The hows and whys

Laurence D. Hurst (University of Bath, UK) **and Csaba Pál** (Institute of Biochemistry, Szeged, Hungary)

In the majority of organisms, most gene knockouts have little or no affect on viability. Does this mean that most genes are dispensable to an organism? What are the underlying mechanisms of dispensability and why did dispensability evolve?

Over the last few years, important advances have been made in addressing these issues, with analysis of dispensability in metabolic networks taking centre stage. It has indeed been established that laboratory assay of dispensability is not a measure of the importance of the gene to an organism, not only because laboratory assays are not sensitive enough, but also because under laboratory conditions many enzymes are neither needed nor expressed. Leaving this trivial but overlooked explanation aside, the role of duplicate genes and alternative routings (metabolic flux re-arrangement or use of alternative signalling pathways) have emerged as the two leading explanations for dispensability. Evidence from the workings of yeast's metabolism suggests the former to be more important than the latter. Whatever the mechanism, the great bulk of data suggests that dispensability is a side consequence of selection for some other property, such as high flux favouring retention of isoenzymes.

Knock out a gene in yeast and four-fifths of the time the cell is able to grow. In a nematode worm, you will see a phenotype on only one in every ten occasions. In *Bacillus subtilis* the figure is lower still. These observations prompt many related questions. Does the fact that a gene can be knocked out and leave no phenotype really mean that the gene is 'dispensable' to the organism? Why do species differ in the proportion of genes that appear to be essential to growth? How is it that most knockouts can grow well? Why is it that a knockout can grow well?

The latter two questions are importantly different. The 'how' question demands a mechanistic answer. For example, it could be that an enzyme is non-essential because there is a paralog in the genome coding for an isozyme. The 'why' question demands an evolutionary answer. We can ask whether the gene has a paralog because selection favoured the duplicate as it provides a back-up. In the present article, we briefly address recent advances in addressing these issues, paying particular attention to the central role of analysis of metabolism.

Key words: *enzymes, gene knockouts, isozymes, mutation rates, paralogs*

'Dispensable' genes are not really dispensable

If a gene appears to be dispensable in the laboratory, is it really dispensable for the organism? The latter would require that a deletion of the gene would not be under purifying selection and hence could spread by drift (neutral evolution) alone. There is a big gap between laboratory assessments of knockout fitness and this more pertinent evolutionary definition; first, because laboratory assays don't have the ability to measure fitness on the necessary scale and second, because they miss genes that are essential under conditions not seen in the laboratory.

To understand the magnitude of the first problem, let us suppose that the wild-type has fitness of unity and the deletion mutant has fitness 1-*s*. How small can *s* be for the deletion to be really dispensable and hence not eliminated from the population by selection? A classical result from population genetics is that for a weakly deleterious allele in a diploid to reach fixation by drift alone would require it to have a fitness effect less than $1/(2N_e)$ (where N_e is the effective population size). Even for organisms with small effective populations (e.g., humans), this means a fitness effect no greater than about 10^{-5}. For bacteria and yeast a more realistic figure might be 10^{-9}. High-throughput laboratory-based fitness measures, even if they provide quantitative assessment of fitness, rather than essential/non-essential calls, are simply not sensitive enough to detect such effects, the current limit[1,2] being around s=0.01. Although genes that are essential in the laboratory are likely to be indispensable, it would be an error to suppose that viability of a knockout strain in the laboratory equates to evolutionary dispensability. This conclusion is supported by analyses of rates of evolution of 'dispensable' proteins. From the very first[3], all analyses of knockout dispensable/nonessential genes suggest that most proteins evolve much slower than expected given the background rate (that expected for neutrally evolving sequences), supporting the thesis that they are doing something useful for the organism.

How important is the second issue, the lack of relevance of laboratory conditions? Both *in silico* and empirical analysis of yeast's metabolism suggest it to be the most important explanation for apparent dispensability. We, for example, addressed the issue using an *in silico* genome-scale metabolic network

# of dispensability

model of baker's yeast (*Saccharomyces cerevisiae*)[4]. With a network of 809 metabolites connected by 851 different biochemical reactions we defined a solution where fluxes of all metabolic reactions in the network satisfy the relevant constraints, given the nutrients available in the environment. Next, we used various mathematical protocols to find the optimal use of the metabolic network to produce major biosynthetic components for growth.

With such a network, we then asked whether a given gene appears to be non-essential simply because the enzyme the gene encodes isn't doing anything under laboratory conditions. The model indicates that condition-specificity explains the majority of apparent dispensability. Similarly, analysis of the growth-phenotypes of *Escherichia coli* mutants showed that most genes have severe fitness defects only under a small fraction (10%) of the 282 different growth conditions investigated.

Direct measurements of metabolic flux in yeast support this conclusion[5]. Indeed about 50 percent of apparently dispensable genes are simply inactive under laboratory conditions. More generally, a recent large-scale study tested the performance of nearly 5000 apparently viable single knock-out yeast strains in the presence of over 1000 different chemical and environmental stress conditions[6]. Nearly all (97%) of the genes exhibited low growth under at least one condition, and deleterious phenotypes are generally restricted to a small fraction of the environments tested. It is an open question as to whether multicellular organisms have high rates of dispensaibility for similar reasons.

Mechanisms of dispensability

Aside from genes not expressed in the laboratory, there remains a set of genes that are not vital for growth in the laboratory, yet are still being expressed and doing something. To see the alternative explanations for such dispensability consider a lift as being like a cellular function. We use the lift to get from one floor of a building to another. What if we cut one of the wires holding the lift – the knockout? If there is a duplicate wire to take the strain, the lift can still work. This would be like an isoenzyme. Alternatively, we could take the stairs and still get to floor we wanted to go to. The cellular equivalent has been termed distributed robustness[7]. Alternatively, and rarely discussed,

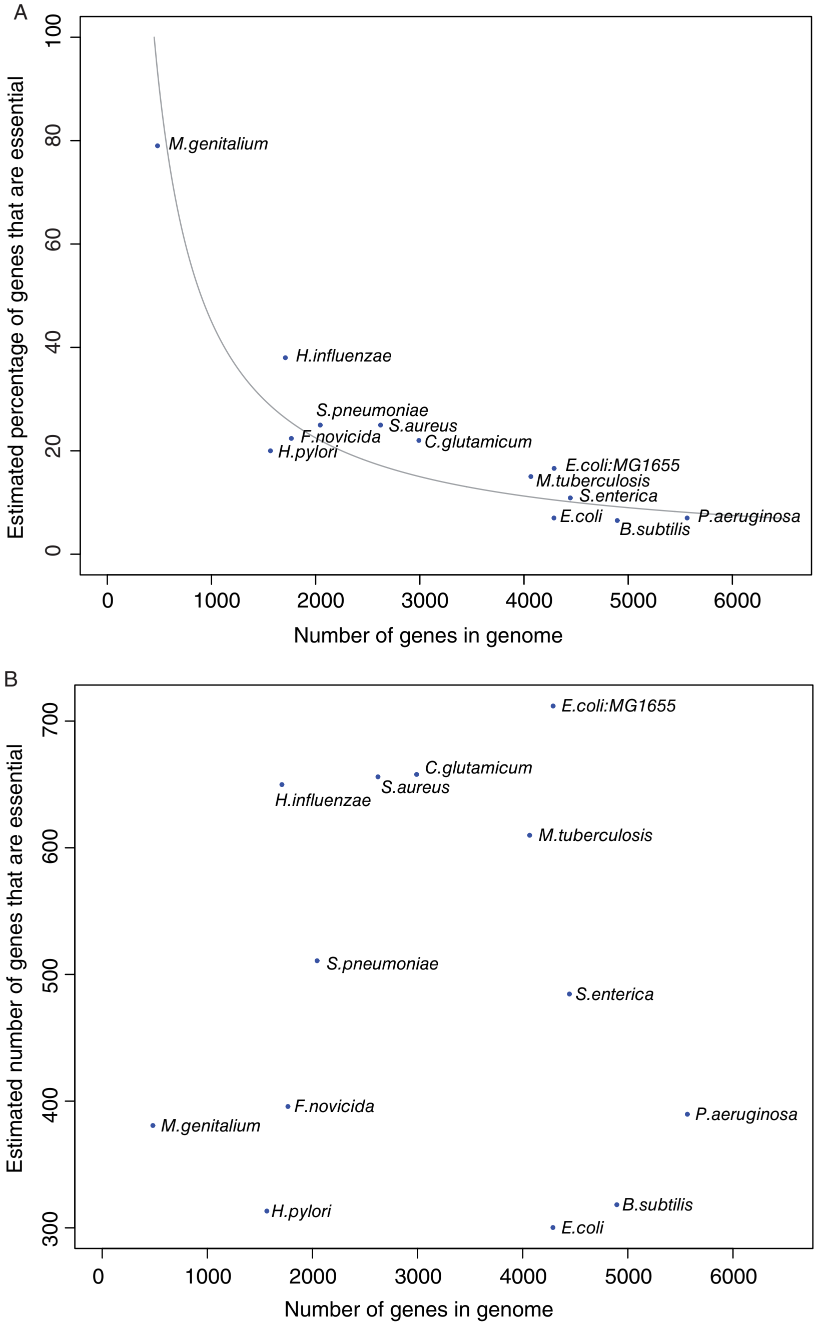


Figure 1. (A) The proportion of protein-coding genes that are essential in a bacterial genome as a function of the number of genes in the genome. The grey curve indicates the percentage expected if a constant 450 genes are essential in every genome. (B) The estimated number of genes that are essential as a function of number of genes in the genome. These estimates are derived from estimates of the proportion of genes that are essential and the number of protein-coding genes in the genome. For data sources see [17]

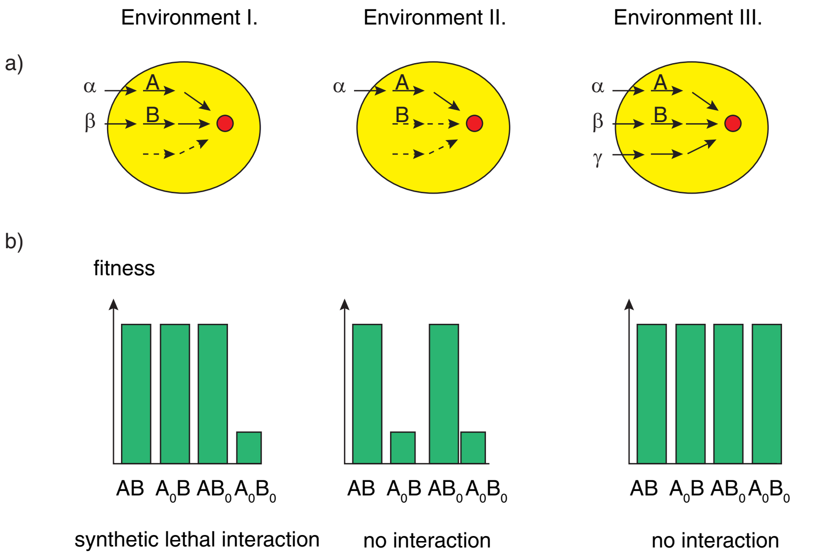

Environment I.
Environment II.
Environment III.
a)
α
β
A
B
α
A
B
α
β
γ
A
B
b)
fitness
AB A_0B AB_0 A_0B_0
AB A_0B AB_0 A_0B_0
AB A_0B AB_0 A_0B_0
synthetic lethal interaction
no interaction
no interaction

Figure 2. A model to explain plasticity of genetic interactions across the environment. Differences in the availability of external nutrients (α, β, and γ) across three hypothetical environments have a large effect on the activity of parallel metabolic pathways (see diagrams) and hence the impact of single (A_0B, AB_0) and double gene deletant (A_0B_0) genotypes (bar graphs). A key metabolite (red circle) can be synthesized through three different routes. Genes A and B, sitting on parallel pathways, show synthetic lethal interactions in environment I, where starting nutrients (α and β) of both pathways are present in the environment. However, B is unable to compensate for deletion of A in environment II, where only a is available. The double mutant (A_0B_0) is rescued in environment III, where a third starting nutrient is present (γ)

we could just fail to get to the floor we wanted to get to, but that this is of little consequence. It could just be that there are lots of things organisms do if they can, but that are of little impact if they don't. A mutation affecting mid-early development (i.e., the so called phylotypic stage in vertebrates), for example, is much more likely to be lethal than one affecting some rather precise bit of brain function. Most attention has been given to the first two explanations.

The first systematic evidence that the presence of a paralogue can be important came from studies showing that in yeast[8] and nematode worms[9], genes with a paralogue were less likely to be essential than singleton genes. While consistent with paralog based backup, the same fact could be explained by a duplication bias in favour of non-essential genes[10]. The active backup model, however, uniquely predicts that if both genes in the duplicate pair are deleted the phenotype should be more severe than expected given the knockout phenotype of the two genes singly. This appears to be the case for many gene pairs in yeast that were left over since the whole-genome duplication[1]. However, the duplication bias model was not intended to be applied to whole-genome duplications. For single gene duplicates the issue is unresolved.

Although distributed robustness is harder to demonstrate some possible cases are described. In mammals, bile acids are the degradation products of cholesterol. The same bile acids act to repress the genes for cholesterol degradation by activating at least two mutually redundant pathways that involve different transcription factors [based on activation of the xenobiotic receptor PXR (pregnane X receptor) or JNK (c-Jun N-terminal kinase)]. Some cases appear paradoxical at first sight. In *E. coli*, glucose-6-phosphate dehydrogenase loss-of-function mutants grow at near wild-type rates. How can this be when this enzyme is part of the pentose phosphate shunt, which produces two-thirds of a cell's NADPH? By contrast, most of the cell's NADH is produced by the tricarboxylic acid cycle and this holds the key. When the pentose phosphate shunt is blocked by a loss-of-function mutation, there is increased flux through the tricarboxylic acid cycle, generating more NADH. A massively increased flux through the transhydrogenase reaction permits this NADH to be converted to NADPH. The route may be substantially adjusted, but the end-product (high NADPH) is the same[7].

Distributed robustness models would predict a relationship between network architecture and dispensability. This is, for example, supported by analysis of carefully curated protein–protein interaction datasets that find that hubs do tend to be essential in yeast[11]. Moreover, those hubs that partner many other hubs are even more likely to be essential[12]. However, it is unlikely that essential genes or reactions could be defined purely by consideration of their topological position in the network alone. Dynamical metabolic networks with a strong empirical footing attempt to handle the limitations. The most successful correctly predict close to 90 percent of the knockout phenotypes of metabolic genes in yeast[13] and *E. coli* [14].

For yeast metabolism, dispensability by paralogues is more important than distributed robustness

Given two mechanisms that ensure dispensability, which is more important? To approach this we compared the proportions of experimentally verified essential genes that encode single-copy enzymes with those of duplicated isoenzymes[4]. Only genes predicted to encode essential reactions were considered. We found that very few essential enzymes are isoenzymes, consistent with previous claims that dispensability results partially from redundant gene duplicates[8]. Two exceptions were thioredoxin reductase, which has two isoenzymes (TRR1 and TRR2), and inorganic pyrophosphatase, also with two isoenzymes (IPP1, IPP2). Their failure to compensate might reflect a lack of duplicate enzyme activity in the same subcellular compartment; in both cases, one isoenzyme is cytoplasmic, while the other is mitochondrial. Overall, we estimate that duplicates account for between 14.6 and 27.8 percent of incidences of gene dispensability.

We also examined the effect of flux reorganization on *in vivo* gene dispensability. To avoid the complication of duplicate gene copies, we confined our analysis to single-copy, experimentally verified essential genes, comparing the proportions of these that encode essential versus dispensable (but non-zero flux) reactions. We hypothesized that essential and dispensable reactions should differ in the network's ability to compensate for loss. From our analysis we estimate that this mode of compensation can only explain 3.8–17% of gene dispensability.

That there are two or three times as many examples of dispensability owing to duplication as to distributed robustness agrees with ^{13}C-tracer experiments[5], in which flux is measured directly. These experiments suggest that, for 207 viable mutants of active reactions, network redundancy through duplicate genes is the major (75%) and, in alternative pathways, the minor (25%) molecular mechanism of genetic network robustness in yeast. The minor contribution of alternative pathways/distributed robustness may well be because the yeast metabolic network has difficulties tolerating extensive flux reorganization. However, it may be wise not to generalize too much from metabolic analyses, because in other systems (e.g., signalling systems) distributed robustness may be more common[7].

Evolution of dispensability: probably a side consequence

Why are some genes dispensable, others not? Could it be that the spread and retention of a duplicate (or alternative pathway) was selected because it provided backup against mutations? Or might backup simply be a possibly fortuitous side product of a duplicate (or alternative pathway) that was retained for other reasons?

One way to approach this problem is to ask about the conditions under which we expect selection to favour the evolution of dispensability and then ask whether such models explain between-organism variation in redundancy levels. The mutation rate is a key parameter. If deleterious mutations are rare, then compensatory (backup) mutations are unlikely to evolve. This is supported by computer experiments. Wilke et al.[15] examined the evolution of populations of digital organisms exposed to high mutation rates. As the mutation rate was increased, competition favours the genotype with the lower replication rate. These slow dividers, although at lower fitness peaks, were also located in flatter regions of the fitness surface so that each mutation had less impact on fitness. It is notable that one of the few cases of empirical evidence for the evolution of robustness, at least to point mutations, comes from an organism with a very high mutation rate. Wagner and Stadler[16], examining RNA viruses, looked at predicted RNA secondary structures to identify those that were conserved and those that were not. They then asked about the fate of point mutations in the two classes of RNA structure. Strikingly, they report that conserved structures are more robust to mutation than non-conserved ones.

If most organisms don't have high enough mutation rates, how then do we account for dispensability and variation between organisms? As regards the latter there is one strikingly good correlate[17]: a higher proportion of essential genes is seen in prokaryotic genomes with fewer genes (Figure 1A). Consequently, the absolute number of genes thought to be essential in different bacteria is relatively invariant, averaging about 450–500 genes (see Figure 1B).

The most obvious explanation for this result is that as genomes shrink they are less likely to lose essential genes and thus the genomes become enriched in such genes[18]. Equally, a higher number of duplicate genes in larger genomes may provide a parallel explanation. However, the rarity of dispensable genes in the small genomes of *Mycoplasma* species may not only reflect the rarity of duplicates in small genomes, but also a higher incidence in larger genomes of genes required only under specific environmental conditions.

This analysis still leaves open the question of why duplicates are in the genome. Are they there to provide dispensability or for other reasons? Population genetical models indicate that is much more likely that a duplication spreads in a population because it provides a direct advantage, rather than because it confers buffering, again, in no small part because the mutation rate is so low[19]. This conclusion is supported by empirical evidence. Flux balance analysis of the yeast metabolic network has shown that essential reactions are not more likely than nonessential reactions to be catalyzed by isoenzymes[4]. Instead, isoenzymes appear at positions in the network where a high flux is needed. This suggests that duplicates were retained to permit a selectively advantageous increase in flux rates, a secondary consequence of which can be buffering. Detailed analysis of the role of duplicates in yeast's glycolysis supports this notion[20].

There have been relatively few attempts to consider how networks might evolve to permit distributed robustness. *A priori*, it is unlikely that distributed robustness is itself selected as it is hard to see how it can evolve in small steps: the stairs cannot function as an alternative to the lift if the staircase only goes part way. An analysis of hub architecture in metabolic networks[21] supports the alternative side-consequence model. In these simulations of the evolution of metabolic networks, hubs and scale-free structures emerged simply under selection for enhanced growth rate rather than for enhanced robustness against mutations.

Another way to ask about the evolution of distributed robustness in networks is to ask about the evolution of synthetic lethal pairs that are not sequence-related. Consistent with earlier findings showing that genetic interactions depend on environmental conditions, at least 51 percent of synthetic lethal interactions are restricted to particular environmental conditions[22]. These results are compatible with a side effect model, where the enzymes are essential under nutrient-specialist conditions (Figure 2), not because they provide back-up under nutrient-diverse conditions.

The idea that gene dispensability is not a directly selected trait is supported further by the failure to detect the evolution of buffering in laboratory studies. Notably, Elena and Lenski[23] examined interactions between random insertion mutations and genetic background in *E. coli*. Each of the mutations was transduced into two genetic backgrounds, one ancestral and the other having evolved in, and adapted to, a laboratory environment for 10,000 generations. Importantly, the mutations were no less harmful in the derived background, suggesting that the derived bacteria had not evolved buffering mechanisms against the harmful effects of mutations. ■

References

1 DeLuna, A., Vetsigian K., Shoresh N. et al. (2008) Nature Genet. **40**, 676-681

2 Breslow, D.K., Cameron D.M., Collins S.R. et al. (2008) Nature Methods **5**, 711-718

3 Hurst, L.D. and Smith N.G.C. (1999) Curr. Biol. **9**, 747-750

4 Papp, B., Pal C., and Hurst L.D. (2004) Nature **429**, 661-664

5 Blank, L.M., Kuepfer L., and Sauer U. (2005) Genome Biol. **6**, R49

6 Hillenmeyer, M.E., Fung E., Wildenhain J. et al. (2008) Science **320**, 362-365

7 Wagner, A. (2005) Bioessays **27**, 176-188

8 Gu, Z.L., Steinmetz L.M., Gu X. et al. (2003) Nature **421**, 63-66

9 Conant, G.C. and Wagner A. (2004) Proc. R. Soc. Lond. Ser. B-Biol. Sci. **271**, 89-96

10 He, X. and Zhang J. (2006) Mol. Biol. Evol. **23**, 144-151

11 Batada, N.N., Hurst L.D., and Tyers M. (2006) Plos Computational Biology **2**, 748-756

12 Batada, N.N., Reguly T., Breitkreutz A. et al. (2006) PLoS. Biol. **4**, 1720-1731

13 Forster, J., Famili I., Palsson B.O. and Nielsen, J. (2003) Omics **7**, 193-202

14 Edwards, J.S. and Palsson B.O. (2000) Proc. Natl. Acad. Sci. U. S. A. **97**, 5528-5533

15 Wilke, C.O., Wang J.L., Ofria C., Lenski, R.E. and Adami, C. (2001) Nature **412**, 331-333

16 Wagner, A. and Stadler P.F. (1999) J. Exp. Zool. **285**, 119-127

17 Hurst, L.D. and Pal C., in Evolutionary Genomics and Proteomics, edited by M. Pagel and A Pomiankowski (Sinauer Associates, Sunderland, Mass, 2008), pp. 141-167.

18 Pal, C., Papp B., Lercher M.J., Csermely, P., Oliver, S.G. and Hurst, L.D. (2006) Nature **440**, 667-670

19 Clark, A.G. (1994) Proc. Natl. Acad. Sci. U.S.A. **91**, 2950-2954

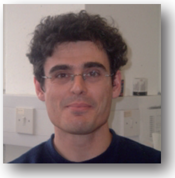
20 Conant, G.C. and Wolfe K.H. (2007) Molecular Systems Biology **3**

21 Pfeiffer, T., Soyer O.S., and Bonhoeffer S. (2005) PLoS. Biol. **3**

22 Harrison, R., Papp B., Pal C., Oliver, S.G. and Delneri, D. (2007) Proc. Natl. Acad. Sci. U.S.A. **104**, 2307-2312

23 Elena, S.F. and Lenski R.E. (2001) Evolution **55**, 1746-1752

Laurence Hurst did his first degree in Zoology at Cambridge University and his D.Phil in evolutionary genetics at Oxford University. He was subsequently a Royal Society University Research Fellow in the Department of Genetics, Cambridge. For the last 10 years he has been the Professor of Evolutionary Genetics at the University of Bath. He is a Royal Society Wolfson Research Merit Award Holder and was elected an EMBO member in 2004. email: l.d.hurst@bath.ac.uk



Csaba Pál earned his Ph.D. degree in Theoretical Biology from Eotvos Lorand University (Budapest, Hungary) in 2002. He worked as a research fellow at the University of Bath, European Molecular Biology Laboratory (EMBL, Heidelberg), and University of Oxford. He currently is a group leader at the Biological Research Center (Szeged, Hungary). His main focus is on evolutionary systems biology using theoretical modeling, comparative genomics, and microbial evolutionary experiments. email:cpal@brc.hu