The complex relationship of gene duplication and essentiality

Takashi Makino, Karsten Hokamp and Aoife McLysaght

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

Corresponding author: McLysaght, A. (aoife.mclysaght@tcd.ie).

In yeast and worm, duplicate genes overlap in function so that deleting one of a pair from the genome is less likely to be lethal than deleting a singleton gene. By contrast, previous analyses showed that mouse duplicate genes were as essential as singletons. We show that the relationship between gene duplication and essentiality is complex in multicellular organisms, with developmental genes and genes that were duplicated by whole genome duplication being more essential than other duplicated genes.

The 'essentiality' of duplicated genes

A gene is considered 'essential' if its removal results in a lethal or sterile phenotype. Gene duplication is frequent in eukaryotic genomes and is the primary source of new genes [1–3]. Duplicate genes can have a backup role and can functionally compensate for the loss of their duplicated copies [4]. This concept was verified by genome-wide gene knockout or knockdown experiments in yeast and worm demonstrating that the essentiality of duplicate genes is significantly lower than that of singletons [4,5]. In addition, double knockout experiments in yeast of paralogs derived from whole genome duplication (WGD) strongly support functional compensation by duplicated genes [6,7]. By contrast, recent studies in mouse reported no significant difference in essentiality between duplicated genes and singletons [8,9]. This surprising result indicated that duplicate genes in mammals do not carry out a backup role and indicated that the factors governing the evolution and retention of duplicate genes differ between mammals and less complex eukaryotes.

Mouse gene knockout dataset is enriched for developmental genes

The data leading to the conclusions on essential genes in yeast and worm were based on whole-genome studies; however, the mouse studies [8,9] relied on data from <4000 genes available from MGI. The patchiness of the Mouse Genome Informatics (MGI; http://www.informatics.jax.org/) collected from many individual studies. The patchiness of the mouse gene knockout dataset is enriched for developmental genes and the genes involved in organismal development, such as GO:0007525 (multicellular organismal development) and GO:0030154 (cell differentiation), were highly over-represented in the knockouts compared to the entire genome.

Corresponding author: McLysaght, A. (aoife.mclysaght@tcd.ie).

Many individual studies. The patchiness of the mouse gene knockout dataset is enriched for developmental genes and the genes involved in organismal development, such as GO:0007525 (multicellular organismal development) and GO:0030154 (cell differentiation), were highly over-represented in the knockouts compared to the entire genome.

By contrast, recent studies in mouse reported no significant difference in essentiality between duplicated genes and singletons [8,9]. This surprising result indicated that duplicate genes in mammals do not carry out a backup role and indicated that the factors governing the evolution and retention of duplicate genes differ between mammals and less complex eukaryotes.

The data leading to the conclusions on essential genes in yeast and worm were based on whole-genome studies; however, the mouse studies [8,9] relied on data from <4000 genes available from MGI. The patchiness of the mouse gene knockout dataset is enriched for developmental genes and the genes involved in organismal development, such as GO:0007525 (multicellular organismal development) and GO:0030154 (cell differentiation), were highly over-represented in the knockouts compared to the entire genome.
The influence of whole genome duplication on the essentiality of duplicate genes

Two rounds of WGD occurred early in the vertebrate lineage [12–18] and duplicate developmental genes created by these events were preferentially retained in vertebrate genomes [19–21]. Interestingly, developmental genes were also preferentially retained after WGD in plants [22], thus indicating particular evolutionary dynamics after WGD in multicellular organisms. Recent analysis of yeast WGD duplicated genes indicated that they are less essential than small-scale duplication (SSD) duplicated genes [23,24]. We investigated the essentiality of WGD and SSD duplicated genes in mouse. We identified 1669 WGD duplicated genes [17] and 2039 SSD duplicated genes with GO ID and knockout data (see methods in the supplementary material online). We confirm that duplicate developmental genes are preferentially generated by WGD rather than SSD, even when we consider only genes from the knockout dataset ($P = 3.0 \times 10^{-6}$, $\chi^2$ test; Figure 1a). Furthermore, the $P_e$ of WGD duplicated genes (45.4%) was significantly greater than SSD duplicated genes (38.1%; $P = 3.1 \times 10^{-6}$, $\chi^2$ test; Figure 1a). This result is true even when we control for age differences between WGD and SSD duplicates (see methods in the supplementary material online). We found there was no difference in essentiality between WGD duplicated genes (45.4%) and singletons (42.2%; $P = 0.10$, $\chi^2$ test) in the entire mouse gene knockout set, but that the $P_e$ of SSD duplicated genes (38.1%) was significantly lower than that of singletons (42.2%; $P = 0.0027$, $\chi^2$ test). This is contrary to the findings in yeast [23,24].

Correlation between sequence divergence from closest paralog and essentiality of duplicated genes

Previous studies reported that there is a positive correlation between sequence divergence from the closest paralog and essentiality of duplicated genes in yeast and worm [4,5]; that is, the greater the sequence similarity between duplicated genes, the greater the propensity for mutual functional compensation. By contrast, in mouse there is a negative correlation between sequence divergence from the closest paralog and essentiality of duplicated genes [9], or no correlation [25]. We examined the relationship between sequence divergence from the closest paralog and essentiality of duplicated genes [9], or no correlation [25]. We examined the relationship between sequence divergence from the closest paralog and essentiality of duplicated genes [9], or no correlation [25].

The possibility of mutual functional compensation in vertebrates was therefore suggested in [25].

The distribution of species divergence from the closest paralog and essentiality of duplicated genes in mouse is shown in Figure 1a. The potential for mutual functional compensation by developmental genes in mouse is lower than that of non-developmentally duplicated genes. However, the distribution of species divergence from the closest paralog and essentiality of duplicated genes in mouse is not significantly higher than that of developmental singletons in mouse ($P = 0.098$, $\chi^2$ test; Table 1), and there was no difference in essentiality between developmental duplicated genes and singletons in mouse ($P = 0.00051$, $\chi^2$ test; Table 1). Thus, developmental genes are likely to be essential irrespective of gene duplication.
development of complex organisms (especially complex development than fly or mammals [29]). Similarly, worm, with only ~1000 cells, has less development process of this unicellular organism.

Concluding remarks

The relationship between gene essentiality and gene duplication is complex in mouse owing to the constraints on the developmental process and the history of genome duplications in the vertebrate lineage. Many transcription factors, members of protein complexes and developmental genes are sensitive to their relative dosage to other genes (i.e., they are dosage-balanced) [26–28]. Dosage-balanced genes are not robust to gene loss and gene duplication [27,28]. WGD duplicates all genes simultaneously and therefore does not perturb relative dosages. Whereas SSD of dosage-balanced genes is likely to be deleterious, WGD should be neutral. Furthermore, subsequent loss of dosage-balanced genes after WGD will be deleterious unless contemporaneous loss is somehow achieved. Therefore, the only opportunity to duplicate dosage-balanced genes might be when WGD occurs [27,28].

Our finding that developmental genes and genes duplicated by WGD are more essential than expected could be explained by dosage-balance constraints. Subunits of a protein complex are particularly likely to be dosage-balanced [27]. We found significant enrichment for protein complex membership for both WGD duplicated genes and all the members of the McLysaght laboratory for helpful discussions. This work is supported by Science Foundation Ireland.

Acknowledgements

We would like to thank Yoichiro Nakatani for supplying lists of the WGD duplicates, as is predicted by functional compensation models.

References

15 Vandepoele, K. et al. (2004) Major events in the genome evolution of vertebrates: parhome age and size differ considerably between ray-finned fishes and

\[ R = 0.90, P = 0.039; \text{Figure 1b} \] and fly \( R = 0.92, P = 0.027; \text{Figure 1c} \).

\[ K_A, \] the lower the \( P \), for SSD duplicated genes in mouse (Pearson’s product-moment correlation coefficient \( R = 0.94, P = 0.017 \)), but this trend was not observed in other groups of duplicated genes (Figure 1b). However, when we focused on genes with \( K_A > 0.2 \), because highly constrained genes might have unusual properties (e.g. ribosomal proteins) [4,9], we observed a positive correlation for non-developmentally duplicated genes in mouse \( (R = 0.90, P = 0.039; \text{Figure 1b}) \) and fly \( (R = 0.92, P = 0.027; \text{Figure 1c}) \).

Similarly, worm, with only ~1000 cells, has less complex development than fly or mammals [29] and has not experienced WGD.

We suggest that the constraints inherent in development of complex organisms (especially dosage constraints) combined with the unique evolutionary opportunities granted by the simultaneous duplication by WGD of all components of a pathway or complex explains the high essentiality of these genes [30,31].

In yeast, genes duplicated by WGD are less in size differ considerably between ray-finned fishes and

\[ K_A, \] the lower the \( P \), for SSD duplicated genes in mouse (Pearson’s product-moment correlation coefficient \( R = 0.94, P = 0.017 \)), but this trend was not observed in other groups of duplicated genes (Figure 1b). However, when we focused on genes with \( K_A > 0.2 \), because highly constrained genes might have unusual properties (e.g. ribosomal proteins) [4,9], we observed a positive correlation for non-developmentally duplicated genes in mouse \( (R = 0.90, P = 0.039; \text{Figure 1b}) \) and fly \( (R = 0.92, P = 0.027; \text{Figure 1c}) \).

Similarly, worm, with only ~1000 cells, has less complex development than fly or mammals [29] and has not experienced WGD.

We suggest that the constraints inherent in development of complex organisms (especially dosage constraints) combined with the unique evolutionary opportunities granted by the simultaneous duplication by WGD of all components of a pathway or complex explains the high essentiality of these genes [30,31].

Because WGD-duplicated genes and developmental genes together constitute 26% of the mouse genome, but 57% of the knockout dataset, we expect that when the data become available the genome-wide trend in mouse will show that with these notable exceptions, singletons are more essential than duplicates, as is predicted by functional compensation models.
3 Nakatani, Y. et al. (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. 17, 1254–1265
5 Blomme, T. et al. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7, R43
7 Huffn, A.L. et al. (2008) Early vertebrate whole genome duplications were preceded by a period of intense genome rearrangement. Genome Res. 18, 1582–1591
10 Hakos, L. et al. (2007) All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 8, R209
Figure 1. The relationship of proportion of essential genes (P_E) and function, divergence, and origin of duplicated genes. (a) Venn diagram of P_E of developmental, non-developmental, WGD and SSD duplicated genes in the mouse gene knockout dataset. (b, c) Relationship of sequence divergence and proportion of essential genes for mouse (b) and fly (c) duplicate genes. The x-axis indicates the non-synonymous substitution rate (K_A) between a duplicated gene and its closest paralog. The y-axis indicates the P_E in each K_A category. Error bars indicate standard error. Color code: Light blue, developmental genes; dark blue, non-developmental genes; light green, WGD genes; and dark green, SSD duplicated genes in the mouse gene knockout dataset.

Table 1. Proportion of essential genes for mouse and fly genes

<table>
<thead>
<tr>
<th>Species</th>
<th>Developmental genes</th>
<th>Non-developmentl genes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mouse</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singletons</td>
<td>52.7% (187/355)</td>
<td>38.5% (210/546)</td>
<td>44.1% (397/901)</td>
</tr>
<tr>
<td>Duplicated genes</td>
<td>60.5% (912/1508)</td>
<td>30.6% (673/2200)</td>
<td>42.7% (1585/3708)</td>
</tr>
<tr>
<td>Total</td>
<td>59.0% (1099/1863)</td>
<td>32.2% (883/2746)</td>
<td>43.0% (1982/4609)</td>
</tr>
<tr>
<td><strong>Fly</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singletons</td>
<td>79.1% (474/599)</td>
<td>34.3% (522/1520)</td>
<td>47.0% (996/2119)</td>
</tr>
<tr>
<td>Duplicated genes</td>
<td>78.9% (607/769)</td>
<td>25.6% (487/1905)</td>
<td>41.1% (1094/2674)</td>
</tr>
<tr>
<td>Total</td>
<td>79.0% (1081/1368)</td>
<td>29.5% (1090/3425)</td>
<td>43.6% (2090/4753)</td>
</tr>
</tbody>
</table>