7 Chisholm, A.D. and Horvitz, H.R. (1995) Patterning of the *Caenorhabditis elegans* head region by the Pax-6 family member *vab-3*. *Nature* 377, 52–55
8 Zhang, Y. and Emmons, S.W. (1995) Specification of sense-organ identity by a *Caenorhabditis elegans* Pax-6 homologue. *Nature* 377, 55–59
9 Adachi, Y. *et al.* (2003) Conserved cis-regulatory modules mediate complex neural expression patterns of the eyeless gene in the *Drosophila* brain. *Mech. Dev.* 120, 1113–1126
10 Hauck, B. *et al.* (1999) Functional analysis of an eye specific enhancer of the eyeless gene in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 564–569
11 Xu, P.X. *et al.* (1999) Regulation of *Pax6* expression is conserved between mice and flies. *Development* 126, 383–395
12 Kammandel, B. *et al.* (1999) Distinct *cis*-essential modules direct the time-space pattern of the *Pax6* gene activity. *Dev. Biol.* 205, 79–97
13 Anderson, T.R. *et al.* (2002) Differential *Pax6* promoter activity and transcript expression during forebrain development. *Mech. Dev.* 114, 171–175
14 Griffin, C. *et al.* (2002) New 3′ elements control *Pax6* expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.* 112, 89–100
15 Williams, S.C. (1998) A highly conserved lens transcriptional control element from the *Pax-6* gene. *Mech. Dev.* 73, 225–229
16 Chow, R.L. *et al.* (1999) *Pax6* induces ectopic eyes in a vertebrate. *Development* 126, 4213–4222
17 Cai, L. *et al.* (2000) Misexpression of basic helix–loop–helix genes in the murine cerebral cortex affects cell fate choices and neuronal survival. *Development* 127, 3021–3030
18 Scardigli, R. *et al.* (2001) Crossregulation between *Neurogenin2* and pathways specifying neuronal identity in the spinal cord. *Neuron* 31, 203–217
19 Kageyama, R. *et al.* (1997) bHLH transcription factors and mammalian neuronal differentiation. *Int. J. Biochem. Cell Biol.* 29, 1389–1399
20 Kageyama, R. *et al.* (2000) The bHLH gene *Hes1* regulates differentiation of multiple cell types. *Mol. Cells* 10, 1–7
21 Ross, S.E. *et al.* (2003) Basic helix–loop–helix factors in cortical development. *Neuron* 39, 13–25
22 Xu, Z.P. *et al.* (2002) Functional and structural characterization of the human gene *BHLHB5*, encoding a basic helix–loop–helix transcription factor. *Genomics* 80, 311–318
23 Morgan, R. (2002) The circadian gene clock is required for the correct early expression of the head specific gene *Otx2*. *Int. J. Dev. Biol.* 46, 999–1004
24 Marsich, E. *et al.* (2003) The *PAX6* gene is activated by the basic helix–loop–helix transcription factor NeuroD/BETA2. *Biochem. J.* 376, 707–715

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.04.009

A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications

Fyodor A. Kondrashov[1] and Eugene V. Koonin[2]

[1]Section of Evolution and Ecology, University of California at Davis, Davis, CA 95616, USA
[2]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20892, USA

The dominance of wild-type alleles and the concomitant recessivity of deleterious mutant alleles might have evolved by natural selection or could be a by-product of the molecular and physiological mechanisms of gene action. We compared the properties of human haplosufficient genes, whose wild-type alleles are dominant over loss-of-function alleles, with haploinsufficient (recessive wild-type) genes, which produce an abnormal phenotype when heterozygous for a loss-of-function allele. The fraction of haplosufficient genes is the highest among the genes that encode enzymes, which is best compatible with the physiological theory. Haploinsufficient genes, on average, have more paralogs than haplosufficient genes, supporting the idea that gene dosage could be important for the initial fixation of duplications. Thus, haplo(in)sufficiency of a gene and its propensity for duplication might have a common evolutionary basis.

The contributions of the individual alleles to the phenotype are often non-additive. Ever since Mendel's experiments, it has been recognized that, at many loci, wild-type alleles are dominant and mutant alleles are recessive. Fisher argued that dominance of wild-type alleles evolved by natural selection because dominance shields heterozygous organisms from the adverse effects of deleterious alleles [1,2]. This concept has been criticized by Wright [3] who noted that selection favoring modifiers of dominance would be weak and unable to overcome genetic drift. Wright suggested that dominance of wild-type alleles is not an adaptation but rather a by-product of the 'physiology of the organism' and that, in dosage-sensitive genes, wild-type alleles should be recessive.

A key prediction of Fisher's theory, which states that dominance of wild-type alleles should be rare in a haploid organism artificially induced to be diploid, failed to come true [4]. By contrast, the physiological theory of dominance was supported by a theoretical analysis of metabolic pathways that showed that dominance of wild-type alleles of enzymes could be a simple consequence of flux functions [5]. Consequently, the physiological theory [3,5] is currently the preferred explanation for dominance. According to this theory, if a gene is part of a multi-step pathway, the phenotype associated with mutation of this gene should be insensitive to the gene dosage [3]. Enzymes are thought to

Corresponding author: Fyodor A. Kondrashov (kondrashov@ucdavis.edu).
Available online 19 May 2004

Box 1. The physiological theory of dominance and the relevant terminology

Genes for which wild-type alleles are dominant over loss-of function alleles are haplosufficient. Genes for which loss-of-function alleles strongly affect the phenotype of heterozygotes are haploinsufficient [9]. These terms are less prone to confusion than 'dominant' and 'recessive' because dominance of one allele implies recessivity of the other. Indeed, in human genetics, dominance is described from the point of view of abnormal alleles, which are recessive in haplosufficient genes. By contrast, the evolutionary literature usually describes dominance from the point of view of wild-type alleles, which are dominant in haplosufficient genes. To avoid ambiguity, we use terms haplosufficient and haploinsufficient, however, a change-of-function mutant allele can be dominant over the wild-type allele in a haplosufficient gene.

According to the physiological theory of dominance, loss-of-function alleles are not notably manifest in heterozygotes if the phenotype is a diminishing returns function of gene dosage (Figure I; [3]). Wright has shown that such functions are expected for genes that act in multistep pathways, and Kacser and Burns [5] have shown that, when the intermediates are not saturated, the pathway flux is a diminishing returns function of the concentrations of individual enzymes.

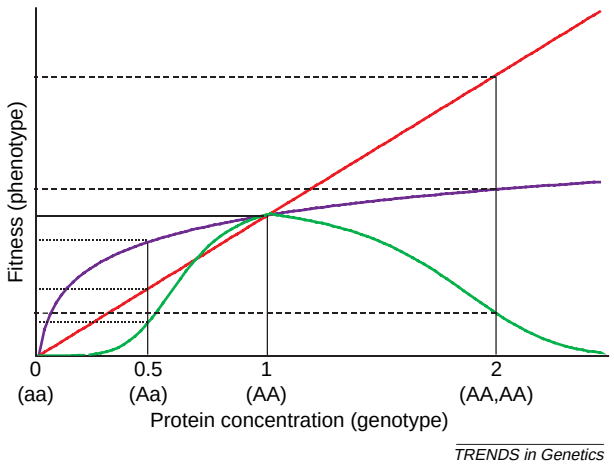

Figure I. The relationship between gene dosage and phenotype under the linear and diminishing returns functions. The difference in phenotype between homozygous wild-type (AA) genotype and the heterozygote or halving protein dosage is small for genes with a diminishing returns relationship between dosage and phenotype (purple line); these are expected to be mostly genes encoding enzymes. By contrast, a decrease of dosage for a protein with a linear relationship between dosage and phenotype has a major effect on the phenotype (red line); these are predicted to be genes encoding structural and regulatory proteins. Similarly, an increase of gene dosage that can be caused by gene duplication (AA, AA) contributes to a change in phenotype for genes with the linear genotype–phenotype relationship but not for genes with a diminishing returns function. Some genes, particularly those that encode the subunits of protein complexes, can show a decrease in fitness for both an increase and decrease of protein dosage (green line). Protein concentration as a comparative measure between different genotypes such that protein concentration of 1 was arbitrarily assigned to homozygous wild-type, single gene copy genotype. The uniformly broken lines show the fitness of a duplicated gene for various fitness functions and the non-uniformly broken lines trace the fitness of heterozygous, unduplicated genotypes. This is a modified version of Figure 7 in Ref. [3].

be particularly dosage-insensitive ([4] but see [6–7]), whereas genes encoding proteins with structural, regulatory, mechanochemical and other non-enzymatic function are more likely to be dosage-sensitive [8–10]. Recently, Hurst and coworkers showed that yeast proteins, which are subunits of multiprotein complexes and whose

Box 2. Materials and methods

We analyzed human genes responsible for mendelian diseases, for which either dominant or recessive, abnormal alleles were listed in the OMIM database (http://www.ncbi.nlm.nih.gov/omim). Haploinsufficient and haplosufficient genes were extracted from the OMIM database with the search terms 'dominant' and 'recessive', respectively, using the Entrez retrieval system (http://www.ncbi.nlm.nih.gov/Entrez/) [22]. Only genes for which the mode of inheritance is considered proven by OMIM were used; this was done by selecting the appropriate conditions in the limits options section of Entrez. Using OMIM identifiers for each gene, gene names were obtained from the complete list of genes in OMIM. Genes that were retrieved as both haploinsufficient and haplosufficient were discarded. The number of paralogs for each gene was determined by clustering all annotated human protein sequences using the blastclust program (ftp://ftp.ncbi.nih.gov/blast/executables/), with the cutoff identity values of 50–90%. The gene functions annotated by the GO annotation system [13] were obtained from SpTrembl [23] and the specific annotations in SpTrembl for each protein were redirected to a more general functional characteristic using the GO annotation tree. We used previously published information on haplosufficiency and haploinsufficiency in *S. cerevisiae* [13] and obtained the GO functional annotation for *S. cerevisiae* genes from the complete genome [24] flatfiles at the National Center for Biotechnology Information.

dosage therefore appears to be tightly regulated [10], tend to be encoded by haploinsufficient (wild-type recessive) genes (Box 1), a result that is best compatible with the physiological theory of dominance [11].

In this article, we test a central prediction of the physiological theory, which states that genes encoding proteins whose functions tend to be protein-dosage insensitive typically should be haplosufficient (dominant wild-type). We validate this prediction by showing that haplosufficient genes (i.e. genes that have dominant wild type alleles) encode enzymes significantly more often than haploinsufficient genes. We also demonstrate that haploinsufficient genes, on average, have more paralogs than haplosufficient ones, probably because the initial fixation of duplications depends on gene dosage effects.

Functional repertoire of haplo(in)sufficient proteins

We compared the functional classification of 685 genes that cause dominant genetic disorders (i.e. haploinsufficient genes) with 422 haplosufficient genes that were responsible for human mendelian diseases using the information extracted from the OMIM database (http://www.ncbi.nlm.nih.gov/omim) and the gene ontology (GO) annotation system (http://www.geneontology.org) (Box 2) [12]. As predicted by Kacser and Burns [5], the proportion of haplosufficient genes is the highest among genes that encode proteins with enzymatic functions (Table 1). By contrast, haploinsufficient genes preferentially encode regulatory and structural proteins, transcription regulators, proteins involved in signal transduction and proteins with various binding function (Table 1). For the abundant functional categories, such as enzymes, binding proteins and transcription regulators, the differences between the fractions of haplosufficient and haploinsufficient genes are highly significant

Table 1. Major functional categories in human and yeast haploinsufficient and haplosufficient genes[a]

Category	Haplosufficient[b]	Haploinsufficient[b]	*P* value[c]
Human			
Binding activity (GO: 0005488)	161 genes (45.1%)	393 genes (64.7%)	< 0.001
Cell adhesion molecule activity (GO: 0005194)	14 genes (3.9%)	16 genes (2.6%)	N.S.
Defense and/or immunity protein activity (GO: 0003793)	14 genes (3.9%)	11 genes (1.8%)	N.S.
Enzyme activity (GO: 0003824)	197 genes (55.2%)	195 genes (32.1%)	< 0.001
Enzyme regulator activity (GO: 0030234)	12 genes (3.4%)	44 genes (7.2%)	< 0.025
Signal transducer activity (GO: 0004871)	54 genes (15.1%)	166 genes (27.3%)	< 0.001
Structural molecule activity (GO: 0005198)	19 genes (5.3%)	63 genes (10.4%)	< 0.025
Transcription regulator activity (GO: 0030528)	18 genes (5.0%)	94 genes (15.5%)	< 0.001
Transporter activity (GO: 0005215)	61 genes (17.1%)	89 genes (14.7%)	N.S.
Yeast			
Binding activity (GO: 0005488)	549 genes (29.0%)	147 genes (28.8%)	N.S.
Chaperone activity (GO: 0003754)	38 genes (2.0%)	18 genes (3.5%)	0.049
Enzyme activity (GO: 0003824)	1042 genes (55.1%)	203 genes (39.8%)	< 0.001
Enzyme regulator activity (GO: 0030234)	77 genes (4.1%)	11 genes (2.2%)	0.048
Signal transducer activity (GO: 0004871)	81 genes (4.3%)	6 genes (1.2%)	< 0.01
Structural molecule activity (GO: 0005198)	89 genes (4.7%)	138 genes (27.1%)	< 0.001
Transcription regulator activity (GO: 0030528)	175 genes (9.2%)	38 genes (7.5%)	N.S.
Transporter activity (GO: 0005215)	252 genes (13.3%)	59 genes (11.6%)	N.S.

[a]Abbreviation: N.S., not significant.
[b]The percentages of haplosufficient or haploinsufficient genes that belong to the corresponding functional category; only genes that have been functionally characterized by gene ontology (GO) annotation system were included. The number of haplosufficient and haploinsufficient genes for human was 357 and 607, respectively. The number of haplosufficient and haploinsufficient genes for yeast was 1892 and 510, respectively.
[c]The *P*-values of the difference between the fractions of haplosufficient and haploinsufficient genes for a given functional category according to the chi-squared test.

statistically (Table 1). A similar pattern was observed in the comparison of the functional classification of 510 haploinsufficient genes with 1892 haplosufficient genes in the yeast *Saccharomyces cerevisiae*; haplosufficient genes were more likely to encode enzymes, whereas genes encoding structural proteins were more likely to be haploinsufficient (Table 1). In contrast to the human data, yeast genes with regulatory and binding functions did not appear to be more common among haploinsufficient genes.

These results should be interpreted with some caution because not all genes that cause dominant disorders are necessarily haploinsufficient [9]. The data on haploinsufficiency in *S. cerevisiae* also might be biased towards revealing haploinsufficient genes with enzymatic functions because the original study concentrated on the identification of fermentation-related phenotypes [13]. These caveats notwithstanding, the present results expand the previous observations showing that enzymes are found commonly among haplosufficient genes, whereas transcription factors are more common among haploinsufficient genes [14].

We then compared the sizes of paralogous gene families of haplosufficient with haploinsufficient genes after partitioning them into functional categories and found that, in all functional categories, haploinsufficient genes had substantially more paralogs in the human genome than happlosufficient genes (Table 2). This observation could be explained by a series of ancient polyploidisation events after which haplosufficient gene copies were lost preferentially, whereas the haploinsufficient genes were retained [11]. However, this seems unlikely because haploinsufficient genes were found to have more paralogs than haplosufficient genes regardless of the similarity threshold used for delineation of paralogous families, which should roughly reflect the time of duplication such that more similar paralogs share a more-recent common ancestor (Figure 1). Thus, it appears that copies of haploinsufficient genes are more likely to be fixed following duplication than copies of haplosufficient genes.

Dosage, dominance and duplication

Fisher and Haldane proposed that gene duplications might act as dominance modifiers such that the extra gene copy shields the original one from new mutations [15,16]. Thus, Fisher's theory of dominance implies that haplosufficient genes should form larger paralogous families than haploinsufficient genes. Theories of gene duplication that assume no fitness difference between genomes with different copy number [17,18] predict an equal rate of gene duplication for all genes. By contrast, Wright's physiological theory [3] predicts that haploinsufficient genes should have more paralogs than haplosufficient genes because selection for increased dosage

Table 2. The size of paralogous families for human haplosufficient and haploinsufficient genes

	Haplosufficient[a]		Haploinsufficient[a]	
	Number of genes	Mean family size	Number of genes	Mean family size
All genes	422	1.77	685	4.65
Enzymes	197	1.83	195	2.89
Binding function	161	1.91	390	3.89
Other function or unknown function	135	1.54	216	6.95

[a]The differences in family size between haplosufficient and haploinsufficient genes were statistically significant ($P < 0.01$) for all categories according to the Mann–Whitney U test.

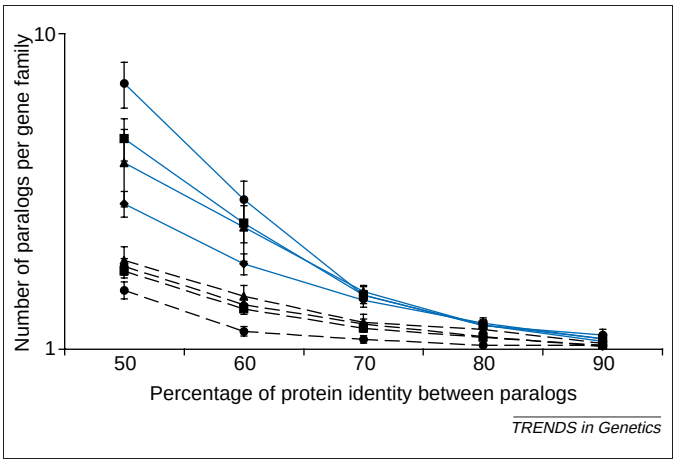

Figure 1. The average number of paralogs for haploinsufficient (solid blue lines) and haplosufficient (broken lines) genes of different functional categories. The different shapes represent different functional categories as follows: the diamonds represent enzymes; the triangles represent binding functions; the circles represent other functions; the squares represent all functions together. The bars on the data points are standard errors.

should be more effective for dosage-sensitive (haploinsufficient) genes. Clearly, the above finding is compatible with Wright's prediction. We showed previously that duplicated genes were subject to purifying selection immediately after duplication and suggested that the initial fixation of duplications was related to the advantage of increased gene dosage [19]. This conclusion is supported by the subsequent observation that recent duplications in *Arabidopsis thaliana* have substantially reduced nucleotide polymorphism, providing evidence of positive selection [20]. The present observations further support the hypothesis that many gene duplications are fixed by positive selection for increased gene dosage [19].

The unification of dominance and duplication theory

The key prediction of the physiological theory of dominance is that genes whose phenotypic effect shows strong dosage-dependence should be haploinsufficient [3,5] (Box 1). Both results reported here are compatible with this prediction but not with the predictions of Fisher's theory. First, we found that human genes encoding enzymes, which typically are required in catalytic amounts, are haplosufficient (dominant wild-type) much more often than genes that encode various structural and regulatory proteins. The difference between these functional categories of proteins, in terms of haplosufficiency, was highly statistically significant but far from all-or-none. However, this is not surprising because some enzymes, particularly those with a low-turnover number, could be associated with a dosage-dependent phenotype [6,7] and, conversely, some structural proteins and, particularly, regulators are likely to be required in low amounts and hence could be dosage-independent. Second, we observed that haploinsufficient genes, on average, belonged to significantly larger paralogous families than haplosufficient genes. The emerging chain of causation goes thus: genes for various non-enzymatic proteins often have strongly dosage-dependent phenotypes, therefore, duplication of such genes tends to be beneficial and they are usually haploinsufficient.

At first glance, the connection between duplication and haploinsufficiency in genes encoding structural and regulatory proteins might seem to be at odds with the observations of Hurst and coworkers who reported that yeast genes encoding complex subunits are most often haploinsufficient and tend not to have paralogs [11]. In a different context, our analysis supports their conclusion by showing that nearly all highly conserved, single-copy eukaryotic proteins are complex subunits [21]. However, we believe that there is no actual discrepancy between these observations because functional categories such as 'binding proteins' are extremely broad and only a small fraction of proteins in these categories are likely to be subunits of complexes with tight stoichiometry.

Taken together, these observations seem to comprise a direct, genome-wide validation of the physiological theory of dominance. Furthermore, these results strongly suggest that the proper theoretical framework for investigating the early phases of gene duplications should be similar to that of the physiological theory of dominance (i.e. based on the effects of gene dosage on fitness).

References

1 Fisher, R.A. (1928) The possible modification of the response of the wild type to recurrent mutations. *Am. Nat.* 62, 115–126
2 Fisher, R.A. (1958) *The Genetical Theory of Natural Selection*, Dover, New York
3 Wright, S. (1934) Physiological and evolutionary theories of dominance. *Am. Nat.* 68, 24–53
4 Orr, A.H. (1991) A test of Fisher's theory of dominance. *Proc. Natl. Acad. Sci. U.S.A.* 88, 11413–11415
5 Kacser, H. and Burns, J.A. (1981) The molecular basis of dominance. *Genetics* 97, 639–666
6 Cornish-Bowden, A. (1987) Dominance is not inevitable. *J. Theor. Biol.* 125, 333–338
7 Hurst, L.D. and Randerson, J.P. (2000) Dosage, deletions and dominance: simple models of the evolution of gene expression. *J. Theor. Biol.* 205, 641–647
8 Fisher, E. and Scambler, P. (1994) Human haploinsufficiency–one for sorrow, two for joy. *Nat. Genet.* 7, 5–7
9 Strachan, T. and Read, A.P. (1999) *Human Molecular Genetics 2*, BIOS Scientific Publishers
10 Veitia, R.A. (2002) Exploring the etiology of haploinsufficiency. *BioEssays* 24, 175–184
11 Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197
12 Gene Ontology Consortium, (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
13 Steinmetz, L.M. *et al.* (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.* 31, 400–404
14 Jimenez-Sanchez, G. *et al.* (2001) Human disease genes. *Nature* 409, 853–855
15 Haldane, J.B.S. (1933) The part played by recurrent mutation in evolution. *Am. Nat.* 67, 5–19
16 Fisher, R.A. (1935) The sheltering of lethals. *Am. Nat.* 69, 446–455
17 Ohno, S. (1970) *Evolution by Gene Duplication*, George Allen and Unwin, London
18 Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473
19 Kondrashov, F.A. *et al.* (2002) Selection in the evolution of gene duplications. *Genome Biol.* 3, (http://genomebiology.com/2002/3/2/research/0008)
20 Moore, R.C. and Purugganan, M.D. (2003) The early stages of

duplicate gene evolution. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15682–15687
21 Koonin, E.V. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5, R7
22 Wheeler, D.L. *et al.* (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* 32, D35–D40
23 Boeckmann, B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370
24 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274 (546), 563–567

doi:10.1016/j.tig.2004.05.001

Erratum

Erratum: Cladogenesis, coalescence and the evolution of the three domains of life☆

***Trends in Genetics* 20 (2004), 182–187**

In the article by Olga Zhaxybayeva and J. Peter Gogarten, which was published in the April issue of *TIG*, there was an error in Figure 3. The *x*-axis in the figure was incorrectly given as years. The correct scale is in time intervals chosen for the simulation, which are defined by one speciation event occurring in the 200 lineages. *TIG* apologizes to the authors and readers for this error. The doi of the original article is 10.1016/j.tig.2004.02.004.

doi:10.1016/j.tig.2004.05.006

☆ DOI of original article: 10.1016/j.tig.2004.02.004
Available online 28 May 2004