Evolution of Gene Deserts in the Human Genome

Advanced article

James Taylor, *New York University, New York, USA*

'Gene deserts' are improbably large regions of the human genome that contain no genes. The evolutionary origin of these regions is still unknown, but experimental and computational evidence suggests there are multiple classes of gene deserts in the human genome, each with different evolutionary origins and functional associations.

Online posting date: 14th March 2008

Introduction

'Gene deserts' are large genomic regions that are completely free of genes. Genes in the human genome (and other vertebrate genomes) are not distributed randomly, and the resulting gene free regions are both longer and more frequent than would be expected by chance.

Human genes are still not exhaustively annotated, and there is no precise definition of the minimum size of intergenic interval that should be considered a gene desert. Thus, exact annotation of gene deserts is impossible. Frequently minimum lengths of 400–700 kb are used, which is consistent with about 20–30% of the human genome being contained in gene deserts. Despite these difficulties with exact annotation, the gene desert phenomenon appears to be robust. This is true even when considering only transcription rather than gene annotations supported by protein evidence. For example, The FANTOM Consortium (2005) performed a comprehensive survey of transcription in the mouse genome using several experimental techniques, and found 18 641 dense clusters of transcription activity, separated by transcription deserts.

The fraction of the genome that is functionally important is still an open question. It was once believed that protein-coding regions made up the full functional complement of the genome, and that the vast majority of the sequence was 'junk DNA (deoxyribonucleic acid)'. Gene deserts in particular, because they lack protein-coding sequence, were likely junk DNA candidates. Comparisons between the genomes of different species are a useful tool to understand the functional complement of genomes. Functional sequences are subject to selection. In particular, purifying selection causes sequences to change more slowly that they would at random, and this signal can be used to identify 'conserved' regions between species. Comparative genomic analyses suggest that *at least* 5% of the human genome is under purifying selection, but that protein-coding regions of genes account for at most 2% of the genome. The remainder is likely involved in other functions, including the regulation of gene transcription. Between-species sequence comparisons also reveal large sets of conserved noncoding regions, and a variety of methods have been described for identifying these sequences. Interestingly, regardless of the specific species comparisons or methods used, conserved noncoding sequences appear to be enriched in gene deserts.

The evolutionary origins and functional significance of gene deserts are still not well understood. Some gene deserts have been shown to contain functional elements (particularly *cis*-regulatory elements), while others have been shown to be unnecessary for healthy animal development. This dichotomy is consistent with results from between-species sequence comparisons, which suggest that there are at least two different classes of gene deserts, one of which tends to contain more highly conserved sequences and shows associations with regulation and development.

Functional Significance of Gene Deserts

Some gene deserts may be disposable

The complete lack of protein-coding regions in gene deserts make then a natural candidate for containing nonfunctional 'junk DNA'. A growing understanding of the complexity of genomic function has raised questions about the amount of (or even existence of) nonfunctional genomic sequence. However, there is evidence that at least some gene deserts contain no sequences that serve necessary functions, and thus can be deleted from the genome with no adverse effects.

Two potentially superfluous gene deserts were identified by Nóbrega *et al.* in 2004 who deleted a 1817-kb gene desert from mouse chromosome 3 (see **Figure 1**) and a 983-kb gene desert from mouse chromosome 19. They found that mice homozygous for both deletions were healthy and showed no detectable differences from wild-type littermates in

ELS subject area: Evolution and Diversity of Life

How to cite:
Taylor, James (March 2008) Evolution of Gene Deserts in the Human Genome. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020830

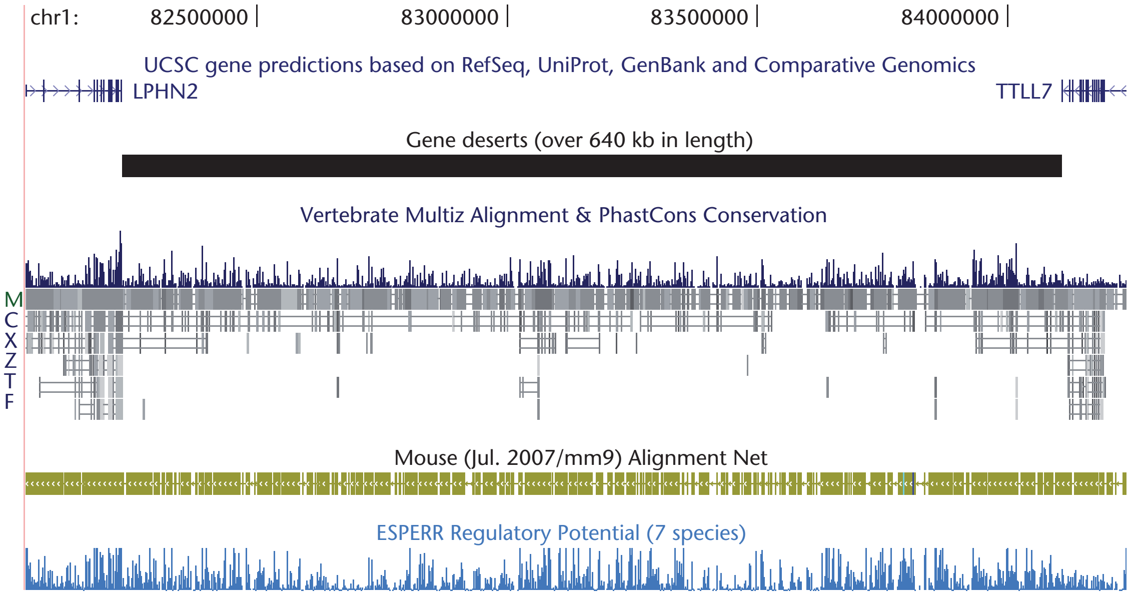
chr1:
82500000
83000000
83500000
84000000
UCSC gene predictions based on RefSeq, UniProt, GenBank and Comparative Genomics
LPHN2
TTLL7
Gene deserts (over 640 kb in length)
Vertebrate Multiz Alignment & PhastCons Conservation
M
C
X
Z
T
F
Mouse (Jul. 2007/mm9) Alignment Net
ESPERR Regulatory Potential (7 species)

Figure 1 Gene desert on human chromosome 1 flanked by *LPHN2* and *TTLL7*. The orthologous gene desert on mouse chromosome 3 is one of two gene deserts that were deleted in mice with no observable phenotypic consequences (by Nóbrega *et al*). View from UCSC Genome Browser (Kent *et al.*, 2002).

overall fitness or for a variety of other phenotypic parameters. They also performed quantitative gene expression assays in 12 tissues for 9 genes flanking the deleted gene deserts, and found measurable expression differences in only two cases. Although it is possible that the mice with deleted gene deserts possess some abnormalities that are not easily measured, these results at least suggest that some gene deserts may contain no functionally important sequences. Intriguingly, the deleted gene deserts contained 1243 long regions conserved between human and mouse (over 100 bp and more than 70% identical). Five of these conserved elements were tested in a transgenic mouse enhancer assay, and one of the five elements tested showed reproducible enhancer activity. Thus the deleted regions contain at least one potential regulatory element; however the effect of this element appears to be either unobservable or redundant.

Some gene deserts contain regulatory elements

In stark contrast to the two gene deserts described earlier, which have no observable phenotypic consequence when deleted in mice, gene deserts have also been shown to contain many regions that act as regulatory enhancers.

Human *DACH1*, a gene which is expressed in many tissues and has been linked to brain, limb and sensory organ development, is flanked by two gene deserts (870 kb upstream and 1330 kb downstream). In 2003 Nobrega *et al.* identified 32 conserved noncoding sequences in the region containing *DACH1* and the two gene deserts. These regions were selected using a stringent interspecies conservation criteria: requiring high conservation with mouse (greater than 70% identity over 100 bp or more) as well as aligning sequence in frog and three different fish. Nine of the 32 elements were sampled and tested for *in vivo* enhancer activity in transgenic mice. Of the nine elements tested, seven were shown to reproducibly drive expression of a reporter gene, and to recapitulate aspects of the endogenous expression pattern associated with *DACH1*. The authors further noted that the synteny of these elements is conserved across the mammals and fish investigated, suggesting the possibility of a functional relationship between these sequences that depends on their relative order. The human *DACH1* region including the putative enhancers tested is shown in **Figure 2**.

Several other studies have identified enhancer elements in gene deserts:

- The *Iroquois* genes are important for patterning during vertebrate embryonic development, and are organized into two clusters in the human genome that surround four gene deserts. De la Calle-Mustienes *et al.* (2005) identified several regions in the human *Iroquois B* gene cluster that act as enhancers in frog and zebrafish.
- The 1400 kb gene desert downstream of human *ISL1* contains at least two highly conserved elements whose zebrafish orthologues act as enhancers (Uemura *et al.*, 2005), as well as an enhancer derived from a transposable element that recapitulates multiple aspects of *Isl1* expression in transgenic mouse embryos (Bejerano *et al.*, 2006).
- A more comprehensive study by Pennacchio *et al.* (2006) tested 167 conserved noncoding regions in a transgenic mouse enhancer assay, and identified 75 active enhancers, of which 36 fall in gene deserts (defined as 400 kb or greater euchromatic intergenic regions based on the UCSC (University of California, Santa Cruz) genes).

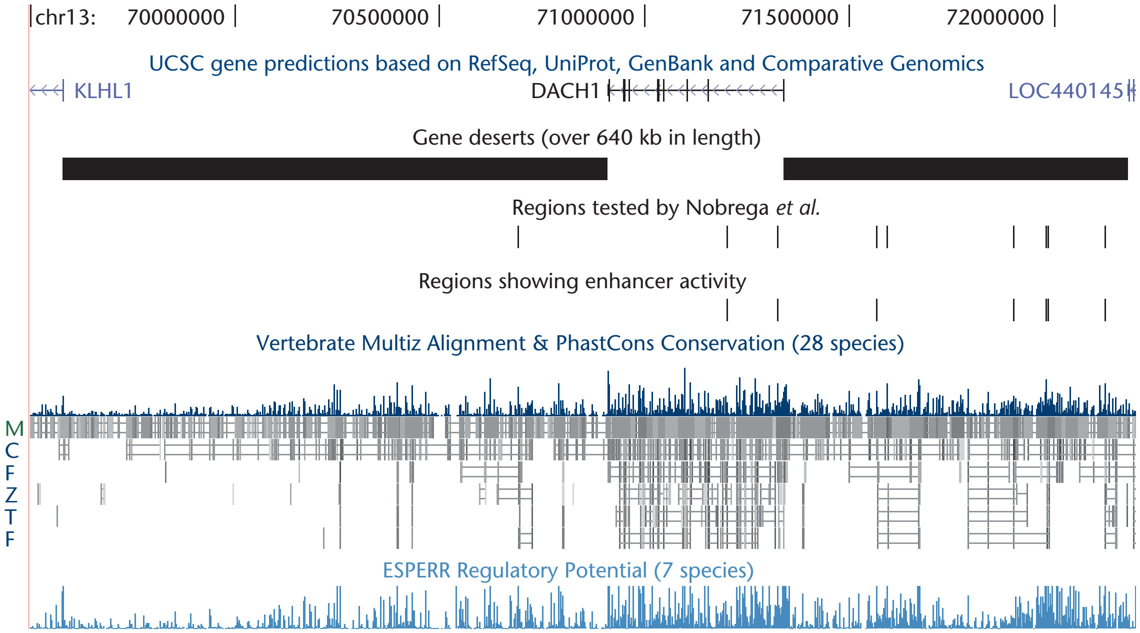

Figure 2 Region on human chromosome 13 showing *DACH1* and the two gene deserts that flank it. The nine highly conserved regions that tested for enhancer activity by Nobrega *et al*. in 2003 are shown, along with the seven regions that tested positive. View from UCSC Genome Browser (Kent *et al*., 2002).

Demonstrating that a sequence can act as an enhancer in a constructed enhancer assay does not conclusively demonstrate that it has an important functional role in its normal context (which is generally much more difficult to assay). Another approach that can be used to identify potentially functional genomic regions is an association study, in which many different individuals are scanned for genetic markers that are associated with different traits. If large numbers of genomic single nucleotide polymorphisms (SNPs) are used as markers, it can be possible to map regions of the genome associated with particular traits at high resolution. Libioulle *et al*. (2007) performed such a study to identify regions associated with Crohn disease, genotyping 302 451 SNPs in 547 patients with Crohn disease and 928 healthy controls. In addition to finding several regions already known to be associated with Crohn disease, they found a 250-kb region on human chromosome 5 that contained 6 SNPs highly associated with the Crohn disease patients ($p < 10^{-6}$). Genotyping 1266 additional Crohn disease patients and 559 additional controls confirmed these strong associations. The 250-kb region identified is contained in a 1.25-Mb gene desert. The region is devoid of CpG islands (which are often associated with transcription start sites) but does contain many highly conserved elements. Further, analysis of 378 individuals with both genome-wide gene expression and genome-wide genotype data available showed an association between 8 of 26 SNPs corresponding to the Crohn disease-associated region and *PTGER4*. These results support the hypothesis that the Crohn disease-associated region within the gene desert contains one or more distal *cis*-regulatory elements.

Many gene deserts contain conserved elements that act as regulatory enhancers and recapitulate aspects of the expression pattern of the flanking genes. Additionally, variations in gene deserts have been shown to be associated with at least one human disease, as well as affecting gene expression in cell lines. Combined, these results strongly suggest the widespread presence of distal regulatory elements in at least some subset of gene deserts.

Gene deserts may be important for higher order chromatin organization

The information encoded in the primary sequence is not the only aspect of the genome that is of functional importance. Regions of the eukaryotic genome are spatially organized within the cell nucleus, and the way in which chromatin is folded and where it is located in the nucleus is an important part of genomic function. In particular, groups of genes that are distantly located along the chromosome are sometimes collocated in specific nuclear compartments. In 2006 Shopland *et al*. studied the spatial organization of a 4.3-Mb region on mouse chromosome 14, which contains 19 genes concentrated in 4 clusters separated by 3 gene deserts of at least 400 kb in length. This cluster and desert architecture is conserved in human (see **Figure 3**). They found that the gene clusters in this region tended to associate spatially – neighbouring gene clusters overlapped in the nucleus much more frequently than neighbouring gene deserts. Analysis of 132 different chromosomes showed that the chromatin for the region of interest tended to be arranged in one of a few specific conformations, typically involving ‘hubs’ of colocated gene clusters. Gene deserts and clusters also showed differential positioning in the nucleus, with gene deserts preferentially aligning in the nuclear periphery. These results suggest that gene clusters and gene deserts

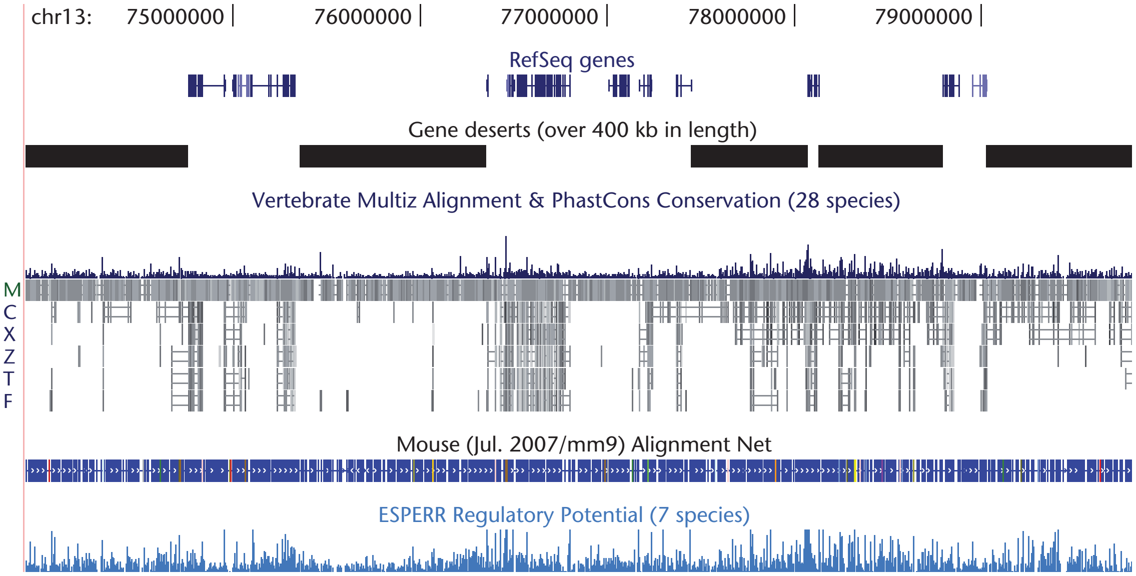

Figure 3 Region of human chromosome 13 containing four gene clusters separated and surrounded by gene deserts. The homologous region on mouse chromosome 14 was examined by Shopland *et al.* in 2006, who suggest a relationship between the cluster/desert architecture of this region with higher order chromatin structure and nuclear positioning.

fold together in different combinations and with different frequencies, yielding certain predominant conformations. Thus, organization of genes into clusters and deserts may be important for determining higher order chromatin structure and nuclear organization.

Evolutionary Comparisons Reveal Distinct Classes of Gene Deserts

Gene deserts separate into stable and variable classes

Experimental support for both function and absence of function in various human gene deserts indicates that there may in fact be multiple distinct classes of gene deserts, some of which contain functionally important sequence elements and some of which do not. Between-species genome comparisons are a natural way to investigate this question, since we expect functionally important sequences in gene deserts to be evolutionarily conserved. In 2005 Ovcharenko *et al.* identified 545 human gene deserts (defined as the largest 3% of intergenic intervals, yielding a minimum size of 640 kb). The conservation of these regions was evaluated by determining the fraction of each gene desert covered by regions highly conserved between human and chicken (again greater than 70% identity over 100 bp or more). Although the average density of conserved regions in gene deserts is similar to that of other intergenic intervals, the distribution of conserved region density is broad, and not explained simply by the density of transposable elements. The gene deserts were separated into two categories: 172 *stable* gene deserts having more than 2% coverage by conserved regions, and 373 *variable* gene deserts have less than 2% coverage. Although this partitioning is arbitrary, it appears to be robust to changes in the underlying conservation metric. The authors also rule out low neutral substitution rates as an explanation for the observed constraint in gene deserts by showing that both classes of gene deserts actually have a neutral mutation rate higher than the genome average, suggesting that conserved regions in stable gene deserts might be under purifying selection.

Stable gene deserts appear strongly associated with regulatory function

The stable gene deserts have a variety of characteristics that suggest functional importance, and particularly association with transcriptional regulation and development. Stable gene deserts flank several genes which have been shown experimentally to have distal regulatory elements, including *DACH1*, *OTX2* and *SOX2*, which suggests that these deserts might contain regulatory elements. Comprehensive experimental identification of regulatory regions is not yet possible, but computational methods can be used to score genomic regions for possible regulatory function. In stable gene deserts, the density of putative regulatory elements predicted computationally by the regulatory potential score (Taylor *et al.*, 2006) is three times higher than in variable gene deserts.

Genes flanking stable gene deserts also are enriched for certain Gene Ontology (GO) categories when compared to all annotated human genes. Several GO categories related to gene regulation and development are substantially enriched, including ‘regulation of metabolism’ (4.4-fold enrichment), ‘transcription factor activity’ (4.2-fold), 'regulation of transcription' (2.7-fold) and 'development' (2-fold).

Synteny of stable gene deserts is highly conserved between human and chicken, with a frequency of synteny breakpoints 10 times lower than the genome average. For almost all stable gene deserts, the linear order of conserved regions and flanking genes is largely unchanged. Among the stable gene deserts, 56 are paired with another stable gene desert by an intervening region of at most 1 Mb containing at most 3 genes. These 'conjoined stable gene deserts' show even stronger functional associations with gene regulation, and flank genes that are known or believed to be involved in key developmental steps or critical processes.

Several of the largest gene deserts in the human genome are located on human chromosomes 2 and 4. Hillier *et al.* (2005) note that a 3.3-Mb stable gene desert on chromosome 2 has an interesting distribution of conserved regions: while the entire interval shows conservation between human, mouse and dog, only the 2 Mb closest to the flanking gene *ZFHX1B* shows conservation with other vertebrates (specifically chicken and fish). This raises the interesting question of whether the sequences in the more deeply conserved portion have a functional significance that is different from the regions conserved only in mammals. Using machine-learning techniques, the authors found that short nucleotide patterns (4–9 bases) in the human sequence alone were sufficient to discriminate between these two sets of sequences with 75% accuracy. The presence of such discriminative patterns is further indicative that these regions may contain functional sequence.

Stable gene deserts may contain critical vertebrate regulatory elements

Stable gene deserts contain highly conserved sequences, frequently flank genes related to transcription regulation and vertebrate development, and show remarkable conservation of synteny over long evolutionary distances. These observations suggest an intriguing hypothesis for the origin of stable gene deserts: these regions may contain highly constrained regulatory sequences that are of critical importance for vertebrate development. As the genome has evolved, expansions between these regions have been tolerated, but integrity of the individual elements, their order along the chromosome and position relative to the genes they control have been preserved.

Variable gene deserts also show some evidence for function

There is also evidence that variable gene deserts may also have some functional importance. Ovcharenko *et al.* also evaluated GO enrichments for genes flanking variable gene deserts, and found strong enrichments for a variety of functional categories. Hiller *et al.* found that the protocadherin gene *PCDH7* is surrounded by two large variable gene deserts, one greater than 5 Mb in length – the largest in the human genome – and one of 2.5 Mb in length. Although the intervening gene deserts are of the variable type, this structure, including flanking genes, is conserved in among human, mouse, dog and chicken. They also identified another protocadherin (*PCDH10*) that is flanked by two multimegabase gene deserts. The protocadherins are a family of membrane proteins important for cell–cell adhesion and are well-conserved among vertebrates. These two genes in particular are believed to have diverged before the separation of mammals and fish. The fact that this gene desert architecture has been maintained over such a long evolutionary distance suggests that these variable gene deserts are of some functional importance.

'Large ancient duplications' created some human gene deserts

Unequal crossing-over during recombination can duplicate large portions of the genomic sequence. Duplication has been a major driving force in genome evolution, from the ancient duplications that have generated human gene families to copy number variation within the current human population.

The lack of conservation with chicken that is characteristic of variable gene deserts suggests that many gene deserts may have evolved more recently in human evolution. In 2005 Itoh *et al.* found that self-similarity in the human sequence of a gene poor region on chromosome 21, was explained by a low-similarity duplication spanning a total of 3.4 Mb. A genome wide search for additional low-similarity duplications identified 14 addition large ancient duplications, all of greater than 1 Mb in length, with duplicons of 0.1–2.6 Mb and copy numbers of 2–6. One large ancient duplication of chromosome 13 is shown in **Figure 4**. All 14 large ancient duplications were associated with gene deserts, and most make up the majority of the sequence in the gene desert they are associated with. Interestingly, all large ancient duplications could be identified in dog as well as human, but none in chicken. The fraction of large ancient duplications covered by regions highly conserved with chicken was at most 1.05%.

Summary

Gene deserts remain a mysterious aspect of the architecture of the human genome. However, the available experimental evidence and computational analyses suggest distinct classes of gene deserts with different evolutionary origins.

One putative class of gene deserts contains those that flank critical developmental genes. These gene deserts contain deeply conserved distal regulatory regions that are crucial for core aspects vertebrate development. Proper development depends on the position of these regions relative to each other and to the genes they regulate. As the vertebrate genome has evolved and expanded these regions have allowed insertions between the crucial elements, but have resisted rearrangements and other evolutionary

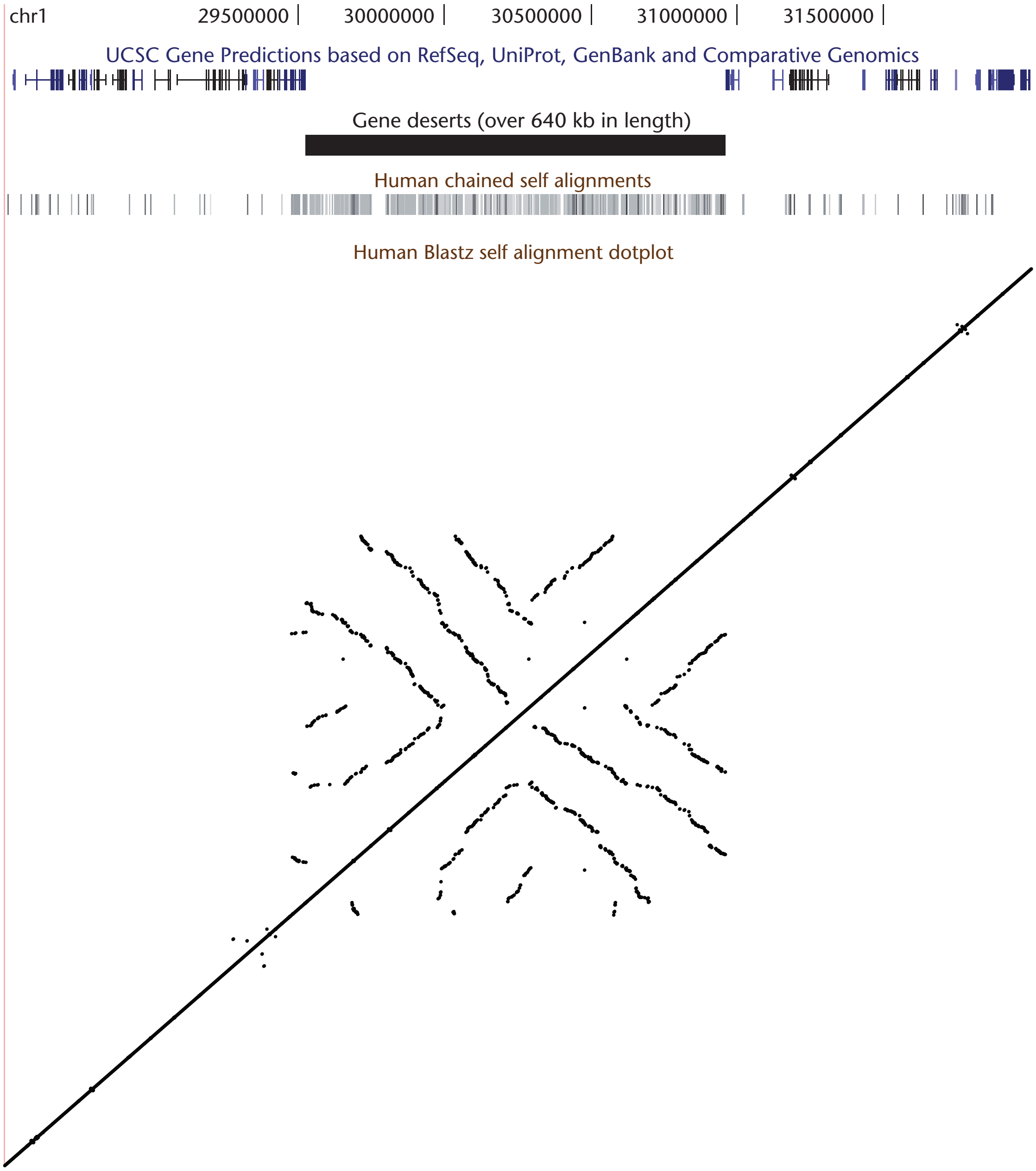

Figure 4 Region of human chromosome 13 containing a gene desert that corresponds to a large ancient duplication (LAD) identified by Itoh *et al*. The top portion of the plot is a view from the UCSC Genome Browser (Kent *et al*., 2002) showing the location of genes, the gene desert and genome-wide chained human self-alignments. The bottom portion shows a dotplot generated from Blastz (Schwartz *et al*., 2003) alignments of this human region with itself (each line corresponds to a local alignment). This representation clearly shows multiple large low-similarity duplications corresponding to the position of this gene desert.

changes that would affect their regulatory function, thus creating gene deserts.

A second putative class of gene deserts contains those that are truly superfluous. These regions contain far fewer highly conserved regions, either because they are evolving neutrally or were generated more after the divergence of birds and mammals. At least one mechanism exists that can explain the generation of these deserts: large ancient duplications have generated multimegabase regions of the human genome, which have since diverged to the point where only limited self-similarity can be observed. These regions in general do not contain important functional elements, tolerate rearrangements and can be removed from the genome entirely without major phenotypic effects.

A third class of putative gene deserts contains those that have a structural role in higher order folding and nuclear positioning of chromatin. These gene deserts have a tendency to locate in the nuclear periphery, and exhibit different frequencies of collocation with gene-rich versus gene-poor regions. Chromatin structure and nuclear location are both important for cellular function, and selection may be acting on the size of these gene deserts or on the sequence of blocks of gene clusters and deserts along the genome. However, it is unclear how this class of gene deserts relates to the other two. It is possible that these regions need to be maintained at a certain size and place but have no sequence constraint; however, there are also primary sequence elements that effect chromatin structure, so the interactions may be more complex.

References

Bejerano G, Lowe CB, Ahituv N *et al*. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**(7089): 87–90.

de la Calle-Mustienes E, Feijóo CG, Manzanares M *et al*. (2005) A functional survey of the enhancer activity of conserved noncoding sequences from vertebrate Iroquois cluster gene deserts. *Genome Research* **15**(8): 1061–1072.

Hillier LW, Graves TA, Fulton RS *et al*. (2005) Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**(7034): 724–731.

Kent WJ, Sugnet CW, Furey TS *et al*. (2002) The Human Genome Browser at UCSC. *Genome Research* **12**(6): 996–1006.

Libioulle C, Louis E, Hansoul S *et al*. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genetics* **3**(4): e58.

Pennacchio LA, Ahituv N, Moses AM *et al*. (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499–502.

Schwartz S, Kent WJ, Smit A *et al*. (2003) Human–mouse alignments with BLASTZ. *Genome Research* **13**(1): 103–107.

Taylor J, Tyekucheva S, King DC *et al*. (2006) ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Research* **16**(12): 1596–1604.

The FANTOM Consortium (2005) The transcriptional landscape of the mammalian genome. *Science* **309**(5740): 1559–1563.

Uemura O, Okada Y, Ando H *et al*. (2005) Comparative functional genomics revealed conservation and diversification of three enhancers of the isl1 gene for motor and sensory neuron-specific expression. *Developmental Biology* **278**(2): 587–606.

Further Reading

Itoh T, Toyoda A, Taylor TD, Sakaki Y and Hattori M (2005) Identification of large ancient duplications associated with human gene deserts. *Nature Genetics* **37**(10): 1041–1043.

Nobrega MA, Ovcharenko I, Afzal V and Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* **302**(5644): 413.

Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V and Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. *Nature* **431**(7011): 988–993.

Ovcharenko I, Loots GG, Nobrega MA *et al*. (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Research* **15**(1): 137–145.

Shopland LS, Lynch CR, Peterson KA *et al*. (2006) Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *Journal of Cell Biology* **174**(1): 27–38.