The Role of Translocation and Selection in the Emergence of Genetic Clusters and Modules

David Newth*
CSIRO Centre for Complex Systems Science
CSIRO Marine and Atmospheric Research
GPO Box 284
Canberra, ACT 2601
Australia
david.newth@csiro.au

David G. Green
Faculty of Information Technology
Monash University
Clayton, VIC 3800
Australia
david.green@infotech.monash.edu.au

Abstract Biomolecular studies point increasingly to the importance of modularity in the organization of the genome. Processes such as the maintenance of metabolism are controlled by suites of genes that act as distinct, self-contained units, or *modules*. One effect is to promote stability of inherited characters. Despite the obvious importance of genetic modules, the mechanisms by which they form and persist are not understood. One clue is that functionally related genes tend to cluster together. Here we show that genetic translocation, recombination, and natural selection play a central role in this process. We distill the question of emerging genetic modularity into three simulation experiments that show: (1) a tendency, under natural selection, for essential genes to co-locate on the same chromosome and to settle in fixed loci; (2) that genes associated with a particular function tend to form functional clusters; and (3) that genes within a functional cluster tend to become arranged in transcription order. The results also imply that high proportions of junk DNA are essential to the process.

Keywords
Genetic modularity, self-organization, feedback, translocation, clustering

1 Introduction

Living systems are highly modular in nature. Plant growth, for instance, consists of repeatedly adding modules in the form of branches, buds, and leaves. Modularity also plays an important part in the organization of the genome. Genetic modules are independent, self-contained suites of genes that act together to achieve some function (Figure 1a). For example, Halder and colleagues [13] were able to produce extra eyes on the antennae, wings, and legs of *Drosophila* by targeting non-expressed instances of the *Drosophila* eyeless gene. This result provides two key insights. First, the eyeless gene acts as a switch for initiating the formation of eyes in *Drosophila*. Second, the genes that code for the eye act as a single self-contained module [12].

Functional modularity in the makeup of an organism appears to be reflected in the organization of the genes that are involved in growth and development [14, 15]. There is considerable evidence that functionally related genes group together within the genome [3, 7, 9, 21]; this suggests a potential mechanism for controlling gene expression [8, 10]. In the lambda phage [6], genes are closely located and even appear in the order they are transcribed (Figure 1b).

Likewise, evidence of gene clustering in eukaryotes is increasing. A study by Boutanaev et al. [4] found substantial numbers of gene clusters in the genome of *Drosophila*. Analysis of the *C. elegans*

* To whom correspondence should be addressed.

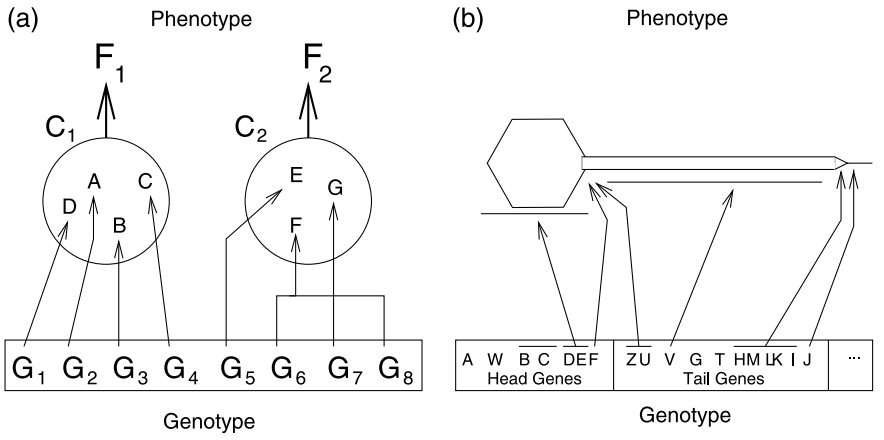

Figure 1. Examples of genetic modularity. (a) A modular representation of the character complexes $C_1 = \{A, B, C, D\}$ and $C_2 = \{E, F, G\}$, which produce the functionality F_1 and F_2. In natural systems, genes can belong to a number of different modules, and modules themselves belong to hierarchies of behaviors, and often overlap. (b) A map of the left arm of the lambda phage chromosomes, schematically showing the order of the morphogenetic genes and the points of action of those genes [3]. The genes for the formation of the left arm of the phage are co-located together. Moreover, they are arranged in the same order in which they are transcribed during development to form the morphology.

genome [16] found evidence for large-scale clustering. There is evidence from the human genome too. Lercher et al. [18] found that the human genome contains clusters of housekeeping genes, and a study of the human transcriptome map by Caron et al. [5] found that highly expressed genes tend to cluster in specific chromosomal regions.

Despite growing evidence of the prevalence and importance of genetic modules, the way they form has remained an open question. To understand genetic modularity, a fundamental question is how the relationship arises between function and location [10, 12, 22–25]. We propose that mechanisms leading to the co-location and clustering of genes provide the key, especially transposable genetic elements [17, 19, 20] and exon shuffling [11]. Lawrence and Roth [17] suggested that horizontal gene transfer drives the evolution of gene clusters.

We have tested the above hypothesis via a series of simulation experiments that address three aspects of modularity. We frame these aspects as the following hypotheses about the combined effects of gene movement and natural selection:

1. that genes essential to the survival of an organism will tend to migrate to a single chromosome and to settle at a fixed locus;
2. that functionally related genes (genes involved in the genetic process) will tend to cluster together in regions where they are active together; and
3. that genes that need to be active in a definite sequence will tend to arrange themselves in transcription order.

2 Methods

The study consists of three simulation experiments to test the above hypotheses in sequence.

2.1 The Basic Model

The model consists of a population of haploid organisms, each represented by its genome. These individuals breed with each other (random mating) over a series of non-overlapping generations.

For the purposes of this study, we represent a chromosome as a contiguous sequence of slots, each of which contains either white space (junk DNA) or a gene (Figure 2a). In reality, the exons coding for

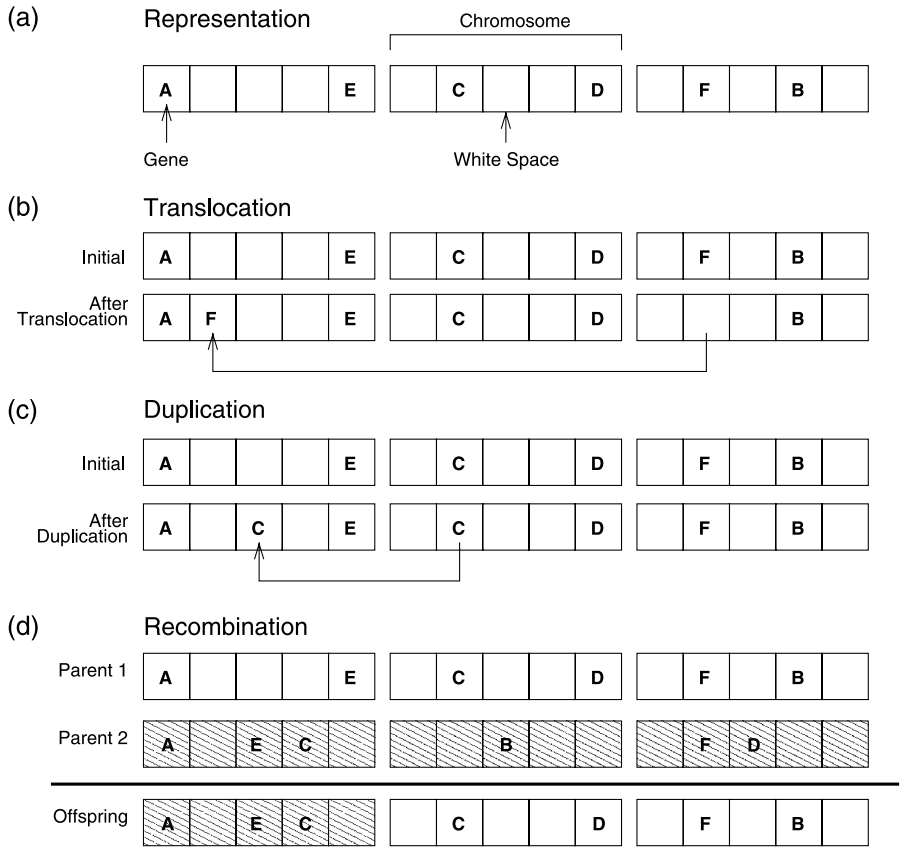

Figure 2. The experimental models. (a) The genes are represented as slots, chromosomes as sequences of slots, and junk DNA as empty slots or white space. (b) The process used in the model to simulate gene relocation (jumping). (b) The process used in the model to simulate gene duplication (copying). (c) Reproduction in the model. Offspring inherit a random combination of chromosomes from their parents.

a gene are often interspersed with introns, but the entire gene sequence is nevertheless confined within a compact region of a chromosome. We represent these compact regions as *slots*. In the model, the entire chromosome is therefore represented as a sequence of slots. At any given stage, each slot in the simulated chromosome either is empty (that is, contains junk DNA) or else contains a viable gene.

The model embodies two processes by which genes are known to migrate from one location on a genome to another:

1. *Translocation* – a gene moves from one slot to another (Figure 2b).
2. *Duplication* – a copy of a gene is inserted at a new slot (Figure 2c).

In both cases, the gene that moves overwrites the content of the slot at its new location.

The model represents time by a sequence of generations in the population. In the turnover from one generation to the next, pairs of individuals reproduce to form offspring. Reproduction consists of the selection of two individuals followed by the random selection of chromosomes to form the offspring (Figure 2d). The mutation operators are then applied.

2.2 The Experiments

The study consists of three experiments (see Figure 3). Experiment 1 addresses the co-location of genes on chromosomes. Experiment 2 examines the clustering of functionally related genes on a

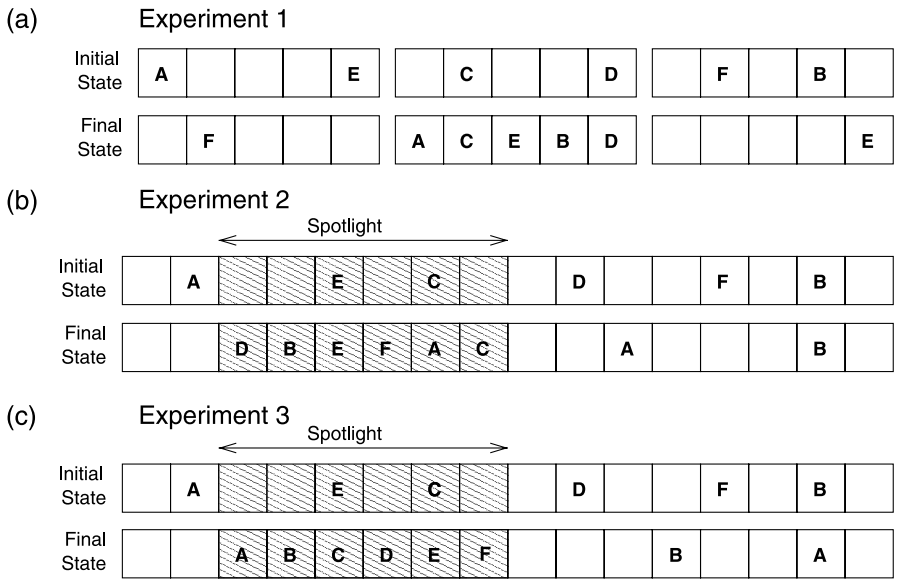

Figure 3. The setup for the experiments. In each case, the bottom row provides an example illustrating the hypothesis that, starting with a random configuration, genes will become concentrated on a single chromosome. (a) The setup for Experiment 1. From a random initial condition, functionally related genes cluster closely together on the same chromosome. (b) The setup for Experiment 2. The line represents a *spotlight* region in which genes are active during some developmental process. The bottom row illustrates the hypothesis that genes associated with the process will cluster together in the spotlighted region. (c) The setup for Experiment 3, showing typical initial and final configurations expected under the hypothesis that genes that act sequentially will tend to arrange themselves in transcription order.

single chromosome. Experiment 3 examines the ability of functional order to promote spatial ordering of genes. The models used in the experiments build on each other. Further details are given in the following sections.

3 Experiment 1

The first experiment addressed the co-location of genes on chromosomes (Figure 3). We begin by investigating the simplest case of sexual reproduction: a population of haploid organisms for which a set of functional genes $G_1,\ldots,G_n$ must all be present in an individual's genotype for it to survive. We make no assumptions about the locus of any gene, merely that it needs to be present. Nor do we consider different alleles. It suffices that any variant of a particular gene be present to perform its function. We assume that two kinds of shuffling can occur: (1) the duplication of a gene (*duplication*), and (2) the relocation of a gene from one slot to another (*translocation*).

In the results reported here, the model genome contained 10 chromosomes, each of 50 slots, and 1–30 genes in total. The model population consisted of 100 individuals, each with a randomly assigned configuration of genes. That is, the genes were scattered at random across several chromosomes, and the arrangement differed from one individual to another. The experiment analyzed the resulting configurations of genomes after 2000 generations.

3.1 Results of Experiment 1

Figure 4 shows that both mechanisms cause genes to move, or to be lost. In models of this system, the genes that are subject to selection migrate to form groupings and often become concentrated on a single chromosome. Starting from a random configuration, the most common outcome is that a single chromosome contains most of the genes, with the other chromosomes containing the few remaining genes. The tendency to form clusters is clear whatever the number of genes (Figure 4a).

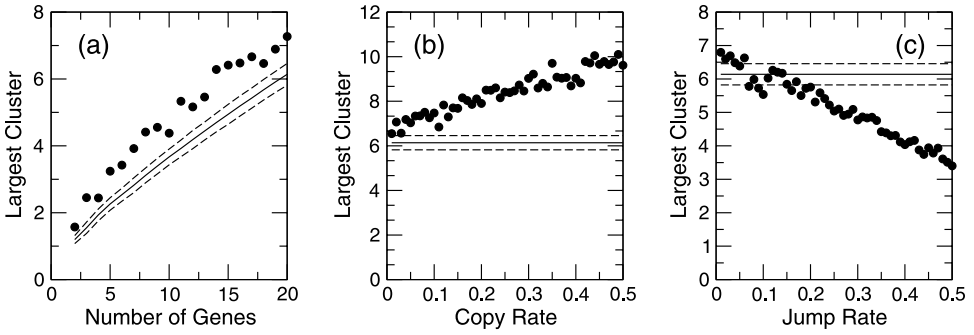

Figure 4. Clustering of genes on chromosomes. (a)–(c) The plotted points show the dependence of the average size of the largest cluster arising in the model on (a) number of genes, (b) duplication (copying) rate, and (c) relocation (jumping) rate. The solid lines indicate the results that would be expected under random arrangement, with 95% confidence intervals (dashed lines).

The duplication (Figure 4b) tends to enhance clustering, whereas relocation (Figure 4c) tends to break clusters up faster than they form. The model is sensitive to the amount of junk DNA (white space): the greater the amount of white space, the more genes the system can support.

4 Experiment 2

This experiment considers functional clustering (Figure 3). Experiment 1 dealt with convergence to a single chromosome for any functional genes. What modularity requires is a suite of genes that act together. For instance, if a set of genes all contribute to a particular process, then they may need to be active at the same stage of growth and development. The rationale for this is that as chromosomes undergo folding and unfolding, genes become exposed and can be transcribed at different times. The *spotlight model*, which simulates this phenomenon, is based on the basic model with the following added assumptions:

1. There is some gene X (or equivalently a preexisting group of genes) that is required for some developmental process to occur.
2. The process is associated with some phenotypic character that contributes to the fitness of individuals. We assume that this fitness is measurable and given by an individual's probability of reproducing; its initial value for every individual in the population is the same and is denoted by f. The population evolves under selection for fitness associated with that character.
3. When gene X is active, a fixed region of the chromosome also becomes active (the region under the spotlight). This assumption is based on the understanding that genes are active when the chromosome unwinds to expose particular regions. The exact shape of this region is immaterial to the experiment, so for simplicity we used a contiguous region in our experiments.
4. A number of genes A, B, C, D,… are located elsewhere on the chromosome, and each could enhance the process in some way provided that it is active at the same stage of development as gene X. We express this by assuming that each of the genes increases the fitness of the individual if it lies within the spotlight region.

We represent the genome as in Experiment 1 (Figure 3), but with just a single "chromosome," consisting of 40 gene slots. Each gene (say B) that lies within the spotlight region increases an individual's initial fitness f by $(1 - f)\ \Delta f_B$, where Δf_B is the intrinsic increase in fitness that can be ascribed to gene B.

Our hypothesis is that under the spotlight model, genes that contribute to the process associated with gene A will gradually migrate to lie under the spotlight. As a test of the hypothesis, we carried out a sensitivity analysis for each parameter. For each combination of parameter settings, averages were taken over 100 trials. Because there were so many possible combinations, we kept the values of every parameter at a fixed *base* value except for the parameter being tested. That is, the sensitivity results are slices in different dimensions through a base run with a spotlight of size 10 slots, 10 genes, an initial fitness of 0.3, an increase in fitness of 0.01 per gene, a duplication rate of 0.01 (1 in 100 genes), and a relocation rate of 0.01.

To provide an indication of statistical significance, confidence intervals were calculated from the distribution predicted by the null hypothesis. The null hypothesis in this case assumes that genes are spread randomly over the entire chromosome. So for g genes in a chromosome of length c, and a spotlight of size s, the number of genes under the spotlight would follow a binomial distribution with mean gs/c and variance $gs(c - s)/c^2$.

The points plotted were the average numbers of genes (over 100 trials) that lie within the spotlight in a population of 100 individuals after 2000 generations.

4.1 Results of Experiment 2

The results of this experiment (Figure 5) confirm that a marked clustering tendency does exist. That is, related genes do tend to aggregate in the neighborhood of genes that control a process. Neither (Figure 5a) the size of the spotlight region, nor (Figure 5b) the number of genes available has any effect on the clustering tendency. Under the conditions used, the final numbers of genes within the spotlight region were always significantly greater than those expected by chance. The clustering effect

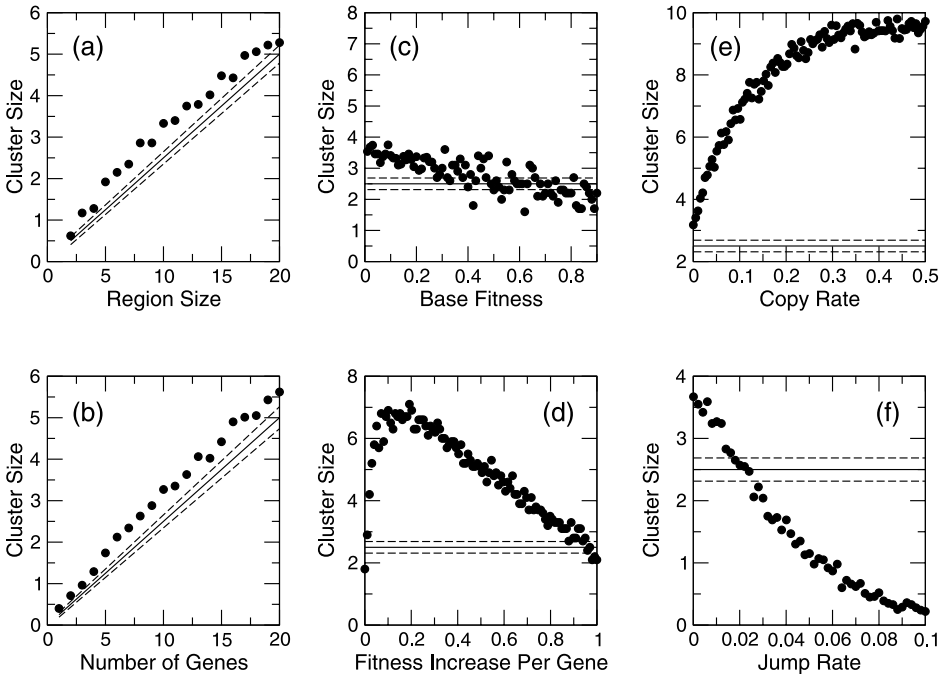

Figure 5. Experimental results for the spotlight model (Experiment 2). The graphs show sensitivity analyses (points) for each parameter, with the other parameters all held at the fixed base values given below. Solid lines indicate the results expected by chance alone, and are surrounded by 95% confidence intervals (dashed lines). The parameters varied are: (a) size of the spotlight region (base = 10 slots), (b) number of genes (base = 10), (c) initial fitness of the individual (base = 0.3), (d) fitness increase per gene (base = 0.01), (e) rate of gene duplication (copying) (base = 0.01), and (d) rate of gene relocation from one slot (jumping) to another (base = 0.01).

is greatest in individuals with low initial fitness (Figure 5c), with the number of genes in the spotlight falling to chance levels when the value exceeds 0.4. In contrast, the clustering effect is highly sensitive to the intrinsic increase in fitness (Figure 5d) for each gene added to the spotlight, the cluster size being several times the chance level except at very low or very high values. At low intrinsic fitness values, any additional genes contribute very little to the fitness of an individual, and clustering is unlikely to occur. At very high intrinsic fitness values, the individual only requires a few genes to gain a high fitness; as a result only small clusters form. Maximum clustering occurs when all genes carry a high fitness contribution and all are required to contribute to the fitness of the individual. Finally, the duplication rate (Figure 5e), and the relocation rate (Figure 5f) have opposite effects. Duplication greatly enhances the clustering effect, but relocation rates above about 0.02 tend to break up clusters faster than they form.

5 Experiment 3

The final experiment considers the emergence of transcription order (see Figure 3). The transcription order model extends the previous model by adding the following assumptions:

1. There is some physical relationship between the phenotypic products of genes. That is, having the product of gene A available before the product of gene B conveys a selective advantage.
2. There exists a sequence of genes that is optimal for sustaining some process. The fitness of an individual is greater the closer its genotype is to this optimal configuration.

The model setup was similar to that in Experiment 2 (Figure 3). The single chromosome was 50 slots in length, and there were 10 genes in all. The spotlight size was fixed at 10 slots. The model was run for 2000 generations in each case. An individual's fitness (probability of reproducing) is determined by the number of genes within the spotlight and number of genes in optimal order:

$$f(\vec{x},\vec{y}) = \sum_{i=1}^{L} \begin{cases} \Delta f x_i & \text{if } x_i = y_i, \\ 0 & \text{otherwise,} \end{cases}$$

where $\vec{x}$ is the optimal sequence, $\vec{y}$ is the sequence of genes contained within the spotlight, and $\Delta f x_i$ is the intrinsic increase in fitness due to having gene x_i in the correct order. In the sensitivity analysis for each parameter, averages were taken over 100 trials. To provide an indication of statistical significance, 95% confidence intervals were calculated from random sequences.

5.1 Results of Experiment 3

The results (see Figure 6) confirm that suites of co-adapted genes are likely to assemble in the order of transcription within a compact region on a chromosome. Starting from a random initial configuration, after 2000 generations the genes had arranged themselves in the optimal order (Figure 6a). Solid lines indicate the results expected by chance alone, and are surrounded by 95% confidence intervals (dashed lines).

Over time, the genetic operators (duplication and relocation) manipulate the organization of the genome and slowly approach the optimal arrangement (Figure 6a). As the number of genes within the system increases (Figure 6b), the ability to organize into the optimal arrangement is impeded; however, the genes do tend to cluster within the spotlight and partially arrange themselves. As found in the previous experiment, sequence completeness is very sensitive to the duplication rate (Figure 6c), with near-optimal sequence formation at low duplication rates. The formation of optimal transcription

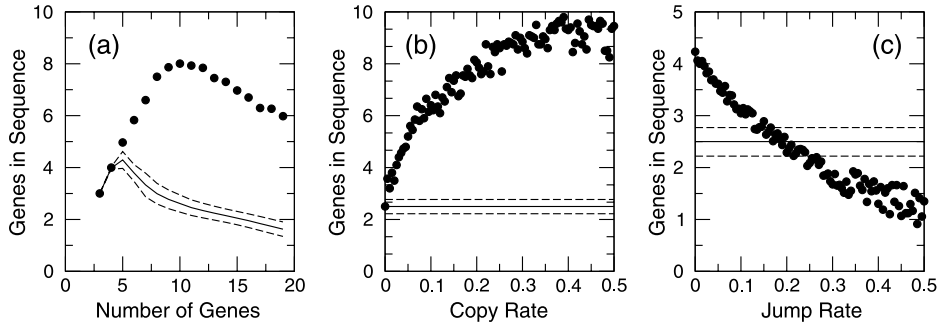

Figure 6. Experimental results on the evolution of the transcription order model (Experiment 3). The graphs present results from sensitivity studies for each of the parameters involved in the transcription order model. They show the changes in completeness of the optimal sequence against: (a) number of genes, (b) duplication (copy) rate, and (c) relocation (jump) rate.

orders is also sensitive to the relocation rate (Figure 6d). Very low rates enhance the clustering effect, but (as in Figure 5f) relocation rates above about 0.04 tend to destroy sequences faster than they form.

6 Discussion

In each of the experiments, the selection process leads to a race between gene clustering (genes packed close together) and genetic convergence (all individuals in a population having the same genetic makeup). Convergence in a large population effectively freezes the genetic organization, as mutations tend to get swamped and disappear. However, convergence happens slowly in a large population. On the other hand, a small population (less than about 30 individuals) converges rapidly, but there is still a finite chance of clustering through the mechanism proposed by our model [6].

The clustering arises from the following positive feedback process. A chromosome that acquires more genes than others conveys to its owner a greater chance of breeding viable offspring. Survival allows it to acquire still more genes via random shuffling, which further enhances the effect. Conversely, organisms inheriting a chromosome that has lost genes are less likely to have a full complement of genes. Therefore, the genetically enriched chromosomes quickly flood the population, resulting in convergence to one of a small number of genetic patterns.

The results of Experiment 1 imply that junk DNA (noncoding regions) plays a role in the above process. If the density of white space (junk DNA) is too low in the experiment, then genes are lost faster than they cluster and the entire population collapses. However, because there are no fixed slots in real DNA, we cannot conclude that junk DNA necessarily plays this role in real populations.

The assumptions that we have made in the above models need to be considered. We represent a chromosome as a sequence of slots of fixed length. The number of slots therefore limits the number of genes that can cluster there. In reality, the translocation process can simply splice a string into an existing sequence. That is, a chromosome can grow as it acquires genes. This process would only enhance the clustering effect. Also, we assume that each gene occupies a compact region of a chromosome, as implied by the slot model. Note that if, in Experiments 2 and 3, we replace genes by exons, then the results (Figure 5) actually confirm that scattered exons would tend to cluster in the way that we assume. The experiments reported here only dealt with haploid organisms. To test the limitations imposed by that assumption, we implemented a diploid version of the model. Results from this model (not shown) show that the diploid case still leads to cluster formation in the same way as the haploid case.

Finally, our results have a bearing on a growing debate about the nature and distribution of gene clustering within different groups of organisms. Clustered groups of co-expressed genes (operons)

are well known in prokaryotes, but few examples are known from eukaryotes [1], which are "not considered to contain operons" [2]. Although functional clustering does occur, it appears to be restricted to functionally important homeobox genes that play an important role in development.

The implications of our results on the above debate about gene clustering are clear from the assumptions underlying our models. In Experiment 1, we assume that organisms cannot survive without all of the genes considered. In the other experiments, we assume that the genes convey a selective advantage. The more crucial genes are to individual survival, the more applicable the result of Experiment 1. Therefore, our results predict that genes that are essential to an individual's survival and reproduction are those most likely to co-locate to the same chromosome or region within a chromosome. This conclusion is consistent with the findings of Caron et al. [5] mentioned above. It is also supported by observations of Blumenthal and Gleason [2], who noted that functionally important homeobox genes are often clustered.

Acknowledgments

David Newth was supported by an Australian Postgraduate Award. Adrian Gibbs, Terry Bossomaier, and Merlin Crossley made helpful comments on draft versions of the manuscript.

References

1. Blumenthal, T. (1998). Gene clusters and polycistronic transcription in eukaryotes. *BioEssays*, *20*(6), 480–487.
2. Blumenthal, T., & Gleason, K. S. (2003). *Caenorhabditis elegans* operons: Form and function. *Nature Reviews*, *4*, 110–118.
3. Botstein, D. A. (1980). Theory of modular evolution for bacteriophages. *Annals of the New York Academy of Sciences*, *354*, 484–490.
4. Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y., & Nurminsky, D. I. (2002). Large clusters of co-expressed genes in the *Drosophila* genome. *Nature*, *420*, 666–669.
5. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., & Versteeg, R. (2001). The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science*, *291*(5507), 1289–1292.
6. Casjens, S., & Hendrix, R. (1974). Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. *Journal of Molecular Biology*, *90*, 20–23.
7. Dandekar, T., Snel, B., Huynen, M., & Bork, P. (1998). Conservation of gene order: A fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, *23*, 324–330.
8. Davidson, E. H., Peterson, K. J., & Cameron, R. A. (1995). Origin of bilaterian body plans: Evolution of developmental regulatory mechanisms. *Science*, *270*, 1319–1325.
9. Doolittle, W. F., & Spaienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*, 601–603.
10. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., & Carucci, D. J. (2002). A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, *419*, 520–526.
11. Gilbert, W. (1978). Why genes in pieces. *Nature*, *271*, 501.
12. Hansen, T. F. (2003). Is modularity necessary for evolvability? Remarks on the relationship between pleiotropy and evolution. *Biosystems*, *69*, 83–94.
13. Halder, G., Callerts, P., & Gehring, W. J. (1995). Induction of ectopic eyes by targeted expression of the eyeless gene in *Drosophila*. *Science*, *267*, 1788–1792.
14. Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–1166.
15. Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*, 318–356.

16. Kamath, R. S., Fraser, A. G., Yan, D., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P., & Ahringer, J. (2003). Systematic functional analysis of the *C. Elegans* genome using RNA interference. *Nature*, *421*, 231–237.

17. Lawrence, J. G., & Roth, J. R. (1996). Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, *143*, 1843–1860.

18. Lercher, M. J., Urrutia, A. O., & Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, DOI:10.1038/ng887.

19. McClintock, B. (1951). Chromosomal organization and genetic expression. *Cold Spring Harbor Symposia on Quantitative Biology*, *16*, 13–47.

20. McClintock, B. (1956). Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*, 197–216.

21. Orgel, L. E., & Crick, F. H. C. (1980). Selfish DNA: The ultimate parasite. *Nature*, *284*, 604–607.

22. Pepper, J. W. (2000). The evolution of modularity in genome architecture. In C. C. Maley & E. Boudreau (Eds.), *Proceedings of the Artificial Life 7 Workshop*.

23. Pepper, J. W. (2003). The evolution of evolvability in genetic linkage patterns. *Biosystems*, *69*(2–3), 115–126.

24. Wagner, A. (1996). Does evolutionary plasticity evolve? *Evolution*, *50*(3), 1008–1023.

25. Wagner, G. P., & Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. *Evolution*, *50*, 967–976.