Genomics 105 (2015) 282–287

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Maruyama's allelic age revised by whole-genome GEMA simulations

Shuhao Qiu, Alexei Fedorov *

Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, Toledo, OH 43614, USA
Department of Medicine, University of Toledo, Health Science Campus, Toledo, OH 43614, USA

ARTICLE INFO

Article history:
Received 31 July 2014
Accepted 16 February 2015
Available online 21 February 2015

Keywords:
Genomics
Computational biology
Allelic age
SNP
Haplogroups

ABSTRACT

In 1974, Takeo Maruyama deduced that neutral mutations should, on average, be older than deleterious or beneficial ones. This theory is based on the diffusion approximation for a branching process, which considers mutations independently of one another and not as multiple groups of interconnected mutations with strong linkage disequilibrium (haplotypes). However, mammalian genomes contain thousands of haplotypes, in which beneficial, neutral, and deleterious mutations are tightly linked to each other. This complex haplotype organization should not be ignored for estimation of allelic ages. We employed our GEMA computer simulation program for genome evolution to re-evaluate Maruyama's phenomenon in modeled populations that include haplotypes approximating real genomes. We determined that only under specific conditions (high recombination rates and abundance of neutral mutations), the deleterious and beneficial mutations are younger than neutral ones as predicted by Maruyama. Under other conditions, the ages of negative, neutral, and beneficial mutations were almost the same.

1. Introduction

Investigations of "allelic age" began in 1970s. This term was defined as the number of generations a mutant allele has persisted in the population since its first occurrence [7,10–12]. Initially, the prediction of allelic age relied upon mathematical modeling—a diffusion approximation for a branching process. In 1973, Kimura and Ohta [7] inferred that the "average ages of neutral alleles, even if their frequencies are relatively low, are quite old." Specifically, they demonstrated that a neutral mutation whose current frequency is 10% has an expected age (measured in generations) roughly equal to the effective population size N_e. This result complicates experimental verification of allele age predictions. Thus, allelic age estimates currently come from either mathematical modeling or indirect experimental hints about the distribution patterns of mutations with various population frequencies. In 1974, Takeo Maruyama [11] modeled semidominant mutations and made a principal prediction that neutral mutations, on average, are significantly older than both deleterious and beneficial alleles. This prediction has been widely accepted and became an important landmark in this field. A year after Maruyama's paper, Wen-Hsiung Li [10] inferred the age of deleterious mutations having various degrees of dominance. He demonstrated that the mean age decreases with increasing selection coefficients against heterozygotes. Allelic age has been nicely reviewed in the late 1990s [5] and early 2000s [18]. The allelic age has been indirectly estimated in several independent experimental studies that statistically examined the distributions of multiple mutant alleles. Slatkin and Rannala [17] estimated the allelic age by use of intra-allelic variability. Further, Rannala and Reeve applied high-resolution multipoint linkage-disequilibrium mapping [14], while Genin and colleagues analyzed shared haplotypes of rare disease mutations [4]. Last year, Kiezun and co-authors [6], concluded from analysis of large-scale population sequencing studies and computer simulations, that deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. However, the allelic ages for neutral, deleterious and beneficial mutations are still unclear because the direct measurement of the age is impossible.

Recent whole-genome sequencing of numerous individuals revealed that each human individual bears millions of mutations [1]. These millions of mutations form intricate patterns of haplotypes, where neutral, beneficial, and deleterious mutations are tightly linked with each other and strongly influence the ages of their neighbors. A haplotype structure for a gene strongly depends on the local recombination rate, which may vary thousands of times from one chromosomal location to another [2].

In order to examine the role of haplotypes on the allelic age, we applied whole-genome computer simulations of SNP dynamics using our GEMA program package. A "naturally occurring" intense influx of 40 novel mutations per person has been applied in this computer modeling. Such intense mutation influx generated thousands of SNPs in each modeled individual. The time of the arrival for each mutation has been recorded and used for the calculation of its age. These simulations allow the direct measurement of the average age of mutations with high accuracy. In these computational experiments, we changed various parameters such as recombination rate, degrees of dominance, and distributions of mutations by their selection coefficients. These various

* Corresponding author at: Department of Medicine, University of Toledo, Health Science Campus, Toledo, OH 43614, USA. Fax: +1 419 383 3102.
E-mail address: Alexei.fedorov@utoledo.edu (A. Fedorov).

http://dx.doi.org/10.1016/j.ygeno.2015.02.005
0888-7543/

conditions drastically altered the patterns of haplotype ensembles in the modeled genome. We demonstrated that Maruyama's effect appears only for specific sets of parameter ranges and quantitatively described its variation under different conditions.

2. Materials and methods

Computer simulations were performed using a new v3 release of our Perl program GEMA_v3.pl, named Genome Evolution with Matrix Algorithms (GEMA). The previous release (GEMA_v2.pl) has been described in detail [13]. Both v2 and v3 versions are freely available from our web site: http://bpg.utoledo.edu/~afedorov/lab/GEMA.html. The V3 release of GEMA has only a small addition compared to v2, which, upon creation of a new mutation, records the time of its arrival (measured in generations, as $g variable inside a multidimensional array @matrix). Finally, the age of every SNP is periodically recorded into a new fifth column of the GEMA backup file.

In the described simulations with GEMA_v3.pl, we always used the following parameters: (1) unsaturated mode; (2) duration: 10,000 generations; (3) population size ($N = 100$); (4) number of offspring per mating pair ($\alpha = 5$); (5) mutation rate per gamete ($u = 20$); (6) recombination rate ($r = 1$ or $r = 48$); (7) dominance coefficient ($h = 0$, $h = 0.5$, or $h = 1$); (8) MatingScheme: permanent random male–female pairs; and (9) upon generation of a random mutation, a random number generator imbedded into GEMA program assigned a selection coefficient to it either according to the "experiment B" or "experiment C" distributions demonstrated in Fig. 1. Experiments B and C were first described in our paper [13], and we kept their original names in this paper for clarity. Those two experiments were chosen for the ease of interpretation of the results. The effects of all deleterious mutations in these experiments are equal to each other since their selection coefficients (s) always equal to -1. Consequently, all beneficial mutations are also equal to each other ($s = +1$ for all beneficial mutations).

Our GEMA modeling approximates natural conditions in a way in which we consider thousands of genes in genome of virtual individuals and the real influx of novel mutations (which is about 40 new mutations per individual). As we demonstrated in Qiu et al. [13], several hundreds of genes in the modeling genome have approximately the same effect on SNP dynamics as 25,000 genes observed in humans. In addition, the length of modeling genes does not significantly influence the SNP dynamics. Due to these reasons and for the speed of computations, we used a 0.6 Mb long DNA segment with a random nucleotide sequence as the genome for modeled individuals. Thousands of nucleotide-long segments of this sequence were used to model 600 genes. The simplification of our modeling, compared to real conditions, is that all genes in our simulations have the same properties. This includes the same recombination rate, same frequencies of deleterious, beneficial, and neutral mutations and the same dominance coefficient. In real human genes, these parameters vary significantly from gene to gene. However, these simplifications allow us to evaluate the influence of each parameter on the dynamics of SNP in the population.

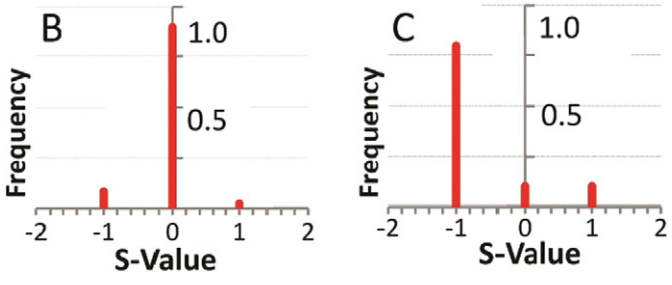

Fig. 1. Distribution of computer-generated mutations by their selection coefficients (s-values). B—"Experiment B" models a discrete distribution of mutations characterized predominantly by neutral mutations, occurring at a frequency of 90% within the population, while the remaining 10% is characterized by deleterious and beneficial mutations occurring in a ratio of 9:1. C—In "Experiment C," the ratio of deleterious to beneficial mutations occurs again in the ratio of 9:1. However, this model is characterized by a preponderance of mutations with deleterious effects (81%). Neutral mutations in this case comprise 10% and beneficial - 9% of overall nucleotide changes occurring within the population.

The snapshot of all SNPs in all modeled individuals was recorded after every 1000 generations as backup files. These backup files contain the following information on each SNP: position; selection coefficient; mutant nucleotide; modeled individuals bearing this SNP including location on a maternal or paternal DNA; and the time of SNP arrival (in generations). Backup files was processed with our Perl scripts AllelicAge_10bin.pl and AllelicAge_csv.pl, that calculate the frequency of each SNP, its selection coefficient and the time of its arrival, and present this information in an output table in Excel format (Supplementary Materials, Tables S1 and S2). These tables were used to calculate the distribution of SNPs by their population frequency, the number of SNPs with particular selection coefficient within a designated range of population frequencies (from 10% to 30% range or in 40%–60% range), and the distribution of SNPs within a particular range of population frequency by their age. The SNP frequency stands for the frequency of the mutant alleles in the entire modeled population.

3. Results

Computer simulations of whole-genome SNP dynamics were performed using the program GEMA_v3.pl. In these computations, the following three parameters were always the same for every experiment: (1) population size was 100 modeled individuals ($N = 100$); (2) every modeled individual had 40 novel mutations ($\mu = 20$ mutations per gamete); and (3) the mating scheme was a default GEMA choice—permanent random male–female pairs (MatingScheme = 1) with 5 offspring per mating pair ($\alpha = 5$). Also, genomes of modeled individuals always consisted of 600 genes each 1000 nucleotide long. [As we discussed previously, the exact number of genes above a certain threshold (~200) does not significantly influence SNP dynamics [13]]. Variable parameters for each computational experiment were the following: (1) number of recombination events per gamete (r) was either $r = 1$ or $r = 48$; (2) gene dominance coefficient (h) for every gene was either $h = 0$ (dominant genes), $h = 0.5$ (co-dominant genes), or $h = 1$ (recessive genes); and (3) distribution of mutations by their selection coefficients corresponded to the "Experiment B" or "Experiment C" shown in Fig. 1. We specifically used $r = 48$, because it represents the average number of pieces of paternal and maternal genomes in a human gamete [13]. The alternative $r = 1$ settings model the regions with low recombination rate frequency, which are abundant in the human genome.

The distribution of SNPs by their age for different modeled parameters is shown in Fig. 2. This distribution has been combined for 12 independent experiments. The total number of all SNPs in specific experiment varied from 152,582, for simulations with $r = 1$, $h = 0$, and "experiment C", to the 505,970 SNPs for $r = 1$, $h = 1$, and "experiment B" simulations. Since the number of SNPs varies from one experiment to another, we performed their normalization by division by the total number of SNPs in each experiment. Hence, the results in Fig. 2 are presented as relative SNP frequencies counted within 10-generation bins. The details for every SNP from these data are provided in the supplementary Table S1. In all experiments the youngest SNPs were the most numerous ones, as expected from population genetics. We observed that, when the recombination rate was high ($r = 48$), the older SNPs were more abundant than when the recombination rate was low ($r = 1$). A special case that does not follow this rule is provided by the combination of low recombination rate ($r = 1$) with recessive dominance coefficient ($h = 1$). As we explained previously [13], these specific conditions may result in an un-stable number of SNPs in the population, periodically producing gigantic peaks of SNP numbers.

The calculated mean age of SNPs, for which population frequencies belong to a particular range (10%–30% or 40%–60%) is shown in Fig. 3.