Separating derived from ancestral features of mouse and human genomes

Chris P. Ponting1 and Leo Goodstadt
MRC Functional Genomics Unit, University of Oxford, Department of Physiology, Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, U.K.

Abstract
To take full advantage of the mouse as a model organism, it is essential to distinguish lineage-specific biology from what is shared between human and mouse. Investigations into shared genetic elements common to both have been well served by the draft human and mouse genome sequences. More recently, the virtually complete euchromatic sequences of the two reference genomes have been finished. These reveal a high (~5%) level of sequence duplications that had previously been recalcitrant to sequencing and assembly. Within these duplications lie large numbers of rodent- or primate-specific genes. In the present paper, we review the sequence properties of the two genomes, dwelling most on the duplications, deletions and insertions that separate each of them from their most recent common ancestor, approx. 90 million years ago. We consider the differences in gene numbers and repertoires between the two species, and speculate on their contributions to lineage-specific biology. Loss of ancient single-copy genes are rare, as are gains of new functional genes through retrotransposition. Instead, most changes to the gene repertoire have occurred in large multicopy families. It has been proposed that numbers of such ‘environmental genes’ rise and fall, and their sequences change, as adaptive responses to infection and other environmental pressures, including conspecific competition. Nevertheless, many such genes may be under little or no selection.

Introduction
During the Cretaceous Period, approx. 90 million years ago, a line of small mammals, probably similar in appearance to modern tree shrews, split into two lineages. One of these led to the primates, including Homo sapiens, the other to rodents such as the laboratory mouse, Mus musculus. Despite the obvious morphological differences between humans and mice, close observation over 100 years of mouse genetics has demonstrated the many close anatomical and physiological affinities between humans and mice [1,2].

Yet how are these species similar or different in their genes and in genomes? How much has changed in these genomes in the last 90 million years, and how much has resisted change to preserve essential ancestral functions? If Darwinian adaptation has been responsible for the divergence between rodents and primates, would this have left a corresponding signature in their genomes? To answer these and other questions the sequencing of human and mouse genomes was required, followed by detailed comparisons pinpointing those nucleotides or exons or genes or chromosomal segments that have remained intact in both lineages since their last common ancestor.

The draft genome sequences of humans in 2001 [3] and mice in 2002 [4] were great boons for research in genomics, genetics and evolution. Yet, these draft sequences were never intended to be comprehensive. It was always realized that sequencing of the extremely repetitive and transposon-rich heterochromatin would be too technically challenging. However, because heterochromatin is believed to have low gene density and is largely silenced, its absence has been met with little complaint.

On the other hand, it was of great concern that a large amount of euchromatic sequence (>10% of human and >6.5% of mouse) was also found to be absent from the initial drafts. The process by which this additional sequence has been painstakingly added to genome assemblies is known as ‘finishing’, although even ‘finished’ euchromatin is not perfect and will contain a low level of missing and inaccurate sequence for some time to come [5].

The ‘finished’ human genome sequence was reported in 2004 [5] and a similar sequence for mice will be reported elsewhere soon [6]. The purpose of the present review is to provide a brief overview of the similarities and differences between these two genome sequences in the light of newly found sequence previously absent from draft assemblies.

Mind the gaps
The human genome assembly extends to 3.09 Gb (Table 1), containing 99% of the euchromatin and 94% of the entire genome [5]. This is a slightly larger genome assembly than for the mouse, which comprises approx. 2.66 Gb. Similarities between chromosome numbers and genome sizes belie the substantial rearrangement of the ancestral genome, particularly in the rodent lineage, with over 300 large (>300 kb) blocks of genes being reordered, and many more smaller scale rearrangements observed [4]. Across the whole
of the human genome, only 40% of sequence can be aligned to the mouse genome, with much of the remainder representing remnants of transposable elements that have been frequently inserted and deleted in each lineage, much of which has decayed beyond recognition.

Finishing the genome assemblies revealed that the draft assemblies were particularly deficient in segmental duplications, defined as $\sim$1 kb fragments of genomic sequence with high sequence identity (>90%) that map to multiple locations [7]. The repetitive nature of this sequence explains its recalcitrance to assembly, especially via the whole genome shotgun approach. Segmental duplications are now known to cover approx. 5% of both human and mouse euchromatin [8]. These newly discovered duplicated sequences would perhaps be only of passing interest, except that they contain a high density of duplicated protein-coding genes (see below).

These same regions also tend to be highly variable in structure, including copy number, among unrelated human or mouse individuals [9–11]. The mouse sequence is from a single highly inbred individual female from the laboratory black 6 (C57BL/6) strain. However, the human reference genome represents the agglomeration of contributions from an anonymous panel of outbred individuals [3], although more than half the assembly stems from a single contributor. Each individual carries $\sim$0.5% copy number variant sequence, often in a heterozygous state [12].

The human genome sequence is thus not a consensus representing the most frequent variants. Instead, of selection for size in choosing insert clones for sequencing, parts of the human reference genome may contain a systematic bias towards the incorporation of variable regions, especially those with high copy numbers. In any case, even for the mouse where copy number variants are seen in abundance between different inbred mouse strains [8], structural variation between individuals highlights the limitations of a single reference genome in representing an entire population.

Gene repertoires

Interest in relating gene numbers to organismal complexity has waned considerably since the discovery that the genome of the simple nematode Caenorhabditis elegans contains almost as many protein-coding genes as humans [13–15]. Nevertheless, the reduction of the human gene count from an initial 32000 in the draft human genome publication [3] to its current level of approx. 19000 shows the many inherent difficulties in gene predictions as well as the great progress that has been made since.

Initial gene counts were greatly inflated by the inclusion of fragmentary gene models, many of which represent the debris of non-functional pseudogenes. Spurious ORFs (open reading frames) present by chance in RNA transcripts were also misidentified even in the absence of either protein-coding potential or evolutionary conservation [15]. Disruptions to putative ORFs such as small insertions, deletions and frameshifts were difficult to distinguish from sequence errors that are often found among even bona fide gene predictions in a draft genome sequence. Finally, it has never been straightforward to distinguish one gene from its genomic neighbour, particularly as (non-coding or chimaeric) transcripts can often be found to span the two [16,17].

The current low estimates of gene number in the human genome [13–15] result from the exclusion of erroneous gene models that fail criteria for protein-coding potential and conservation. These rely upon three widely held observations: that intron positions and especially phase are very well conserved within coding sequence, that confirmed cases of mammalian protein-coding genes recruited de novo from non-coding sequence are very rare, and that retrotransposition generally results in non-functional sequence with immediate loss of constraint. This last assumption appears to be the most fallible, since promoter elements are sometimes fortuitously present at the 5′-end of the integration site of retrotransposons. However, empirically, because only about one functional retrogene arises per million years, this results

<table>
<thead>
<tr>
<th>Property</th>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembled genome size (Gb)</td>
<td>3.091</td>
<td>2.661</td>
</tr>
<tr>
<td>Segmentally duplicated sequence (Mb)</td>
<td>159.2 (5.52%)</td>
<td>126.0 (4.94%)</td>
</tr>
<tr>
<td>Interspersed repeats (Gb)</td>
<td>1.406</td>
<td>1.091</td>
</tr>
<tr>
<td>Number of gaps</td>
<td>118</td>
<td>1218</td>
</tr>
<tr>
<td>Sequence in gaps (Mb)</td>
<td>9.343</td>
<td>6.088</td>
</tr>
<tr>
<td>Number of gene models</td>
<td>19042</td>
<td>20210</td>
</tr>
<tr>
<td>Coding sequence (%)</td>
<td>1.07</td>
<td>1.27</td>
</tr>
</tbody>
</table>

**Table 1** | Properties of finished human and mouse reference genome assemblies

NCBI Builds 36.1 and 36 respectively.

**Figure 1** | Mouse and human gene duplication as a function of age

These are extant genes from the two reference genome assemblies. Their age of duplication has been estimated phylogenetically from the ratio of the duplicates’ branch length to the distance to the rodent primate speciation node, and by assuming that the split occurred 90 million years ago.
in only a small underestimation of \(\sim 100\) genes since the human mouse divergence [18,19].

The great majority of genes are thus gained in each lineage not by de novo creation or retrotransposition, but instead by whole gene duplication, often in tandem copies; simultaneously, functional genes are lost by deletion or after disruptive mutations (‘pseudogenization’). The fates of genes on each of the human and mouse lineages are thus best described by gene births and deaths in the time since their last common ancestor approx. 90 million years ago. Figure 1 shows the rate of accumulation of extant genes in both Mus musculus and Homo sapiens lineages. The large numbers of the most recently duplicated genes largely reflect copy number variable genes: many of these are unlikely to be fixed in future populations (see below).

**Correspondence between mouse and human genes**

To a Well-Connected Mouse

(Upon reading of the genetic closeness of mice and men.)

_Wee, sleekit, cow’rin, tim’rous beastie,_
_Braw science says that at the leastie_
_We share full ninety-nine per cent_
_O’genes, where’er the odd ane went._

© 2003 John Updike. Originally published in The New Yorker. All rights reserved.

This reworking of the 18th Century Robert Burns poem (“To a Mouse”) was sparked by the Nature editorial commentary that 99% of mouse genes have “direct counterparts in humans” [20]. This statement unfortunately gives the misleading impression that human and mouse genes mostly correspond, and that divergence in their genomes is thus largely either within the sequence of each corresponding gene, or in regulatory regions.

In fact, the mouse and human gene repertoires have been dramatically remodelled since the divergence of the lineages \(\sim 90\) million years ago: 20% (3852 of 19042) of human genes and 24% (5020 of 20210) of mouse genes are duplicate copies which have arisen since their last common ancestor (Figure 2). As for “where’er the odd ane went”, the Nature commentary reflects that in fewer than 1% of genes has every homologue become extinct in the other lineage. These are the genes expected to convey much of the functional repertoire that is conserved among mammals and, more broadly, among other animals. It is thus appropriate that these simple orthologues are being specifically targeted for disruption in large-scale phenotypic screens in the mouse to illuminate conserved mammalian biology.

The estimated gene count is substantially higher in mice (20 210) than in humans (19 042) because of the larger number of rodent-specific gene duplicates. This is almost entirely due to genes with roles in chemosensation, such as olfactory and vomeronasal receptors and pheromone genes. The cull of primate-specific chemosensation genes relative to murid rodents and other mammals seems to have accelerated in the old-world primates in the last \(\sim 25\) million years [23,24]. The contrast in gene repertoires perhaps reflects the divergent sensory requirements for mammals that differ in diet and behaviour.

Gene duplications in mice are predominantly found in tandem copies of genomic sequence [8], and most probably arise via unequal crossover or non-allelic homologous recombination. In contrast, human duplicates are more often dispersed to different chromosomes in recombinations that were mediated by the primate-specific Alu-SINE (short interspersed element) retrotransposons [25,26].
Large genomic duplications are less frequent than smaller ones. Thus compact genes are more likely to be duplicated with a complete and intact ORF, as well as the accompanying promoter and other necessary non-coding regulatory sequence. Median sizes of human- or mouse-lineage-specific genes are thus approx. 3- and 6-fold smaller than those of simple orthologues (6.6 or 3.3 kb compared with 26.6 or 21.2 kb respectively); they have fewer exons (medians of five and three compared with nine for both mice and humans); and they are 5- and 12-fold (in humans or mice) more likely to contain only a single coding exon.

Unequal crossing-over can also generate novel functional genes, either via the formation of chimaeric genes or as partial duplicates (e.g. [27]). For example, a duplication of a fragment of the rodentDlg5 gene spawned an extended family of testis-specific genes that have become extremely widespread in the mouse and rat genomes [28]. Partial duplications have also occasionally been observed within duplicated genes. The primate-specific LPA gene, for example, was formed initially by a complete duplication of PLG with a subsequent succession of further internal duplications of individual domains [29].

**Innovative functions of lineage-specific genes**

Duplicated genes form a far from random selection of all genes. Aside from the bias in gene sizes mentioned above, several functional categories are overrepresented among genes specific to either primate or rodent lineages. These include (i) chemosensation genes which are more numerous in rodents, and genes associated with (ii) immunity or host defence (e.g. T-cell receptor genes), (iii) with detoxification (e.g. cytochrome P450), and (iv) with reproduction (e.g. pheromone or cancer-testis antigen genes [30]).

Many of these genes are of considerable biological interest and are found in regions of the genome rich in segmental duplications and interspersed repeats, and are thus only present in finished genome assemblies. The extreme repetitive nature of the Y chromosome has required particularly focused efforts to complete its sequence. Indeed, further copies of rapidly evolving gene families can be expected to be discovered as the remaining gaps in the human and mouse genomes are closed.

These functional biases might imply that gene duplication events are largely adaptive: expansions in the gene repertoire may be a response to environmental challenges from pathogens and parasites, or confer reproductive advantages over other individuals from the same species. Evidence of positive selection at specific codons \(d_N/d_S\) ratio (ratio of non-synonymous to synonymous substitutions) significantly greater than 1) has often been cited for duplicated gene families (e.g. [31,32]).

However, with the small mammalian (especially primate) effective population sizes, chance gene duplications may remain in the population for some time or proceed to fixation when they are not advantageous or even mildly deleterious. This would explain the recent provenance and hence short lifespans of gene duplicates (Figure 1), some of which are unfixed and copy-number-variable within human or mouse populations. The observed overrepresentation of functional categories may simply be because duplications of genes in other classes tend to be deleterious (in part due to stoichiometric constraints) and thus are preferentially purged from the population [33,34].

Even where evidence for positive selection is cited, this may instead be the result of loss of constraint in redundant gene copies after duplication. Other cases of apparent adaptive change can be due to biased gene conversion [35]. This is a bias in the recombination-associated repair process following double-strand breaks and can lead to the erroneous inference of positive selection [36,37]. It is suggested that biased gene conversion among adjacent tandemly duplicated genes may drive their accelerated evolution [38].

On the other hand, the short lifespan of many gene duplicates may point to the short period of advantage that novelty confers in a constantly changing adaptive landscape. A high turnover in the gene repertoire with many births and deaths would be seen in an evolutionary ‘arms race’ with pathogens [39]. There can also be positive selection among genes expressed in germine or stem cells [40–42] which confers no advantage on either the individual or population. Instead, gain-of-function mutations that, for example, favour clonal expansion of mutant spermatogonia [43], may underlie much strong diversifying selection.

Almost 40 years after Ohno [44] first proposed that gene duplications are the principal forces for the generation of novelty in evolution, the relative contributions of adaptation and genetic drift in maintaining our gene repertoires thus remain to be determined. The availability of further primate and rodent assembled genomes of high quality on one hand, and extensive allele frequency data for copy number variants in natural primate and rodent populations on the other, may be required to decide between these two evolutionary scenarios [45].

**Concluding remarks**

Strenuous efforts have now provided near-complete euchromatic genome sequence assemblies of humans and mice. These have revealed large numbers of lineage-specific genes lying within segmentally duplicated genomic regions that were previously recalcitrant to sequencing and assembly. Many of these novel genes are studied only rarely, partly because of their relatively late arrival into genome sequence assemblies, and partly due precisely to the absence of single orthologous genes in the other species: inferences from model organisms about lineage-specific human genes have to be treated with the greatest caution, while a case has to be made for the relevance of each mouse-specific gene to human physiology and disease. Nevertheless, if we are to fully appreciate the functional repertoire of our own species, and
that of the mouse, our most important model organism, then these genes demand much further scrutiny in the future.

Funding

We gratefully acknowledge support from the Medical Research Council.

References


Received 10 February 2009
doi:10.1042/BST0370734