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 Mouse Genome Informatics (MGI; http://www.informatics.jax.org/) collected from many individual studies. The patchiness of the dataset makes it susceptible to potential data biases because individual researchers might preferentially report a gene with a discernable phenotype in the knockout experiment. Therefore, reports of gene knockouts with no phenotypic change are likely to be dramatically under-represented even in cases in which the requisite experiment has actually been carried out. By contrast, the stronger the knockout phenotype, the more likely it is that the observations are reported.

Liao and Zhang [9] investigated potential data bias by comparison of their estimate of the proportion of embryonic lethal genes from the knockout dataset (14.0%) with an estimate from a random mutagenesis study (13.7%) [10]. The consistency of these two estimates led them to conclude there was no significant data bias. However, we found that 1523 out of 5078 knockout genes (30.0%) cause prenatal–perinatal lethality in the most recent knockout dataset (see methods in the supplementary material online), strongly indicating that the knockout dataset is not a representative sample. We considered the possibility that there might be a functional bias in the genes selected for knockout experiments, and in particular genes involved in development are likely to have a prenatal–perinatal lethal knockout phenotype.

We tested the hypothesis of a functional bias in knockout gene datasets for mouse and fly (see methods in supplementary material online). Out of 5078 knockout genes in mouse, 4609 genes were annotated with at least one Gene Ontology (GO) ID. We found that 18 GO terms are over-represented in the knockout dataset with respect to their frequency in the entire genome (Table S1). Notably, GO terms related to early development, such as GO:0007525 (multicellular organismal development) and GO:0030154 (cell differentiation), were highly over-represented in reported knockout genes in mouse (genes with either of these GO terms are hereafter referred to as 'developmental genes'). Even though only 11% of genes in the genome are annotated as

developmental (2682/23727), they constitute 37% of the knockout dataset (1863/5078). We also found a similar bias in fly (Table S2 and methods in the supplementary material online). Thus, there is a large bias in the reported knockout set towards genes that function in development.

Are developmental genes essential in mouse and fly?

If there is a large difference in essentiality between developmental genes and others, then this knockout dataset might give a misleading impression of the genome-wide trend. To investigate whether developmental genes are more essential than other genes, we compared the essentiality of developmental genes with non-developmental genes. Using the same approach as Liang and Li [8], and Liao and Zhang [9], we defined an essential gene in mouse as one with the knockout phenotype of sterility or lethality before maturity [8,9]. The proportion of essential genes (P_E) of developmental genes was significantly higher than that of non-developmental genes (mouse, $P < 2.2 \times 10^{-16}$; fly, $P < 2.2 \times 10^{-16}$; χ^2 test; Table 1). These results are consistent with a recent report that showed greater essentiality of genes highly expressed in early development [11]. The greater likelihood of fly and mouse developmental genes being essential is understandable given the importance of the developmental process.

The essentiality of developmental and non-developmental duplicates and singletons

Given their overall high essentiality, we wondered whether developmental genes were subject to less functional compensation by duplicate copies and whether the abundance of developmental genes in the knockout dataset had the potential to mask functional compensation in other genes. Therefore, we subdivided the developmental and non-developmental genes into duplicates and singletons (see methods in the supplementary material online). We found that the essentiality of non-developmental duplicated genes was significantly lower than that of non-developmental singletons in mouse and fly (mouse, $P = 0.00051$; fly, $P = 2.7 \times 10^{-8}$; χ^2 test; Table 1), following the trend observed in yeast and worm [4,5]. Interestingly, the essentiality of developmental duplicated genes was significantly higher than that of developmental singletons in mouse ($P = 0.0086$, χ^2 test; Table 1), and there was no difference in essentiality between developmental duplicated genes and singletons in 0.0051 fly $P = 0.98$, χ^2 test; Table 1. Thus, developmental genes are likely to be essential irrespective of gene duplication.

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 ($P = 3.0 \times 10^{-10}$, χ^2 test; Figure 1a). Furthermore, the P_E of WGD duplicated genes (45.4%) was significantly greater than SSD duplicated genes (38.1%; $P = 3.1 \times 10^{-6}$, χ^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%; $P = 0.10$, χ^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%; $P = 0.0027$, χ^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 (most similar protein sequence) 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 used in above analyses (see methods in the supplementary material online). We found that the lower the divergence from the closest paralog (i.e. lower

Comment [J1]: Author: correct?

K_A), the lower the P_E 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-developmental duplicated genes in mouse ($R = 0.90$, $P = 0.039$; Figure 1b) and fly ($R = 0.92$, $P = 0.027$; Figure 1c).

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 (21.8%; 388/1781) and developmental genes (20.0%; 372/1863) compared with the total dataset (17.9%; 906/5068; see methods in the supplementary material online). In addition, the WGD-duplicated genes and developmental genes in our dataset are significantly enriched for the functional category GO:0030528 'transcription regulator activity' (data not shown), which are likely to be dosage-balanced [27,28].

In yeast, genes duplicated by WGD are less essential than those duplicated by SSD [23,24]. The contrast with observations in mouse can be explained by the comparatively simple development process of this unicellular organism. 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.

Acknowledgements

We would like to thank Yoichiro Nakatani for supplying lists of the WGD duplicated genes and all the members of the McLysaght laboratory for helpful discussions. This work is supported by Science Foundation Ireland.

References

1 Ohno, S. (1970) Evolution by gene duplication. *Springer-Verlag*

2 Long, M. *et al.* (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875

3 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404

4 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66

5 Conant, G.C. and Wagner, A. (2004) Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans. Proc Biol Sci* 271, 89–96

6 DeLuna, A. *et al.* (2008) Exposing the fitness contribution of duplicated genes. *Nat. Genet.* 40, 676–681

7 Musso, G. *et al.* (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18, 1092–1099

8 Liang, H. and Li, W.H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* 23, 375–378

9 Liao, B.Y. and Zhang, J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.* 23, 378–381

10 Wilson, L. *et al.* (2005) Random mutagenesis of proximal mouse chromosome 5 uncovers predominantly embryonic lethal mutations. *Genome Res.* 15, 1095–1105

11 Roux, J. and Robinson-Rechavi, M. (2008) Developmental constraints on vertebrate genome evolution. *PLoS Genet.* 4, e1000311

12 McLysaght, A. *et al.* (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31, 200–204

13 Hokamp, K. *et al.* (2003) The 2R hypothesis and the human genome sequence. *J. Struct. Funct. Genomics* 3, 95–110

14 Panopoulou, G. *et al.* (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* 13, 1056–1066

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

land vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1638–1643

16 Dehal, P. and Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3, e314

17 Nakatani, Y. *et al.* (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17, 1254–1265

18 Putnam, N.H. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453, 1064–1071

19 Blomme, T. *et al.* (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7, R43

20 Brunet, F.G. *et al.* (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* 23, 1808–1816

21 Hufton, A.L. *et al.* (2008) Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res.* 18, 1582–1591

22 Maere, S. *et al.* (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5454–5459

23 Guan, Y. *et al.* (2007) Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics* 175, 933–943

24 Hakes, L. *et al.* (2007) All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* 8, R209

25 Su, Z. and Gu, X. (2008) Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. *J Mol Evol* (in press)

26 Veitia, R.A. (2002) Exploring the etiology of haploinsufficiency. *Bioessays* 24, 175–184

27 Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197

28 Wapinski, I. *et al.* (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449, 54–61

29 Nelson, C.E. *et al.* (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* 5, R25

30 Freeling, M. and Thomas, B.C. (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16, 805–814

31 Otto, S.P. (2007) The evolutionary consequences of polyploidy. *Cell* 131, 452–462

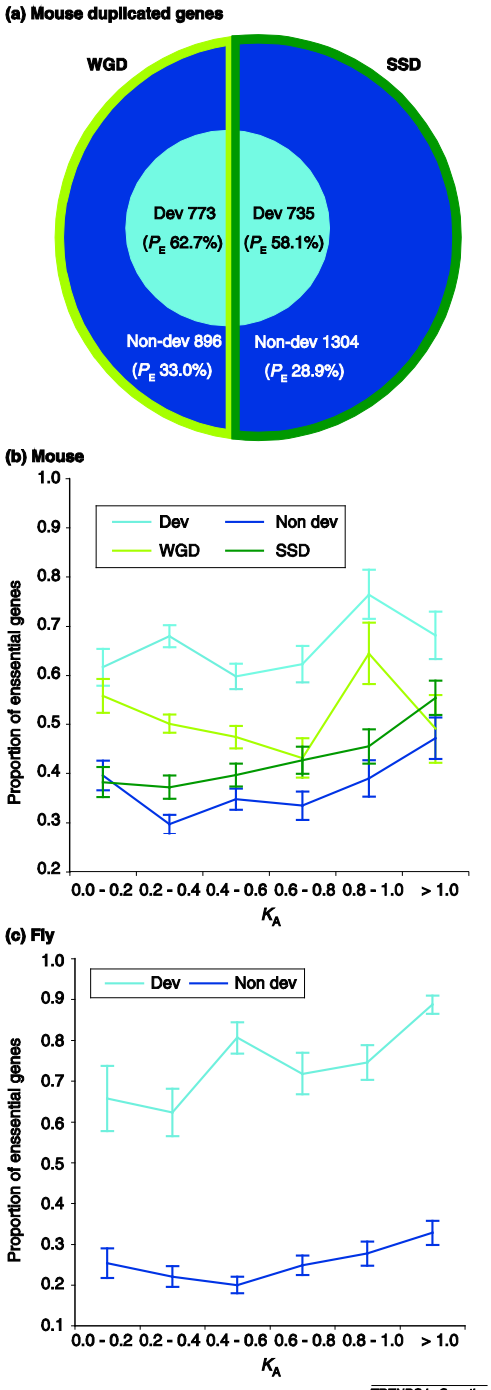
(a) Mouse duplicated genes
WGD
SSD
Dev 773
(P_E 62.7%)
Dev 735
(P_E 58.1%)
Non-dev 896
(P_E 33.0%)
Non-dev 1304
(P_E 28.9%)
(b) Mouse
Dev
Non dev
WGD
SSD
Proportion of enssential genes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0 - 0.2
0.2 - 0.4
0.4 - 0.6
0.6 - 0.8
0.8 - 1.0
> 1.0
K_A
(c) Fly
Dev
Non dev
Proportion of enssential genes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0 - 0.2
0.2 - 0.4
0.4 - 0.6
0.6 - 0.8
0.8 - 1.0
> 1.0
K_A

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

Species		Developmental genes	Non-developmental genes	Total
Mouse	Singletons	52.7% (187/355)	38.5% (210/546)	44.1% (397/901)
	Duplicated genes	60.5% (912/1508)	30.6% (673/2200)	42.7% (1585/3708)
	Total	59.0% (1099/1863)	32.2% (883/2746)	43.0% (1982/4609)
Fly	Singletons	79.1% (474/599)	34.3% (522/1520)	47.0% (996/2119)
	Duplicated genes	78.9% (607/769)	25.6% (487/1905)	41.1% (1094/2674)
	Total	79.0% (1081/1368)	29.5% (1009/3425)	43.6% (2090/4793)