0.014706$. If we assume that nucleotides must also be identical in ancestors, this probability drops to 0.01234.

The binary indicators of ultra-conservation are independent and identically distributed according to a Bernoulli distribution with success probability p_{cons}. The probability of seeing an ultra-conserved element of length at least ℓ starting at a given position in the alignment therefore equals p_{cons}^{ℓ}. Moreover, the probability of seeing an ultra-conserved element of length at least ℓ anywhere in a genome of length N can be bounded above by Np_{cons}^{ℓ}. Recall that the length of the human genome is roughly 2.8 Gbp and the length of *D. melanogaster* is approximately 120 Mbp. Table 22.8 contains the evaluated probability bound for different values of ℓ.

	Nine-vertebrate (human)			*Drosophila* (*D. melanogaster*)	
	Jukes–Cantor	HKY85		Jukes–Cantor	HKY85
p_{cons}	0.0456	0.0147	p_{cons}	0.1071	0.05969
10	0.0001	$1.3 \cdot 10^{-9}$	15	$7.8 \cdot 10^{-6}$	$1.2 \cdot 10^{-9}$
20	$4.1 \cdot 10^{-18}$	$6.2 \cdot 10^{-28}$	75	$4.6 \cdot 10^{-64}$	$4.3 \cdot 10^{-83}$
125	$6.0 \cdot 10^{-159}$	$2.4 \cdot 10^{-220}$	209	$4.3 \cdot 10^{-194}$	$4.1 \cdot 10^{-247}$

Table 22.8. *Probabilities of seeing ultra-conserved elements of certain lengths in an independence model with success probability* p_{cons} *derived from two phylogenetic tree models.*

However, 46% of the ungapped columns in the nine-vertebrate alignment are actually ultra-conserved. This fraction is far greater than the 5% we would expect with the JC model and the 1% under the HKY85 model. This suggests

that the model of independent alignment positions is overly simplistic. If we collapse the alignment to a sequence of binary indicators of ultra-conserved positions, then a very simple non-independence model for this binary sequence is a Markov chain model (cf. Section 1.4 and Chapter 10).

In a Markov chain model, the length of ultra-conserved elements is geometrically distributed. That is, the probability that an ultra-conserved element is of length ℓ equals $\theta^{\ell-1}(1-\theta)$, where θ is the probability of transitioning from one ultra-conserved position to another. The expected value of the length of an ultra-conserved element is equal to $1/(1-\theta)$. The probability that an ultra-conserved element is of length ℓ or longer equals

$$\sum_{k=\ell}^{\infty} \theta^{k-1}(1-\theta) = \theta^{\ell-1}.$$

Therefore, the probability that at least one of U ultra-conserved elements found in a multiple alignment is of length at least ℓ is equal to

$$1-(1-\theta^{\ell-1})^U \approx U \cdot \theta^{\ell-1} \quad \text{for large } \ell.$$

Restricting ourselves to the nine-vertebrate alignment (computations for the *Drosophila* alignment are qualitatively similar), we used the mean length of the ultra-conserved elements described in Section 22.3.1 to estimate the transition probability θ to 0.4785. Then the probability that at least one of the 1,513,176 ultra-conserved elements of the nine-vertebrate alignment is of length 25 or longer equals about 3%. The probability of seeing one of the U ultra-conserved elements being 30 or more bp long is just below 1/1000. However, the dependence structure in a Markov chain model cannot explain the longest ultra-conserved elements in the alignment. For example, the probability of one of the U elements being 125 or more bp long is astronomically small (0.3×10^{-33}). This suggests that the Markov chain model does not capture the dependence structure in the binary sequence of ultra-conservation indicators. At a visual level, this is clear from Figure 22.3. Were the Markov chain model true then, due to the resulting geometric distribution for the length of an ultra-conserved element, the log-scale frequencies should fall on a straight line, which is not the case in Figure 22.3. Modeling the process of ultra-conservation statistically requires more sophisticated models. The phylogenetic hidden Markov models that appear in [McAuliffe *et al.*, 2004, Siepel and Haussler, 2004] provide a point of departure.

Despite the shortcomings of the calculations, it is clear that it is highly unlikely that the ultra-conserved elements studied in this chapter occur by chance. The degree of conservation strongly suggests extreme natural selection in these regions.