Divergence of Spatial Gene Expression Profiles Following Species-Specific Gene Duplications in Human and Mouse

Lukasz Huminiecki and Kenneth H. Wolfe

*Genome Res.* 2004 14: 1870-1879
doi:10.1101/gr.2705204

**Supplementary data**

“Supplemental Research Data”
[http://www.genome.org/cgi/content/full/14/10a/1870/DC1](http://www.genome.org/cgi/content/full/14/10a/1870/DC1)

**References**

This article cites 42 articles, 25 of which can be accessed free at:
[http://www.genome.org/cgi/content/full/14/10a/1870#References](http://www.genome.org/cgi/content/full/14/10a/1870#References)

Article cited in:
[http://www.genome.org/cgi/content/full/14/10a/1870#otherarticles](http://www.genome.org/cgi/content/full/14/10a/1870#otherarticles)

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://www.genome.org/cgi/content/full/14/10a/1870#otherarticles)

**Notes**

To subscribe to *Genome Research* go to:
[http://www.genome.org/subscriptions/](http://www.genome.org/subscriptions/)

© 2004 Cold Spring Harbor Laboratory Press
Divergence of Spatial Gene Expression Profiles Following Species-Specific Gene Duplications in Human and Mouse

Łukasz Huminiecki¹,² and Kenneth H. Wolfe

Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland

To examine the process by which duplicated genes diverge in function, we studied how the gene expression profiles of orthologous gene sets in human and mouse are affected by the presence of additional recent species-specific paralogs. Gene expression profiles were compared across 16 homologous tissues in human and mouse using microarray data from the Gene Expression Atlas for 1575 sets of orthologs including 250 with species-specific paralogs. We find that orthologs that have undergone recent duplication are less likely to have strongly correlated expression profiles than those that remain in a one-to-one relationship between human and mouse. There is a general trend for paralogous genes to become more specialized in their expression patterns, with decreased breadth and increased specificity of expression as gene family size increases. Despite this trend, detailed examination of some particular gene families where species-specific duplications have occurred indicated several examples of apparent neofunctionalization of duplicated genes, but only one case of subfunctionalization. Often, the expression of both copies of a duplicated gene appears to have changed relative to the ancestral state. Our results suggest that gene expression profiles are surprisingly labile and that expression in a particular tissue may be gained or lost repeatedly during the evolution of even small gene families. We conclude that gene duplication is a major driving force behind the emergence of divergent gene expression patterns.

[Supplemental material is available online at www.genome.org.]
titative, form of subfunctionalization can also occur through de-
generative promoter mutations that reduce the level of expres-
sion of the daughter genes to the point where both of them are
needed to supply enough protein product, even without any
change in the gene’s function or tissue specificity of expression.

In complex organisms subfunctionalization can perhaps
most readily be detected if it causes changes in the expression
profiles of genes. In the long run, this type of subfunctionaliza-
tion should lead to a decrease in expression breadth and the
development of tissue-specific genes, as has been demonstrated
in some previous studies on individual genes where subfunc-
tionalization has occurred (Force et al. 1999; Prince and Pickett 2002).

Over evolutionary time, duplicated copies of genes will accumu-
late amino acid substitutions at the same time as they are diverg-
ing in expression pattern. Furthermore, new tissue and cell types
can be created during the course of evolution while others may
disappear. The molecular basis of phenotypic differences be-
tween species are beginning to be understood through genome
sequence information, and there is a growing awareness that
gene duplications are responsible for some of these differences
(Enes et al. 2003).

A more global study of the relationship between gene du-
plication and gene expression changes was recently carried out
by Makova and Li (2003), who used microarray data to study the
divergence in transcription profiles of large numbers of paralo-
gous genes within the human genome. They showed that the
correlation in the spatial expression profiles (across different tis-
sues) of paralogs decreases with increasing age of the paralogs.
They also showed that divergence in expression profiles occurred
quite rapidly, consistent with a subfunctionalization model of
duplicate gene preservation. However, because they only consid-
ered expression data from humans, Makova and Li were not able
to put the levels of transcription profile change seen in dupli-
cated genes (paralogs) into the context of the changes in tran-
scription profile that occur during evolution even without gene
duplication, for example, in comparisons of orthologous genes
between species.

Here, we have used microarray data from the Gene Expres-
sion Atlas (GEA; Su et al. 2002) to study divergence of transcrip-
tion profiles of genes between human and mouse, compared
across homologous tissues in the two species. We focus on loci
where recent species-specific gene duplications have occurred
within human or mouse, creating pairs of young paralogs whose
transcription profiles can be compared with that of the single-
copy ortholog in the other species. We show that the presence of
a species-specific gene duplication accelerates the rate of expres-
sion divergence between human and mouse, and also that these
recent duplicates are subject to reduced constraints on their pro-
tein sequences.

Finally, we observe that expression domains become pro-
gressively narrower as the number of paralogs increases. This
finding is consistent with the subfunctionalization model. How-
ever, detailed examination of several gene families in which re-
cent duplications have occurred suggests that in most cases mul-
tiple changes in the spatial expression profile have occurred, and
neofunctionalization is suggested by the appearance of expres-
sion in a new tissue.

RESULTS

Linking Gene Expression Data to Orthologs
and Paralogs in Human and Mouse

Gene expression data from human and mouse were obtained
from the Gene Expression Atlas (GEA) of Su et al. (2002). These
data comprise 101 human (microchip U95A) and 89 mouse (mi-
crochip U74A) Affymetrix experiments. From these, we used data
from 16 homologous tissues that had been studied in both hu-
man and mouse (see Methods). Annotations of the human and
mouse genomes were obtained from the Ensembl database using
the EnsMart tool (Hubbard et al. 2002). We were able to confi-
dently assign 5261 out of the total of 24,848 human genes (21%),
and 4522 of 24,950 mouse genes (18%) to Affymetrix probes,
after omitting probes that mapped to multiple genomic locations
and other possible artifacts as described in Huminiecki et al.
(2003).

A list of putative human/mouse orthologs was also down-
loaded from Ensembl. We retained only orthologs that were
listed in a simple one-to-one relationship between human and
mouse. A pair of orthologs was regarded as linked to expression
data if both the human and mouse gene were mapped to Af-
ymetrix probes. Of the 13,341 pairs of orthologs in the list, 1833
(14%) were mapped to probes in both human and mouse.
Among these, 1575 (94%) were expressed in at least one of the 16
homologous tissues we used for human/mouse comparisons. The
minimal signal accepted for expression was an Affymetrix aver-
age difference (AD) value of 200 (see Methods). Paralogs within
each genome were identified using TRIBE (Enright et al. 2002).
Both genes could be linked to Affymetrix probes for 2697 (4%) of
the 67,666 paralog pairs in human, and for 2680 (4%) of the
73,843 paralog pairs in mouse.

Orthologs Are More Conserved in Amino Acid
Sequence Than Paralogs of Similar Evolutionary Age

Mean and median values of synonymous ($K_s$) and nonsynonym-
ous ($K_a$) sequence divergence were calculated for 13,341 or-
thologs between human and mouse, and for paralogs within
each species, using the method of Yang and Nielsen (2000;
Supplemental Table S1). The median values for orthologs were
$K_a = 0.61$ and $K_s = 0.08$, in good agreement with the medians
reported in the mouse genome sequencing paper (0.602 and
0.071, respectively; Waterston et al. 2002) which used the same
method and a similar number of orthologs, but our median $K_a$
is significantly higher than the value of 0.46 reported in the earlier
study by Makalowski and Boguski (1998), which used Ina’s
method (ina 1995) and 1138 orthologs.

To compare the $K_a/K_s$ ratio between orthologs and paralogs,
we only looked at gene pairs whose $K_s$ values were in the interval
0.51–0.71 (i.e., within a narrow range around the median value
for orthologs) because comparisons of $K_a/K_s$ ratios are only mean-
ful if the $K_s$ values are similar (Nembaware et al. 2002). This
filtering meant that only 29% of orthologs and 3% of paralog
pairs were considered, but the sample sizes remained large
(Supplemental Table S1). Among these, there is a very substantial
difference in $K_a/K_s$ with mean ratios of 0.71 and 0.88 seen in
paralogs of mouse and human, respectively, compared with only
0.17 in orthologs. This fourfold to fivefold increase in the total
cumulative amount of nonsynonymous substitutions calculated
for relatively old paralogs indicates even more dramatic relax-
ation in selective constraints in the period immediately follow-
ing the duplications event, which took place approximately at the
same time as the mouse/human speciation date. Similar con-
clusions were reached by Lynch and Conery (2000), Kondrashov
et al. (2002), and Seoighe et al. (2003), using different approaches
and data sets.

Identification of Ortholog Pairs With Species-Specific
Duplication in Human or Mouse

To investigate whether the relaxation of constraint on protein
sequence evolution seen in paralogs is correlated with a tendency
to change their spatial patterns of gene expression, we focused on
paralogs that have been formed recently by species-specific gene duplication in either human or mouse. The Ensembl list of human/mouse orthologs was merged with human and mouse paralog data sets to identify lineage-specific duplications. Lineage-specific duplications were defined as those where the synonymous substitution distances between paralogs was lower than 0.70 and, in any case, lower than that between the corresponding human/mouse orthologs. We were satisfied that the quality of ortholog assignment was equally high regardless of whether a human or mouse duplication occurred. The average K, between orthologs was 0.68 ± 0.42 for the set with human-specific duplications and 0.60 ± 0.45 for the set with mouse-specific duplications, values that are actually lower than the 0.71 ± 0.43 calculated for the set of 1325 one-to-one orthologs.

For this analysis, we considered only ortholog sets for which gene expression data were available from GEA. Expression information was available for 1575 human/mouse ortholog sets including 250 with species-specific paralogs. Among these, there were 1325 simple one-to-one human/mouse orthologs (those without any duplication in the human or mouse lineage), 163 sets with recent duplications in human and 139 with recent duplications in mouse. For most of these sets, GEA expression data were only available for one of the species-specific duplicates, as well as for the ortholog in the other species. Several sets included multiple species-specific duplications: there were 192 sets with two lineage-specific paralogs, 44 with three, 27 with four, 8 with five, 11 with six, and 20 with more than six. In 52 ortholog sets there were both human and mouse lineage-specific duplications.

Orthologs With Lineage-Specific Gene Duplications Tend Not to Have Highly Correlated Expression Profiles

One-to-one orthologous gene pairs are generally expected to be correlated in their expression patterns across different tissues. For example, a human liver-specific gene is expected to have a mouse ortholog that also has liver-specific expression. Figure 1A shows the histogram of expression correlation coefficients (R) for the set of 1325 human/mouse one-to-one orthologs. The distribution of R-values appears bimodal with one peak showing orthologs with highly correlated expression profiles (close to the value of R = 1), and a second peak centered on R = 0. The latter group of orthologs presumably includes those that are primarily expressed in tissues other than the 16 that we were able to compare between human and mouse. A similarly bimodal plot of ortholog expression profiles was obtained in the original analysis of GEA data by Su et al. (2002) with 799 ortholog pairs.

In contrast, there is no peak at the high end of correlation values when analogous histograms are drawn for ortholog sets that also have lineage-specific duplications in either human or mouse (Fig. 1B,C). In other words, the presence of a species-specific paralog increases the likelihood that two species’ expression profiles differ. This result is not an artifact of the choice of bin sizes chosen for the histograms, as shown by the cumulative curves in Figure 1.

A nonparametric randomization procedure was used to estimate the statistical significance of the underrepresentation of highly correlated pairs in orthologs with lineage-specific paralogs. Specifically, we estimated the probability that R-value distributions that are skewed similarly to those in Figure 1, B and C, could be obtained by chance alone. Random samples of K ortholog pairs were chosen from the total set of expressed orthologs, where K was 163 and 139 for human and mouse, respectively, and the sampling was repeated 10,000 times. For each random sample, the number of ortholog pairs in four overlapping intervals (R > 0.9, R > 0.8, R > 0.7, and R > 0.6) on the right-hand side of the distribution was counted. The proportion of samples that produced equal or lower numbers of orthologs in a given R-interval than observed in the real data defined the p-value of the observation. All intervals with R > 0.7 were significant at α < 0.01 level in both human and mouse, except for one that was significant at α < 0.05 (Table 1).

Orthologs With Multiple Duplications Show Even Weaker Correlations of Expression Profiles

From Table 1, we chose a value of R > 0.8 as a threshold for defining “similar” expression profiles and compared the proportions of genes with similar expression among groups of orthologs with different amounts of lineage-specific gene duplication. In the set of 1325 one-to-one orthologs, 181 pairs (13.7%) had similar expression profiles by this criterion. The proportion of pairs having similar profiles was lower in all categories of orthologs with lineage-specific duplications. Among 148 ortholog sets with exactly one human- or mouse-specific duplication, only 7.4% (11
For human/mouse orthologs with tissue-specific, intermediate, and housekeeping gene expression profiles. Kruskal-Wallis rank test, p analyzed, and using expression in one to two tissues as a working of both the human and the mouse gene in the 16 tissues ana-

mouse orthologs, defined on the basis of the expression breadth

ded independently in both human and mouse are even less likely to be highly correlated in expression than those where only one duplication event occurred. This result is suggestive of either sub-

functionalization or neofunctionalization after each gene dupli-

pated in both human and mouse (1.9%; 1 pair out of 52).
The reduction in each of these three groups relative to the group of one-to-one orthologs is statistically significant in $\chi^2$ tests (cal-

culated as two-by-two contingency tables in pairwise comparis-

; $p < 0.05$ for each of the three tests). Thus, orthologs with multiple lineage-specific duplications or that duplicated inde-

pendently in both human and mouse are even less likely to be highly correlated in expression than those where only one duplication event occurred. This result is suggestive of either sub-

functionalization or neofunctionalization after each gene dupli-

Tissue-Specific Genes Are More Likely to Belong to Large Gene Families

It was previously reported that tissue-specific genes evolve faster than ubiquitously expressed genes, based on analyses of ex-

pressed sequence tag data (Duret and Mouchiroud 2000) and a microarray data set of 1581 genes (Zhang and Li 2004). We con-

firmed this observation using the larger GEA data set (Fig. 2A).

Average $K_a/K_s$ ratios were calculated for three subsets of human/ mouse orthologs, defined on the basis of the expression breadth of both the human and the mouse gene in the 16 tissues ana-

lyzed, and using expression in one to two tissues as a working tissues studied in which a given gene was expressed above the threshold level. $\text{PEMM}_\text{MAX}$ is the maximal value of the Preferential Expression Measure ($\text{PEM}$; Humanniecki et al. 2003) for a gene, which measures the extent to which the gene’s transcription pro-

ile is concentrated into one tissue. $\text{PEM}$ is $\log_{10}(S/A)$, where $S$ is the Affymetrix signal for a given gene in a specific tissue, and $A$ is the arithmetic mean signal for the gene across all tissues. $\text{PEMM}_\text{MAX}$ is the maximal value of $\text{PEM}$ among all tissue scores for a gene. The more tissue-specific a gene’s expression is, the higher its $\text{PEMM}_\text{MAX}$ score. It can be seen that, in both human and mouse, larger gene families tend to have decreased expression breadth and increased tissue specificity (Fig. 2B,C), consistent with sub-

functionalization. This observation differs from Zhang and Li’s conclusion that there is little difference in the average family size of housekeeping and tissue-specific genes in mammals. The dif-

ference between these results may hinge on the fact that Zhang and Li (2004) grouped genes into only two classes of family size (single-copy and those with two or more copies), whereas we see differences in expression breadth and specificity between smaller and larger multigene families (Fig. 2B,C).

Statistical tests confirmed that each of the three size sub-

classes of genes in families (Fig. 2B,C) is narrower in its expres-

sion domain than singleton genes. The distribution of $\text{PEMM}_\text{MAX}$ values was transformed to a normal distribution using a classic Box–Cox transformation with the $\lambda$ parameter of 0.1, according to the formula $f(x) = (x^\lambda - 1)/\lambda$, where $x$ is the $\text{PEMM}_\text{MAX}$ value,
and the Welch one-sided two-sample t-test was applied. The difference between the singleton genes and the families is highly significant in all cases. Pairwise comparisons show singletons to have less-specialized expression than genes in families of two to three, four to five, and more than five members with $P = 3e - 06$, $1e - 09$, and $6e - 14$, respectively, for human, and $P = 8e - 12$, $7e - 08$, and $1e - 08$ for mouse. For percentage breadth of expression values, the distribution was very different from normal, with most genes being far toward either the tissue-specific or the housekeeping end of the distribution. Therefore, the nonparametric Fligner-Killeen test for homogeneity of means was applied (http://www.r-project.org/; “The R Reference Index”). P-values were <0.001 in both human and mouse. Additionally, pairwise Wilcoxon tests were applied following a transformation according to the formula $f(x) = 1/(50.01 - x)$, which was used to center the distribution of percentage breadth of expression around zero. The following $P$-values were obtained in order of increasing family size: 0.002, 0.001, and 7e - 10 for human, and 0.024, 0.007, and 1.05e - 06 for mouse.

Examples of Expression Divergence in Sets of Orthologs With Lineage-Specific Duplications

To examine how gene expression changes following species-specific gene duplications, we studied four gene families in detail. In particular, we looked at cases in which GEA expression data were available for two or more paralogs in one species as well as the ortholog in the other species. We used literature and EST data, in addition to GEA, to gain further information about these genes and their expression patterns.

Seven In Absentia Homolog 1

SIAH1 is a broadly expressed human ortholog of the *Drosophila Seven in Absentia* gene involved in the development of the R7 photoreceptor (Hu et al. 1997). SIAH1 has undergone a duplication in the rodent lineage giving rise to two murine orthologs: Siah1a and Siah1b (Della et al. 1993). Siah1b is broadly expressed like its human ortholog, but Siah1a has a dramatically narrowed expression pattern, being expressed only in the lung. Figure 3A shows the relationships among the spatial expression profiles for the three genes in this ortholog set. The profile for Siah1b is positively correlated with the human ortholog, whereas Siah1a has a negative correlation coefficient. Lung expression is seen in both of the mouse co-orthologs but not in the human gene, thus it is not clear whether the ancestral expression profile included lung (tissue #11 in Fig. 3A). For other tissues, the data in Figure 3A allow parsimony arguments to be made in favor of both subfunctionalization and neofunctionalization of gene expression patterns. For example, the gain of amygdala (tissue #2) expression by mouse Siah1b is inferred to be neofunctionalization, whereas the loss of expression of Siah1a in many other tissues (#1, #3–#9) could be viewed as subfunctionalization. However, this is not a classic case of subfunctionalization in which ancestral functions have been divided up among daughters, because Siah1a has not retained any ancestral expression sites (at least, not among the 16 tissues considered here).

Glucose–6–Phosphate Dehydrogenase

G6PD is also a gene with broad expression in human tissues. It is an X-linked gene, whose main function is in the pentose phosphate pathway (Martini et al. 1986). Mouse has two co-orthologs of this gene: the X-linked *G6pd* and the autosomal *G6pd2*, which lacks introns and was formed by reposition (Zollo et al. 1993; Hendriksen et al. 1997). Like the SIAH1 example, the profile correlation coefficient of the human gene with one of the mouse co-orthologs is much higher than with the other, and one of the $R$-values is negative (Fig. 3B).

*G6pd2* was reported to be transcribed in testis in spermatogenic cells where the X-linked gene is not expressed (Hendriksen et al. 1997). Indeed, this is confirmed in the GEA data set, where *G6pd2* shows expression in testis (tissue #3 in Fig. 3B), but *G6pdx*...
and G6PD do not. Hence, expression in testis (as well as in liver; tissue #15) could be regarded as a neofunctionalization of G6pd2 following its retroposition. Remarkably, none of the five tissues in which mouse G6pd2 is expressed is the same as any of the six tissues in which human G6PD is expressed. Viewed from the human perspective, the expression pattern of G6pd2 could be described as complete neofunctionalization. However, mouse G6pdx shows a very broad expression pattern, covering all six tissues where the human gene is expressed, as well as three of the five where G6pd2 is expressed, and four tissues that are unique to G6pdx. Hence, the two mouse co-orthologs show extensive neofunctionalization (gaining expression in two tissues for G6pd2 and four tissues for G6pdx), but there is no evidence of subfunctionalization of the human gene’s expression pattern (it has all been retained by G6pdx).

Cysteine-Rich Secretory Proteins

There are three CRISP genes in both human and mouse, but their nomenclature is misleading because none of the genes are simple one-to-one orthologs between human and mouse. Figure 4 illustrates the complex evolutionary history of the CRISP family, including sequences from rat, Fugu, and Drosophila. A single ancestral gene at the base of the vertebrate lineage was most likely subject to two rounds of gene duplication before the human/rodent split, a human-specific duplication (resulting in CRISP2 and CRISP3), a mouse-specific duplication (resulting in Crisp1 and Crisp3), differential gene losses, and multiple shifts in expression patterns.

Northern blots indicated that human CRISP1 is epididymis-specific, CRISP2 is expressed in testis, and the most prominent expression site for CRISP3 is in salivary gland (Kratzschmar et al. 1996). Using information from the literature and EST expression patterns as well as data from GEA, and interpreting expression patterns in the context of the phylogenetic tree in Figure 4, we suggest that the ancestral vertebrate CRISP gene was primarily expressed in the epididymis. This is agreement with what is reported to be the classic biological role of CRISP: facilitating plasma membrane recognition and fusion between sperm and egg (Brooks et al. 1986; Evans 2002). This role is known to be carried out by CRISP1 in human and AEG in rat.

Testicular expression, as seen in the ortholog trio of human CRISP2, mouse Crisp2, and rat Tpx1, appears to be a neofunctionalization following duplication (node A in Fig. 4) of a gene expressed in the salivary gland shortly before the human/rodent split. These testis-specific genes play a role in sperm maturation and are responsible for interactions between spermatogenic precursors and Sertoli cells (Maeda et al. 1998). The precise function of CRISP proteins in thyroid and prostate remains to be established, but it is interesting that there are several homologous cysteine-rich peptides in snake venoms, which are products of modified salivary glands (Chang et al. 1997; Yamazaki et al. 2002, 2003). The human-specific CRISP2/CRISP3 duplication is a putative subfunctionalization event. The ancestral human gene was expressed in both testis and salivary gland. Following duplication, CRISP2 and CRISP3 subfunctionalized to become expressed specifically in testis and salivary gland, respectively. The CRISP phylogenetic tree also shows several lineage-specific gene losses. If all duplicate genes were preserved following duplications, there should be four CRISP proteins in human and mouse. Gene losses are inferred, for example, in the case of the mouse ortholog of human CRISP1, and human ortholog of the murine Crisp1/Crisp3 pair.

The overall impression from Figure 4 is that a surprisingly high number of changes in gene expression patterns have occurred during the evolution of the CRISP family. Using a parsimony approach and assuming ancestral expression in epididymis alone, we infer that there have been at least nine changes in expression site during the evolution of this family in vertebrates (Fig. 4), including several reversals of state. Alternative assumptions about the ancestral expression profile require even more changes and so are less parsimonious.

Figure 4 Phylogenetic relationships and expression profiles in the cysteine-rich secretory protein (CRISP) family. The phylogenetic tree was constructed by the minimum evolution method after CLUSTALW sequence alignment of protein sequences. Bootstrap values are shown. Human CRISP1, CRISP2, CRISP3, and mouse Crisp1, Crisp2, Crisp3, were compared with three rat genes: testis-specific protein 1, Tpx1 (Maeda et al. 1998; O’Bryan et al. 1998); epididymal glycoprotein, AEG (Charest et al. 1988); and an Ensembl novel gene prediction, as well as outgroups from Fugu rubripes and Drosophila melanogaster (used to root the tree). Where GEA data were not available (CRISP1 and all rat genes), expression evidence from the literature (L) or EST databases (E) was used. The numbering and color scheme for gene expression data is the same as in Figure 3.

Divergence of Gene Expression Profiles
Eosinophil-Associated Ribonucleases

These are the classic example of multiple species-specific duplications, and have been the subject of intense investigation in the field of molecular evolution. In human, the eosinophil-expressed subfamily expanded through two lineage-specific duplications to give rise to RNASE2 (eosinophil-derived neurotoxin), RNASE3 (eosinophil cationic protein; Hamann et al. 1990), and ECRP (GenBank X55989). Multiple mouse-specific duplications produced at least six genes (Ear-1, Ear-2, and mR-3 through mR-6; Batten et al. 1997). Zhang and co-workers have previously shown that the eosinophil-associated subfamily has been subject to very strong positive selection (Zhang et al. 1998, 2000; Zhang and Rosenberg 2002). To investigate whether expression pattern shifts are likely to occur in a family whose evolution is well documented to be driven by positive selection at the protein sequence level, we examined the family of eosinophil-associated ribonucleases using expression data from GEA, ESTs, and literature (Fig. 5). We found that this family contrasts sharply with the CRISPs because of few expression profile changes. Almost all the expression profiles were consistent with expression in single adult cell-type eosinophils (immune cells infiltrating other tissues). However, we also found expression in immune and blood progenitor tissues such as bone marrow, CD34+ cells from cord blood, and fetal liver and spleen. mR-5 was a sole exception, not being expressed in any of the immune tissues.

DISCUSSION

After a gene duplication, the daughter sequences tend to undergo a period of accelerated protein sequence change, as evidenced by the markedly increased $K_a/K_s$ ratio in paralogs as compared with orthologs of similar ages (Supplemental Table S1). In addition, paralogs tend to diverge in their gene expression patterns, which we documented in several ways using transcription profiles from 16 homologous tissues from the GEA data set (Su et al. 2002). One persistent concern with microarray experiments is that cross-hybridization between closely related transcripts and probes may affect the results. However, we do not think that cross-hybridization is a serious problem in our analysis, because the primary data come from Affymetrix oligonucleotide arrays as opposed to cDNA arrays, and we excluded oligonucleotide probes that did not map to a unique gene in human or mouse. Also, our detailed examination of the GEA data showed that, although a small number of probes that were flagged by Affymetrix as being susceptible to cross-hybridization did, indeed, show evidence of cross-hybridization, the great majority of the data was free of this artifact and there was no correlation between the sequence similarity of a pair of genes and the correlation of their expression profiles (see Methods).

We found that among orthologs, a relatively small number of genes show strongly conserved expression profiles between human and mouse. For the one-to-one ortholog class (those lacking recent species-specific duplications in either human or mouse), only 14% of pairs show a strong correlation coefficient of $R > 0.8$ across the 16 tissues compared. However, this is significantly more conservation than in the class of orthologs that have recent gene duplications. Because the ortholog pairs are all the same age (corresponding to the date of the human/mouse speciation), the increased divergence in expression profiles in ortholog sets having recent gene duplications can be directly attributed to the presence of those duplications. Moreover, the divergence in expression profiles becomes more pronounced in gene families with increasing numbers of duplications.

We propose that in many cases both genes of the duplicated pair have changed their expression pattern in comparison to the ancestral state. The overall trend is toward expression subfunctionalization and the development of tissue-specific expression in large gene families (Fig. 2B,C). Despite this trend, examining the expression profiles of three gene families in detail uncovered several examples of apparent neofunctionalization, and only one case of putative subfunctionalization (the CRISP2/CRISP3 duplication). It should be noted, however, that possible neofunctionalization between closely related transcripts and probes may affect the results. However, we do not think that cross-hybridization is a serious problem in our analysis, because the primary data come from Affymetrix oligonucleotide arrays as opposed to cDNA arrays, and we excluded oligonucleotide probes that did not map to a unique gene in human or mouse. Also, our detailed examination of the GEA data showed that, although a small number of probes that were flagged by Affymetrix as being susceptible to cross-hybridization did, indeed, show evidence of cross-hybridization, the great majority of the data was free of this artifact and there was no correlation between the sequence similarity of a pair of genes and the correlation of their expression profiles (see Methods).

We found that among orthologs, a relatively small number of genes show strongly conserved expression profiles between human and mouse. For the one-to-one ortholog class (those lacking recent species-specific duplications in either human or mouse), only 14% of pairs show a strong correlation coefficient of $R > 0.8$ across the 16 tissues compared. However, this is significantly more conservation than in the class of orthologs that have recent gene duplications. Because the ortholog pairs are all the same age (corresponding to the date of the human/mouse speciation), the increased divergence in expression profiles in ortholog sets having recent gene duplications can be directly attributed to the presence of those duplications. Moreover, the divergence in expression profiles becomes more pronounced in gene families with increasing numbers of duplications.

We propose that in many cases both genes of the duplicated pair have changed their expression pattern in comparison to the ancestral state. The overall trend is toward expression subfunctionalization and the development of tissue-specific expression in large gene families (Fig. 2B,C). Despite this trend, examining the expression profiles of three gene families in detail uncovered several examples of apparent neofunctionalization, and only one case of putative subfunctionalization (the CRISP2/CRISP3 duplication). It should be noted, however, that possible neofunctionalization between closely related transcripts and probes may affect the results. However, we do not think that cross-hybridization is a serious problem in our analysis, because the primary data come from Affymetrix oligonucleotide arrays as opposed to cDNA arrays, and we excluded oligonucleotide probes that did not map to a unique gene in human or mouse. Also, our detailed examination of the GEA data showed that, although a small number of probes that were flagged by Affymetrix as being susceptible to cross-hybridization did, indeed, show evidence of cross-hybridization, the great majority of the data was free of this artifact and there was no correlation between the sequence similarity of a pair of genes and the correlation of their expression profiles (see Methods).
alizations are relatively easy to detect (i.e., the appearance of gene expression by one paralog in a tissue where neither the other paralog, nor the ortholog in the other species, is expressed—although this does not demonstrate that the change was fixed by positive selection). In contrast, inferring expression subfunctionalization requires that the tissues of expression of a gene in one species become more-or-less perfectly divided up among multiple recent paralogs in the other species, which is a much more complicated set of conditions. Moreover, partitioning-out of the tissues of expression is only one of several ways by which two daughter genes could become subfunctionalized (Lynch 2004). We should also bear in mind that the present-day expression profiles may not be the same as those shortly after gene duplication. We feel that at present there are insufficient data to attempt to quantify the relative rates of neo- versus subfunctionalization, and we note that these processes are not mutually exclusive. However, our observations do suggest that the concepts of subfunctionalization and neofunctionalization may be too simple a vocabulary to describe what has actually happened at many loci after duplication, and that it may be possible for two paralogs to simultaneously become subfunctionalized as regards expression in one group of tissues, and neofunctionalized in other tissues.

Microarray data have further potential to enhance our understanding of the patterns of subfunctionalization and neofunctionalization of duplicated genes. Furthermore, neofunctionalization does not necessarily have to be secondary to a gene duplication event. The existence of many one-to-one orthologs that are poorly correlated in expression between human and mouse suggests that expression diversification also occurs independently of duplication. This may be a neutral evolutionary phenomenon, or it could reflect genuine physiological differences between human and mouse, such as in the biology of reproduction or olfaction. Further investigation is also warranted into the evolution and interdependency of arbitrary expression domains (tissue categories) investigated in microarray studies, such as somatic versus reproductive tissues, or tissues derived from the same embryonic layer. Finally, from a cell biology perspective, tissue expression is simply a cumulative function of heterogeneous cellular transcriptomes. The expression states of a tissue may therefore vary, depending on the proportions and interactions of the constituent cell types. On the other hand, organs and tissues with very different biological roles might share a surprisingly high proportion of cellular elements, such as vascular cells (endothelial cells, vascular smooth muscles), infiltrating white cells, or cells of the connective tissue.

METHODS

Sequence Data

The Ensembl-confirmed gene protein and nucleotide set were extracted using the EnsMart tool (Hubbard et al. 2002). There were 24,848 human genes, and 24,950 mouse genes. In case of multiple transcripts being mapped to one gene, only the transcript resulting in the longest protein product was accepted. A list of human/mouse putative orthologs was downloaded from the Ensembl database. Only reciprocally unique pairs were retained resulting in 13,341 pairs of high-quality orthologs. Human and mouse paralogs were determined within the Ensembl data set using the TRIBE algorithm described by Enright et al. (2002).

Proteins coded for by ortholog and paralog gene pairs were aligned using CLUSTALW with default parameters. Protein-coding nucleotide sequences were then aligned using protein alignments as a guide with the program TRANALIGN from the Emboss suite (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/tranalign.html). Pairwise Ks and Kp distances were calculated using the method of Yang and Nielsen implemented in the program Yn00 from the PAML suite (version 3.13).

Expression Data

Human gene and mouse expression data were derived from the Gene Expression Atlas (GEA) of Su et al. (2002), downloaded from http://expression.gnf.org. This data set comprises 101 human U95A and 89 mouse U74A Affymetrix microchip experiments. Many of the measurements were performed in duplicate or triplicate on the identical or the same tissue sample. These replicates show very good correlation (Huminiecki et al. 2003) with Pearson R-squared values of 0.94 ± 0.04 for the same RNA sample (N = 45 pairwise comparisons) and of 0.87 ± 0.06 (N = 17 pairwise comparisons) for repeat hybridizations of different RNA preparations from the same tissue type. We calculated arithmetic means of the average difference (AD) values and used these in further analyses.

For orthology comparisons, 16 tissue categories shared by the mouse and human GEA data sets were used (namely, adrenal gland, amygdala, cerebellum, heart, kidney, liver, lung, ovary, placenta, prostate, salivary gland, spleen, testis, thyroid, trachea, and uterus).

An average difference (AD) value of 200 was used as a threshold for a gene to be regarded as “on” in a given tissue. Several methods of data transformation were experimented with. For example, two-state (on/off) and four-state (off/weak/strong/very strong) expression scales were investigated. However, these data transformations resulted in a lower overall Pearson R-value for the total set of human/mouse orthologs and were not used in further analysis. Pearson R was chosen as a preferred pairwise distance measure of expression correlation because it is sensitive to shape (not absolute values) of expression profiles. As such, it corresponds well to the biologist’s intuitive understanding of what coexpressed or coregulated genes are. This is particularly true for tissue-specific genes, which are most interesting from the biological point of view, being widely used as markers in tissue staining or clinical diagnosis, as drug targets, or homing molecules for drug delivery systems.

Mapping Expression Data to Genes

Affymetrix tag/UniGene cluster mapping was extracted from the Affymetrix probe consensus sequence file, which lists UniGene clusters used to design microarray probes (http://www.affymetrix.com/analysis/downloadcenter.affx). In some cases, two or more Affymetrix tags were targeted against the same UniGene cluster. In these cases, only the tag with strongest average expression across all libraries from the human Gene Expression Atlas data set was retained. UniGene clusters were mapped to the human genome (NCBI build 31) using the LocusLink and Ensembl databases. Firstly, UniGene to LocusLink mapping was extracted from the UniGene release file Hs.data (human build U150). Secondly, LocusLink/Ensembl gene id mapping was extracted from the Ensembl database (release 14.31.1) using the EnsMart tool (http://www.ensembl.org/EnsMart/). LocusLink clusters mapping to multiple UniGene clusters or multiple Ensembl genes were discarded to ensure that the resulting mapping was unique and nonredundant. A similar procedure was followed for the mouse genome (NCBI build 30).

Cross-Hybridization in Microarray Data

The primary data in this study are derived from the Affymetrix oligonucleotide array experiments of Su et al. (2002). In these arrays, 16 oligonucleotide probes are used for each gene, and probes likely to cause cross-hybridization are omitted from the chip during the design stage. In addition, Affymetrix chips have a very stringent built-in internal subtraction mechanism in which every antisense probe is coupled with a single mismatch probe (Wodicka et al. 1997). Target abundance is estimated by the difference between the hybridization signals from the perfect match probe and a single mismatch probe. Because the probes are 25 nt in length, we can make the approximation that even 1 nt difference in 25 (96% identity) is sufficient to safeguard differential hybridization for a single probe, although this does depend on the location of the mismatching base (mismatches in
the middle of the target/probe duplex are most effective). When all Affymetrix tags were compared against each other using BLAST, the average percentage identity between probes for pairs of paralogs with K < 0.2 was only 93%, and 88% for K between 0.2 and 1. Furthermore, we detected no correlation between the measured R-values (expression profile correlation coefficients) for pairs of paralogs and their percentage sequence identity, number of matching, or number of mismatching base pairs.

Affymetrix probes with name suffixes _f.at (“sequence family”) and _s.at (“similarity constraint”) are thought to be more prone to cross-hybridization than others (see Affymetrix manual, Data Analysis Fundamentals, Appendix B). We refer to these probes collectively as suboptimal probes. To investigate whether inclusion of data from these probes was affecting our results, we performed a separate analysis of ortholog and paralog pairs that mapped to suboptimal probes. Out of 1575 human/mouse orthologs, only in two cases did both the human and the mouse gene map to suboptimal probes. In addition, there were 155 cases in which one of the orthologs linked to a suboptimal probe, but the average R for this subset was not different from that calculated for the total set (0.296 and 0.283, respectively; P = 0.686).

In the analysis of ortholog sets with duplication, there was only one mouse set in which both the human and the mouse gene were mapped to suboptimal probes. Additionally, there were 24 and 20 sets, in groups with human-specific and mouse-specific duplications, respectively, in which one of the orthologs linked for the total set (0.296 and 0.283, respectively; P = 0.686).

In the analysis of ortholog sets with duplication, there was only one mouse set in which both the human and the mouse gene were mapped to suboptimal probes. Additionally, there were 24 and 20 sets, in groups with human-specific and mouse-specific duplications, respectively, in which one of the orthologs linked for the total set (0.296 and 0.283, respectively; P = 0.686).

ACKNOWLEDGMENTS

This study was supported by Science Foundation Ireland. We thank Greg Singer and Andrew Lloyd for many helpful discussions and lively exchange of ideas, and are grateful for helpful comments from anonymous referees.

REFERENCES


**WEB SITE REFERENCES**

http://expression.gnf.org; Gene Expression Atlas.

http://www.affymetrix.com/analysis/downloadcenter.affx; Affymetrix.


http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/tranalign.html; Emboss.

http://www.r-project.org/; The R Reference Index.

Received April 23, 2004; accepted in revised form July 21, 2004.