Orientation, distance, regulation and function of neighbouring genes

*Adrian Gherman,[1] Ruihua Wang[2] and Dimitrios Avramopoulos[1,2]**

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, School of Medicine, 733 North Broadway, Baltimore, MD 21205, USA
[2]Department of Psychiatry, Johns Hopkins University, School of Medicine, 600 North Wolfe Street, Baltimore, MD 21287, USA
**Correspondence to:* Tel: +1 410 955 8323; Fax: +1 410 955 7397; E-mail: adimitr1@jhmi.edu

Date received (in revised form): 4th August 2008

Abstract

The sequencing of the human genome has allowed us to observe globally and in detail the arrangement of genes along the chromosomes. There are multiple lines of evidence that this arrangement is not random, both in terms of intergenic distances and orientation of neighbouring genes. We have undertaken a systematic evaluation of the spatial distribution and orientation of known genes across the human genome. We used genome-level information, including phylogenetic conservation, single nucleotide polymorphism density and correlation of gene expression to assess the importance of this distribution. In addition to confirming and extending known properties of the genome, such as the significance of gene deserts and the importance of 'head to head' orientation of gene pairs in proximity, we provide significant new observations that include a smaller average size for intervals separating the 3′ ends of neighbouring genes, a correlation of gene expression across tissues for genes as far as 100 kilobases apart and signatures of increasing positive selection with decreasing interval size surprisingly relaxing for intervals smaller than ~500 base pairs. Further, we provide extensive graphical representations of the genome-wide data to allow for observations and comparisons beyond what we address.

Keywords: *genome*, *gene orientation*, *gene expression*, *gene function*, *phylogenetic conservation*

Introduction

With the sequencing of the human genome,[1,2] the identification of most human genes, a goal of geneticists for many years, has become a reality. Additionally, most human protein coding genes have been placed in the genome through sequence alignments, and their localisation and direction of transcription are known. The positioning of genes within the genome expands what we know for each gene to include the neighbouring genes and genomic context.

The distance and relative transcriptional direction of adjacent genes is known to be important in some organisms, but has not been studied intensively in humans. For example, in prokaryotes, genes are often arranged in operons, transcribed in a single transcript and thus co-regulated. Such polycistronic transcription has been described in eukaryotes, yet its extent and importance remains unclear.[3–5] Co-regulation of genes transcribed on opposite strands, with their transcription start sites in proximity, has been described in humans and the existence of common regulatory elements has been shown experimentally in some cases.[6–11] The literature refers to this orientation as 'head to head' (HH) and we will use this nomenclature here, naming the three possible orientations as shown in Figure 1. To date, a number of studies have addressed the importance of HH orientation for genes that are close to each other. Adachi and Lieber,[12] examining DNA repair genes, housekeeping genes but also a functionally unbiased set, first observed that among genes that are

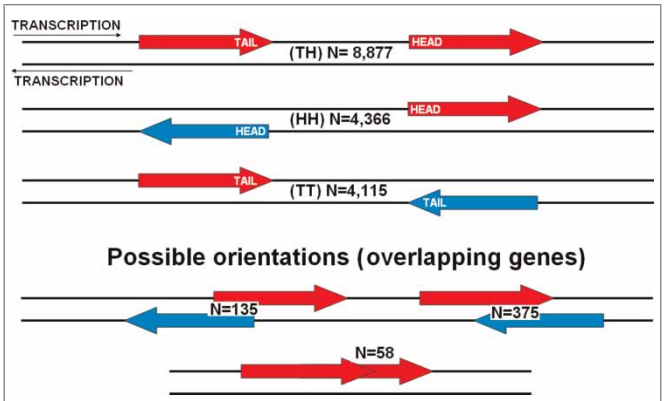

Figure 1. Possible orientations of neighbouring genes. The possible orientations of neighbouring genes, how they are referred to in the text (in parentheses) and the number of such pairs we observed in the genome are shown.

in close proximity, HH genes are more common. Trinklein *et al.*[13] greatly expanded the number of genes studied and showed that these HH pairs also show correlated expression, that many involve shared regulatory elements and that their arrangement is conserved in the mouse genome. Koyanagi *et al.*[14] expanded the analysis to many species, showing that this is a property specific to mammalian genomes. A study by Li *et al.*[15] further supported these results, showing conservation of the HH arrangement, correlation of expression and similarity of function. Studies confined to other organisms have also provided interesting data. Cho *et al.*[16] and Kruglyak and Tang[17] showed that adjacent genes are co-regulated in yeast, while Williams and Bowles[18] showed the same in *Arabidopsis thaliana*, with HH genes showing higher correlations but longer average distances than tail to tail (TT) genes. Similarly, Roy *et al.*[19] showed clustering of co-expressed genes in *Caenorhabditis elegans*. Finally, Fukuoka *et al.*[20] compared gene distance and co-expression in six eukaryotes and found a correlation in all six, although with significant differences between them. In contrast to nearby HH genes, little research has focused on longer intergenic distance and other orientations. Some reported work on the TT-oriented gene has focused on how antisense transcription might play a role in their regulation.[21–24]

Compared to previous work, in this study we expanded the search for evidence of functional importance to all non-overlapping gene orientations and distances and investigated the properties of the intergenic intervals, as well as the genes themselves.

Materials and methods

Gene location data

Our primary data source was the University of California, Santa Cruz (UCSC) genome database and browser (UCSC Genome Bioinformatics, http://genome.ucsc.edu),[25,26] and we used scripts written in Perl (http://www.perl.org) for data parsing and analysis. We used the March 2006 assembly of the human genome, which was annotated at the time of data acquisition using RefSeq version 21 (National Center for Biotechnology Information, Bethesda, MD)[26] We downloaded information for all genes in RefSeq and used exon coordinates to define their start and end locations. We excluded from the analysis all genes located entirely within other genes. For overlapping genes, we searched for shared exons and, if present, we concatenated the genes into one. If no shared exons were identified, we analysed each gene only in relation to its non-overlapping neighbours. Of the remaining 17,531 intergenic intervals, 173 were removed from the analysis because they contained sequence gaps and/or were across centromeres, leaving us with a final set of 17,358 intervals ranging from one base pair (bp) to ~ 4.9 mega bp (Mbp). The distribution of intervals if genes were positioned at random was calculated using a random number generator to assign 17,358 random points on the total intergenic length (1.7 Mbp) and examining the distribution of the distances between them in multiple iterations.

Phylogenetic conservation data

We retrieved coordinates of phylogenetically conserved elements from the UCSC genome browser in a table generated by the PhastCons algorithm, which uses sequences from 28 species and a phylogenetic hidden Markov model to identify varying lengths of sequence likely to be conserved.[27] Using these and the intergenic interval coordinates, we calculated the fraction of conserved bases for each

interval, as well as separately for each half of each interval, closer to each of the neighbouring genes. The latter was performed to investigate whether in tail to head (TH) intervals there is a difference in conservation towards the end of one gene or the beginning of the next, the latter being expected to carry more regulatory elements.

Conserved transcription factor binding sites (TFBSs)

We retrieved the coordinates of predicted TFBSs from three-species (human/mouse/rat) comparisons[28] from the UCSC genome browser. The scores and threshold for identifying TFBSs are computed using the Transfac matrix database (v7.0). A binding site is considered to be conserved across the alignment if its score meets the threshold score for its binding matrix in all three species.[28] We mapped the predicted TFBS locations to the intergenic intervals and also calculated TFBS statistics for each half of each interval.

Expression data from the Genomic Institute of the Novartis Research Foundation – Gene Express Atlas 2 (GNF2)

We obtained the GNF2 data[29] from the UCSC genome database. This dataset consists of genome-wide gene expression (mRNA) measurements from 79 different human healthy and diseased tissues. We excluded seven tissues because they were neoplastic and could confound our results. Data were available for both genes of the pair for 10,397 pairs. For these, we calculated the correlation coefficient r of the 72 expression level measurements. The calculated r was Fisher transformed to make the correlation coefficients normally distributed and allow the use in statistical tests for comparisons.[30]

Microarray gene expression data

Data on RNA from the superior temporal lobe of 23 brain donors with no gross brain pathology was generated in our laboratory using the Illumina Sentrix HumanRef-8 Expression BeadChips (Illumina, San Diego, CA, USA), containing 24,000 genes recognised by the National Center for Biotechnology Information (NCBI) at the time of production. We extracted total RNA using Trizol (Invitrogen, Carlsbad, CA, USA) with additional purification on RNAeasy columns (Qiagen, Valencia, CA, USA). We assessed the quality of total RNA on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and 0.5 μg of total RNA from each sample was labelled using the Illumina TotalPrep RNA Amplification Kit (Ambion, Austin, TX, USA) in a process of cDNA synthesis and *in vitro* transcription. We generated and labelled single-stranded RNA (cRNA) by incorporating biotin-16-UTP (Roche Diagnosics GmbH, Mannheim, Germany) and hybridised (16 hours) a total of 0.85 μg of biotin-labelled cRNA to the BeadChips. The hybridised biotinylated cRNA was detected with streptavidin-Cy3 and quantified using the Illumina's BeadStation 500GX Genetic Analysis Systems scanner. The primary Illumina data were returned from the scanner in the form of an '.idat' file, which contains single intensity data values per gene following the computation of a trimmed mean average for each probe type, represented on the array by a variable number of bead probes. We performed preliminary analyses of the scanned data using Illumina BeadStudio software, which returns a detection call, D, based on a comparison between the intensity of a single probe and the intensities of a large number of negative control beads. Genes with calls consistently below $D = 0.98$ were eliminated from further analysis, leaving data for 11,328 named genes expressed in the temporal lobe for analysis. Normalisation by Z-transformation was performed on each sample/array on a stand-alone basis.[31]

This study was carried out in accordance with the Declaration of Helsinki (2000) of the World Medical Association and was approved by the appropriate institutional review board. Appropriate consent was obtained from human subjects.

Functional relatedness based on gene ontology (GO) and pathway membership data

We used the DAVID Bioinformatics website tools[32] (DAVID Bioinformatics Resources, http://david.

abcc.ncifcrf.gov/home.jsp; National Cancer Institute, Frederick, MD, USA) to generate functional annotation clusters for all human RefSeq genes. We selected two annotation categories for our analysis: GO Biological Processes and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways. For each annotation category, all clusters of size two or larger were considered and similarity scores for each pair of neighbouring genes were calculated. To take account of the fact that some clusters are more broadly defined than others, and thus contain more genes, we increased the score by one divided by the cluster size each time the gene pair was found together in a cluster, thus giving higher scores to pairs in smaller, narrowly defined functional clusters. Binary scores were also used for analysis, only considering whether the pair did or did not co-occur in any cluster.

Single nucleotide polymorphisms (SNPs)

We obtained SNP data from the UCSC genome database, querying for all SNPs available in the NCBI Single Nucleotide Polymorphism database (dbSNP) located within each interval. SNP density was measured in SNPs/kilobase (kb).

Tajima's D

Tajima's D values[33] across the genome have been pre-calculated using data from the Perlegen dataset[34] and integrated into the UCSC browser.[35] The data are only available for the May 2004 version of the human reference sequence, so this version had to be used for the present analysis. Values are calculated in adjacent 10 kb blocks and are likely to be inflated owing to the SNP selection criteria applied by Perlegen; however, they remain useful for comparisons between regions. We assigned values to the intergenic intervals by calculating the average Tajima's D values across each interval. When the 10 kb block covered the entire intergenic interval, the value of D for that interval was used for the region; when it contained more than one 10 kb block, the D values were weighted according to the size included in the interval and averaged.

Results

We exported 18,123 mapped RefSeq genes from the March 2006 release of the UCSC genome database, which defined a total of 17,358 intergenic intervals between non-overlapping genes, excluding intervals containing centromeres or sequence gaps. A total of 568 gene pairs were overlapping without sharing exons, 510 of them transcribed from opposite strands and most often (375/510) overlapping at their 3′ ends (Figure 1). It should be noted that, based on the results of the ENCODE project,[36] there are probably more overlapping pairs than we have identified using the current annotations. For consistency, as the ENCODE project is limited to a small fraction of the genome, we did not incorporate ENCODE data in our analyses. Therefore, our results must be viewed bearing in mind that more intervals should probably be excluded.

Orientation and size distribution

There were 8,877 TH, 4,366 HH and 4,115 TT intergenic intervals. The apparent deficit in TT intervals is because most (375/568) of the excluded overlapping gene pairs were orientated in TT fashion. Taking this into account, the number of neighbouring genes transcribed in the same orientation was no different from the number of those transcribed in opposite orientations (8,935 versus 8,991; chi-squared test, $p = 0.7$), and, overall, those transcribed in opposite orientations were equally divided between those with adjacent start sites and those with adjacent 3′ untranslated regions (4,490 versus 4,501; chi-squared test, $p = 0.9$).

The size distribution for the three types of intervals and the expected distribution if the genes were positioned across the genome at random are shown in Figure 2. This graph is similar to a histogram, with the y-axis showing the percentage of all observations in each size bin. The bins start at the size shown on the x-axis and extend up to 1.78 times that size, a tenfold increment every five bins. Due to the logarithmic scale of the x-axis, the bins are of equal width on the axis but not in base pairs. The sum of fractions of intervals over all bins (the area under the curve)

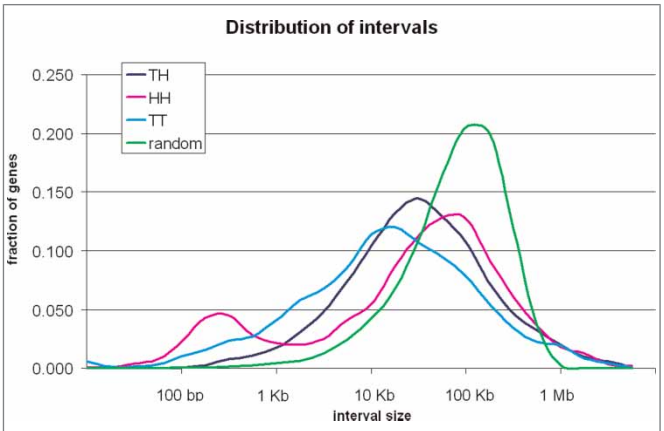

Figure 2. Distribution of intervals. The size distribution of the three types of interval is shown in different colours. The expected distribution for random positioning across the genome is shown in green. The y-axis shows the percentage of observations of each type and the x-axis marks size bins. The size bins start at the size shown on the axis and extend to 1.78 times that size. Due to the logarithmic scale of the x-axis, the bins are of equal width on the axis but not in base pairs.

represents all intervals (100 per cent) of the specific type, allowing for direct comparisons. The size distribution of the intergenic intervals is strikingly different from random in all three categories. There are more intervals smaller than 50 kb than expected, but also more intervals greater than 500 kb (gene deserts), where the three categories are equally represented. In smaller intervals, there are remarkable differences between orientation categories. First, as reported previously,[13,14] HH gene intervals show a bimodal length distribution, with one peak at 20 kb and a second peak at 300 bp, arguing for the significance of this orientation, especially at distances greater than 1,000 bp. The abundance of genes with this arrangement prompted us to use the bioinformatics tools provided by PANTHER (PANTHER classifications of genes and proteins, http://www. pantherdb.org/tools/ ; SRI International, Menlo Park, CA), to test them for enrichment of specific functional annotations. We found a highly significant excess of genes involved in DNA metabolism and repair (Table 1), confirming previous observations[13] and supporting a biological importance of this gene arrangement. A new observation is that HH intervals are over-represented at larger sizes, 50–500 kb (chi-squared test, $p = 5.2\times10^{-8}$ compared with the expected result if the three categories had a single distribution), and under-represented between 2–50 kb; thus, their distribution is not only bimodal, but also biased to the extremes.

Table 1. Biological processes and molecular function enrichment in HH genes closer than 500 base pairs. Biological processes and molecular functions enriched in the genes that are HH orientated and closer than 1 kb from each other. The number of genes observed, with annotation for each process, out of the total of 942 genes in the group is shown, as well as the expected number based on the frequencies of annotations for all RefSeq genes. All values are calculated using the PANTHER bioinformatics engine and *p* values are Bonferroni corrected.

Biological process	**Observed**	**Expected**	***p* Value**
DNA metabolism	39	13.14	6.0×10^{-7}
Nucleoside, nucleotide and nucleic acid metabolism	181	121.99	1.2×10^{-6}
DNA repair	24	6.17	6.9×10^{-6}
rRNA metabolism	12	2.41	1.2×10^{-3}
Other intracellular protein traffic	11	2.26	3.6×10^{-3}
Protein biosynthesis	40	21.57	2.9×10^{-2}
Chromatin packaging and remodelling	21	8.65	3.5×10^{-2}
Protein complex assembly	10	2.48	3.7×10^{-2}
Molecular function	**Observed**	**Expected**	***p* Value**
Nucleic acid binding	170	104.0	3.51×10^{-9}
Histone	12	3.14	1.67×10^{-2}
Dehydrogenase	21	8.21	1.97×10^{-2}
Oxidoreductase	38	22	3.06×10^{-2}

Our result suggests that the HH arrangement might be avoided or reserved for special gene pairs at distances of 2–100 kb. Interestingly, pairs at 2–100 kb were significantly enriched for cell adhesion ($p = 0.002$) and developmental process ($p = 0.018$) genes, suggesting the latter scenario. HT and TT intervals also showed very significant differences from a random size distribution, the most striking being a large excess of small sizes for TT intervals, compared not only with a random distribution, but also with the other two interval types (45 per cent of TT intervals <10 kb compared with 30 per cent of other types of intervals; $p < 10^{-68}$). The average size of TT intervals (101 kb) was significantly smaller than TH (124 kb) or HH (128 kb) intervals (all $p < 10^{-14}$; log transformed sizes were compared to achieve normality), which has not been reported previously for humans and is reminiscent of the results of Williams *et al.* in *A. thaliana*.[18] Genes around TT intervals <10 kb showed a highly significant overrepresentation of genes coding for protein-modifying enzymes, hydrolases, kinases and transferases (Table 2), which further argues for the importance of this arrangement.

Table 2. Biological processes and molecular functions enriched in the 3,734 genes forming TT pairs closer than 10 kb, a size range significantly enriched in TT gene pairs. Methods and column labels are as in Table 1.

Biological process	Observed	Expected	*p* Value
Protein modification	236	167.74	2.87×10^{-5}
Protein metabolism and modification	536	440.74	4.44×10^{-5}
Chromatin packaging and remodelling	64	34.36	5.13×10^{-4}
Cell structure and motility	221	166.44	6.28×10^{-4}
DNA repair	48	24.5	3.19×10^{-3}
Nucleoside, nucleotide and nucleic acid metabolism	561	484.67	4.48×10^{-3}
Intracellular protein traffic	189	146.14	9.02×10^{-3}
Protein phosphorylation	134	95.69	2.01×10^{-2}
Molecular function	**Observed**	**Expected**	***p* Value**
Select regulatory molecule	248	172.53	5.07×10^{-7}
Hydrolase	166	106.71	1.16×10^{-6}
Kinase	155	99.17	2.46×10^{-6}
Transferase	180	128.16	1.78×10^{-4}
Nucleic acid binding	491	413.19	1.18×10^{-3}
Other enzyme regulator	27	10.44	2.10×10^{-3}
Transporter	129	93.95	8.47×10^{-3}
Protein kinase	112	76.69	1.25×10^{-2}
Histone	28	12.47	1.63×10^{-2}
Non-receptor serine/threonine protein kinase	71	43.93	1.91×10^{-2}

Phylogenetic conservation

Conservation over evolutionary time is considered to be an indication of functional significance. Conservation by interval type and size is shown in Figure 3 (PhastCons conserved bases per 1,000). Because of the large variability in conservation among intervals, the y-axis shows a sliding average of 100 intervals consecutive in size. For example, the conservation at size 1 kb in Figure 3 corresponds to the average conservation of 100 intervals, of which the median interval has a size of 1 kb. This was done in this and other figures to reduce the variance between individual points and make the trends visible. Although useful and necessary, it should be noted that this illustration can generate some artefacts at the smallest intervals, where the inclusion of a single small interval with, for example, high conservation leads to a sudden increase in the average, which then slowly declines as more intervals without conservation are added (Figure 3). The reported statistical analyses always use individual, not average, values.

We observed an increase in conservation for all types of intervals larger than 50 kb as they increased in size, peaking at around 500 kb. Higher conservation in gene deserts has previously been reported.[37,38]. Here, we showed that the increase is gradual, starting well below the conventional gene desert threshold

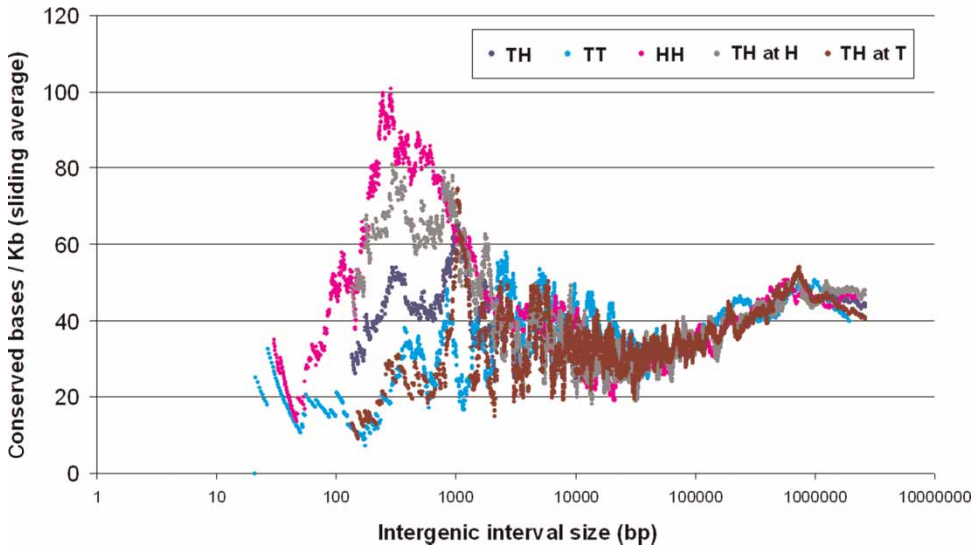

Figure 3. Phylogenetic conservation. Conservation in the three types of intervals arranged by size, measured as the number of bases within conserved elements identified by the PhastCons algorithm per 1,000 base pairs. The y-axis shows the sliding average of the conservation of 100 intervals consecutive in size, while the x-axis shows the average size of the same 100 intervals. This artificially reduces the variance between individual points and is done for illustration only. All reported statistics use the individual values, not these sliding averages.

size of 500 kb. The correlation between distance and conservation was strong for intervals greater than 50 kb ($r = 0.21$; $p = 1 \times 10^{-55}$; distance in logarithmic scale), even when we excluded gene deserts (size 50–500 kb, $r = 0.17$; $p = 6 \times 10^{-32}$). At around 50 kb, conservation was at a minimum and increased again for all types of interval with decreasing size. Below 1 kb, we observed differences between the three interval types, with TT intervals showing the least conservation (for intervals <1 kb, conservation in HH > TH, $p = 0.003$; TH>TT, $p = 7.4 \times 10^{-5}$; HH > TT, $p = 6.4 \times 10^{-13}$). When we examined the conservation on the two halves of TH intervals separately, we found the side adjacent to the 5′ end of the next gene (which we term 'TH at H') to have the same degree of conservation as HH intervals, while the side adjacent to the 3′ end of the previous gene ('TH at T') was about as conserved as TT intervals, which suggests that most of the conservation is likely to be due to elements regulating the downstream gene in this type of arrangement. Finally, we observed an unexpected reduction in conservation in intervals smaller than 250 bp, which was obvious in all three types, providing the most statistical evidence in HH intervals (HH 0–250 bp compared with HH 250–1,000 bp intervals, $p = 3.1 \times 10^{-5}$; for TT intervals, $p = 0.003$; and for HH intervals, $p = 0.07$). The possible implications of this novel observation will be addressed in our discussion.

Density of predicted TFBSs

The density of predicted TFBSs for the three types of intervals relative to their size is shown in Figure 4. It is interesting to view this in comparison with the phylogenetic conservation. First, it is obvious and probably expected that the increased conservation in gene deserts is not related to TFBS density. It is also expected that TFBS density increases in shorter TH and HH intervals, as these are in proximity to the 5′ ends of genes. The apparent increase in TT intervals below 100 bp is the result of two outliers and is not significant. The proximity of two 5′ ends of genes seems to account for the large increase in TFBSs in HH intervals. The reduction in phylogenetic conservation for intervals <250 bp is not reflected in TFBS density.

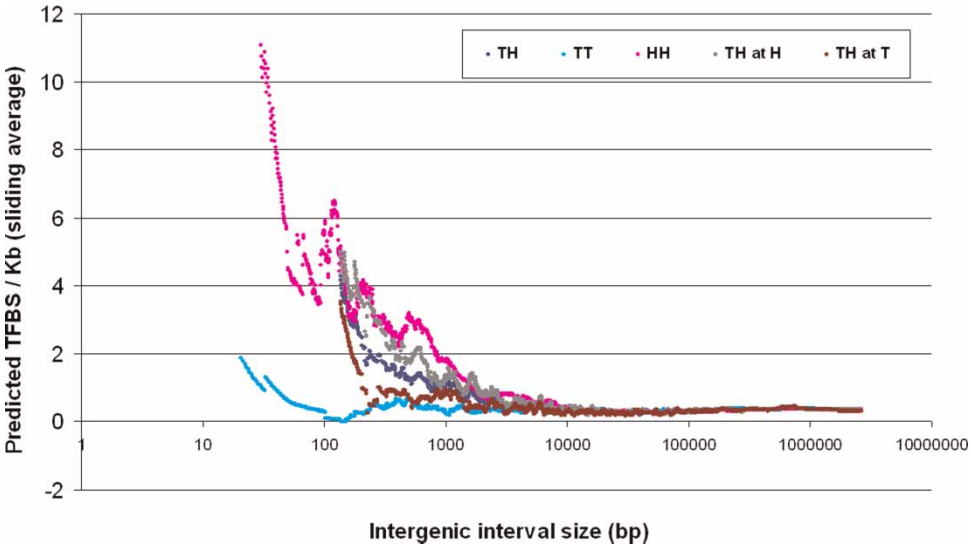

Figure 4. Predicted transcription factor binding site (TFBS) density for the three types of interval according to size. The sliding average technique used in Figure 3 is also used here. TH at H notes the density in the half of the TH interval that is closer to the gene beginning adjacent to the interval, while TH at T notes the density toward the gene that ends there.

Gene ontology

Joint membership of neighbouring genes in functional clusters based on biological processes in the GO database was observed more frequently than expected by chance for all three categories of neighbours and for all distances between adjacent genes (Figure 5). TT and HH genes showed the same increased functional relatedness with each other across all interval sizes. TH genes showed the same increase at distances >100 kb but significantly greater relatedness at smaller distances. Very similar results were obtained using clusters based on the KEGG pathways database. This observation, which has not been reported previously, could reflect the importance of the physical proximity of related genes, which facilitates their co-regulation, the generation of functionally related genes through tandem duplications or a combination of both phenomena.

Gene expression

Correlation of expression across tissues suggests shared tissue-specific regulatory elements. Based on GNF2 expression data from multiple tissues, the expression of neighbouring genes was more correlated than for randomly selected genes across all types of intervals (average Pearson's $r = 0.1$ for all pairs versus 0.03 for 10,000 randomly selected pairs of genes) (Figure 6). This correlation was stronger with decreasing distance, a phenomenon most pronounced for HH intervals below 1 kb. This is in agreement with the results of Trinklein *et al.*,[13] who only studied short HH intervals, but expands their observations, showing that the correlations are present across distances and orientations. Figure 6 includes both positive and negative correlations (anti-correlations; Pearson's $r < 0$). When anti-correlations are examined separately, there is no change with distance, and they are no stronger than randomly observed anti-correlations.

Correlation of expression across individuals within the same tissue type (Figure 7) examines a different aspect of co-regulation, suggesting polymorphic sequence variants acting on both genes and/or common regulation in response to exposures or sample properties (age, sex, cause of death etc). We examined 13 males and ten females with an average age of 76 years (range 35–95). Both genes were expressed in the brain and had available data for

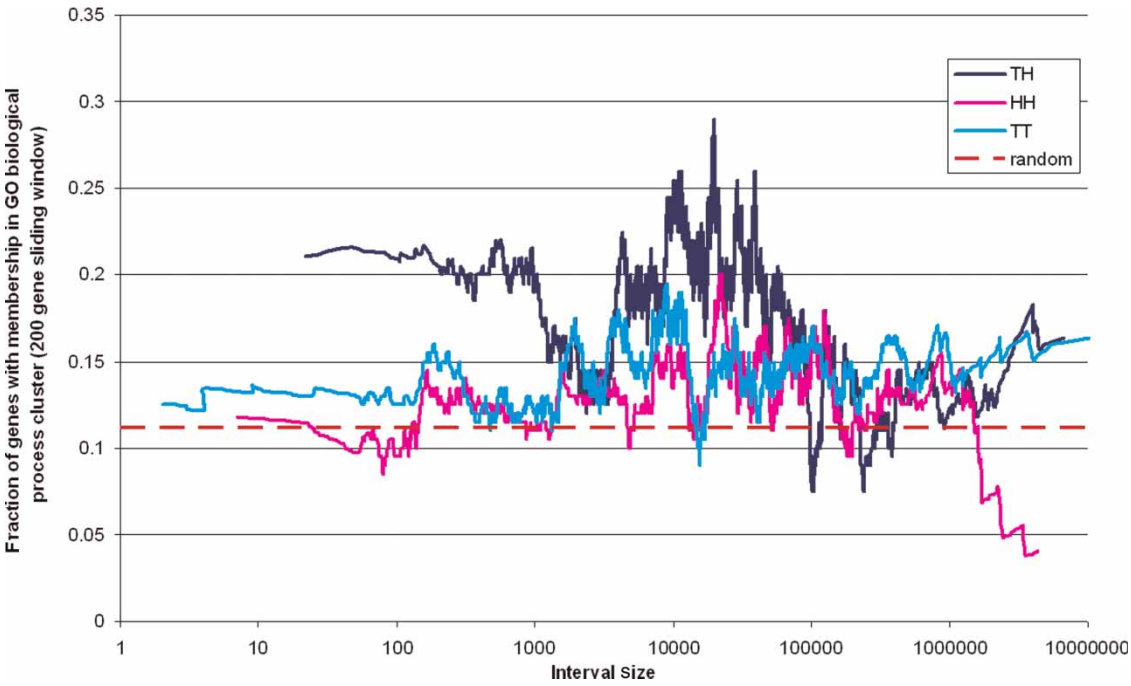

Figure 5. Gene Ontology clusters. The fraction of gene pairs with membership in common in biological proces clusters for the three types of gene pairs and according to distance. The sliding average approach used in Figure 3 is also applied here. The random expectation is shown with a dashed red line. Abbreviation: GO, gene ontology.

4,837 pairs (2,372 TH, 1,267 HH and 1,198 TT). For all pairs, we observed higher interindividual expression correlation (average $r = 0.086$) than for 20,000 random gene pairs ($r = 0.004$; $p < 10^{-22}$). This remained constant up to interval sizes of about 100 kb, at which size it was no longer increased for TT or TH intervals. The examination of interindividual correlation of expression for neighbouring genes is novel and it is interesting that the results are different from analyses across tissues, with increased correlation but a lack of change with distance. This suggests that gene distance between co-regulated genes is only important for factors determining tissue specificity, and not so important for variation of expression within a tissue.

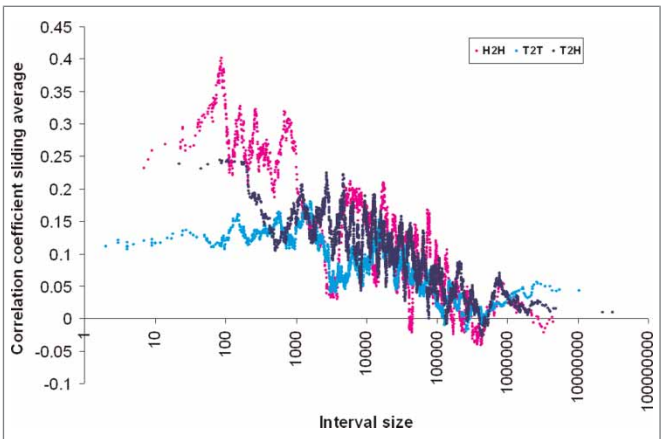

Figure 6. Spatial correlation of gene expression. Correlation of neighbouring gene expression across multiple tissue types and according to gene pair type and distance. The sliding average approach used in Figure 3 is also applied here.

Gene ontology and gene expression

Using the GNF2 data, we found that neighbouring genes with joint membership in any GO cluster (referred to as GO(+) genes) had stronger expression correlation than GO(−) genes (mean $r = 0.19$ versus 0.084; $p < 10^{-39}$). This was most pronounced in TH-orientated genes (mean $r = 0.2$ versus 0.076; $p < 10^{-34}$) but was significant in all three categories. It was also more pronounced in short intervals (<50 kb) but remained significant in

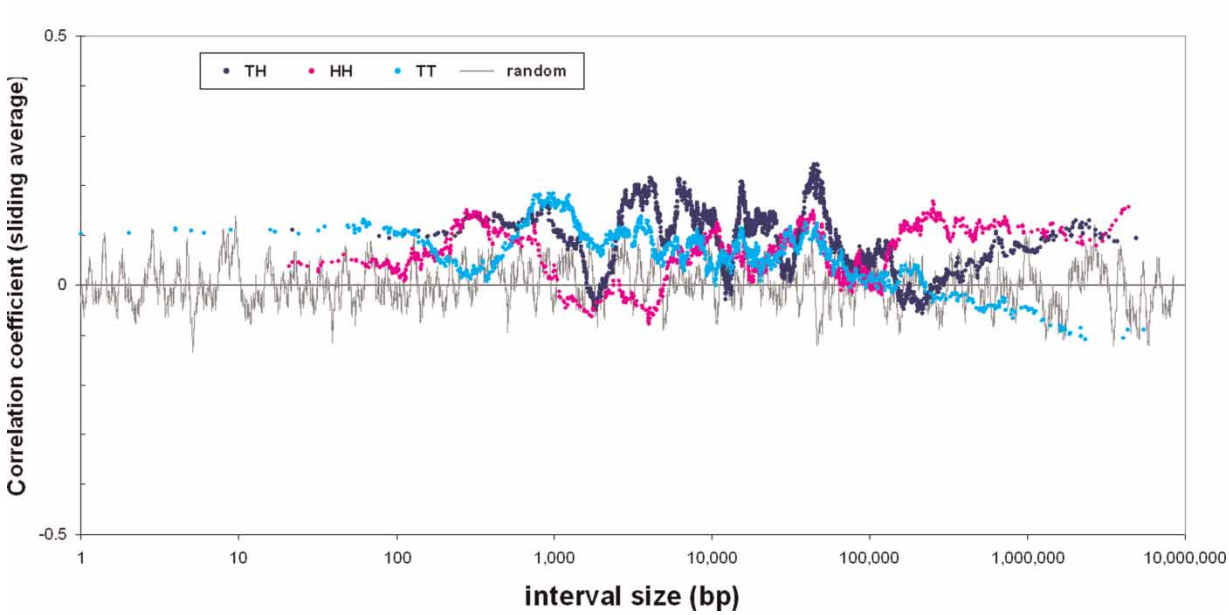

Figure 7. Brain expression across individuals. Correlation of neighbouring gene expression in the temporal lobe of the brain across multiple individuals and according to gene pair type and distance. The sliding average approach used in Figure 3 is also applied here.

long intervals. The level of functional relatedness (see Methods section for scoring algorithm) and the correlation of expression were correlated with each other. This was weak in short intervals (<50 kb) ($r = 0.05$; $p = 0.056$), but stronger in long intervals ($r = 0.18$; $p = 5 \times 10^{-5}$), especially for HH and TH pairs. Finally, negative expression correlations were not stronger or weaker for GO(+) pairs, suggesting no co-regulations in opposite directions. The microarray data of expression across individuals gave similar results. Neighbouring GO(+) genes had stronger expression correlation than GO(−) genes (mean $r = 0.13$ versus 0.07; $p < 0.02$), which was entirely due to TH-orientated genes (mean $r = 0.17$ versus 0.08; $p < 0.004$). This correlation of expression was, again, more pronounced in short intervals while in long intervals (50 kb) its strength correlated (but not significantly) with the level of functional relatedness. Negative expression correlations were not stronger or weaker for GO(+) pairs. The increased expression correlations in gene pairs with known functional relationships further supports that their coregulation is likely to reflect their functional relatedness.

SNP density

The density of dbSNP SNPs at large interval sizes was very similar in the three intergenic interval categories (Figure 8). For intervals between 5 kb and 100 kb it increased steadily (regression $r = 0.06$; $p = 2 \times 10^{-8}$) from 4.0 to 4.5 SNPs/kb for all types of intervals, and then remained constant ($r = 0.025$; $p = 0.14$) for larger intervals up to the size of gene deserts. Below 5 kb, all intervals increased in SNP density with decreasing size ($r = 0.08$; $p = 4 \times 10^{-7}$), most pronounced below 500 bp. This observation, which has not been previously reported, suggests the interesting possibility of relaxed conservation in human lineage. The sharp reduction in SNP density observed in the smallest TT intervals is due to 44 intervals, all smaller than 50 bp, free of SNPs.

Tajima's D

This statistic, which measures nucleotide diversity and compares the heterozygosity distribution of randomly selected SNPs against the expected distribution under selective neutrality and constant population size, was examined in order better to

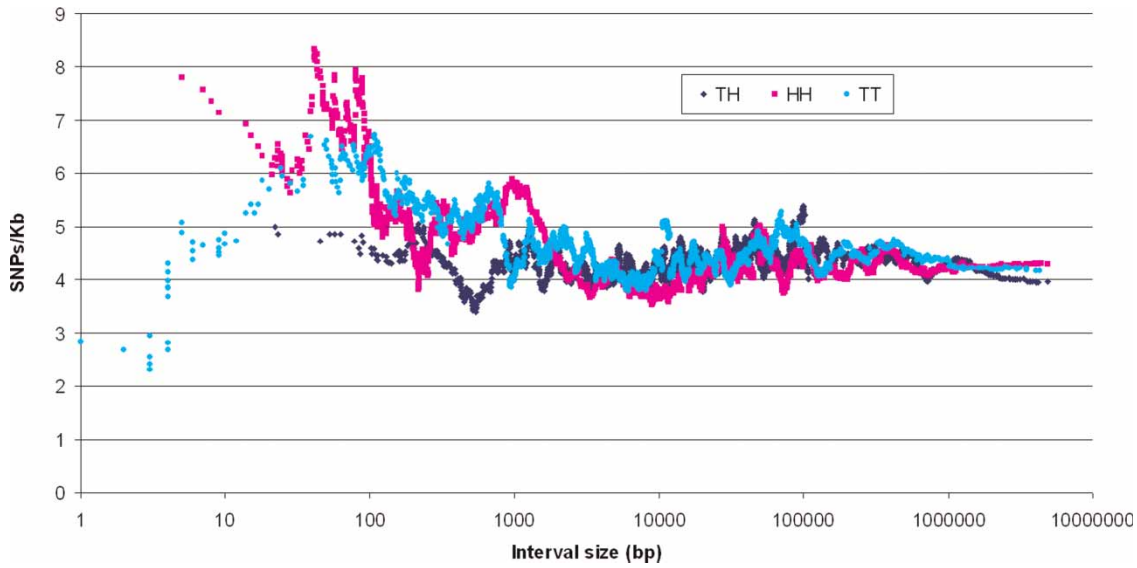

Figure 8. Single nucleotide polymorphism (SNP) density. SNP density from the Single Nucleotide Polymorphism database (dbSNP) in SNPs/kb for the three types of interval according to their size. The same sliding average technique as in Figure 3 is used here.

understand the significance of the observed differences in SNP density and conservation. High values of D are considered as evidence of balancing selection due to an advantage of heterozygosity in the region or the result of a reduction in population size. Low values indicate an excess of rare variation and are consistent with purifying selection, positive selection or population growth.[35] We found a significant trend for an increase in D with increasing intergenic interval size above 500 bp and up to 50 kb, and no change thereafter for all intervals (Figure 9). This was reversed below 500 bp, following the SNP density results discussed above. This is consistent with the relaxed conservation in the human lineage for these intervals, as suggested by the SNP data.

Discussion

We have described the properties of the three types of intergenic intervals with regard to their size distribution, phylogenetic conservation, predicted TFBSs and content of polymorphisms across sizes, as well as the expression and function of the flanking genes. Our data strongly reject the hypothesis of random gene arrangement in the genome.

We found that the size distribution of intergenic intervals is markedly different from the random expectation for all three types of intervals. In agreement with previous reports,[13,14] we saw a bimodal distribution that clearly argues for the functional significance of the HH gene orientation, especially at short distances. This was further supported by the strong correlation of the expression of genes thus arranged, as shown in Figure 6. Among these genes, we replicated the finding of a very significant enrichment for products involved in DNA metabolism and repair.[13,14] A novel finding was that TT intervals also have a markedly different size distribution than expected (Figure 2), with many more intervals shorter than 10 kb and a smaller average size than the other types. The strong over-representation of protein-modifying genes coding for hydrolases, kinases and transferases (Table 2) further argues for a biological significance that needs further investigation.

Our analysis confirmed that there are far more large intergenic intervals (gene deserts) than expected under a random gene distribution, and added to previous work by showing that the

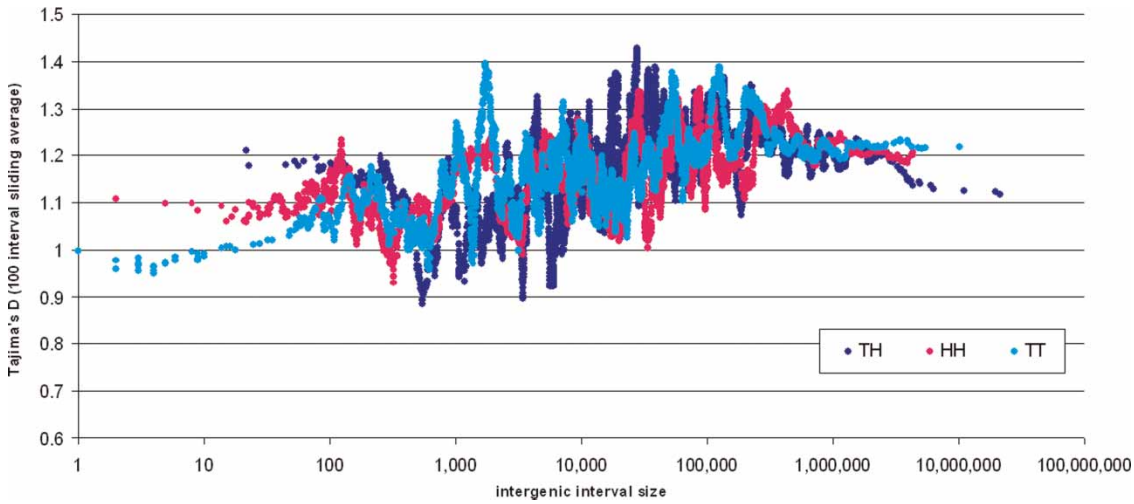

Figure 9. Tajima's D. Tajima's D for the three types of interval according to their size. The same sliding average technique as in Figure 3 is used here. Tajima's D is calculated from the available 10 kb bin data by weighing and averaging, as described in the text.

transition from small to large intergenic sizes is gradual, with no obvious threshold for defining a gene desert. The increase in phylogenetic conservation as interval sizes approach the size of gene deserts was also gradual. The increasing values of Tajima's D suggest that the positive conservation in large intervals and gene deserts is more likely due to balancing rather than purifying selection, a suggestion that is further supported by the lack of change in SNP density in contrast to increased conservation. It would be interesting to examine conservation within primates when enough genome sequences become available to investigate further the natural history of conservation in gene deserts. We found that gene deserts are equally represented in the three types of intervals, and there is no significant correlation of expression of the flanking genes across tissues or among individuals. This suggests that if control is exerted by deserts on the expression of immediate neighbours, it is probably not bidirectional. It is possible that the functional elements located in deserts are there because they need to be far from genes. To explore this possibility, we tested whether conservation near the ends of gene deserts (50 and 100 kb from next gene) is different from conservation in the middle of the desert. We did not observe significant differences, suggesting no strong bias against the presence of putative functional elements close to genes. The answer to the function (or not) of gene deserts remains unclear and further advances in this field of research are necessary.

We found significant correlation of expression of neighbouring genes closer than 100 kb, increasing with decreasing distance for all types of intervals and becoming strongest for HH below 1,000 bp (Figure 6). The similarly increased phylogenetic conservation suggests that part of this co-regulation might be due to DNA sequences. Conservation and co-regulation, however, do not reflect predicted TFBS density, which does not increase except for intervals below 5 Kb, not including TT intervals. Other types of shared regulatory elements and/or the propagation of chromatin states are possible explanations. It is tempting to speculate that, whatever the mechanism, the size distribution favouring smaller intervals and gene deserts might be related to the co-regulation of genes.

We found a stronger correlation of expression when gene pairs were functionally related, suggesting that their positioning and functional relatedness are relevant to their co-expression. The correlation of expression is strongest for intervals shorter than

50 Kb and most pronounced in TH gene pairs, which raises the possibility that it might, in part, be the result of tandem duplications generating paralogs. Such paralogs would be expected to have similar functions and to carry similar regulatory elements in their vicinity, providing an explanation for our observations. The well-discussed HH gene arrangement and co-regulation in short distances appears to be a special case that remains of interest. Our results on TFBS density and conservation show that there are either no shared TBFSs and other regulatory elements in HH pairs or that, if there are such elements, more of them are necessary in these types of pairs.

We observed a reduction in phylogenetic conservation in intervals shorter than ~250 bp in all three types. The opposite pattern was observed in SNP density data, which is consistent with reduced selective pressure at this distance. Tajima's D values[33] further supported this, as they increased in intervals smaller than 500 bp. The congruence of the three types of data is intriguing, yet it is important to consider some important limitations of the SNP and D data. The increase in SNP density in smaller intervals might reflect increased sequencing efforts in areas close to genes by independent researchers. It is hard to tease apart the number of SNPs detected by large-scale sequencing projects from those derived from smaller-scale gene-directed sequencing and reported in the databases. In our study, SNPs in areas flanking transcripts (often not included in gene sequencing projects) could bias toward higher SNP density, a possibility we cannot exclude. An increased SNP density might also reflect higher mutation rates in CG-rich areas close to gene promoters. We examined the fraction of SNPs forming a CpG with either allelic nucleotide in intervals smaller than 1 kb. We found that 32 per cent were in CpG dinucleotides, with 60 per cent of these (19.2 per cent of total) representing transitions from CG to TG. Among 9,000 random SNPs, 27 per cent were in CpG dinucleotides, with 80 per cent of these (21.8 per cent of total) representing transitions from CG to TG. Therefore, although there were more SNPs involving CpG in the short intervals, as expected by the higher CG content, there were fewer transitions from CG to TG, possibly reflecting the lack of CpG methylation in promoter sequences. Another limitation pertains to the available Tajima's D data that have been generated using the Perlegen dataset,[34] which is biased toward high allele frequency SNPs, so higher values are expected. Although the significance of the values *per se* cannot be evaluated, the bias is the same across the genome, allowing for comparisons across regions, which is the strategy we used. Perlegen's ascertainment scheme[34] might introduce an additional bias for Tajima's D close to genes because a fraction of SNPs were selected to be in transcribed sequences, regardless of allele frequency. These SNPs would influence the entire 10 kb block, and when that extended outside a gene, it would cause a reduction in D, especially in smaller intergenic intervals. This bias, however, is opposite to our observation of increased D in small intervals. We cannot fully address the limitations regarding SNP density; we feel, however, that the congruence of conservation, SNP density and Tajima's D data argues for the existence of a very interesting phenomenon of selective relaxation in very small intergenic intervals.

In conclusion, our systematic survey of the distribution and orientation of genes and the properties of intergenic intervals has revealed previously unknown properties of the genome, which might reflect unexplored regulatory mechanisms and evolutionary forces. As expected for this type of research, our observations have generated more questions than were answered. We hope that, as better annotation of the genome becomes available, our results will provide insight for the exploration and interpretation of the new information, and a stimulus for specific questions to be addressed experimentally.

Acknowledgments

This work was supported, in part, from an NIA award (R01-AG022099) to D.A., a NARSAD young investigator award to D.A. and an award from the Neurosciences Education and Research Foundation to D.A. We thank Drs

David Valle and Andrew McCallion for critical suggestions on the manuscript.

References

1. Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001), 'The sequence of the human genome', *Science* Vol. 291, pp. 1304–1351.
2. Lander, E.S., Linton, L.M., Birren, B. *et al.* (2001), 'Initial sequencing and analysis of the human genome', *Nature* Vol. 409, pp. 860–921.
3. Ben-Shahar, Y., Nannapaneni, K., Casavant, T.L., Scheetz, T.E. and Welsh, M.J. (2007), 'Eukaryotic operon-like transcription of functionally related genes in Drosophila', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 222–227.
4. Blumenthal, T. (1998), 'Gene clusters and polycistronic transcription in eukaryotes', *Bioessays* Vol. 20, pp. 480–487.
5. Blumenthal, T. (2004), 'Operons in eukaryotes', *Brief. Funct. Genomic Proteomic* Vol. 3, pp. 199–211.
6. Braastad, C.D., Leguia, M. and Hendrickson, E.A. (2002), 'Ku86 autoantigen related protein-1 transcription initiates from a CpG island and is induced by p53 through a nearby p53 response element', *Nucleic Acids Res.* Vol. 30, pp. 1713–1724.
7. Connelly, M.A., Zhang, H., Kieleczawa, J. and Anderson, C.W. (1998), 'The promoters for human DNA-PKcs (PRKDC) and MCM4: Divergently transcribed genes located at chromosome 8 band q11', *Genomics* Vol. 47, pp. 71–83.
8. Galgoczy, P., Rosenthal, A. and Platzer, M. (2001), 'Human-mouse comparative sequence analysis of the NEMO gene reveals an alternative promoter within the neighboring G6PD gene', *Gene* Vol. 271, pp. 93–98.
9. Platzer, M., Rotman, G., Bauer, D. *et al.* (1997), 'Ataxia-telangiectasia locus: Sequence analysis of 184 kb of human genomic DNA containing the entire ATM gene', *Genome Res.* Vol. 7, pp. 592–605.
10. Shimada, T., Fujii, H. and Lin, H. (1989), 'A 165-base pair sequence between the dihydrofolate reductase gene and the divergently transcribed upstream gene is sufficient for bidirectional transcriptional activity', *J. Biol. Chem.* Vol. 264, pp. 20171–20174.
11. Xu, C.F., Chambers, J.A. and Solomon, E. (1997), 'Complex regulation of the BRCA1 gene', *J. Biol. Chem.* Vol. 272, pp. 20994–20997.
12. Adachi, N. and Lieber, M.R. (2002), 'Bidirectional gene organization: A common architectural feature of the human genome', *Cell* Vol. 109, pp. 807–809.
13. Trinklein, N.D., Aldred, S.F., Hartman, S.J. *et al.* (2004), 'An abundance of bidirectional promoters in the human genome', *Genome Res.* Vol. 14, pp. 62–66.
14. Koyanagi, K.O., Hagiwara, M., Itoh, T., Gojobori, T. and Imanishi, T. (2005), 'Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system', *Gene* Vol. 353, pp. 169–176.
15. Li, Y.Y., Yu, H., Guo, Z.M. *et al.* (2006), 'Systematic analysis of head-to-head gene organization: Evolutionary conservation and potential biological relevance', *PLoS Comput. Biol.* Vol. 2, p. e74.
16. Cho, R.J., Campbell, M.J., Winzeler, E.A. *et al.* (1998), 'A genome-wide transcriptional analysis of the mitotic cell cycle', *Mol. Cell* Vol. 2, pp. 65–73.
17. Kruglyak, S. and Tang, H. (2000), 'Regulation of adjacent yeast genes', *Trends Genet.* Vol. 16, pp. 109–111.
18. Williams, E.J. and Bowles, D.J. (2004), 'Coexpression of neighboring genes in the genome of Arabidopsis thaliana', *Genome Res.* Vol. 14, pp. 1060–1067.
19. Roy, P.J., Stuart, J.M., Lund, J. and Kim, S.K. (2002), 'Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans', *Nature* Vol. 418, pp. 975–979.
20. Fukuoka, Y., Inaoka, H. and Kohane, I.S. (2004), 'Inter-species differences of co-expression of neighboring genes in eukaryotic genomes', *BMC Genomics* Vol. 5, p. 4.
21. Robb, G.B., Carson, A.R., Tai, S.C. *et al.* (2004), 'Post-transcriptional regulation of endothelial nitric-oxide synthase by an overlapping antisense mRNA transcript', *J. Biol. Chem.* Vol. 279, pp. 37982–37996.
22. Katayama, S., Tomaru, Y., Kasukawa, T. *et al.* (2005), 'Antisense transcription in the mammalian transcriptome', *Science* Vol. 309, pp. 1564–1566.
23. Yelin, R., Dahary, D., Sorek, R. *et al.* (2003), 'Widespread occurrence of antisense transcription in the human genome', *Nat. Biotechnol.* Vol. 21, pp. 379–386.
24. Dahary, D., Elroy-Stein, O. and Sorek, R. (2005), 'Naturally occurring antisense: Transcriptional leakage or real overlap?', *Genome Res.* Vol. 15, pp. 364–368.
25. Kent, W.J. (2002), 'BLAT —the BLAST-like alignment tool', *Genome Res.* Vol. 12, pp. 656–664.
26. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007), 'NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res.* Vol. 35, pp. D61–D65.
27. Siepel, A., Bejerano, G., Pedersen, J.S. *et al.* (2005), 'Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes', *Genome Res.* Vol. 15, pp. 1034–1050.
28. Matys, V., Fricke, E., Geffers, R. *et al.* (2003), 'TRANSFAC: Transcriptional regulation, from patterns to profiles', *Nucleic Acids Res.* Vol. 31, pp. 374–378.
29. Su, A.I., Cooke, M.P., Ching, K.A. *et al.* (2002), 'Large-scale analysis of the human and mouse transcriptomes', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 4465–4470.
30. Fisher, R.A. (1915), 'Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population', *Biometrika* Vol. 10, pp. 507–521.
31. Cheadle, C., Vawter, M.P., Freed, W.J. and Becker, K.G. (2003), 'Analysis of microarray data using Z score transformation', *J. Mol. Diagn.* Vol. 5, pp. 73–81.
32. Dennis, G., Jr., Sherman, B.T., Hosack, D.A. *et al.* (2003), 'DAVID: Database for Annotation, Visualization, and Integrated Discovery', *Genome Biol.* Vol. 4, p. P3.
33. Tajima, F. (1989), 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics* Vol. 123, pp. 585–595.
34. Hinds, D.A., Stuve, L.L., Nilsen, G.B. *et al.* (2005), 'Whole-genome patterns of common DNA variation in three human populations', *Science* Vol. 307, pp. 1072–1079.
35. Carlson, C.S., Thomas, D.J., Eberle, M.A. *et al.* (2005), 'Genomic regions exhibiting positive selection identified from dense genotype data', *Genome Res.* Vol. 15, pp. 1553–1565.
36. Birney, E., Stamatoyannopoulos, J.A., Dutta, A. *et al.* (2007), 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature* Vol. 447, pp. 799–816.
37. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003), 'Scanning human gene deserts for long-range enhancers', *Science* Vol. 302, p. 413.
38. Ovcharenko, I., Loots, G.G., Nobrega, M.A. *et al.* (2005), 'Evolution and functional classification of vertebrate gene deserts', *Genome Res.* Vol. 15, pp. 137–145.