Genet. Res., Camb. (2007), **89**, *pp*. 337–339. © 2008 Cambridge University Press
doi:10.1017/S001667230800952X Printed in the United Kingdom

Genetic variability and neutral mutations: a commentary on 'Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles' by Motoo Kimura

DAMIAN D. G. GESSLER*
National Centre for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA

In 2007, the ENCODE consortium confirmed and expanded previous studies highlighting that the vast majority of the human genome is transcribed (ENCODE Project Consortium, 2007). This is in marked contrast with our earlier worldview that as much as 95% of the genome is 'junk' and of little or dubious consequence to the everyday functioning of the cell or organism. For example, recent reviews on RNA have shown it to be ubiquitous in cellular functioning (Amaral *et al*., 2008), which, combined with the startling discovery in 2001 that there are perhaps only 30 000 human protein-coding genes, is drastically changing the way we understand genes – and ultimately – genetic variation itself.

Thirty years ago, when we had a simpler view of the genome, Kimura (1968*b*) published his work on neutral isoalleles, thereby contributing to a true upheaval in our understanding of genetic variation no less than that which we are experiencing today. Just 2 years earlier, Lewontin & Hubby (1966) published an influential paper examining electrophoretic variation. The surprising result was that there was much more protein variation than expected, and specifically more variation than expected at individual genes or loci. This presented a paradox that Motoo Kimura sought to unravel.

Classical population genetics had two main classes of models to explain genetic variation: in the first model, it was known that new deleterious mutations will segregate in a population for an average of $1/s$ generations before they are removed by selection, where s is a measure of the selective effect, or reduction in fitness, experienced by an individual as a result of carrying the mutation. So while mutation pressure is introducing new variation, selection is removing it. If these mutations occur at a rate of μ per generation, the forces equilibrate at an average of μ/s mutations per individual – and indeed Kimura and Maruyama (Kimura & Maruyama, 1966) had derived the entire distribution of the number of mutations per individual just 2 years earlier. Yet parameterizing μ (the genome-wide mutation rate) and s (the average selection coefficient) based on experimental studies gave values either too low to account for the observed high levels of variation or too high to be sustained under models of genetic load. Supposing new mutations to be beneficial was even more unhelpful in reconciling the data, since beneficial mutations were expected to occur at an even lower rate, and those mutations not lost due to chance when rare would sweep to fixation and in the process supplant all variants with their superior type, thereby reducing variation. The second class of models gave selection a more nuanced role in maintaining variation. Perhaps if some middle or optimal combination of variants was most fit – essentially selection for variation itself – then the observed high levels of segregating variation could be understood. For example, models on overdominance or balancing selection could indeed stably segregate large amounts of variation. The only problem was that the data were not kind to these hypotheses either. Impressive cases of heterozygote advantage and/or balancing selection could be documented in areas such as mate recognition and disease resistance, but the special nature of these examples never allowed them to shoulder the burden of a general solution. Additionally, did not the presence of an optimum generate a 'cost' or 'load' experienced by all lesser fit individuals harbouring variants displaced from the ideal? Would the underlying genetic architecture not be susceptible to invasion by load-reducing mechanisms? Indeed, decades earlier, Haldane had given an elegant model of how heterozygote advantage is susceptible to gene duplication, whereby all offspring are invariant but produce the optimal intermediate genotype (one copy of the gene is fixed for one allele, while the second copy is fixed for the other allele).

Against this backdrop, Kimura (1968*b*) shows himself at the height of his powers. Kimura used the toolbox of diffusion modelling to analyse the relative forces of mutation and drift. These techniques

* e-mail: ddg@ncgr.org

(borrowed and expanded from their use in physics and chemistry) were well known to the great triumvirate of twentieth century population genetics: Ronald A. Fisher, Sewall Wright and J. B. S. Haldane. But Kimura raised them to a new level both in refinement of technique and as an investigative tool.

Kimura (1968*b*) first re-establishes the estimation of the effective number of alleles

$$n_e = \bar{H}_0^{-1} = 4N_e\mu + 1.$$

The effective number of alleles (n_e) in a population is the reciprocal of the average homozygosity and is a function of the effective population size and the mutation rate. Thus, n_e is the number of equi-frequent alleles in an idealized population that would produce the same level of homozygosity as in an observed population. So as populations become larger, or the mutation rate increases, so does their segregating variation. Alternatively, if the population (more correctly, $4N_e$) falls much below the reciprocal of the mutation rate, then drift will overwhelm mutation, and most variation will be rare or lost at a locus. This result is for neutral alleles, i.e. alleles with no selection coefficient s, in a model that allows an infinite number of allelic states at a site. Equations for a finite number of possible states are also presented.

Kimura then solves for the distribution of allele frequencies, again without selection, but with special consideration to loss and fixation. Because there is continual mutation, a locus is never in a permanent state of monomorphism, so Kimura's diffusion approximation on the boundary states can be seen as the probability that a locus within a population (or a single-locus population within an ensemble) transitions from zero variability at a given locus to some variability (i.e. variant frequency $1/(2N)$). Kimura then closes the loop on the extended analysis by showing that in the limit of an infinite number of possible allelic states, the original estimate of n_e and $\bar{H}_0$ is recovered.

An idealized number of alleles and the average number of alleles are not the same concept. The former (n_e) represents the number of equi-frequent alleles in a population that could explain the observed average level of homozygosity (and therefore by extension the average level of heterozygosity or genetic variation). The latter (n_a) is simply the reciprocal of the mean frequency of alleles observed in a population – very much a function of variance in allele frequencies: a locus with four alleles, each with a frequency of 0·25, will contribute to genetic variation and harbour the consequences of genetic variance much more than a locus with one allele at 0·97 and three at 0·01. The variables n_e and n_a capture these two concepts, respectively. Kimura solves for n_e as well as n_a and partakes in a number of computer simulations to check theory with implementation. This is classic Kimura; he checks his analytical derivations not just for mathematical competency, but also with independent model implementations. This uncovers a slight underestimation of the diffusion theory for both n_e and n_a, which he attributes to the theory's inability to completely capture the effect of rare alleles (specifically alleles represented only once in the population).

The relevance of Kimura's paper in *Genetical Research* is in interpreting and supporting his even more influential paper: 'Evolutionary rate at the molecular level' (Kimura, 1968*a*). The elegance of the argument for neutrality espoused in Kimura (1968*a*) was underwritten by much of the mathematics in Kimura (1968*b*). In the latter, he ties the concept of neutral mutations together with early estimates on the prevalence of synonymous mutations and their consequences in terms of high amounts of genetic variance and low load. Kimura further expands the discussion to nearly neutral mutations ($|2N_e s| \ll 1$), correctly noting that for many mutations their neutrality is not an absolute property of the phenotype, but a relative property of the phenotype marginal on a given effective population size. Other authors had raised the possibility of neutral mutations as a component of the paradox of high segregating genetic variance and low load, but it was Kimura (1968*a*, *b*) who made an effort to quantify the repercussions of such a model.

Kimura's appreciation of neutralism is eminently intuitive. But intuition can be both a weak ally and a deceptive foe in population genetics. Soon after Kimura's mathematical exposition on neutralism, his brilliant young colleague, Tomoko Ohta, significantly expanded its application, in her analysis of slightly deleterious, nearly neutral alleles. More recently, selectionism has regained evidence in its favour (Bustamante *et al.*, 2005, for example, is one of many papers). According to M. Takahata (in Kimura, 1994), later in life Motoo Kimura tended to re-favour a more dichotomous strictly neutral/strictly beneficial view of genetic variation. Today, our new glimpses of RNA and forays into massive population resequencing enabled by the latest sequencing technologies suggest that no sweeping generalization is likely to be meaningful across the many types of genetic variation we discover. So Kimura's (1968*b*) paper is likely to remain relevant as a historical anchor for many years.

References

Amaral, P. P., Dinger, M. E., Mercer, T. R. & Mattick, J. S. (2008). The eukaryotic genome as an RNA machine. *Science* **319**, 1787–1789.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D.,

https://doi.org/10.1017/S001667230800952X Published online by Cambridge University Press

Civello, D., Adams, M. D., Cargill, M. & Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157.

ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816.

Kimura, M. (1968*a*). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kimura, M. (1968*b*). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

Kimura, M. (1994). *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers* (ed. N. Takahata). Chicago, IL: Chicago University Press.

Kimura, M. & Maruyama, T. (1966). The mutational load with epistatic gene interactions in fitness. *Genetics* **54**, 1337–1351.

Lewontin, R. C. & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595–609.

https://doi.org/10.1017/S001667230800952X Published online by Cambridge University Press