Genome Biology and Evolution Advance Access published April 10, 2014 doi:10.1093/gbe/evu075

Genome Evolution by Matrix Algorithms (GEMA): Cellular Automata Approach to Population Genetics

Shuhao Qiu[1,2,§], Andrew McSweeny[1,2,§], Samuel Choulet[3], Arnab Saha-Mandal[1], Larisa Fedorova[2], and Alexei Fedorov[1,2,*]

[1]Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, Toledo, OH 43614, USA;

[2]Department of Medicine, University of Toledo, Health Science Campus, Toledo, OH 43614, USA;

[3]University of Toledo College of Medicine and Life Sciences, University of Toledo, Health Science Campus, Toledo, OH 43614, USA;

[§]S.Q. and A.M. contributed equally to this work

[*]Corresponding Author: E-mail: Alexei.fedorov@utoledo.edu

Author contributions: SQ, AM, SC, and AF wrote ***GEMA*** programming codes. SQ, AM, and ASM performed computational experiments and analyzed the data. AF and LF designed and supervised the study and paper writing.

Abstract

Mammalian genomes are replete with millions of polymorphic sites, among which those genetic variants that are co-located on the same chromosome and exist close to one another form blocks of closely linked mutations known as haplotypes. The linkage within haplotypes is constantly disrupted due to meiotic recombination events. Whole ensembles of such numerous haplotypes are subjected to evolutionary pressure, where mutations influence each other and should be considered as a whole entity – a gigantic matrix, unique for each individual genome. This idea was implemented into a computational approach, named Genome Evolution by Matrix Algorithms (*GEMA*) to model genomic changes taking into account all mutations in a population. *GEMA* has been tested for modeling of entire human chromosomes. The program can precisely mimic real biological processes that have influence on genome evolution such as: 1) authentic arrangements of genes and functional genomic elements; 2) frequencies of various types of mutations in different nucleotide contexts; 3) non-random distribution of meiotic recombination events along chromosomes. Computer modeling with *GEMA* has demonstrated that the number of meiotic recombination events per gamete is among the most crucial factors influencing population fitness. In humans, these recombinations create a gamete genome consisting on an average of 48 pieces of corresponding parental chromosomes. Such highly mosaic gamete structure allows preserving fitness of population under the intense influx of novel mutations (40 per individual) even when the number of mutations with deleterious effects is up to ten times more abundant than those with beneficial effects.

Key words: Fixation, Gene, Genomics, Linkage, Neutral Theory, SNPs

Introduction

Humans have modest intra-species genetic variations among mammals (Kaessmann et al. 2001; Zhang and Plastow 2011). Even so, the number of genetic variations between two persons from the same ethnic group (e.g. Japanese, Finnish) is in the range of 3.4-5.2 millions as demonstrated by the "1000 Genomes" International Sequencing Project (Abecasis et al. 2012). This gigantic pool of nucleotide variations is constantly updating by 40-100 novel mutations arriving in each person (Kondrashov and Shabalina 2002; Conrad et al. 2011; Li and Durbin 2011). Closely located mutations on the same DNA molecule are linked together forming haplotypes that are inherited as whole units and span over a considerable portion of a gene or several neighboring genes (Consortium 2005). An intense intermixture of millions of mutations occurs in every individual due to frequent meiotic recombinations during gametogenesis. On an average, a haploid genome of a human gamete is comprised of 48 pieces of parental chromosomes (see section 2 of the Supplementary file S1 (***GEMA***_Instructions.pdf)). DNA recombination process causes gradual change of haplotype structures from generation to generation. Several mathematical theories have attempted to describe the intricate dynamics of genetic variations in populations. These models often conflict with each other and there is no universally acknowledged perception of genomic nucleotide dynamics (Wagner 2008; Nei et al. 2010). Population Genetics mathematical theories often consider mutations individually despite that natural selection, a major player in evolution, occurs simultaneously on entire ensembles of mutations in an organism. It should be acknowledged that background selection and genetic hitchhiking deals with groups of neighboring mutations from the same locus (Hill and Robertson 1966; Stephan 2010), while Fisher, Wright and later researchers considered interactions of mutations in a few loci (Fisher 1930; Wright 1965; Bodmer and Felsenstein 1967; Gavrilets and

Hastings 1994). We suppose that mutations should be treated as a whole entity – a gigantic matrix of all genetic variations, unique for every individual genome. With this aim we developed a computer program to process such matrices, named Genome Evolution by Matrix Algorithms (***GEMA***). Application of ***GEMA*** has already revealed new insights in population genetics presented in this paper. This public program can be used for a broad range of investigations in the field of Genomics.

A key question in population genetics that has been investigated for decades is: What is the probability of a certain mutation with a selection coefficient ***s*** to be fixed in a population? For a trivial case of a neutral mutation (when ***s***=0), there exists an undisputable solution to the problem inferred by the Neutral theory of evolution. This theory predicts that the ultimate fixation probability of a novel neutral mutation (which is initially present as a single copy) is equal to $1/(2N_e)$, where N_e is the effective size of the population (Kimura 1983). Lately, Tomoko Ohta demonstrated that nearly neutral mutations ($2\boldsymbol{Ns} \ll 1$) behave as if they are neutral (Ohta and Gillespie 1996). However, the general solution of this problem (when $\boldsymbol{s} \neq 0$ and $2\boldsymbol{Ns}$ product is not close to zero) is very convoluted and depends on a number of parameters/variables characterizing organisms, populations, and environment. Moreover, these parameters have significant synergistic/antagonistic effects, making it impossible to infer fixation probability even with the most advanced mathematical approaches. As we discuss further, even for the trivial case of neutral mutations ($\boldsymbol{s}=0$), the probability of fixation of a novel mutation, for a particular combination of parameters characterizing organism and population, might significantly deviate from Kimura's $1/(2N_e)$ formula, obtained using diffusion theory of stochastic processes (Kimura 1962). Mathematical theories in population genetics deal only with oversimplified models considering no more than two or three parameters at a time and predominantly examining

a single or a few loci. Thus, the profound query by John Sanford in "Genetic Entropy" (Sanford 2008) – "What will happen with mankind in the nearest future when each person has a hundred of novel mutations?" – remains unanswered. Instead of mathematical modeling, this problem can be approached more fruitfully from a different dimension, taking advantage of the enormous power of contemporary supercomputers. Computer modeling of genetic processes may be considered as an advanced branch of cellular automata, named by Stephen Wolfram as "A New Kind of Science" (Wolfram 2002). On numerous examples Wolfram demonstrated that any system of interacting elements creates patterns within their arrangements, which are hard to predict mathematically yet trivial to reproduce and study computationally. Here, we implemented such computational approach and present our program, named Genome Evolution by Matrix Algorithms (or ***GEMA***). This program models the evolution of genomic sequences in a population under the influx of numerous mutations at multiple loci and can take into account dozens of parameters/variables simultaneously. It belongs to a forward-time simulation category (Carvajal-Rodriguez 2010) and implies a Wright-Fisher population modeling where generations do not overlap (Hartl and Clark 2006). ***GEMA*** has many features similar to previously published programs (GENOMPOP (Carvajal-Rodriguez 2008), SFS_CODE (Hernandez 2008), FREGENE (Chadeau-Hyam et al. 2008), Mendel's Accountant (Sanford J 2007) among others, reviewed by Carvajal-Rodriguez (Carvajal-Rodriguez 2010)). However, ***GEMA*** is designed specifically to answer important questions that have not been addressed with previous programs. Specifically, here we present a core program ***GEMA***_*r1.pl* that models the influx of ~50 novel point mutations per individual (the real rate observed in the human genome) in order to determine the genetic parameters most crucial for maintaining population fitness. We also introduce the advanced version ***GEMA***_*r01.java* that can precisely mimic real biological processes influencing

genome evolution. It is designed to perform computational experiments to understand non-randomness in genomic nucleotide compositions such as GC-isochores (Bernardi 2007), codon usage bias (Plotkin and Kudla 2011), and mid-range inhomogeneity regions (Bechtel et al. 2008; Prakash et al. 2009).

Results

Several examples of ***GEMA*** computations are shown in Figure 1. These graphs illustrate the modeled dynamics for the influx of mutations, 12% of which have positive selection coefficient ($s>0$) while the rest 88% have a negative effect ($s<0$). The distribution of mutations by ***s***-parameter has been modeled according to a decay curve, shown in the Figure 2A. When the number of meiotic recombination events was low ($r=1$, recombinations per gamete) and the rate of mutations were approximated to the one naturally observed for humans ($\mu=20$, mutations per gamete), the relative fitness of individuals declined with generations. Yet, a higher degree of purifying selection pressure (corresponding to a larger number of offspring per individual -- ***α***-parameter) caused the decline of fitness to be less sudden with respect to increasing number of generations (see Figure 1A). These parameters are thoroughly explained in the User Guide for GEMA (in Supplementary file S1, pages 6-9) and also in the ***GEMA*** web site (http://bpg.utoledo.edu/~afedorov/lab/GEMA.html) including GEMA_video_presentation.m4v, GEMA.java pseudocode, and other supporting materials.

Figure 1B illustrates the model with two fixed parameters: $\mu=20$ and $\alpha=5$ (offspring per individual). The only variable parameter in this experiment was the number of meiotic recombination events per gamete (***r***). The increase of ***r*** to 48 prevented the declining of fitness. We specifically used $r=48$, because it represents the average number of pieces of paternal and

maternal genomes in a human gamete (on an average, 35.2 pieces result from meiotic recombinations in autosomes and 11.5 pieces result from the existence of 23 pairs of chromosomes).

The variations of total number of SNPs in generations are shown in the Figures 1C and 1D. The latter picture exemplifies some peculiarities in the SNP dynamics under certain conditions. The gigantic peaks in the number of SNPs in the population were observed when the meiotic recombination rate was low ($r \leq 1$) and genes had a recessive mode (gene fitness of heterozygote is close to the maximal fitness of maternal and paternal alleles; dominance coefficient $h=1$). This effect is also discussed below.

We computed the probability of fixation of a novel mutation with the selection coefficient s, which we denote as π_s. To make these results immediately understandable, we simplified the distribution of mutations by their selection coefficients to trivial cases, where a mutation has only three options for a possible s value: -1, 0, or +1. Two of such distributions, used in our modeling and named as experiments B and C, are shown in the Figures 2B and 2C. In both the experiments B and C, mutations with $s=-1$ are nine times more abundant than those with $s=+1$. However, in the experiment B, a majority of mutations (90%) are neutral ($s=0$) while in experiment C, neutral mutations represent a minor fraction (10%).

By taking advantage that we can trace the fate of each mutation in the simulation experiments, we computed the probability of fixation of a novel mutation with the selection coefficient s, which we denote as π_s. The probability of fixation was calculated as

(1) $$\pi_s = F_s / C_s ,$$

where C_s is the number of novel mutations with selection coefficient s that occurred from generation 2,000 to 10,000 in all offspring, while F_s is the number of fixed mutations with selection coefficient s within the same period of 8,000 generations. (After the first 2,000 generations, the population reaches equilibrium in the number of SNPs and subsequent consideration of the next 8,000 generations allows us to acquire sufficient statistics for fixed mutations.) Figures 3, 4 and 5 show values of π_s for six different parameters: 1) N – size of the population (24, 50 or 100 individuals); 2) μ– number of novel mutations per gamete (1, 5, 10, or 20); 3) r – number of meiotic recombinations events per gamete (1 or 48); 4) h – dominance coefficient (0, 0.5, or 1); 5) α - number of offspring per individual (2, 5, or 10); and 6) D – distribution of novel mutations by selection coefficients (experiments B or C shown in Fig 2B and 2C respectively). The original tables with these complete datasets are provided in the supplementary Tables S1 and S2. These Figures 3 and 4 and Tables S1 and S2 demonstrate intricacies in variations of π_s as a function of six arguments: $\pi_s = \pi_s(N, \mu, r, h, \alpha, D)$. We detail below some of the major consequences of these dependencies.

In our computer experiments the level of selection pressure is measured as the number of offspring per individual (α). The ***GEMA*** settings in all the described experiments were always based on "survival of only the fittest" and a constant size of population (N is fixed for a particular computational experiment). Thus, when α=2, the selection is completely turned off even for beneficial and deleterious mutations with $s \neq 0$ (because no offspring are removed). The setting with α=2 serves as a good control for the computational algorithm because in every experiment with α=2, we observed that $\pi_s(\alpha=2)$ was very close to $1/2N$ for every value of s in accordance with Kimura's law for neutral mutations (Kimura 1983)). Importantly, Kimura did not consider variations in the number of offspring per individual. His probability of ultimate

fixation π_s^{kimura} is calculated based on the number of novel mutations in adults (a subset of offspring that reach adulthood and subsequently create next generation of offspring). For nearly neutral mutations π_s^{kimura} can be calculated from our π-value from formula (1) by simple normalization: $\pi_s^{kimura} = \pi_s \times \alpha/2$. (Observe that this normalization formula may not be correct for beneficial mutations, where fixation probability might turn out to be greater than 1 post the normalization). In a majority of ***GEMA*** computations when selection is turned on (number of offspring is >2), the $\pi_{s=0}^{kimura}$, denoting the probability of fixation of neutral mutations (s=0), follows Kimura's law and is very close to $1/2N$. In other words, the product of three of our parameters $\pi_{s=0}$, α, and N approximates to 1 ($\pi_{s=0} \times \alpha \times N \cong 1$). However, for a specific set of parameters, $\pi_{s=0}^{kimura}$ significantly deviates from 1/2N. For example, for (α =10, h=0, r=1, D=expC, μ=1, N=50) the product of $\pi_{s=0} \times \alpha \times N$ equals 2.13 instead of being equal to 1. For another set of conditions (α =10, h=1, r=1, D=expC, μ=1, N=50) the product of $\pi_{s=0} \times \alpha \times N$ equals 0.85 (for details see Tables S1 and S2). The anomalies from Kimura's law resulted from numerous mutations in individuals being linked together as multiple haplotypes from various genomic loci (because r is low). Neutral mutations are linked with non-neutral ones and all mutations within a haplotype are selected as a whole unit. The length of haplotypes is in the reverse proportion to the recombination rate (r). The data from Figure 5 demonstrate that the highest deviations of $\pi_{s=0}$ from neutrality law were observed for the lowest recombination rate when r=1.

The size of a population considerably influences the fixation probabilities π_s in such a way that the average fitness of the population always improves via increasing its size (N). Tables S1 and S2 demonstrate how a growth of N changes π_s values for deleterious and beneficial mutations. ***GEMA*** simulations are in concordance with the well-known observations

that deleterious and neutral mutations have a higher chance to be fixed in small populations due to random drift (Small et al. 2007). We also observed that the rate of fixation for beneficial mutations depends on **N**. Yet, the change of $\boldsymbol{\pi}_{s>0}$ with respect to ***N*** was much lower than that observed for deleterious and neutral mutations.

After Haldane publication in 1927, it is generally accepted that the probability of fixation of beneficial mutations ($\boldsymbol{\pi}_{s>0}$) in large populations should be twice greater than ***s*** ($\boldsymbol{\pi} \cong 2\boldsymbol{s}$) (Haldane 1927; Patwa and Wahl 2008; Charlesworth and Charlesworth 2010; Chelo et al. 2013). This formula was mathematically derived through consideration of branching Galton-Watson process of chance extinction of a new mutation in a stationary population where individuals have Poisson distributed number of offspring with variance equal to 1. However, our ***GEMA*** results demonstrate that the probability of fixation of a beneficial mutation $\boldsymbol{\pi}$ also notably depends on the combination of the six aforementioned parameters (***N, μ, r, h, α, D***). This phenomenon can be explained by the linkage of deleterious, beneficial, and neutral mutations within haplotypes and selecting them as whole units. The most dramatic example of such linkage is presented in the Figure 1D. It shows a computational experiment for recessive genes ($\boldsymbol{h}$=1) with low level of recombination ($\boldsymbol{r} \leq 1$). In this model, beneficial mutations happen to occur spontaneously in a small fraction of genes. Let's consider one of such genes, denoted by ***A***. We further assume that ***A*** acquired beneficial mutations by chance that are on their way for a rapid fixation. At the same time, neighboring genes (let's call them ***B*** and ***C***) are likely to gradually accumulate deleterious mutations (which are more abundant than beneficial ones in our experiments). The mutations in all neighboring genes ***A***, ***B***, and ***C*** are linked together within a single haplotype because recombination rate is set to be low. Under a recessive mode of dominance, the effect of deleterious mutations in ***B*** and ***C*** is negligible until their frequency is low. These linked

beneficial and deleterious mutations are long trapped as clustered SNPs that can neither be easily fixed nor drifted away. The increase of this haplotype frequency causes a prevalence of negative effects on fitness from genes ***B*** and ***C***, averting the fixation of all mutations within this haplotype. On the other hand, a decrease of frequency of this haplotype causes a significant decline in the negative effects from genes ***B*** and ***C***. So in this case, the positive effects of beneficial mutations in gene ***A*** start prevailing and thereby forestall the complete loss of this particular haplotype. Thus, such specific combinations of parameters ($r \leq 1$, $h = 1$) can cause a dramatic instability of the number of SNPs as observed in ***GEMA*** computations. Peculiarities of such unstable SNP dynamics can be observed in either the gigantic peaks of SNPs numbers (Figure 1D) or in the gradual accumulation of SNPs with severe fluctuations (the latter occurs when ***r*** is significantly lower than 1 recombination per gamete).

Finally, using ***GEMA*** modeling, we investigated the **K/μ** ratio of the number of fixed mutations per generation (***K***) to the number of novel mutations per gamete (***μ***). Moto Kimura demonstrated that under neutral selection conditions the **K/μ** - ratio is equal to 1 (Kimura 1983). In 2008, Chen, Chi, and Sawyer (Chen et al. 2008) advanced the mathematical apparatus for the Neutral theory generalizing it for incomplete dominance ($0<h<1$), over-dominance ($h<0$) and under-dominance($h>1$) modes and characterized the effects of dominance on the probability of fixation of a mutant allele. However, mathematical models do not consider the following problems: 1) the linkage between nearly neutral mutations and beneficial/deleterious ones through formation of haplotypes that may be not neutral, 2) the selection that is carried out simultaneously on the entire pool of genes. ***GEMA*** computations have revealed that even under the influx of predominantly neutral mutations (90%, experiment B), a significant deviation of the **K/μ** - ratio from 1 may be observed. Figures 6 and 7 demonstrate that the **K/μ** - ratio depended

on all of the considered parameters (***N, μ, r, h, α, D***). These results show that the ***K/μ*** ratio varied from 2.5 to 0.78, under realistic conditions for human population. In experiment C, with less neutral mutations, the deviations of ***K/μ*** ratio from 1 are significantly higher.

Discussion

The ultimate goal of our ***GEMA*** project is to make a computational model for the evolution of human genome at as close to natural conditions as possible. A major challenge for such simulations is the gigantic size of the genome. Processing this entity of more than three billion nucleotides is possible only on advanced supercomputers running for many days. Hence, at this initial stage of ***GEMA*** project we take a portion of the human DNA sequence (which may be a considerable section or even a whole chromosome) and assume it to be the entire genome of our virtual individuals. Other computation simulations have also conceived a large chromosomal segment modeling an entire genome (Chadeau-Hyam et al. 2008; Kiezun et al. 2013). During these previous simulations the authors considered the same number of mutations and meiotic recombinations in the modeling genome as their particular chromosomal segment has in reality. Such an approach ignores the existence of vast majority of other mutations that constantly occur in other chromosomes and which may interfere with the modeling chromosomal segment. In this respect, our ***GEMA*** approximations are completely opposite. Inside the modeling chromosomal segment, which we consider as the genome of virtual individuals, we introduce the entire influx of mutations and meiotic recombination events that are observed for the whole human genome. Our approach ignores the existence of a majority of genes. However, in numerous computational experiments we demonstrated that the exact number of genes (like 600 versus 6,000 genes) or gene length (like 1000 nts versus 10,000 nts) do not influence the main results in our focus such

as the fitness of individuals and the number of SNPs in the population during evolution. This observation inclines us to think about the fruitfulness of our approach for the assessment of the recombination and mutation rates on the fitness and mutation dynamics. In ***GEMA*** simulations the selection and evolution are implemented simultaneously on gigantic ensembles of mutations that are regrouped in every individual due to multiple meiotic recombination events. Such modeling may reveal unknown features in dynamics of mutations, which we plan to present in the next publications.

In this paper we primarily focused on the impact of meiotic recombinations on the population fitness. Our computer simulations demonstrated that an increase in the number of recombination events per gamete considerably improves the fitness of the population via increasing the probability of fixation of beneficial mutations and simultaneously decreasing the probability of fixation of deleterious mutations. This behavior is in accordance with the fundamentals of classical population genetics that acknowledge "the evolutionary advantage of recombination" (Felsenstein 1974) and, in particular, the Hill-Robertson effect. However, the Hill-Robertson effect is rather a qualitative estimation showing recombination driven enhancement of a population's ability of fixation of favorable mutations. Textbooks on this topic do not provide quantitative estimations on how a specific change in recombination frequency impacts the probability of fixation of favorable mutations (Hartl and Clark 2006; Durrett 2008; Charlesworth and Charlesworth 2010). The advantage of ***GEMA*** simulations lies in its ability to precisely measure the effect of a particular recombination rate (***r***-parameter) on the population fitness and probability of fixation of mutations with different selection coefficients. For example, let's consider the results from the supplementary Table S2 for a chosen set of parameters: (N=100; α=5; h=0.5; μ=5; and the distribution of selection coefficients

as in Fig 2C;). When the recombination rate was set to r=1, the probability of fixation of neutral mutations was $\pi_{(s=0)}$ =0.0022, beneficial ones was $\pi_{(s=+1)}$ =0.0082, and deleterious - $\pi_{(s=-1)}$ =0.00048. The increase of the recombination rate up to r=48 elevated the probability of fixation of beneficial mutations 2.7 times to $\pi_{(s=+1)}$ =0.022 and simultaneously reduced the probability of fixation of deleterious mutations 40 times to the $\pi_{(s=-1)}$ =0.000012, while the probability of fixation of neutral mutations was marginally changed ($\pi_{(s=0)}$ =0.00205). Moreover, ***GEMA*** simulations demonstrated that the elevation of the influx of mutations also had a dramatic effect on the probability of fixation. For instance, doubling mutation rate to μ=10 while keeping the same parameters as described above (N=100; α=5; h=0.5; and r=48) caused the decrease in probability of fixation of beneficial mutations 1.7 times to $\pi_{(s=+1)}$ =0.013 and simultaneously increased the probability of fixation of deleterious mutations 6.2 times to $\pi_{(s=-1)}$ =0.000074. When we quadrupled the mutation rate to μ=20, the $\pi_{(s=+1)}$ became equal 0.0078 while $\pi_{(s=-1)}$ equaled to 0.00027. This example illuminates the ability of GEMA simulations to evaluate the total effect of thousands of deleterious, beneficial and neutral mutations under different conditions (gene dominance modes, recombinaition rates, population size, mating schemes, selection pressure, and various distribution of mutations by selection coefficients).

Intricate dynamics of mutations in genomes depends on numerous parameters of a different nature including those that determine the following biological processes: 1) Level of selection pressure (number of offspring per individual and non-randomness in formation of next generation from these offspring); 2) Genetic drift (mainly determined by the population size); 3) Population structure (e.g. mating schemes, population subdivision, migrations, inbreeding); 4) Genome structure and functioning (number and arrangement of genes, number of meiotic

recombinations per gamete, distribution of dominance coefficients among genes, etc.); and 5) Mutation characteristics (number of novel mutations per gamete, distribution of these mutations by their selection coefficients, arrangement of mutations along genome, possible "mechanistic" fixation bias, etc.). In this introduction paper on ***GEMA***, we considered only six parameters (***N, μ, r, h, α, D***) and demonstrated that their specific combinations intricately and dramatically affect the fixation probability and fitness. Our multiple experiments with ***GEMA*** have confirmed that the probability of ultimate mutation fixation π_s, fitness of individuals, and the number of SNPs in the modeling population practically do not depend on the length of genes (***L***) and the number of genes (N_{genes}) in the genomes when $N_{genes} >> \mu$ and $N_{genes} >> r$. To increase the speed of computations, our presented data were obtained for N_{genes} =600 and ***L*** = 1000 nucleotides settings. Yet, these results should be the same as for the entire human genome (N_{genes} ≈25,000 and ***L***≈35,000 bp). In other words the total number of mutations and recombinations per individual and not the density of those mutations and recombinations per genomic length are important for dynamics of numerous mutations in population.

We performed our computations using the core version of ***GEMA*** (***GEMA***_*r1.pl*). In these simulations we did not use real mutation distributions in respect to the local nucleotide context or real gene sequences because they do not influence the main focus of this paper, which is towards finding important parameters that preserve population fitness under intense influx of mutations. For other queries that require mimicking biological reality with much closer proximity, the extended version of ***GEMA*** (***GEMA***_*r01.java*) should be used. It has many advanced features described on the web (http://bpg.utoledo.edu/~afedorov/lab/GEMA.html). For example, the input of our program is real chromosomal DNA of mammals, on which positions of genes and functional elements are tabulated in input matrices. Then, mutations that are modeled

by the program have the same frequencies and distributions as those observed in nature and computed from the SNP databases. Positions and frequency of modeled meiotic recombinations are also taken from the public databases describing these events (HapMap, NCBI (Frazer et al. 2007)). ***GEMA***_*r01.java* has several advanced features already build in including the availability of multiple environment option where each mutation is assigned a selection coefficient vector $\vec{S}$ with coordinates representing scalar ***s***-values specific for each environment. We provide extensive training web pages regarding the usage of ***GEMA*** programs and have a strong commitment to help the scientific community in maximizing their preferred workflows. However, the usage of ***GEMA***_*r01.java* is computationally consuming and often requires supercomputer power, which we are unable to provide. ***GEMA***_*r01.java* can be applied to the investigation of many specific questions related to the fields of Genomics and Population Genetics. Our lab is focused to use ***GEMA*** for verification of alternative ideas about the evolution of specific genomic regions (isochores, third codon positions, etc.) and for investigation of genomic pattern formation and evolution.

Materials and methods

The simplest scheme of ***GEMA*** is demonstrated in Figure 8 and its major steps are outlined below.

A) Genomes and individuals. A large portion of a real genomic sequence (even whole chromosomes of human or other species) can be assigned as a reference genome for a model population. A user specifies the number of individuals in the population (***N***). Each individual is constructed as a diploid genome that descended as two haploid gamete genomes from its parents.

B) Mutations. Taking a user-defined parameter **μ** (number of novel mutations per gamete) ***GEMA*** creates mutations in the genomic sequences using random number generator for choosing mutation positions. The relative frequencies of different types of mutations (e.g. T -> C, or G -> C) can be defined in an input table that approximates the observed frequencies in nature and can also take into account the local nucleotide context (option available for ***GEMA***_*r01.java*). Upon generation of a mutation, ***GEMA*** assigns a selection coefficient (***s*** parameter) to the mutation using a user-defined ***s***-distribution. Note that ***s***-values are not normalized (see also ***GEMA*** user guide in Supplementary file S1). In the advanced version of the program (***GEMA***_*r01.java*) different genomics elements (exons, introns, ncRNA, conserved elements, etc.) may have their own specific ***s***-distributions.

C) Meiotic recombination and gametogenesis. Haploid genomes of gametes are generated for each virtual individual from its parents' chromosomes. The number of meiotic recombinations between parents' chromosomes is an input parameter (***r***). The recombination sites are defined by a random number generator, which can take into account the "hot-spots" and "cold-spots" for recombination events from the International HapMap Consortium genetic maps (option available for ***GEMA***_*r01.java*).

D) Computation of a new generation of virtual individuals. Different mating schemes for virtual individuals are possible as input options. By default we use random permanent pairings between male and female virtual individuals. Their offspring have diploid genomes formed by two randomly chosen parents' gametes. The number of offspring per individual (***α***) is a user-defined input parameter.

E) Selection. The overall fitness of every created virtual individual in the ***GEMA*** algorithm is computed by taking into account all the mutations possessed by this individual. The current

version of ***GEMA*** does not take into account mutual effects of mutations such as compensatory mutations and epistasis. ***GEMA*** calculates fitness for each gene by summing all the ***s***-values of mutations within that gene. For example, assume that for a human gene, its maternal allele is composed of a particular haplotype containing ***x*** number of SNPs and its paternal allele is composed of a different haplotype comprising ***y*** number of SNPs. The fitness of the maternal allele for the given gene (w_m) will be a sum of ***s***-values for all the ***x*** SNPs within this gene, while the fitness of the paternal allele (w_p) will be a sum of s-values for all the ***y*** SNPs. The fitness of the gene in this example is calculated from the w_m and w_p values and also another input parameter, the dominance coefficient (***h***). In a co-dominance mode (***h***=0.5), the gene fitness is the average of the fitness of maternal and paternal alleles. Under a recessive mode (***h***=1), which corresponds to recessive genes, the fitness is the maximum between w_m and w_p values (heterozygotes with one deleterious allele are healthy), while for a dominant mode (***h***=0), which corresponds to dominant genes, the gene fitness is the minimum between w_m and w_p values. For a general case, the gene fitness is calculated by the formula: $w = \min(w_m, w_p) + \boldsymbol{h}*\mathrm{abs}(w_m - w_p)$. Finally the overall fitness of the virtual individual is the sum of fitnesses of all genes inside the genome. In the selection phase of ***GEMA*** algorithm, the program picks the ***N*** fittest offspring and forms from them the new generation. This new generation replaces the previous one and the entire cycle repeats for a user-defined number of generations.

GEMA regularly outputs the following major parameters: Current generation, total fitness of the population, number of SNPs, total number of fixed mutations ($\boldsymbol{F_s}$) and total number of mutations ($\boldsymbol{C_s}$) with selection coefficient ***s***. In addition, all genotypes of each individual are stored in the backup files A and B and can be easily retrieved (see Supplementary file S1).

A detailed description of ***GEMA*** algorithm is presented in the "***GEMA***_Instructions.pdf" available from our web page: http://bpg.utoledo.edu/~afedorov/lab/GEMA.html while a copy of it is presented in the Supplementary file S1.

The programming codes for ***GEMA***_*r01.java* ***GEMA***_*r1.pl* and pseudo-codes are freely available from our Lab web site http://bpg.utoledo.edu/~afedorov/lab/GEMA.html. Our Lab web pages also have extensive explanations via video demonstrations. A discussion board has also been arranged for a broader public community to share experiences and concerns.

ACKNOWLEDGEMENTS

We are grateful to Dr. Ashwin Prakash, Johns Hopkins School of Medicine, for his insightful discussion of the project. The computations were performed in Oakley supercomputer with support from Ohio Supercomputer Center. This work is supported by the National Science Foundation Grant MCB-0643542 (to A.F).

References.

Abecasis GR, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

Bechtel JM, et al. 2008. Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. Bmc Genomics 9: 284.

Bernardi G. 2007. The neoselectionist theory of genome evolution. Proceedings of the National Academy of Sciences of the United States of America 104: 8385-8390.

Bodmer WF, Felsenstein J. 1967. Linkage and selection: theoretical analysis of the deterministic two locus random mating model. Genetics 57: 237-265.

Carvajal-Rodriguez A. 2008. GENOMEPOP: A program to simulate genomes in populations. Bmc Bioinformatics 9.

Carvajal-Rodriguez A. 2010. Simulation of Genes and Genomes Forward in Time. Current Genomics 11: 58-61.

Chadeau-Hyam M, et al. 2008. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9.

Charlesworth B, Charlesworth D. 2010. Elements of Evolutionary Genetics: Roberts and Company Publishers.
Chelo IM, Nedli J, Gordo I, Teotonio H. 2013. An experimental test on the probability of extinction of new genetic variants. Nature communications 4: 2417.
Chen CT, Chi QS, Sawyer SA. 2008. Effects of dominance on the probability of fixation of a mutant allele. Journal of mathematical biology 56: 413-434.
Conrad DF, et al. 2011. Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714.
Consortium TIH. 2005. A haplotype map of the human genome. Nature 437: 1299-1320.
Durrett R. 2008. Probability models for DNA sequence evolution. New York: Springer.
Felsenstein J. 1974. The evolutionary advantage of recombination. Genetics 78: 737-756.
Fisher RA. 1930. The Genetic Theory of Natural Selection. Dover: Oxford University Press.
Frazer KA, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
Gavrilets S, Hastings A. 1994. Dynamics of genetic variability in two-locus models of stabilizing selection. Genetics 138: 519-532.
Haldane J. 1927. A Mathematical Theory of natural and artificial selection, part V: selection and mutation. Math. Proc. Cambridge Phil. Soc. 23: 838-844.
Hartl D, Clark A. 2006. Principles of Population Genetics, Fourth Edition: Sinauer Associates, Inc.
Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786-2787.
Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genetical research 8: 269-294.
Kaessmann H, Wiebe V, Weiss G, Paabo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nature genetics 27: 155-156.
Kiezun A, et al. 2013. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS genetics 9: e1003301.
Kimura M. 1983. The neutral theory of molecular evolution.
Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47: 713-719.
Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Human molecular genetics 11: 669-674.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.
Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. Annual review of genomics and human genetics 11: 265-289.
Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theoretical Population Biology 49: 128-142.
Patwa Z, Wahl LM. 2008. The fixation probability of beneficial mutations. Journal of the Royal Society, Interface / the Royal Society 5: 1279-1289.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 12: 32-42.
Prakash A, et al. 2009. Evolution of Genomic Sequence Inhomogeneity at Mid-range Scales. Bmc Genomics 10: 513.
Sanford J. 2008. Genetic entropy and the mystery of the genome: FMS Publications.

Sanford J BJ, Brewer W, Gibson P, Remine W. 2007. Mendel's Accountant: A biologically realistic forward-time population genetics program. SCPE 8: 147-165.
Small KS, Brudno M, Hill MM, Sidow A. 2007. Extreme genomic variation in a natural population. Proceedings of the National Academy of Sciences of the United States of America 104: 5698-5703.
Stephan W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365: 1245-1253.
Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nature reviews. Genetics 9: 965-974.
Wolfram S. 2002. A New Kind of Science. Champain, IL: Wolfram Media Inc. .
Wright S. 1965. Factor Interaction and Linkage in Evolution. Proc. R. Soc. Lond. B 162: 80-104.
Zhang C, Plastow G. 2011. Genomic Diversity in Pig (Sus scrofa) and its Comparison with Human and other Livestock. Current Genomics 12: 138-146.

Figures

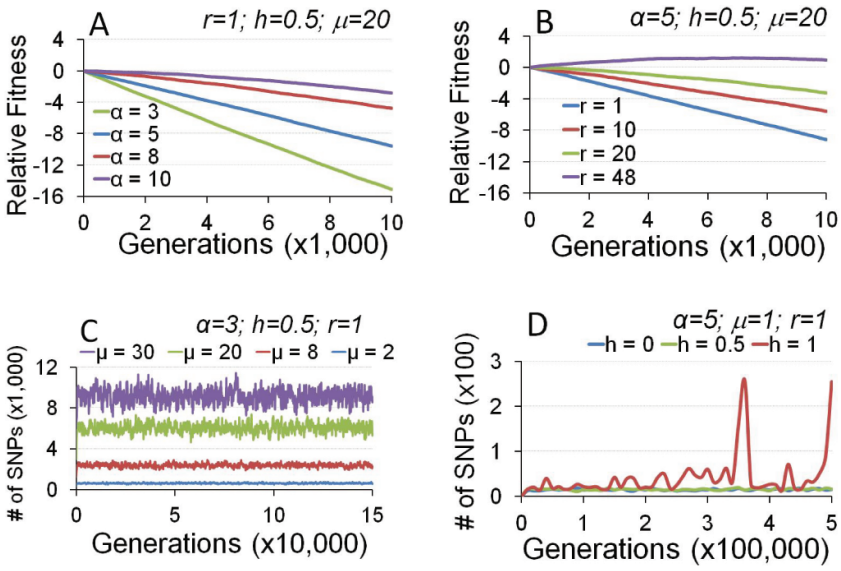

Figure 1. Exemplification of results from ***GEMA_r1.pl*** and ***GEMA_r01.java***, illustrating evolutionary computations for 50 virtual individuals, each of whose genome is represented by human chromosome 22. **A** and **B** represent the change of relative fitness of individuals in population with respect to time (generations). In this modeling, we defined the distribution of mutations as a decay curve of selection coefficient (***s***), where 88% of mutations have negative ***s***-values and only 12% have positive ***s***-values (see Figure 2A). We do not normalize selection coefficient values, so the illustrated fitness of individuals is presented in relative units. Negative values of relative fitness show a decline in organism adaptability while positive values indicate improvement. In these computational experiments, genes were assigned co-dominance mode (***h***=0.5). Figure **A** demonstrates how different numbers of offspring per individual (***α*** = 3, 5, 8, or 10 offspring) influence the relative fitness, under the same recombination rate (***r***=1). Figure **B** demonstrates how different numbers of recombination events per gamete (***r*** = 1, 10, 20, or 48) affect the relative fitness while the number of offspring remained constant (***α***=5). **C** and **D** illustrate the dynamics of number of SNPs in the population. Figure **C** shows variations in the

number of SNPs with respect to generations for four different values of novel mutations per gamete (μ =2, 8, 20 or 30). Figure **D** demonstrates smoothed number of SNPs (by taking averages for extended number of generations) in addition to emphasizing that under specific conditions (e. g. recessive genes in which the dominance mode h is close to 1) there may be considerable and long-lasting spikes in the number of SNPs when recombination rate is low ($r \leq 1$).

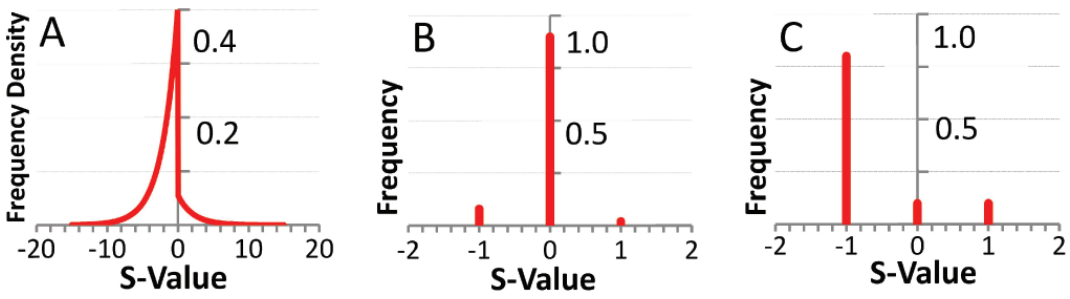

Figure 2. Distributions of mutations by user-assumed selection coefficients (***s***-values), which were used for modeling analysis. **A -** Represents a continuous distribution of mutations by ***s*** that can range from -20 to +20 depending on their deleterious (negative ***s***-values) or beneficial (positive ***s***-values) effects. This curve represents 88% deleterious and 12% beneficial mutations. **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 -** illustrates another discrete distribution for mutations, where 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.

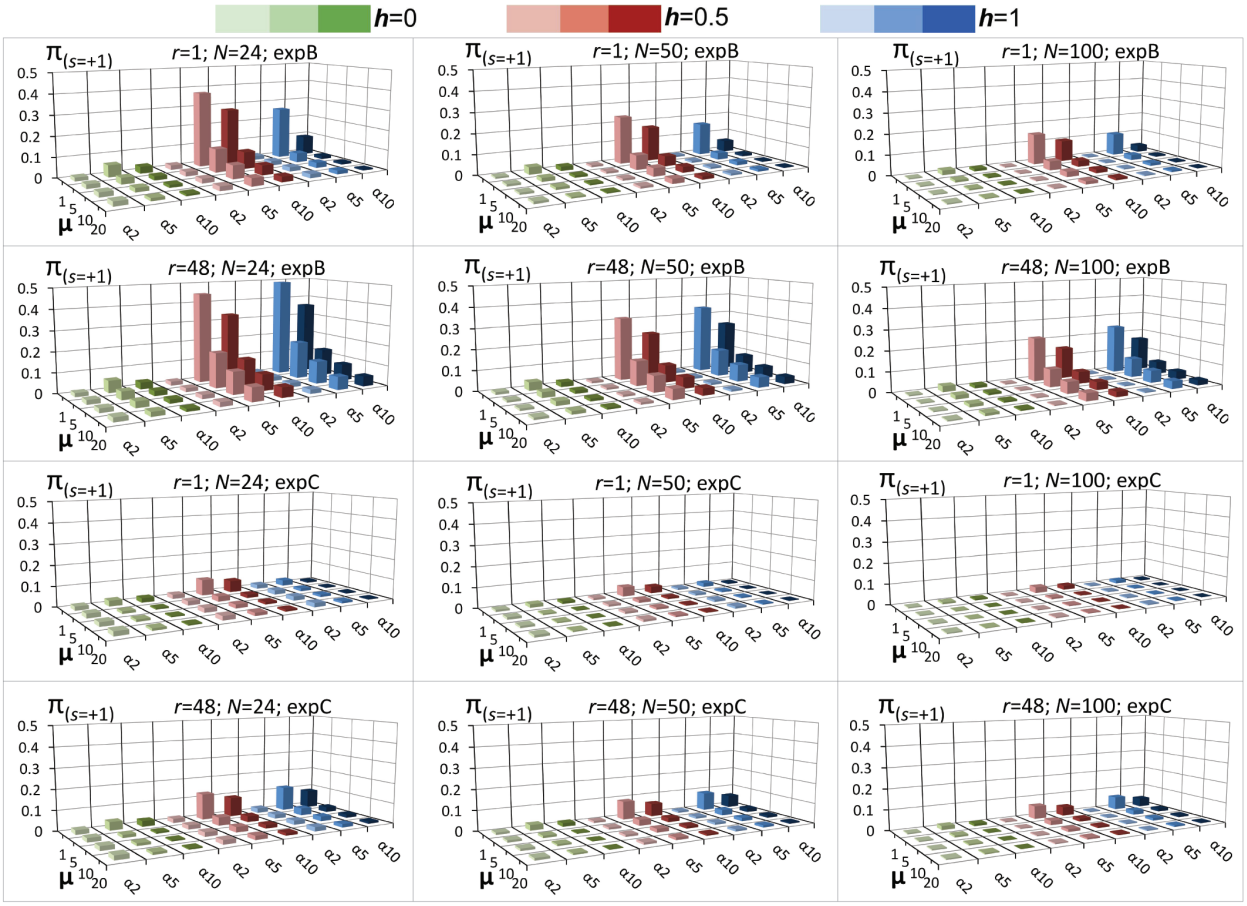

Figure 3. Dependence of the probability of fixation π_s of mutations with beneficial effects. The effects of mutations have been illustrated in our model according to selection coefficient **s** exemplified by values of +1, 0 and -1 for beneficial, neutral and deleterious mutations respectively. Individual 3D plots demonstrate the quantitative behavior of fixation of mutations as interplay of different parameters represented by population size (**N**), recombination rate (**r**), variations in influx of novel mutations (**μ**), mode of dominance (**h**), number of off springs (**α**) and predominance of either neutral mutations (according to Figure 2B) or deleterious mutations (according to Figure 2C). Exact values of all parameters are provided in Supplementary Tables S1 and S2.

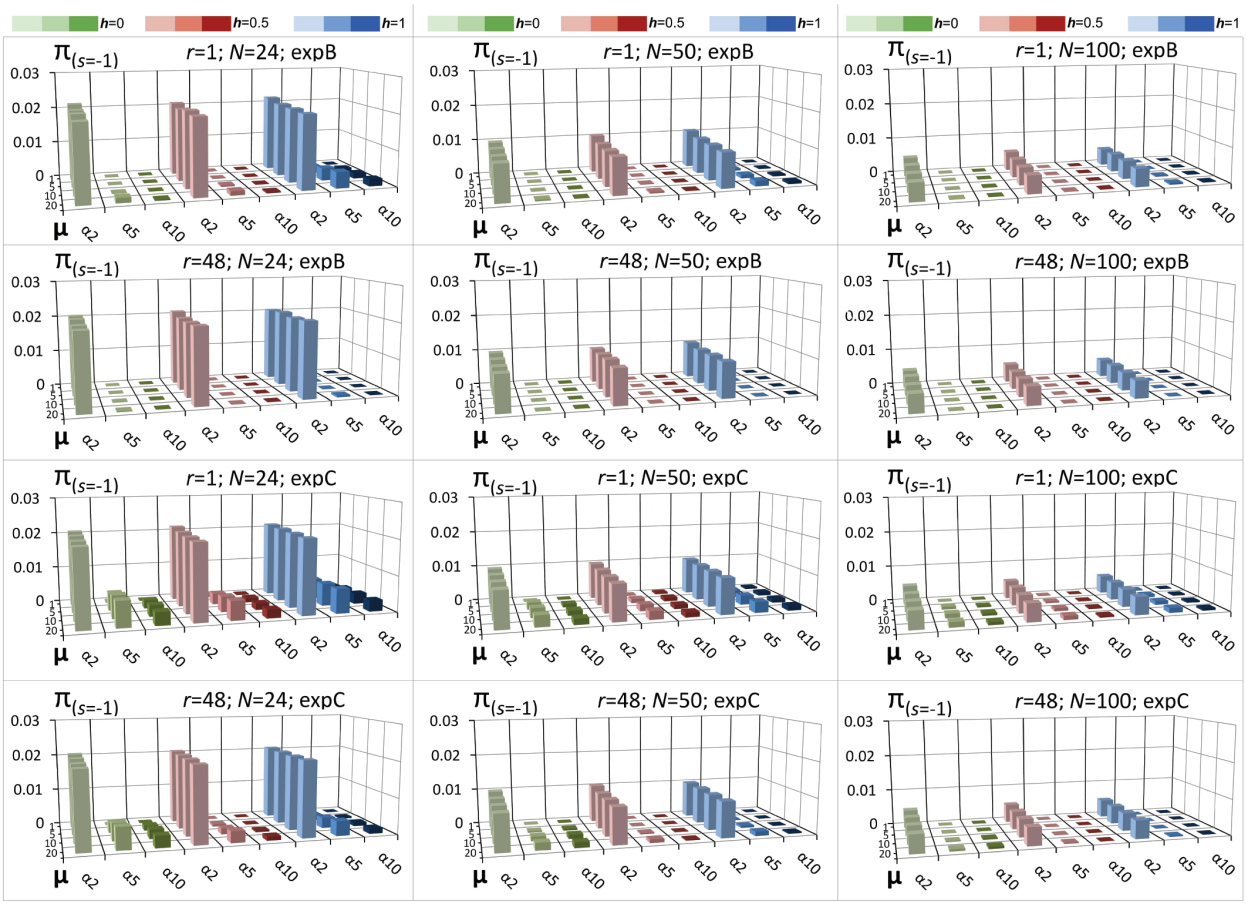

Figure 4. Dependence of the probability of fixation π_s of mutations with deleterious effects (s=-1). All parameters are the same as in Figure 3.

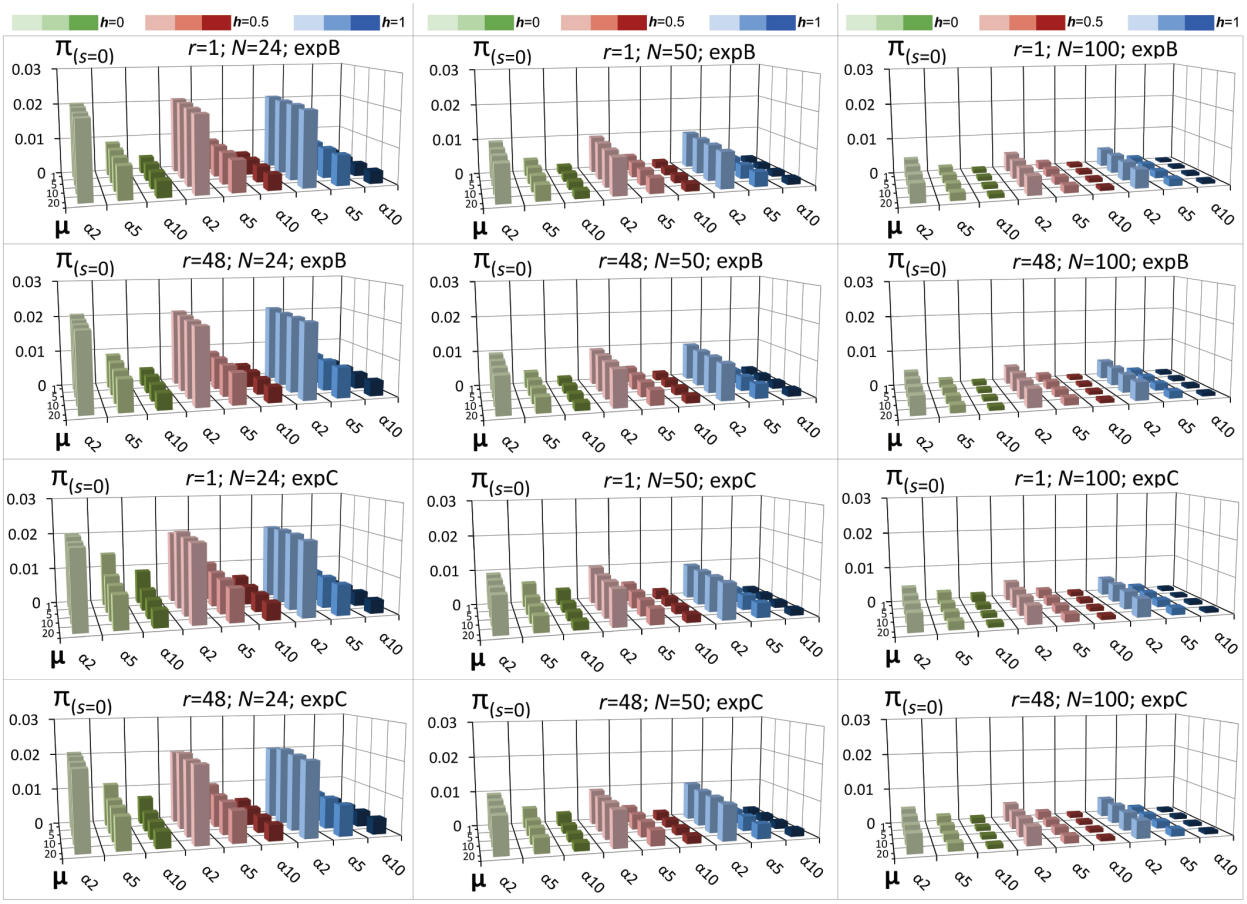

Figure 5. Dependence of the probability of fixation π_s of mutations with neutral effects (s=0). All parameters are the same as in Figure 3. Note that for comparison of these π values with Kimura's law, they should be normalized by taking into account the number of offspring per individual as described in the Results section ($\pi_s^{kimura} = \pi_s \times \alpha/2$).

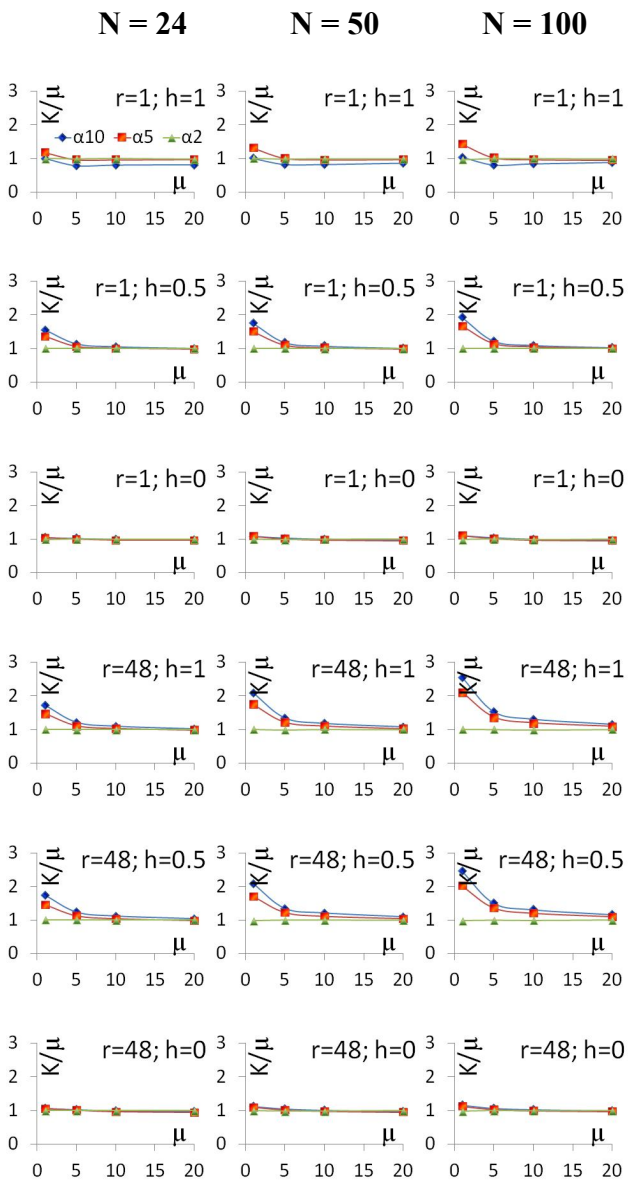

Figure 6 Graphical illustrations of deviations of ***K/μ*** ratio from 1 with respect to change of number of novel mutations per gamete (μ) for particular sets of parameters (***N, r, h, α, D***). ***K*** stands for the number of fixed nucleotides in each generation while ***μ*** is the number of novel mutations per gamete. The graphs are obtained on the basis of predominant pool of neutral mutations, modeled by experiment B for ***s***-distribution (see Figure 2B). Within each graph, variations in the ratio of ***K/μ*** have been calculated for varying number of offspring (***α***) within the

population (green α=2; red α=5; blue α=10). ***In toto***, the interplay of various parameters such as recombination rate (***r***), dominance coefficient (***h***), population size (***N***), novel mutations per gamete (μ), number of offspring (α) and overall effect of mutation pool (deleterious, beneficial or neutral) have been represented as causal factors for deviations from previously assumed unitary ratio of K/μ.

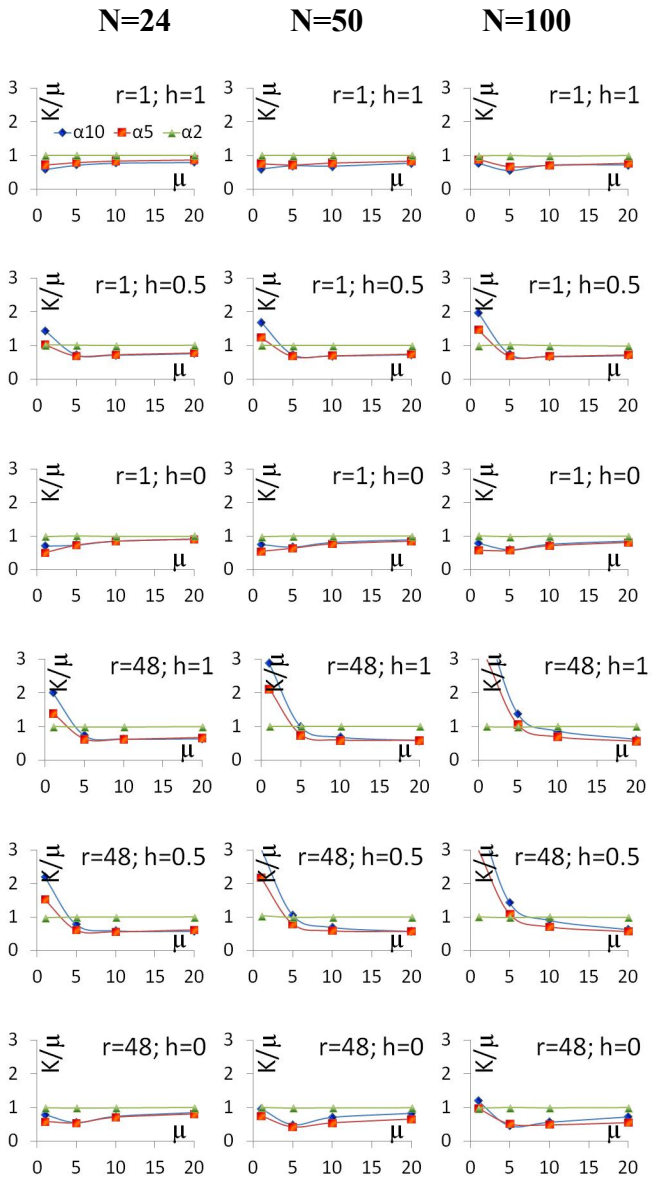

Figure 7 Graphical illustrations of deviations of ***K/μ*** ratio from 1 with respect to change of number of novel mutations per gamete (μ) for particular sets of parameters (***N, r, h, α, D***). The graphs are obtained on the basis of a prevalence of deleterious mutations, quantified by experiment C (see Figure 2C). All parameters are the same as in Figure 6.

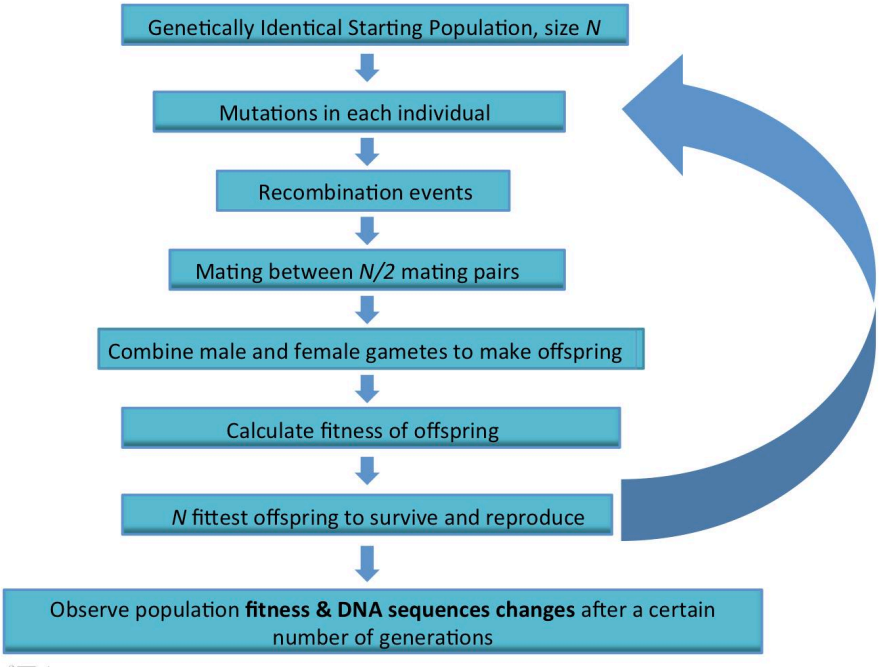

Figure 8. ***GEMA*** begins with a genetically identical population of size ***N***. Genomic mutations occur in each individual, which are passed onto offspring. According to the mutations inherited, fitness is calculated for each offspring. The ***N*** fittest offspring become the next generation and the process repeats for thousands of generations. Additional details on ***GEMA*** are provided in the Materials and Methods section, Supplementary file S1 (GEMA User Guide), and our ***GEMA*** web page.