Interacting Gene Clusters and the Evolution of the Vertebrate Immune System

Takashi Makino and Aoife McLysaght

Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland

Unraveling the "code" of genome structure is an important goal of genomics research. Colocalization of genes in eukaryotic genomes may facilitate preservation of favorable allele combinations between epistasic loci or coregulation of functionally related genes. However, the presence of interacting gene clusters in the human genome has remained unclear. We systematically searched the human genome for evidence of closely linked genes whose protein products interact. We find 83 pairs of interacting genes that are located within 1 Mbp in the human genome or 37 if we exclude hub proteins. This number of interacting gene clusters is significantly more than expected by chance and is not the result of tandem duplications. Furthermore, we find that these clusters are significantly more conserved across vertebrate (but not chordate) genomes than other pairs of genes located within 1 Mbp in the human genome. In many cases, the genes are both present but not clustered in older vertebrate lineages. These results suggest gene cluster creation along the human lineage. These clusters are not enriched for housekeeping genes, but we find a significant contribution from genes involved in "response to stimulus." Many of these genes are involved in the immune response, including, but not limited to, known clusters such as the major histocompatibility complex. That these clusters were formed contemporaneously with the origin of adaptive immunity within the vertebrate lineage suggests that novel evolutionary and regulatory constraints were associated with the operation of the immune system.

Introduction

Genes are held together on chromosomes in an arrangement that has often been likened to "beads on a string," but this simple description belies great organizational complexity. In bacteria, functionally related genes are often clustered on the genome into operons and are transcribed into a single messenger RNA. Similarly, 15% of *Caenorhabditis elegans* genes are also cotranscribed polycistronically in gene clusters similar to bacterial operons (Blumenthal et al. 2002). Operon structures have not been found in other eukaryotes; however, many recent studies have reported nonrandom gene order on eukaryotic genomes (Hurst et al. 2004). It has been reported that neighboring genes are coexpressed in *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *C. elegans* (Cohen et al. 2000; Spellman and Rubin 2002; Lercher et al. 2003). In addition, housekeeping or highly expressed genes have often been found in coexpressed gene clusters in mouse and human (Lercher et al. 2002; Williams and Hurst 2002; Singer et al. 2005).

The transcription of a gene may however affect the transcription of its neighbors even if the coexpression is not intended (Spellman and Rubin 2002; Hurst et al. 2004). Transgenes in plants can assume the expression profile of regions into which they insert in the genome by chromatin-mediated effects (Finnegan et al. 2004). Many of these coincidental transcripts are likely to be suppressed by posttranscriptional regulation, and the correlation between transcription and translation is weak in *S. cerevisiae* (Ghaemmaghami et al. 2003). In other words, there is no guarantee that the transcripts are translated.

Protein–protein interactions (PPIs) are relationships among translated products. Identification of gene clusters in the genome based on PPIs avoids the bias caused by chromatin structure that is inherent to the analysis of coexpressed genes. Therefore, PPIs constitute a suitable biological relationship to study gene clusters in the genome. In *S. cerevisiae*, genes in the same protein complex (Teichmann and Veitia 2004) as well as genes with other interactions (Poyatos and Hurst 2006) tend to be both colocalized and coexpressed in the genome.

Coexpressed yeast gene pairs are well conserved in *Candida albicans* (Huynen et al. 2001; Hurst et al. 2002). In addition, there are fewer breakpoints within coexpressed gene clusters in both human and mouse than expected (Singer et al. 2005). These results indicate that coexpressed gene clusters have been conserved during evolution.

It is more difficult to identify interacting gene clusters in large genomes such as those of vertebrates than in the more compact yeast genomes because of the increased size of the proteome and the relative deficit of experimental data. To date, it has remained unclear whether there are interacting gene clusters in vertebrate genomes (Hurst et al. 2004; Hurst and Lercher 2005; Poyatos and Hurst 2006). Here, we search for evidence of the colocalization of interacting genes within the human genome using PPI data from the Human Protein Reference Database (HPRD). We investigate the evolution of interacting gene clusters by comparative genomics in vertebrates and test the hypothesis that interacting gene clusters in human should be well conserved in other vertebrate genomes.

Materials and Methods

Data

The 22,555 protein-coding genes having known genomic locations in Ensembl release 46 were used in this study (Hubbard et al. 2007). We used PPI data in HPRD release 7 (Peri et al. 2003) excluding self-interactions. From the HPRD data, we used 34,653 interactions involving 9,228 genes whose genomic locations could be traced in Ensembl.

Identification of Tandem Duplicated Genes

An all-against-all BlastP searches were conducted for human protein sequences from Ensembl, using the longest sequence for genes with multiple isoforms. Duplicated

Key words: protein–protein interaction, gene cluster, vertebrate evolution, comparative genomics.

E-mail: aoife.mclysaght@tcd.ie.

Mol. Biol. Evol. 25(9):1855–1862. 2008
doi:10.1093/molbev/msn137
Advance Access publication June 23, 2008

genes were defined as those with $E < 0.2$ (Lercher et al. 2002). Out of the 34,653 interacting pairs, 139 were tandem duplicates within 10 Mbp. We also identified interacting gene pairs sharing protein domains (except for low complexity sequence such as proline-rich region), in any of the alternative isoforms, in InterPro (http://www.ebi.ac.uk/interpro/) or structural classification of proteins (SCOP); (http://scop.mrc-lmb.cam.ac.uk/scop/).

Simulations

We conducted simulations to assess whether the number of interacting gene pairs at a particular genomic distance was larger than expected. We randomly shuffled locations of all genes in the genome 1,000 times and identified interacting gene clusters on each shuffled genome. We counted the number of interacting gene pairs within a range of 1 Mbp each toward 10 Mbp to test the statistical significance of our results.

Conservation of Genes and Gene Clusters

We obtained orthologs of human genes from Ensembl compara release 46 for 13 species (Hubbard et al. 2007): chimpanzee (*Pan troglodytes*), macaque (*Macaca mulatta*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*), cow (*Bos taurus)*, opossum (*Monodelphis domestica*), chicken (*Gallus gallus*), zebrafish (*Danio rerio*), tetraodon (*Tetraodon nigroviridis*), stickleback (*Gasterosteus aculeatus*), medaka (*Oryzias latipes*), and ascidian (*Ciona intestinalis*). Genes were classified into 9 categories: human, hominid, primate, primate/rodent, eutherian, mammal, tetrapod, vertebrate, and chordate according to their most recent common ancestor (fig. 1). We excluded genes that were only annotated in human from the comparative analysis because many such absences are caused by annotation issues.

Similarly, we classified gene clusters into 9 categories based on their conservation during chordate evolution. If both genes in a gene cluster in the human genome have orthologs within 1 Mbp in another vertebrate, we considered the gene cluster as a conserved cluster on the vertebrate genome (fig. 1). To test the statistical significance of gene cluster conservation during chordate evolution, we randomly selected as many gene pairs within 1 Mbp on the human genome as observed interacting gene pairs 1,000 times. In addition, we sampled genes having the same degree of evolutionary conservation as genes in observed gene clusters to normalize the effect of difference in the degree of conservation between randomly sampled and observed interacting genes.

Gene Ontology

The number of each Gene Ontology Annotation (GOA) term assigned into genes in gene clusters was counted (http://www.ebi.ac.uk/GOA/). We calculated the *P* value for each GOA term by comparison of the number of observed GOA term with that of expected GOA term based on a hypergeometric distribution using all genes in the PPI network. The estimated *P* values were adjusted by Bonferroni correction and Benjamini–Hochberg correction.

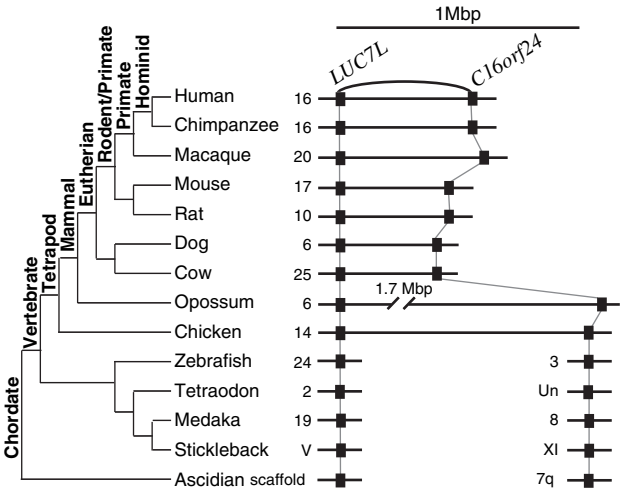


FIG. 1.—Using comparative genomics to assess the degree of conservation of a gene cluster. Rectangles represent the interacting gene pair *LUC7L–C16orf24*, which in human are located <1 Mbp apart. Horizontal lines indicate chromosome segments and are labeled with the chromosome number. Both genes in the interacting gene pair are conserved in all chordate species used in this study, but they are not within 1 Mbp in opossum and chicken and not on the same chromosome in teleost fishes and ascidian.

Results

Exclusion of Interactions between Tandem Duplicate Genes

Tandem gene duplication may create interacting gene clusters in a given region of chromosome because they were derived from the same ancestral gene (Lercher et al. 2002). Hurst et al. (2004) pointed out that many studies of coexpressed gene clusters did not control for this tandem duplication effect. Let us consider the effect of tandem duplication on the origin of gene clusters when PPI data are used. In the case where a gene interacts with another gene on a different chromosome, following tandem gene duplication the duplicated gene pair may share interacting partners (Wagner 2001; Makino et al. 2006), but no new interaction between products of closely linked genes is created (fig. 2*A*). That is, a gene cluster is not created even if tandem duplication has occurred. Therefore, the effect of tandem duplication on the identification of biologically and evolutionarily significant gene clusters is expected to be weaker when using PPI data than when using expression data. However, when a gene having a self-interaction is duplicated, the gene would possibly interact with its duplicated copy (fig. 2*B*; Wagner 2001; Pereira-Leal et al. 2007). So it is necessary to exclude tandemly duplicated gene pairs to identify gene clusters, even when using PPI data.

We found 137 PPIs between duplicated genes located less than 10 Mbp apart in the human genome. Approximately 80% of PPIs between duplicated gene pairs were within 1 Mbp. We excluded the PPIs between duplicated gene pairs in the remainder of our analysis.

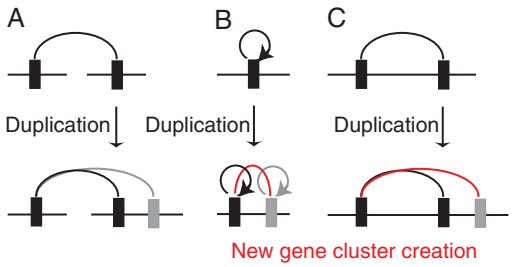


FIG. 2.—Effect of tandem duplication on interacting gene clusters. Rectangles, horizontal lines, and curved lines represent genes, chromosomes, and PPIs, respectively. Red lines indicate PPIs of gene clusters newly created after gene duplication. (*A*) Duplication of a gene interacting with another gene on a different chromosome. (*B*) Duplication of a self-interacting gene. (*C*) Duplication of a gene in an existing gene cluster.

Table 1
Number of Interacting Gene Clusters within 10 Mbp

Interval Size (Mbp)	Observations	Mean	Standard Deviation	Z Score	*P* Value
0–1	100	47.12	7.29	7.25	4.17×10^{-13}
1–2	46	37.54	6.14	1.38	0.17
2–3	39	33.24	5.96	0.97	0.33
3–4	32	30.64	5.84	0.23	0.82
4–5	29	29.14	5.36	−0.03	0.98
5–6	31	28.06	5.25	0.56	0.58
6–7	24	27.18	5.04	−0.63	0.53
7–8	31	25.72	5.26	1.00	0.32
8–9	28	25.36	5.05	0.52	0.60
9–10	27	24.16	4.92	0.58	0.56

NOTE.—The number of observations and the mean of the simulations decrease as the interval increases because the search range must lie within the chromosome.

In another possible scenario, where a gene in an existing gene cluster duplicates tandemly, the gene cluster would be amplified (fig. 2*C*). The amplified gene clusters are likely to have biological meaning because they were derived from existing gene clusters. Therefore, we treated the amplified gene clusters the same as the other interacting gene clusters, although they were not many (5 in our data set).

Identification of Interacting Gene Clusters

We identified interacting pairs and measured the genomic distance between each pair. For distances of 0–10 Mbp, we sorted the gene pairs into groups separated by 0–1, 1–2 2–3 Mbp, etc. (table 1). For each group, we compared the observed number of interacting pairs with the expected numbers, estimated by simulations where gene order was randomized (see Materials and Methods). Only the interval 0–1 Mbp shows a significant excess of interacting pairs; there are 100 such pairs. One mega base pair is a biologically realistic range for finding gene clusters because it has been reported that distal enhancers can affect the transcription of a neighboring gene from a distance of as much as 1 Mbp of genomic sequence (Kleinjan and van Heyningen 2005).

We measured the effect of gene density differences throughout the genome on the identification of gene clusters by repeating this analysis, considering gene pairs separated by a fixed number of intervening genes rather than a Mbp distance (supplementary table S1, Supplementary Material online). We identified 100 interacting gene pairs located within 20 genes of each other, 73 of which were also found within 1 Mbp. This indicates that gene density bias is not significant and that clusters identified using a base-pair distance threshold are robust.

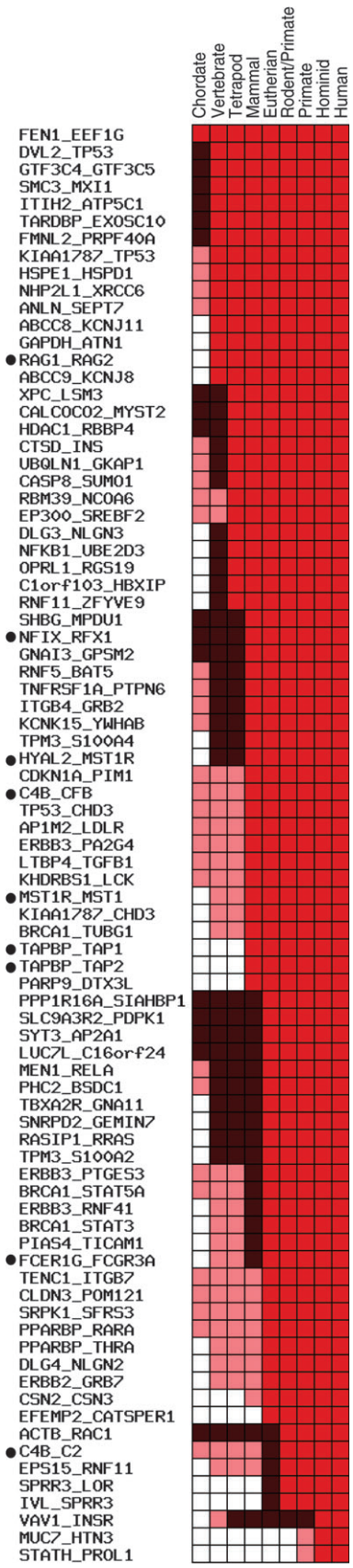
Although we had already excluded interactions between duplicated gene pairs identified by BlastP, we considered the possibility that there might remain some tandemly duplicated genes with divergent sequences that were not detected by this search. The more sensitive program position specific iterated Blast (PSI-Blast) found that 10 of the 100 interacting gene pairs within 1 Mbp have residual sequence similarity (*E* values < 0.2).

In addition, we looked for gene pairs sharing any protein domain as presumable tandemly duplicated gene pairs (see Materials and Methods). Among the 100 gene clusters within 1 Mbp, we identified 17 gene pairs sharing one or more domains in InterPro or SCOP. These 17 gene pairs sharing domains included all 10 duplicated gene pairs identified by PSI-Blast. Even after the removal of these 17 gene pairs, there was still a statistically significant difference in the number of gene clusters within 1 Mbp between observation and simulation ($P = 2.4 \times 10^{-7}$; supplementary fig. S1*A*, Supplementary Material online). The remaining 83 interacting gene pairs identified were used in the following analyses and are named in supplementary table S2 (Supplementary Material online).

Conservation of Interacting Gene Clusters in Vertebrate Genomes

If interacting gene clusters in human have a functional role, we would expect them to be conserved during evolution. We tested this hypothesis by comparing the degree of conservation of the observed clusters to that of other pairs of nearby genes chosen at random. (figs. 1 and 3). The observed gene clusters are significantly more conserved than expected ($P = 1.7 \times 10^{-7}$, 1-tailed Kolmogorov–Smirnov test; fig. 4*A*).

The comparative genomics of each of the gene clusters is summarized in figure 3. Only one cluster is conserved throughout all the species examined (*FEN1–EEF1G*; first row of fig. 3). Other clusters are conserved to varying degrees. In 48 cases, both genes are detected but not clustered in an older vertebrate lineage (indicated by brown squares in fig. 3). Twenty-four of these were present in the genome for a significant time before they were colocalized (*i.e.*, they are unclustered in more than one outgroup lineage). In many of these cases, we are observing the construction of the gene cluster from nonclustered genes within the vertebrate lineage. The fact that the genes are both present but not clustered in multiple outgroup lineages points to the assembly of the cluster on the human lineage. However, the frequency of gene pair "construction" is not more than observed for noninteracting pairs. In the remaining 34 cases, we always observe the cluster given that both orthologs are detected. This may be influenced by a potential failure to identify orthologs in distantly related organisms even when they are present but is also indicative of the biological significance of these clusters.



We speculated on the reason for the high conservation of gene clusters and hypothesized that gene clusters may consist of slowly evolving genes and that this alone might explain the higher conservation of the clusters. We compared the distribution of the evolutionary conservation of genes in clusters (*i.e.,* the phylogenetic range over which they were detected) with expectation and found that genes involved in clusters are slightly more conserved than expected ($P = 0.035$, 1-tailed Kolmogorov–Smirnov test; fig. 4*B*). Therefore, we normalized the degree of evolutionary conservation of randomly sampled gene pairs (see Materials and Methods). We found no difference between the conservation of genes selected in the normalized randomization and the observed conservation of genes indicating that the normalization step was successful. Even after the normalization, we found that the gene clusters were more conserved than randomly sampled pairs of human genes separated by <1 Mbp ($P = 0.024$, 1-tailed Kolmogorov–Smirnov test; fig. 4*C*). This demonstrates that the clusters identified in the human genome are more conserved than expected and that this is not a consequence of the degree of conservation of the constituent genes.

A recent study reported that the physical distance between genes has the greatest contribution to the conservation of gene order in yeast even compared with some known factors such as coexpression (Poyatos and Hurst 2007): that is, the shorter the physical distance, the rarer is gene order rearrangement. Therefore, we considered the possibility that the observed gene clusters were conserved during evolution just due to close physical proximity. This could bias the analysis if the physical distances between the interacting gene pairs within 1 Mbp in the human genome tended to be shorter than those between randomly sampled pairs within 1 Mbp in simulation. To examine this possibility, we compared the distances between interacting gene pairs in the observation with those in the simulation. The mean distances in the observation and simulation were 0.44 and 0.47 Mbp, respectively, which is not a statistically significant difference ($P = 0.34$, Mann–Whitney U test; $P = 0.33$, 1-tailed Kolmogorov–Smirnov test; supplementary fig. S2, Supplementary Material online). Therefore, the high conservation of the interacting gene clusters is not simply a consequence of close physical proximity.

Gene Clusters Are Not Enriched for Housekeeping Genes or for Pairs in the Same Protein Complex

We assessed whether identified colocalized interacting gene pairs in the human genome were in the same protein

←

FIG. 3.—Comparative genomics of 83 identified interacting gene clusters on the human genome. Rows are labeled with the names of the interacting genes as per supplementary table S2 (Supplementary Material online). Coloring indicates the conservation status of an interacting human gene cluster at different phylogenetic depths as indicated by the column labels. Red, conserved pair <1 Mbp apart. Brown, both genes are conserved but are not within 1 Mbp on the same chromosome. Pink, only 1 of the 2 genes was detected. White, neither gene was detected. Analysis is based on the genomes and nodes shown in figure 1, and is based on a most recent common ancestor criterion. Gene clusters marked with a dot are related to the immune response.

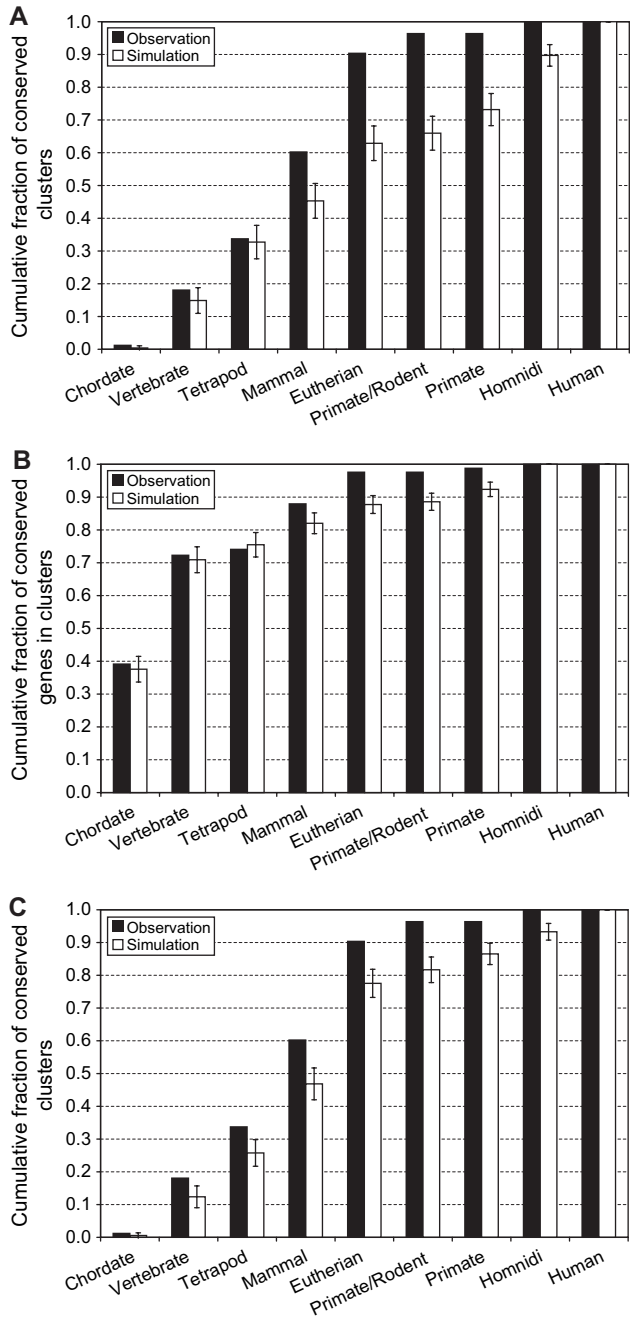


FIG. 4.—Conservation of interacting clusters and their constituent genes. Histograms show the cumulative fraction of conserved clusters or genes at phylogenetic depths indicated by the *x* axis labels (as in fig. 1). Error bars represent standard deviations. (*A*) Conservation of interacting clusters. (*B*) Conservation of genes in interacting clusters. (*C*) Conservation of interacting clusters; the degree of conservation of randomly sampled genes in simulation is normalized (see Materials and Methods).

complex. We found that 793 out of 34,653 interacting gene pairs were in the same complex using protein complex data in HPRD. Only 2 pairs in 83 interacting gene clusters were in the same complex.

Housekeeping and highly expressed genes were found in coexpressed gene clusters on the human genome (Lercher et al. 2002). Whereas housekeeping gene clusters were evolutionarily conserved, highly expressed genes were not (Singer et al. 2005). Therefore, we investigated whether interacting gene clusters identified here were enriched for housekeeping genes. We found 1,505 pairs of housekeeping genes in 34,653 interacting gene pairs, using an accurate set of predicted housekeeping genes (De Ferrari and Aitken 2006). Only 6 pairs of housekeeping genes were found in 83 interacting gene clusters. There was no statistically significant difference in the proportion of housekeeping gene pairs in interacting gene clusters as compared with the entire genome ($P = 0.31$, χ^2 test for contingency table).

Function of Genes in Interacting Gene Clusters

In an attempt to understand the characteristics of genes in interacting clusters, we examined their function using GOA (see Materials and Methods). We did not find any statistically significant difference in the number of GOA terms between observation and expectation. We hypothesized that the power to detect any trend within these clusters may be compromised by the fact that about half of gene clusters within 1 Mbp are possibly artifacts (47 are expected by chance; table 1). Proteins that have many interaction partners in the PPI network are called "hub" proteins. Hubs have more opportunities to colocalize with their PPI partners in the same genome region by chance than others. It has been noted that high-throughput PPI data contained many false positives (von Mering et al. 2002), and hubs are likely to be the central reason. Therefore, interacting gene pairs consisting of at least one hub protein (here defined as a protein with at least 20 PPI partners) were regarded as less reliable pairs, although some of them must be true pairs. Even after the removal of the hubs, there was still a statistically significant difference in the number of gene clusters within 1 Mbp between observation (37 pairs) and simulation ($P = 7.55 \times 10^{-9}$; supplementary fig. S1*B*, Supplementary Material online). We did not exclude all hub proteins from the outset of the project because they must contain some true interactions, and it is instructive to note that even with their inclusion the number of interacting gene clusters in the human genome is highly significant. We examined the function using GOA for these remaining 37 gene pairs and found response to stimulus (GO:0005215) was significantly enriched ($P = 0.031$ after correction for multiple tests, see Materials and Methods). Interestingly, 13 of 21 genes classified into the GOA term response to stimulus were related to immune response (table 2 and fig. 6). Furthermore, many of the immune-related genes had interactions with each other.

Evolutionary History of Interacting Gene Clusters Related to Immune Response

We investigated the conservation of the interacting gene clusters related to immune response in chordate species. All but 2 of the immune-related gene clusters were conserved at least in mammals (fig. 3). The interacting gene cluster *RAG1–RAG2*, which is a well-documented gene

Table 2
Genes Assigned into GO:0005215 Response to Stimulus and Their Interacting Partners

Gene 1		Gene 2	
Name	Function	Name	Function
TAPBP	**Tapasin isoform 3 precursor**	**TAP1**	**Transporter 1, ATP-binding cassette, subfamily**
TAPBP	**Tapasin isoform 3 precursor**	**TAP2**	**Transporter 2, ATP-binding cassette, subfamily**
RFX1	**MHC class II regulatory factor**	*NFIX*	*Nuclear factor I/X (CCAAT-binding transcription)*
FCER1G	**Fc fragment of IgE, high affinity I, receptor**	**FCGR3A**	**Fc fragment of IgG, low affinity IIIa, receptor**
RAG1	**Recombination-activating gene 1**	**RAG2**	**Recombination-activating gene 2**
MST1R	**Macrophage-stimulating 1 receptor**	**MST1**	**Macrophage-stimulating 1**
MST1R	**Macrophage-stimulating 1 receptor**	*HYAL2*	*Hyaluronoglucosaminidase 2*
C4B	**Complement component 4B preproprotein**	**CFB**	**Complement factor B preproprotein**
C4B	**Complement component 4B preproprotein**	**C2**	**Complement component 2 precursor**
SMC3	Structural maintenance of chromosomes 3	*MXI1*	*MAX interactor 1 isoform a*
HSPE1	Heat shock 10 kDa protein 1 (chaperonin 10)	*HSPD1*	*Chaperonin*
CTSD	Cathepsin D preproprotein	*INS*	*Proinsulin precursor*
XPC	Xeroderma pigmentosum, complementation group C	*LSM3*	*Lsm3 protein*
IVL	Involucrin	SPRR3	Esophagin
SPRR3	Esophagin	*LOR*	*Loricrin*
HTN3	Histatin 3	*MUC7*	*Mucin 7, salivary*

NOTE.—Genes in bold are related to immune response. Genes in italics are not classified into response to stimulus, but they interact with genes classified into the GO term. ATP, adenosine triphosphate.

cluster (Oettinger et al. 1990), is conserved throughout vertebrates (fig. 5*A*). The interacting gene pair *RFX1–NFIX* is on the same chromosome among chordate genomes, but the gene cluster within 1 Mbp is conserved only in mammals (fig. 5*B*). This observation suggests that intrachromosomal rearrangements have moved these genes closer together. The pair *FCER1G–FCGR3A* is conserved in mammals, but the intergenic distance is <1 Mbp only in eutherians (except for cow; fig. 5*C*). We found that gene syntenies around both orthologous genes were conserved well among eutherian genomes. Both synteny blocks in the opossum genome were also conserved, but they were not closely linked. The results indicate that genomic rearrangement occurred in a region between the interacting gene pair in either the eutherian or marsupial lineage. If the former is true, then the cluster was created by a genome rearrangement. Unfortunately with present data, it is not possible to distinguish the two scenarios because we cannot identify orthologs of *FCER1G* or *FCGR3A* or their closely linked neighbors in more distantly related species.

Discussion

We have shown that interacting gene clusters are a significant component of the human genome. This is true even after extremely strict removal of tandem duplicates and after eliminating hub proteins from the analysis. There are presumably some false positives present in PPI data, but despite this, gene clusters have proven to be more abundant within 1 Mbp in the human genome by statistical tests. There is no reason to suspect that there is a bias in the database to falsely report interactions between closely linked genes. Furthermore, PPI data are still relatively sparse, and we expect that with more data and with greater accuracy in the data more interacting gene clusters will be identified in the human genome. The biological significance

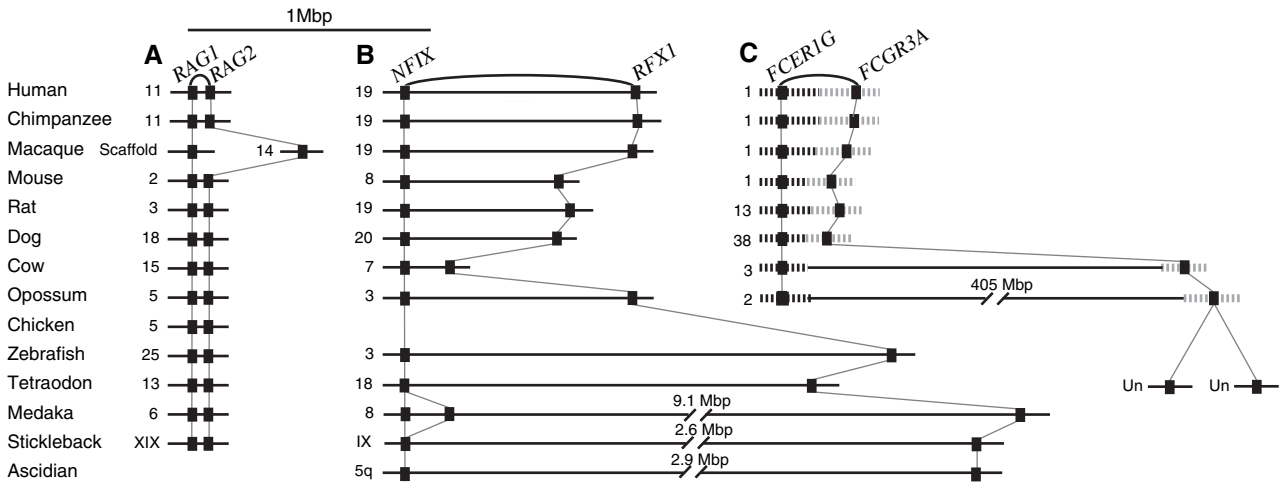


FIG. 5.—Conservation of interacting gene clusters related to the immune response. (*A*) *RAG1–RAG2*. (*B*) *RFX1–NFIX*. (*C*) *FCER1G–FCGR3A*. Dashed black and gray lines represent conserved order of neighboring genes around *FCER1G* and *FCGR3A* among mammalian genomes, respectively.

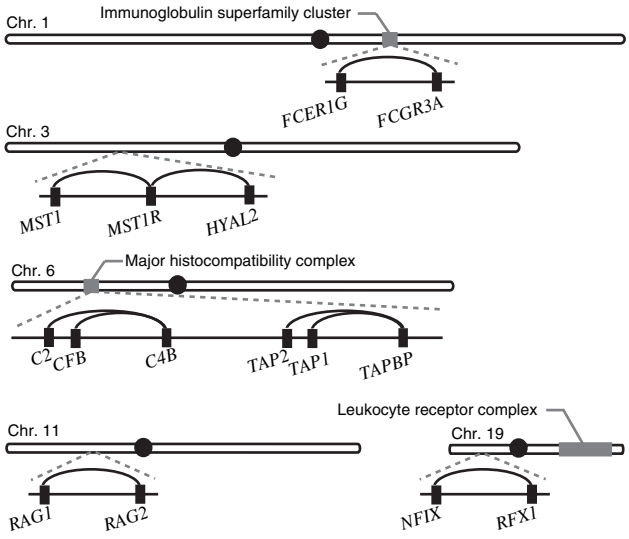


FIG. 6.—Genomic locations of interacting gene clusters related to immune response. Circles represent centromeres. Gray rectangles indicate known immune-related large gene clusters. Note that the gene order is correct but the figure is not to scale, and genes located between the interacting genes are not shown.

of interacting gene clusters might be derived from coregulation by bidirectional promoters (Trinklein et al. 2004), shared enhancers (West et al. 2002), or chromatin-mediated regulation (Robyr et al. 2002; Finnegan et al. 2004). There may also be a selective advantage to retain interacting genes in close linkage because in the case of epistasis, particular, favorable allele combinations may be retained in linkage disequilibrium (Nei 1967). Of the 83 gene pairs, 10 are immediate neighbors and 6 of these have diverging orientations (head-to-head; supplementary table S2, Supplementary Material online), which may allow coregulation by bidirectional promoters.

We found that interacting gene clusters have been conserved during vertebrate evolution. The greatest difference between the observed level of conservation and expectation based on simulations is seen in mammals (fig. 4*A* and *C*). The comparison with primate genomes shows a lesser degree of difference between the observation and the simulations, which probably reflects the short time since they shared a common ancestor. The difference between observation and simulation disappears in more distantly related species. In particular, the interacting gene clusters were rarely conserved in the ascidian (*Ciona*) genome, although approximately 40% of genes in the gene clusters were detected (fig. 4*B*). Under the principle of scientific parsimony, these results suggest that many of the interacting gene clusters have been created in the vertebrate lineage, particularly the mammalian lineage.

We found several important differences between interacting gene clusters in the human genome as compared with what has been reported. Genes in the same protein complex tend to be physically close to each other on the yeast genome (Teichmann and Veitia 2004); however, genes in the interacting gene clusters identified in this study were rarely in the same protein complex. In addition, housekeeping gene clusters were found in the human genome using gene expression data (Lercher et al. 2002), but only very few interacting gene clusters identified here were pairs of housekeeping genes. This indicates that the characteristics of the PPI-based interacting gene clusters identified here are different from those identified in previous studies.

Interestingly, we found that several genes in the interacting gene clusters were related to immune response. The major histocompatibility complex (MHC) region on chromosome 6 is a well-known cluster of genes related to immunity (The MHC sequencing consortium 1999), and it was successfully identified in our search of the human genome (fig. 6). In addition, there are other large clusters such as the leukocyte receptor complex on chromosome 19, the natural killer complex on chromosome 12, and the immunoglobulin superfamily cluster on chromosome 1, although large parts of these gene clusters have been created by tandem gene duplication (Trowsdale 2002; Fukami-Kobayashi et al. 2005; Kelley et al. 2005). In this study, we excluded gene clusters created by tandem gene duplication, but our search did identify immune-related interacting gene clusters existing in large gene clusters such as the MHC and also other local regions (fig. 6). The interacting gene cluster *FCER1G–FCGR3A* is located within the known immunoglobulin superfamily cluster (fig. 6), and *FCGR3A* is a member of the immunoglobulin superfamily. Although *FCER1G* is not a member of this family, the gene was derived from a common ancestor with the zeta chain of the T-cell receptor on the same chromosome (Kuster et al. 1990). Interestingly, we found evidence of genome rearrangements in the eutherian lineage that may have brought *FCER1G* and *FCGR3A* closer together (fig. 5*C*). Notably, the MHC class I genes also experienced gene order rearrangements in the eutherian lineage (Belov et al. 2006).

The adaptive immune system can be traced back to the time of appearance of jawless vertebrates (Cannon et al. 2004; Nagawa et al. 2007). Interestingly, the timing coincided with the first signal of interacting gene cluster creation (fig. 4*C*). The fact that so many clusters include immune genes emphasizes the importance of this system during vertebrate evolution and indicates that selection on immune clusters may have been a major factor affecting vertebrate genome rearrangement. We speculate that new regulatory mechanisms, epistatic factors, or other evolutionary constraints associated with the genes of the adaptive immune system have driven much of the evolution of interacting gene clusters in the human genome.

Supplementary Material

Supplementary tables S1 and S2 and figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Ken Wolfe for a critical appraisal of an early version of this manuscript and all the members of the McLysaght Laboratory for helpful discussions. This work is supported by Science Foundation Ireland.

Literature Cited

Belov K, Deakin JE, Papenfuss AT, et al. (18 co-authors). 2006. Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. PLoS Biol. 4:e46.

Blumenthal T, Evans D, Link CD, et al. (11 co-authors). 2002. A global analysis of Caenorhabditis elegans operons. Nature. 417:851–854.

Cannon JP, Haire RN, Rast JP, Litman GW. 2004. The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. Immunol Rev. 200: 12–22.

Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 26:183–186.

De Ferrari L, Aitken S. 2006. Mining housekeeping genes with a Naive Bayes classifier. BMC Genomics. 7:277.

Finnegan EJ, Sheldon CC, Jardinaud F, Peacock WJ, Dennis ES. 2004. A cluster of Arabidopsis genes with a coordinate response to an environmental stimulus. Curr Biol. 14:911–916.

Fukami-Kobayashi K, Shiina T, Anzai T, Sano K, Yamazaki M, Inoko H, Tateno Y. 2005. Genomic evolution of MHC class I region in primates. Proc Natl Acad Sci USA. 102:9230–9234.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature. 425:737–741.

Hubbard TJ, Aken BL, Beal K, et al. (58 co-authors). 2007. Ensembl 2007. Nucleic Acids Res. 35:D610–D617.

Hurst LD, Lercher MJ. 2005. Unusual linkage patterns of ligands and their cognate receptors indicate a novel reason for non-random gene order in the human genome. BMC Evol Biol. 5:62.

Hurst LD, Pal C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 5:299–310.

Hurst LD, Williams EJ, Pal C. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. 18:604–606.

Huynen MA, Snel B, Bork P. 2001. Inversions and the dynamics of eukaryotic gene order. Trends Genet. 17:304–306.

Kelley J, Walter L, Trowsdale J. 2005. Comparative genomics of natural killer cell receptor gene clusters. PLoS Genet. 1:129–139.

Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet. 76:8–32.

Kuster H, Thompson H, Kinet JP. 1990. Characterization and expression of the gene for the human Fc receptor gamma subunit. Definition of a new gene family. J Biol Chem. 265:6448–6452.

Lercher MJ, Blumenthal T, Hurst LD. 2003. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. 13:238–243.

Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. 31:180–183.

Makino T, Suzuki Y, Gojobori T. 2006. Differential evolutionary rates of duplicated genes in protein interaction network. Gene. 385:57–63.

Nagawa F, Kishishita N, Shimizu K, et al. (15 co-authors). 2007. Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. Nat Immunol. 8:206–213.

Nei M. 1967. Modification of linkage intensity by natural selection. Genetics. 57:625–641.

Oettinger MA, Schatz DG, Gorka C, Baltimore D. 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science. 248:1517–1523.

Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. 2007. Evolution of protein complexes by duplication of homomeric interactions. Genome Biol. 8:R51.

Peri S, Navarro JD, Amanchy R, et al. (52 co-authors). 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13:2363–2371.

Poyatos JF, Hurst LD. 2006. Is optimal gene order impossible? Trends Genet. 22:420–423.

Poyatos JF, Hurst LD. 2007. The determinants of gene order conservation in yeasts. Genome Biol. 8:R233.

Robyr D, Suka Y, Xenarios I, Kurdistani SK, Wang A, Suka N, Grunstein M. 2002. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. Cell. 109:437–446.

Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol. 22:767–775.

Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol. 1:5.

Teichmann SA, Veitia RA. 2004. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. Genetics. 167:2121–2125.

The MHC sequencing consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature. 401:921–923.

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. 2004. An abundance of bidirectional promoters in the human genome. Genome Res. 14:62–66.

Trowsdale J. 2002. The gentle art of gene arrangement: the meaning of gene clusters. Genome Biol. 3:COMMENT2002.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 417:399–403.

Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol. 18:1283–1292.

West AG, Gaszner M, Felsenfeld G. 2002. Insulators: many functions, many mechanisms. Genes Dev. 16:271–288.

Williams EJ, Hurst LD. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. J Mol Evol. 54:511–518.

William Martin, Associate Editor

Accepted June 3, 2008