A Dissertation

entitled

Computational Simulation and Analysis of Mutations: Nucleotide Fixation, allelic age and rare genetic variations in population

by

Shuhao Qiu

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Doctor of Philosophy Degree in

Biomedical Sciences

_____

Alexei Fedorov, PhD, Committee Chair

_____

Robert Blumenthal, PhD, Committee Member

_____

Robert J. Trumbly, PhD, Committee Member

_____

Sadik A. Khuder, PhD, Committee Member

_____

Nikolai Modyanov, PhD, Committee Member

_____

Patricia R. Komuniecki, PhD, Dean
College of Graduate Studies

The University of Toledo

May 2015

An Abstract of

Computational Simulation and Analysis of Mutations: Nucleotide Fixation, Allelic Age and rare Genetic Variations in population

by

Shuhao Qiu

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in
Biomedical Sciences

The University of Toledo

May 2015

In order to investigate the complexity of mutations, a computational approach named Genome Evolution by Matrix Algorithms (*GEMA*) has been implemented. GEMA models genomic changes, taking into account hundreds of mutations within each individual in a population. By modeling of entire human chromosomes, GEMA precisely mimics real biological processes that influence genome evolution, and demonstrates that the number of meiotic recombination events per gamete is among the most crucial factors influencing population fitness. GEMA was further modified and employed in a study of genome evolution to re-evaluate Maruyama's phenomenon in modeled populations, which include haplotypes approximating real genomes. It was determined that only under specific conditions, of high recombination rates and abundance of neutral mutations, were deleterious and beneficial mutations younger than the neutral ones as predicted by Maruyama. Under other conditions, the ages of negative, neutral, and beneficial mutations were almost the same.

After simulating mutations in a population, actual human genome sequence data from the "1000 Genome Project" Phase I was analyzed. All detected nucleotide sequence differences for 1092 people from 14 populations were computed. The distribution of these differences in individuals were then characterized on basis of their origin (European, Asian or African). By analysis of this genetic information of individuals, the very rare genetic variants were found to largely improve the detection of familial relations. Thus, with affordable whole-genome sequencing techniques, very rare SNPs should become important genetic markers for familial relationships and population stratification.

To my dad, Zehong Qiu, you helped me to make a wise decision in coming to US.

To my mum, Lazhuang Qiu, I know you know–your son will get the US doctorate degree. Be proud of it.

To my younger brother, Qunying Qiu, without your love and support I could not finish.

To all my previous English teachers, especially the great ones.

To all the difficulties I have met until now, without you my life would be boring.

To the Goddess of Mercy, everything was possible because of you.

Acknowledgements

I would like to express my gratitude to my advisor, Dr. Alexei Fedorov, who gave me great insights of research with tremendous patience and spending time for fixing my poor writing. You never discourage me. I want to thank my colleagues, Arnab Saha Mandal and Ahmed S Al-Khudhair for all your help. I want to thank my friend, Julie Thomas with great discussion and inspiration.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AMR .....Total Americas Ancestry
AFR .....Total African Ancestry
ASN .....Total East Asian Ancestry
ASW .....African Ancestry in Southwest US

CEU .....Utah residents with Northern and Western European ancestry
CHB .....Han Chinese in Beijing, China
CHS .....Southern Han Chinese, China
CI .....Confidence Interval
CLM .....Colombian in Medellin, Colombia

EUR .....Total European Ancestry
exp .....Experiment

FIN .....Finnish in Finland

GEMA.....Genome Evolution by Matrix Algorithms
GBR .....British in England and Scotland

IBD.....identically by descent
IBS .....Iberian population in Spain

JPT .....Japanese in Tokyo, Japan

LWK .....Luhya in Webuye, Kenya

Mb .....Mega Bases; A million Bases
MXL .....Mexican Ancestry in Los Angeles, California

ncRNA .....non-coding RNA
nts .....Nucleotides

PUR .....Puerto Rican in Puerto Rico

SNPs.....Single Nucleotide Polymorphisms

TSI .....Toscani in Italy

vrGVs.....very rare Genetic Variations

YRI .....Yoruba in Ibadan, Nigeria

List of Symbols

Δl .....the size of the IBD segment
μ.....Number of novel mutations per gamete
π_s .....the probability of fixation of a novel mutation with the selection coefficient ***s***

α .....Number of offspring per individual
C_s .....Total number of mutations with selection coefficient *s*
C% .....Percentage of common genetic materials
h .....Dominance Coefficient
K .....Number of fixed mutations per generation
L .....Length of genes; Size of haploid genome
N .....Number of individuals in the population
N_{genes} .....Number of genes
N_e .....Effective size of the population
F_s .....Total number of fixed mutations
r .....Meiotic Recombination between parents' chromosomes
s.....Selection Coefficient
w_m.....Fitness of maternal allele for a given gene
w_p .....Fitness of paternal allele for a given gene

Chapter 1

Introduction

A mutation in biology refers to any stable change in the nucleotide sequence of the genome of an organism, ranging from one single nucleotide to larger chromosomal changes. Mutations can be silent, or can cause changes in the DNA sequence that affect the expression or activity of a gene or protein product. Mutations can have different effects on an organism, ranging from subtle to drastic. Based on their contribution to fitness, mutations can be viewed as deleterious, beneficial or neutral. Deleterious mutations decrease the fitness of an organism or, in another way of stating it, may cause a disease; beneficial mutations increase the fitness of the organism; while a neutral or nearly neutral mutation will not affect an organism's ability to survive and reproduce (Sawyer et al.; Burrus and Waldor 2004; Aminetzach et al. 2005).

The effect of most mutations on each individual do not happen alone. Many genes and therefore mutations can affect a trait, thus a mutation in a given gene might or might not affect a trait. The quantitative nature of the effects of a mutation are represented finally in the expression of the phenotype or trait. A slightly deleterious mutation may improve net fitness if combined with a strongly beneficial one. A lot of slightly deleterious mutations present in the non-functional area in the genome may result in

subtle or even no difference in the phenotype. Thus, in an individual, it is the combinatorial effects of many of genes and their respective mutational changes that result their unique expression of health or genetic disease. In terms of population, these mutational effects contribute to the different identities with various levels of fitness (Suzanne Estes et al. 2004).

By observation of a trait, it is difficult to estimate how many mutations cause an effect in an individual. In a comparison between two individuals based on their phenotypes, it is impossible to estimate the number of genetic variations between them. A novel mutation in one individual does not necessarily mean it is also new to the other individuals. Thus, when a mutation is viewed in a population (a group of individuals) instead of an organism, the problems regarding mutations become complex. First, the distribution and frequency of a mutation has to be taken into account. Also, other factors, like recombination, population structure, natural selection, and genetic drift will affect mutations in a population. For example, according to one study, two children of different parents had 35 and 49 new mutations relative to the parents. Of these, in one case 92% mutations were from the paternal germ line, and in the other case, 64% were from the maternal germ line (Donald F Conrad 2011).

The role of mutations in a population is thus complex phenomenon. This complexity is further increased by the number of potential mutations. It should be noted that, the amount of genetic variations created by mutations between two individuals is large, even though they might come from the same population. Further it is not difficult to imagine that when we consider a large group of people ---- a population. Then the pool of genetic variations becomes very large. Furthermore, this gigantic pool is constantly

being increased by 40-100 novel mutations with every additional person entering the population (Kondrashov and Shabalina 2002; Conrad et al. 2011; Li and Durbin 2011).

While the effects of a mutation in a population are complex, the latest technologies have made it possible to uncover or understand some of the internal attributes of human diseases in relation to mutations. Although this area of discovery has progressed tremendously in past decade, inevitably there are questions that still need to be explored and answered. We have been able to bridge some gaps towards understanding the correlation between combinatorial effects of mutations and human diseases in the population. The first approach taken was to develop computational software and use it to a simulate mutation in the whole genome (the length of the human genome is about a gigabase; we used a single human chromosome). These simulations also allowed us to determine the probability of fixing a specific mutation. Furthermore, we modified our software and used it to reexamine Dr. Maruyama's allelic age theory (described below). Finally, we focused on the subset of very rare Single Nucleotide Polymorphisms (SNPs) by analyzing existing genetic data from the "1000 Genome Project". That analysis revealed that these very rare SNPs could serve as important genetic markers for familial relationships and population structure/stratification.

1.1 The Fate of Mutations

The study of the fate of a mutation when putting it into a population has long been a central question for population genetics. Basically, there are three possibilities: a new mutation can be kept in a population and maintained for a long time, it can drift away, or it can be fixed which means every individual in the population will have this mutation. A

key question that has been brought up regarding the fixation of a mutation for decades is: what is the probability of a certain mutation becoming fixed in a population?

Several mathematical models have been proposed to try to answer this question. In these models, formulas with multiple parameters are used to investigate the intricate dynamics of mutations arising in populations. However, the results of these models often conflict with one another and, until now, no universally acknowledged perception of genomic nucleotide dynamics has been discovered (Wagner 2008; Nei et al. 2010). One of the reasons for their controversial results may be due to their limitations. Most of the formulas including those very complicated ones, only consider mutations in individuals with little attention to the fact that natural selection, a major player in evolution, occurs simultaneously on entire ensembles of mutations in an organism in the population. The background selection and genetic hitchhiking models deal 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 different loci (Fisher 1930; Wright 1965; Bodmer and Felsenstein 1967; Gavrilets and Hastings 1994). These theories again, deal only with oversimplified models and to some extent, omit the mutation's complexity by not considering the vast number of mutations happening at the same time.

Mathematical modeling of this problem started in the 1980s, but more recently computational approaches began to provide a different dimension to this study. There are several published computational 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, none of these have considered multiple mutations as an entity happening at the same time in an individual as can occur in nature. Further, the simulated genomes in that software are only several thousand nucleotides long at most. Most of the time, these genomes are simulated as being haploid instead of diploid. Recombination is always considered as a Markovian process, and the recombination rate ranges from only 0 to 0.5. All of these settings limit their possibility of answering our questions: what is the fate of a mutation in a given group of individuals?

Here, we have designed and present our program, named Genome Evolution by Matrix Algorithms (or GEMA). Besides many similar features to the previously published programs, 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. Specifically, each mutation itself will have beneficial, deleterious or neutral effects on the individual's fitness, while the final fitness for such an individual is represented as a combinatorial effects of those thousands of mutations (though a simplification is that possible epistatic interactions are ignored). Selections are applied and based on the individual's fitness. The most fit individuals are able to survive and reproduce, with an adjustable bottleneck at each generation. We use our program to simulate the intricate dynamics of mutation, and demonstrate that an increase in the number of recombination events per gamete considerably increases the probability of fixing beneficial mutations; while at the same time decreasing the probability of fixation of deleterious mutations, resulting in an improvement of the overall population fitness.

1.2 The Allelic Age

Presently, with the advancement of next generation sequencing technology, more and more individual genomes are being sequenced. In public databases, there are about 3,000 whole human genomes available (as of March 2015). However, there is still a lack of genomic information across several generations in given families. This kind of information is however necessary for estimating allelic age. Previously, mathematical modeling provided important insights into this problem.

Investigations of "allelic age" can be dated back to 1970s. The term was defined as the number of generations a mutant allele has persisted in the population since its first occurrence (Kimura and Ohta 1973; Maruyama 1974a; Maruyama 1974b; Li 1975). At the beginning, a mathematical model of a diffusion approximation for a branching process was applied to predict the allelic age. In 1973 Kimura and Ohta (Kimura and Ohta 1973) inferred that the "average ages of neutral alleles, even if their frequencies are relatively low, are quite old." From their result, it is difficult for experimental verification of the allele age prediction since many alleles can be relatively ancient.

In 1974, Takeo Maruyama (Maruyama 1974a) predicted that extant neutral mutations are generally older than both deleterious and beneficial ones, based on modeling of semi-dominant mutations. Later, Wen-Hsiung Li (Li 1975) further inferred the age of deleterious mutations by modeling various degrees of dominance. He demonstrated that the mean age decreases with increasing selection coefficients against heterozygotes. In the late 1990s and early 2000s, the topic of allelic age has also been nicely reviewed (Griffiths and Tavare 1999; Slatkin and Rannala 2000).

However, all of these mathematical methods for estimating allelic age consider only one mutation at a time, while ignoring the possible interactions among the mutations

(Kimura and Ohta 1973; Maruyama 1974b; Li 1975). Since mutations never exist alone in an individual, to better study their dynamics they should be modeled and analyzed with other mutations simultaneously. For this purpose, we implemented whole-genome computational simulations to investigate the average age of a mutation under different circumstances. With the help of our GEMA simulations, it is easy to record and trace the dynamics of a mutation while putting it into an integrative network containing thousands of mutations per individual. We demonstrated that Maruyama's effect appears only for specific sets of parameter ranges and quantitatively described its variation under different conditions. Note that while we include additive effects of large numbers of mutations, and this is a significant improvement over previous simulations, we do not attempt to model the huge number of varied potential epistatic interactions.

1.3 Distant Genetic Relations unrevealed by vrGVs (very rare Genetic Variations)

Proper methods for genetic detection of familial relationships are important for forensic identification, in criminal investigations, inheritance claims, and in other areas. Genetic relatedness estimation has been mainly based on the estimated number of alleles shared identically by descent (IBD) on autosomal chromosomes (Browning and Browning 2013; Huff et al. 2011; Weir et al. 2006). Among a number of methods that have been used to detect IBD familial relationships (Boehnke and Cox 1997; Li et al. 2014; Thompson 1975), the most commonly used are GEMLINE, fastBD, ISCA and ERSA. However, these methods either lack sufficient confidence when applied to long distance relatives (Huff et al. 2011; Li et al. 2014), or result in a high false positive rate (Durand et al. 2014). Thus, the efficiency and reliability of such an approach to testing of familial relationships in generations needs improvement.

With the aim of improving identification of distant familial relationships, we examined 1092 genomes from the "1000 Genomes Project" computationally. We demonstrate that by simply counting the total number of genomic differences it is possible to infer familial relations for people that share down to 6% of common IBD genetic material. Furthermore, this detection of familial relations could be significantly improved (by an order of magnitude) when only very rare genetic variants (vrGVs, with frequencies less than 0.2%) are being considered. This is a very simple and powerful method, though it requires whole-genome sequencing. With the advancement and decreasing cost of sequencing technology, vrGVs should become affordable and important genetic markers for familial relationships, and a broad range of other population genetics studies, in the near future.

1.4 Summary

Here, by use of computational simulation methods, we demonstrate the complexity of mutations that existed in the human genomes. Also, by analysis of the SNPs that have been detected and released by "1000 Genome Project Phase I" data, we provide a method to infer the relationship based on those vrGVs.

1.5 Reference

Aminetzach, Y. T., J. M. Macpherson and D. A. Petrov, 2005 Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in Drosophila. Science 309: 764-767.

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

Boehnke, M., and N. J. Cox, 1997 Accurate inference of relationships in sib-pair linkage studies. American journal of human genetics 61: 423-429.

Browning, B. L., and S. R. Browning, 2013 Detecting identity by descent and estimating genotype error rates in sequence data. American journal of human genetics 93: 840-851.

Burrus, V., and M. K. Waldor, 2004 Shaping bacterial genomes with integrative and conjugative elements. Research in Microbiology 155: 376-386.

Carvajal-Rodriguez A. 2008. GENOMEPOP: A program to simulate genomes in populations. Bmc Bioinformatics 9. doi: Artn 223 Doi 10.1186/1471-2105-9-223

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

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008a. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9. doi: Artn 364
Doi 10.1186/1471-2105-9-364

Conrad, D. F., J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang et al., 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714.

Durand, E. Y., N. Eriksson and C. Y. Mclean, 2014 Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree analysis. Mol Biol Evol 31: 2212-2222.

Fisher RA. 1930. The Genetic Theory of Natural Selection. Dover: Oxford University Press.

Gavrilets S, Hastings A. 1994. Dynamics of genetic variability in two-locus models of stabilizing selection. Genetics 138: 519-532.

Griffiths, R. C., and S. Tavare, 1999 The ages of mutations in gene trees. Annals of Applied Probability 9: 567-590.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786-2787. doi: Doi 0.1093/Bioinformatics/Btn522

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genetical research 8: 269-294.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins et al., 2011 Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research 21: 768-774.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics 75: 199-212.

Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Human molecular genetics 11: 669-674.

Li, W. H., 1975 The first arrival time and mean age of a deleterious mutant gene in a finite population. Am J Hum Genet 27: 274-286.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475: 493-496. doi: 10.1038/nature10231

Li, H., G. Glusman, H. Hu, Shankaracharya, J. Caballero et al., 2014 Relationship estimation from whole-genome sequence data. PLoS genetics 10: e1004144.

Maruyama, T., 1974a The age of a rare mutant gene in a large population. Am J Hum Genet 26: 669-673.

Maruyama, T., 1974b The age of an allele in a finite population. Genet Res 23: 137-143.

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. doi: 10.1146/annurev-genom-082908-150129

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.

Sawyer, S. A., Z. Parsch J Fau - Zhang, D. L. Zhang Z Fau - Hartl and D. L. Hartl, Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila.

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu Rev Genomics Hum Genet 1: 225-249.

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. doi: 10.1098/rstb.2009.0278

Suzanne Estes, Patrick C. Phillips, Dee R. Denver, W. Kelley Thomas and Michael Lynch 2004. Mutation Accumulation in Populations of Varying Size: The Distribution of Mutational Effects for Fitness Correlates in Caenorhabditis elegans. Genetics vol. 166 no. 3 1269-1279. doi: 10.1534/genetics.166.3.1269

Thompson, E. A., 1975 The estimation of pairwise relationships. Annals of human genetics 39: 173-188.

Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nature reviews. Genetics 9: 965-974. doi: 10.1038/nrg2473

Weir, B. S., A. D. Anderson and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;10:771-780.

Wright S. 1965. Factor Interaction and Linkage in Evolution. Proc. R. Soc. Lond. B 162: 80-104. doi: 10.1098/rspb.1965.0026

Chapter 2

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, *]

Authors' Affiliations:

[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

[*]To whom correspondence should be addressed. 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.

Published in the Genome Biol Evol, April 10, 2014, doi: 10.1093/gbe/evu075

2.1 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 for (*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

2.2 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 million 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 2002a; Conrad et al. 2011a; 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 $\boldsymbol{s}$ to be fixed in a population? For a trivial case of a neutral mutation (when $\boldsymbol{s}=0$), 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 ($2\boldsymbol{Ns} << 1$) 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 2007). ***GEMA*** has many features similar to previously published programs (GENOMPOP (Carvajal-Rodriguez 2008), SFS_CODE (Hernandez 2008), FREGENE (Chadeau-Hyam et al. 2008a), 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).

2.3 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: $w = \min(w_m, w_p) + \boldsymbol{h}*\text{abs}(w_m - w_p)$. 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.

2.4 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; $h=1$). 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). 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 $\boldsymbol{\pi}_s^{kimura}$ can be calculated from our $\boldsymbol{\pi}$-value from formula (1) by simple normalization: $\boldsymbol{\pi}_s^{kimura} = \boldsymbol{\pi}_s \times \alpha/2$. (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 $\boldsymbol{\pi}_{s=0}^{kimura}$, denoting the probability of fixation of neutral mutations ($\boldsymbol{s}$=0), follows Kimura's law and is very close to $1/2\boldsymbol{N}$. In other words, the product of three of our parameters $\boldsymbol{\pi}_{s=0}$, $\boldsymbol{\alpha}$, and $\boldsymbol{N}$ approximates to 1 ($\boldsymbol{\pi}_{s=0} \times \boldsymbol{\alpha} \times \boldsymbol{N} \cong 1$). However, for a specific set of parameters, $\boldsymbol{\pi}_{s=0}^{kimura}$ significantly deviates from 1/2N. For example, for ($\boldsymbol{\alpha}$ =10, $\boldsymbol{h}$=0, $\boldsymbol{r}$=1, $\boldsymbol{D}$=expC, $\boldsymbol{\mu}$=1, $\boldsymbol{N}$=50) the product of $\boldsymbol{\pi}_{s=0} \times \boldsymbol{\alpha} \times \boldsymbol{N}$ equals 2.13 instead of being equal to 1. For another set of conditions ($\boldsymbol{\alpha}$ =10, $\boldsymbol{h}$=1, $\boldsymbol{r}$=1, $\boldsymbol{D}$=expC, $\boldsymbol{\mu}$=1, $\boldsymbol{N}$=50) the product of $\boldsymbol{\pi}_{s=0} \times \boldsymbol{\alpha} \times \boldsymbol{N}$ 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 $\boldsymbol{r}$ 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 ($\boldsymbol{r}$). The data from Figure 5 demonstrate that the highest deviations of $\boldsymbol{\pi}_{s=0}$ from neutrality law were observed for the lowest recombination rate when $\boldsymbol{r}$=1.

The size of a population considerably influences the fixation probabilities $\boldsymbol{\pi}_s$ in such a way that the average fitness of the population always improves via increasing its size ($\boldsymbol{N}$). Tables S1 and S2 demonstrate how a growth of $\boldsymbol{N}$ changes $\boldsymbol{\pi}_s$ 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; Charlsworth 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 (***h***=1) with low level of recombination (***r***≤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 ($\boldsymbol{r} \leq 1$, $\boldsymbol{h} = 1$) 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 ($0<\boldsymbol{h}<1$), over-dominance ($\boldsymbol{h}<0$) and under-dominance($\boldsymbol{h}>1$) 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.

2.5 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. 2008b; Chadeau-Hyam et al. 2008a; Kiezun et al. 2013b). 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 parameters 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 2007; Durrett 2008; Charlsworth 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 Table S1 for a chosen set of parameters: (***N***=100; $\boldsymbol{\alpha}$=5; ***h***=0.5; $\boldsymbol{\mu}$=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, recombination 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 recombination 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 ($N_{genes} \approx 25{,}000$ and $L \approx 35{,}000$ 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.

2.6 Figure Legends

Figure 2-1 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***≤1).

Figure 2-2 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 2-3 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 2-4 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 2-5 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 $\boldsymbol{\pi}$ 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 ($\boldsymbol{\pi}_s^{kimura} = \boldsymbol{\pi}_s \times \alpha/2$).

Figure 2-6 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 $\boldsymbol{\mu}$ 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 $\boldsymbol{K/\mu}$ have been calculated for varying number of offspring ($\boldsymbol{\alpha}$) within the population (green $\boldsymbol{\alpha}$=2; red $\boldsymbol{\alpha}$=5; blue $\boldsymbol{\alpha}$=10). ***In toto***, the interplay of various parameters such as recombination rate (***r***), dominance coefficient (***h***), population size (***N***), novel mutations per gamete ($\boldsymbol{\mu}$), 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 2-7 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 2-8 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.

2.7 Figures

Figure 2-1

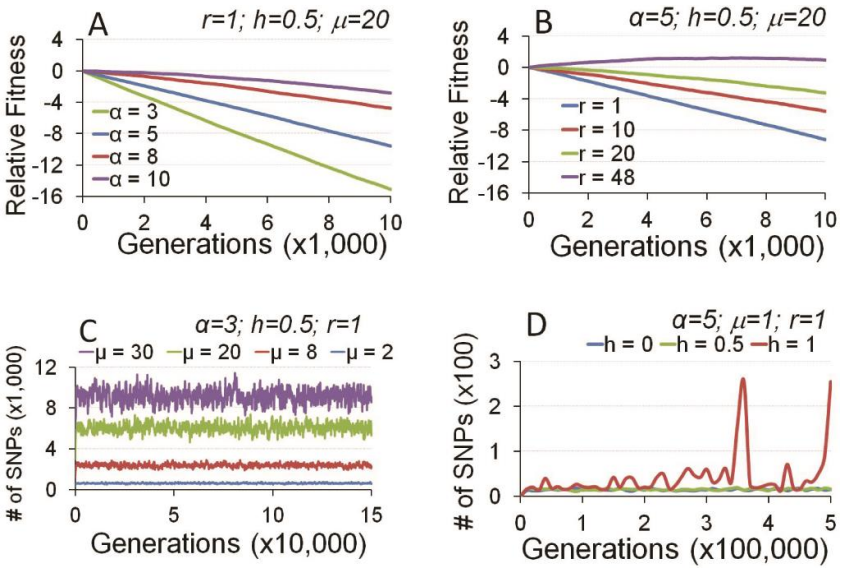


Figure 2-2

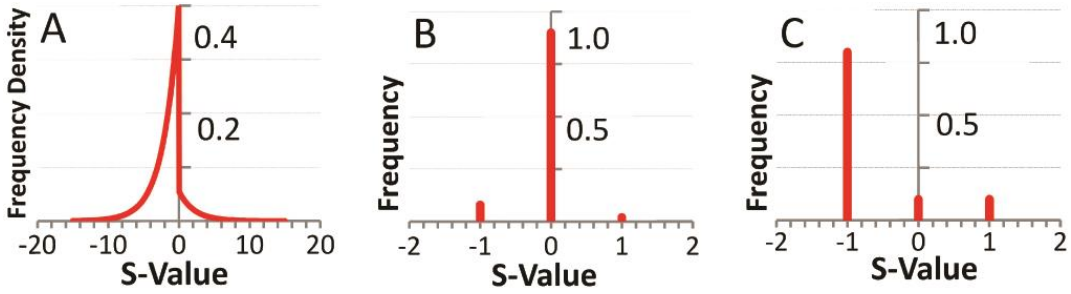

A
Frequency Density
0.4
0.2
-20
-10
0
10
20
S-Value
B
Frequency
1.0
0.5
-2
-1
0
1
2
S-Value
C
Frequency
1.0
0.5
-2
-1
0
1
2
S-Value

Figure 2-3

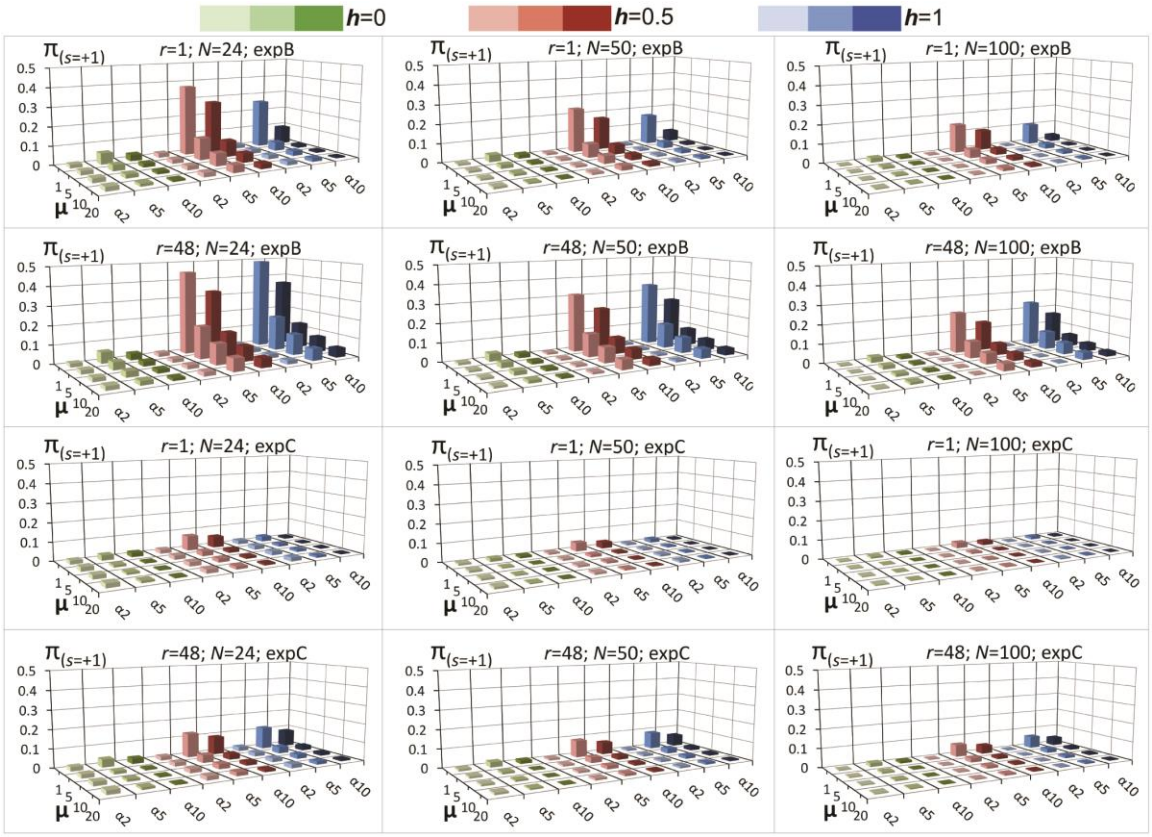

h=0
h=0.5
h=1
π(s=+1)
r=1; N=24; expB
r=1; N=50; expB
r=1; N=100; expB
r=48; N=24; expB
r=48; N=50; expB
r=48; N=100; expB
r=1; N=24; expC
r=1; N=50; expC
r=1; N=100; expC
r=48; N=24; expC
r=48; N=50; expC
r=48; N=100; expC
μ
α2
α5
α10

Figure 2-4

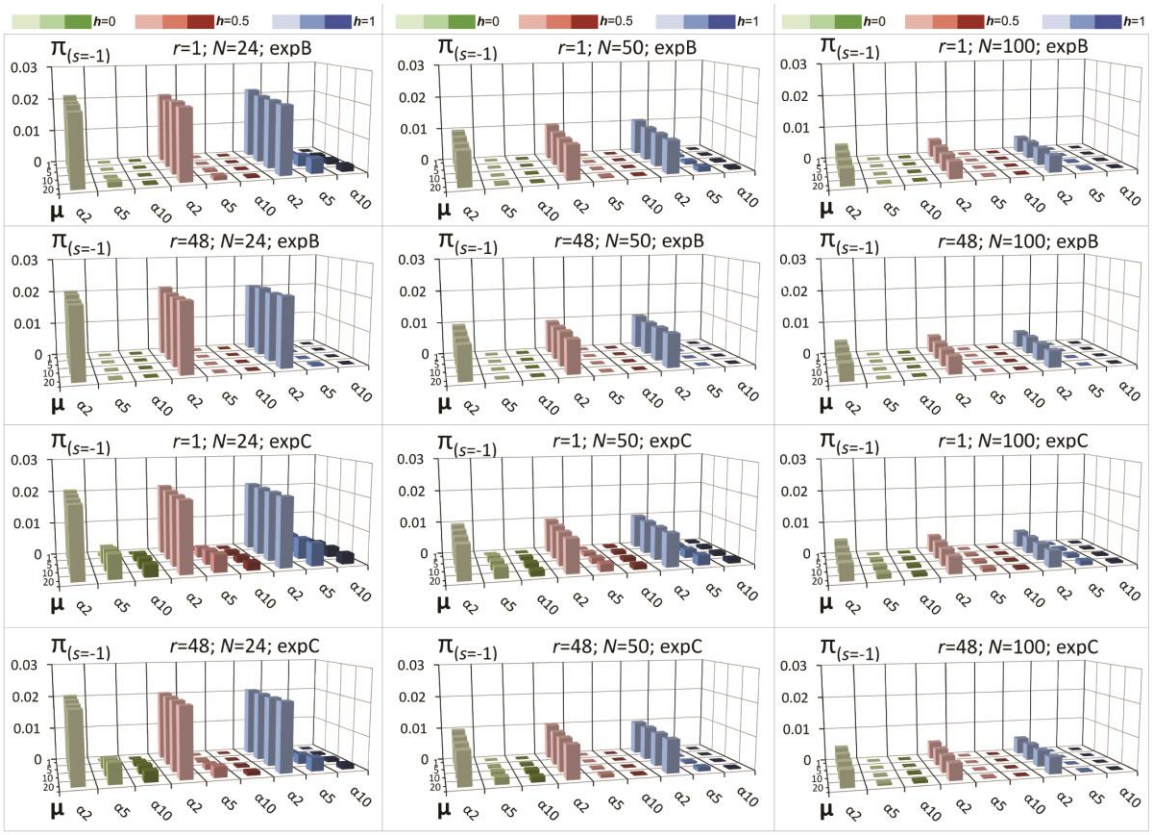

h=0
h=0.5
h=1
π(s=-1)
r=1; N=24; expB
r=1; N=50; expB
r=1; N=100; expB
r=48; N=24; expB
r=48; N=50; expB
r=48; N=100; expB
r=1; N=24; expC
r=1; N=50; expC
r=1; N=100; expC
r=48; N=24; expC
r=48; N=50; expC
r=48; N=100; expC
0.03
0.02
0.01
0
μ
α2
α5
α10

Figure 2-5

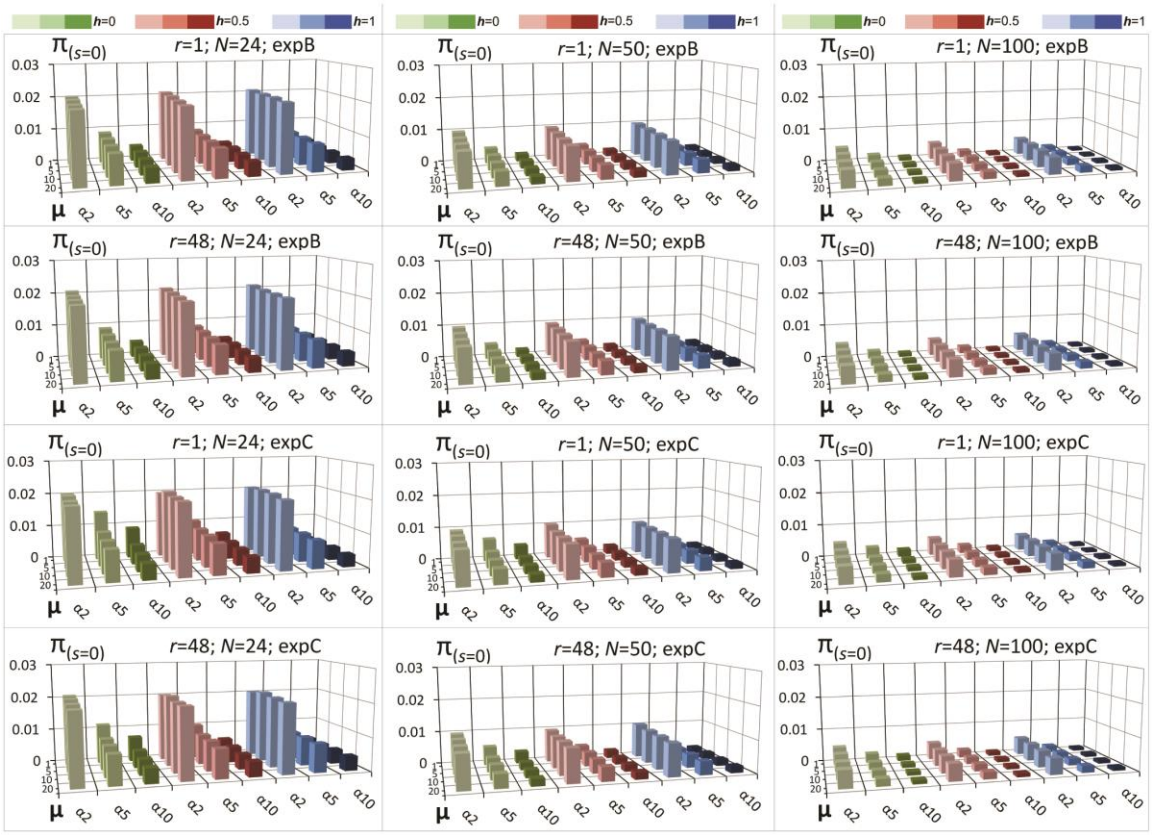

h=0
h=0.5
h=1
π(s=0)
r=1; N=24; expB
r=1; N=50; expB
r=1; N=100; expB
r=48; N=24; expB
r=48; N=50; expB
r=48; N=100; expB
r=1; N=24; expC
r=1; N=50; expC
r=1; N=100; expC
r=48; N=24; expC
r=48; N=50; expC
r=48; N=100; expC
0.03
0.02
0.01
0
μ
α2
α5
α10

Figure 2-6

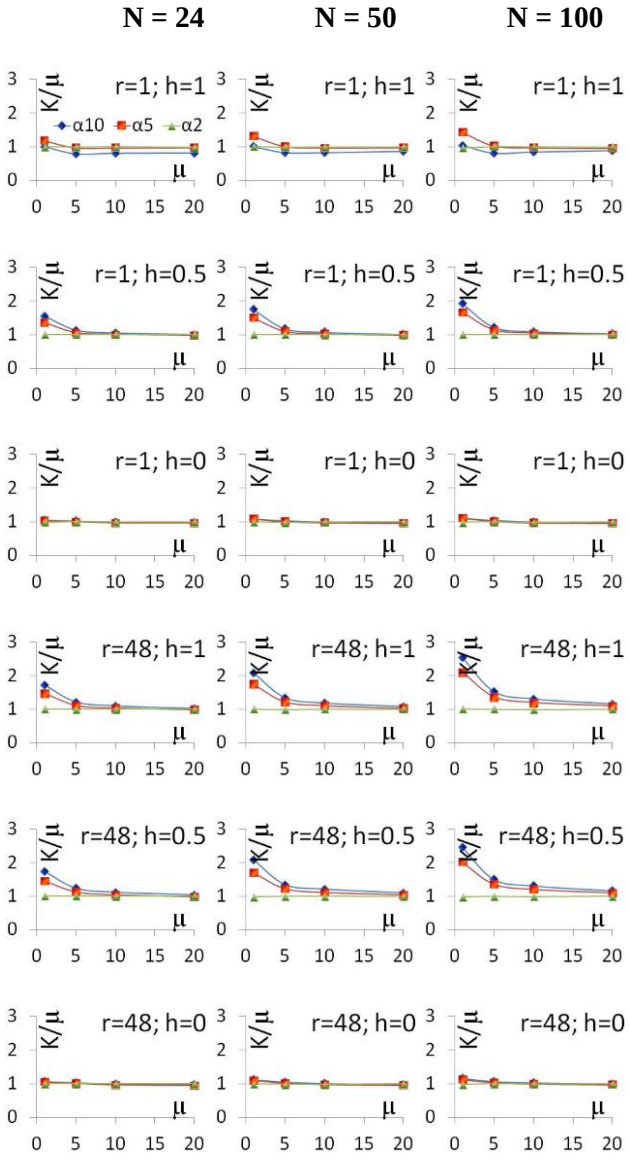


Figure 2-7

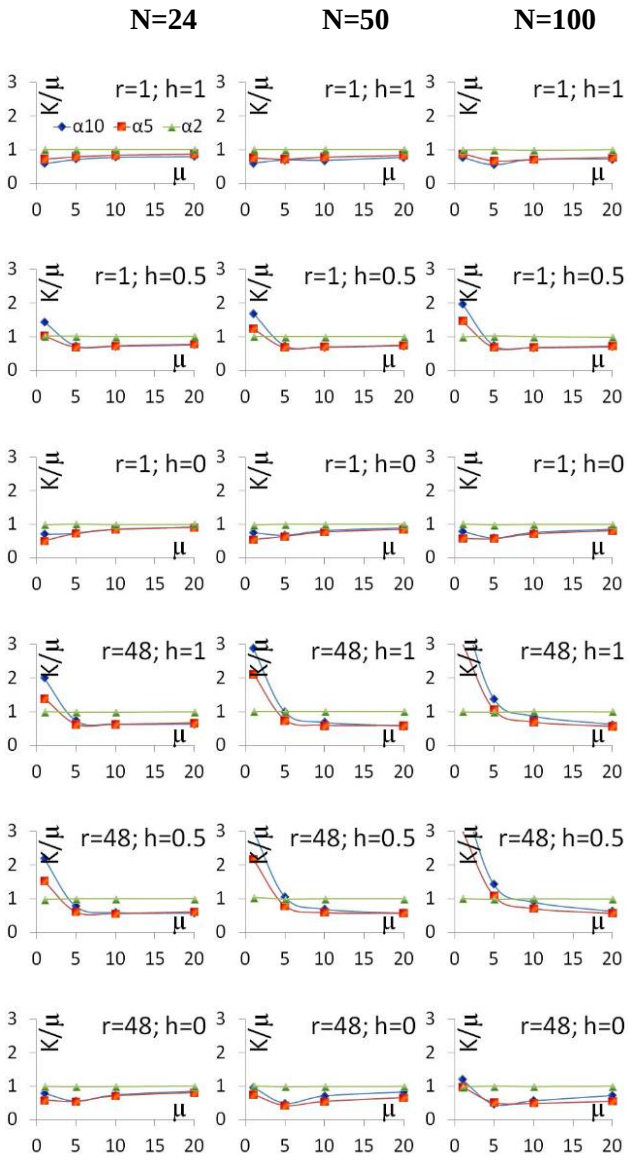
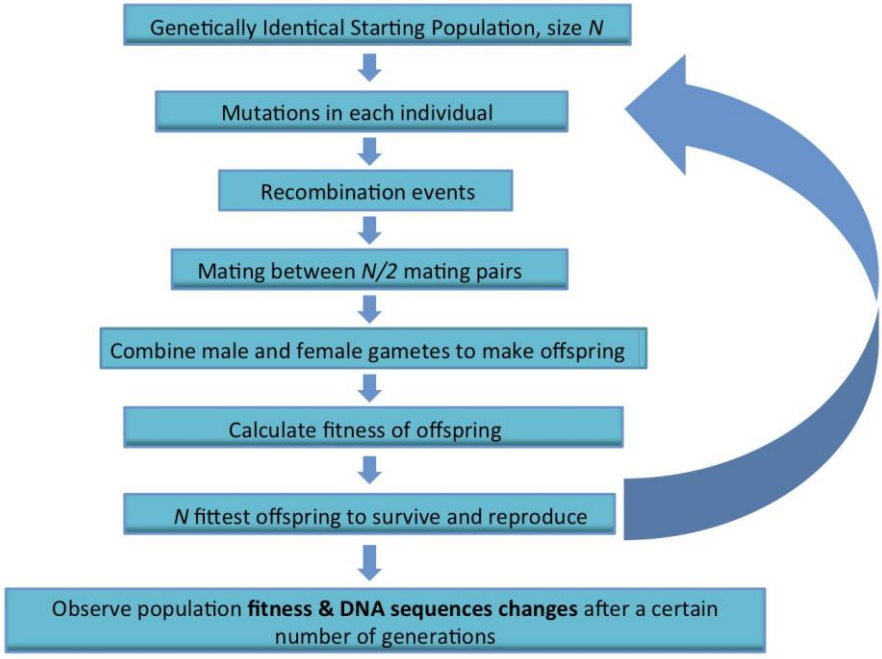


Figure 2-8


Genetically Identical Starting Population, size *N*
Mutations in each individual
Recombination events
Mating between *N/2* mating pairs
Combine male and female gametes to make offspring
Calculate fitness of offspring
N fittest offspring to survive and reproduce
Observe population **fitness & DNA sequences changes** after a certain number of generations

2.8 Supplemental Table Legends

Table S2.1 complete dataset of the GEMA simulation results obtained for the Experiment B (s-distribution curve on Figure 2B)

Table S2.2 complete dataset of the GEMA simulation results obtained for the Experiment C (s-distribution curve on Figure 2C)

2.9 Supplementary Tables

Table S2.1

Input Parameters					Results after 2,000 Generations					Results after 10,000 Generations					Final Results			
N (Pop Size)	***α*** (Offspr per ind)	***μ*** (Muta per gam)	***r*** (recom per gam)	***h*** (Dom coeff)	Ave fit per ind	Ave #SNPs per ind	Number of Fixed Mutations			Ave fit per ind	Ave #SNPs per ind	Number of Fixed Mutations			Fixation Probability (π)			K/μ ratio
							S = 0	S = 1	S = -1			S = 0	S = 1	S = -1	s=0	s=1	s=-1	
24	2	1	1	0	-157.17	34.75	1593	8	146	-837.13	30.875	8700	78	889	0.020564	0.018229	0.021499	0.99
24	2	1	1	0.5	-158.85	37.2917	1621	19	158	-789.1	32.625	8876	104	880	0.020992	0.022135	0.020891	1.00775
24	2	1	1	1	-168.38	33.5833	1625	11	174	-834.96	33.25	8824	89	922	0.02083	0.020313	0.021644	1.003125
24	2	1	48	0	-167.21	33.7083	1616	18	156	-797.04	34.3333	8868	92	867	0.020984	0.019271	0.020573	1.004625
24	2	1	48	0.5	-165.69	34.2917	1612	25	179	-821.9	32.625	8876	99	904	0.021019	0.019271	0.020978	1.007875
24	2	1	48	1	-157.96	34.7917	1641	16	165	-787.25	32.7917	8951	92	875	0.021152	0.019792	0.020544	1.012
24	2	5	1	0	-828.25	170.708	7997	93	800	-4025.3	153.583	43953	490	4422	0.020808	0.020677	0.020961	0.999375
24	2	5	1	0.5	-781.79	152.583	8272	89	808	-3909.2	164.25	44591	494	4319	0.021018	0.021094	0.020318	1.005875
24	2	5	1	1	-769.54	166.792	8207	84	826	-3953.8	181.708	44089	517	4425	0.020765	0.022552	0.020828	0.99785
24	2	5	48	0	-848.04	167.042	8236	96	820	-4050.3	170.167	44138	499	4431	0.020777	0.02099	0.020897	0.9979
24	2	5	48	0.5	-758.17	170.833	8186	82	770	-3893.6	169.458	44394	446	4266	0.020954	0.018958	0.020231	1.0017
24	2	5	48	1	-701.79	170.75	8234	100	768	-3926.3	167.083	44211	505	4408	0.02082	0.021094	0.021065	1.00055
24	2	10	1	0	-1644.1	349.75	16136	193	1602	-8225.5	385.083	87614	1016	8971	0.020682	0.021432	0.021322	0.995875
24	2	10	1	0.5	-1591.6	313.792	16357	182	1640	-8134.7	365.542	88364	1010	8981	0.020835	0.021563	0.021241	1.0022
24	2	10	1	1	-1481.7	341.167	16503	180	1603	-7834.5	310.958	88674	996	8763	0.020883	0.02125	0.020718	1.001838
24	2	10	48	0	-1730.4	332.958	16281	188	1706	-8078.7	332.417	88342	973	8818	0.020851	0.020443	0.020579	0.999475
24	2	10	48	0.5	-1603.5	333.417	16344	176	1648	-7884.4	332	88342	933	8679	0.020833	0.019714	0.020344	0.997325
24	2	10	48	1	-1522.6	340.833	16467	204	1654	-7798.5	327.375	88331	1014	8766	0.020794	0.021094	0.020579	0.997325
24	2	20	1	0	-3199.3	619.458	33251	354	3193	-15928	655.625	176858	2015	17554	0.020776	0.021628	0.020777	0.997681
24	2	20	1	0.5	-3238.5	680.042	32680	359	3283	-16097	681.875	176194	2028	17825	0.020763	0.021732	0.021039	0.998281
24	2	20	1	1	-3152.9	627.167	33352	396	3415	-15898	671.083	176416	1929	17660	0.020698	0.019961	0.020609	0.992763
24	2	20	48	0	-3391.6	664.625	33035	329	3295	-16160	660.792	177147	1882	17616	0.02085	0.020221	0.020719	0.999913
24	2	20	48	0.5	-3217.6	678.25	32863	351	3259	-15980	666.042	176808	1980	17663	0.020825	0.021211	0.020839	0.999863
24	2	20	48	1	-2987.8	665.167	32965	381	3212	-15929	669.042	176651	1998	17778	0.020788	0.021055	0.021073	0.999181
24	5	1	1	0	131	14.4583	1933	131	0	693	12.8333	9684	693	0	0.008971	0.058542	0	1.039125
24	5	1	1	0.5	901.333	9.5	1839	897	0	4501.92	9.41667	9176	4500	1	0.008492	0.375313	1.16E-05	1.367625
24	5	1	1	1	636.792	15.9583	1747	623	3	3130.96	16.75	8689	3128	13	0.008035	0.260938	0.000116	1.182125
24	5	1	48	0	158.25	18.375	1908	158	0	727	18.7083	9768	727	0	0.009097	0.059271	0	1.053625
24	5	1	48	0.5	1045.25	17.25	1818	1041	0	5353.15	17.4583	9213	5347	0	0.008559	0.448542	0	1.462625
24	5	1	48	1	1115.21	17.4167	1712	1101	0	5831.88	17.4167	8783	5815	0	0.008184	0.491042	0	1.473125
24	5	5	1	0	350.25	61.875	9624	351	1	1848.33	57.5	48183	1856	7	0.008926	0.031354	1.39E-05	1.00175
24	5	5	1	0.5	1357.13	53.2917	9158	1366	14	6838.96	49	46115	6905	71	0.008555	0.115396	0.000132	1.063825
24	5	5	1	1	463.917	127.125	8324	609	180	1806.75	158.208	43992	2851	1065	0.008256	0.046708	0.002049	0.969875
24	5	5	48	0	479.458	81.2917	9272	480	0	2446.42	88.2083	47568	2446	1	0.008865	0.040958	2.31E-06	1.006575
24	5	5	48	0.5	2067.5	82.9583	8865	2051	0	10416.2	84.7917	45965	10392	0	0.008588	0.173771	0	1.136025
24	5	5	48	1	2291.83	86.375	8521	2239	0	11167	86.4583	44277	11132	3	0.008277	0.185271	6.94E-06	1.1163
24	5	10	1	0	492.792	121.417	18507	515	20	2292.79	120.375	94219	2410	115	0.008763	0.01974	0.00011	0.971275
24	5	10	1	0.5	1503.4	112	18109	1580	88	7526.4	102.458	92316	7968	452	0.008589	0.066542	0.000421	1.011988

24	5	10	1	1	7.75	296.667	16132	600	618	-456.75	335.083	87837	3177	3619	0.008299	0.026844	0.003473	0.966038
24	5	10	48	0	731.208	170.583	17957	730	0	3576.13	177.292	92358	3589	13	0.008611	0.029781	1.5E-05	0.965913
24	5	10	48	0.5	2637.6	166.958	17577	2626	0	13322.2	173.708	90872	13304	6	0.008483	0.111229	6.94E-06	1.049738
24	5	10	48	1	2702.04	174.333	17144	2658	22	13384.2	160.458	89069	13421	96	0.008325	0.112115	8.56E-05	1.034525
24	5	20	1	0	10.125	183.375	36740	627	597	156.917	187.333	186335	3187	3009	0.008657	0.013333	0.001396	0.966044
24	5	20	1	0.5	1279.75	237.708	36132	1772	496	6678.48	227.583	183654	9220	2545	0.008537	0.038792	0.001186	0.981369
24	5	20	1	1	-903.33	661.375	31509	801	1627	-5368.9	594.375	176234	4158	9465	0.008375	0.017484	0.004536	0.9745
24	5	20	48	0	881.208	257.458	35670	988	89	4647.42	304.75	182248	5097	431	0.008483	0.021401	0.000198	0.943931
24	5	20	48	0.5	3267.69	320.667	34950	3261	25	16420	324.542	180762	16507	120	0.008438	0.06899	5.5E-05	0.994706
24	5	20	48	1	2883.21	333.375	34190	3035	211	14312.8	347.208	178828	15373	1120	0.00837	0.06426	0.000526	0.986781
24	10	1	1	0	145	14.9583	1966	145	0	776	13.9167	9703	776	0	0.004477	0.032865	0	1.046
24	10	1	1	0.5	1332.33	7.5	1710	1331	1	6696.44	6.75	8804	6691	1	0.004105	0.279167	0	1.55675
24	10	1	1	1	775.542	24.9167	1548	733	4	2755.83	119.708	7691	2558	15	0.003555	0.095052	6.37E-05	0.997375
24	10	1	48	0	155.292	13.9167	1922	155	0	784	13.2917	9841	784	0	0.004583	0.03276	0	1.0685
24	10	1	48	0.5	1586.17	13.2083	1759	1582	0	8003.77	13.6667	9263	8000	0	0.004343	0.334271	0	1.74025
24	10	1	48	1	1748.92	13.625	1720	1731	0	8664.75	13.9167	8649	8643	0	0.00401	0.36	0	1.730125
24	10	5	1	0	458	40.5	9712	458	0	2326.08	62.875	48642	2329	3	0.004506	0.01949	3.47E-06	1.0201
24	10	5	1	0.5	2003.52	39.125	9104	2000	12	10200.4	35.4167	46384	10235	41	0.004315	0.085781	3.36E-05	1.1386
24	10	5	1	1	437.458	222.292	7311	564	172	1274.54	858.167	36311	2148	953	0.003356	0.0165	0.000904	0.784125
24	10	5	48	0	580.75	67.0833	9465	579	0	2899.13	76.4167	48404	2899	0	0.004507	0.024167	0	1.031475
24	10	5	48	0.5	3008.27	77.375	8897	2986	0	15197	75.4167	46258	15182	0	0.004324	0.127042	0	1.238925
24	10	5	48	1	3170.42	67.8333	8694	3123	0	15988.3	62.6667	44592	15945	2	0.004155	0.133563	2.31E-06	1.21805
24	10	10	1	0	632.5	92.125	19079	639	7	3191.13	86.3333	95924	3207	16	0.004447	0.013375	5.21E-06	0.992775
24	10	10	1	0.5	2273.46	84.9167	18528	2325	64	11419.6	82	93086	11678	269	0.004315	0.048714	0.000119	1.05145
24	10	10	1	1	59.2917	453.5	14176	639	568	-914.29	1582.17	73730	2616	3210	0.003446	0.010297	0.001529	0.802163
24	10	10	48	0	909.333	118.417	18837	910	1	4454.33	135.583	94949	4465	11	0.004405	0.018516	5.79E-06	0.995963
24	10	10	48	0.5	3879.06	139.25	17609	3858	1	19531.7	140.167	91493	19513	4	0.004276	0.081536	1.74E-06	1.119275
24	10	10	48	1	3884.21	148.292	17417	3814	4	19590.7	144.208	89477	19551	26	0.00417	0.081964	1.27E-05	1.097738
24	10	20	1	0	552.542	114.917	37443	782	218	2907.63	134.625	188764	4008	1094	0.004379	0.008401	0.000253	0.971394
24	10	20	1	0.5	2326.08	165.75	36453	2602	287	11630.2	208.875	185050	13092	1479	0.0043	0.027318	0.000345	1.001744
24	10	20	1	1	-928.33	866.917	29069	826	1582	-6416.7	3080.33	149654	3423	8608	0.003489	0.006763	0.002033	0.8138
24	10	20	48	0	1178.29	256.708	36200	1211	28	6040.75	231.167	186324	6147	103	0.004344	0.012854	2.17E-05	0.969594
24	10	20	48	0.5	4862.15	244.625	35539	4842	7	24278.8	279.208	182614	24277	39	0.004256	0.050612	9.26E-06	1.040888
24	10	20	48	1	4554.79	280.75	34933	4533	77	22858.8	276.875	179007	23185	412	0.004169	0.048573	9.69E-05	1.019131
50	2	1	1	0	-158.08	40.96	1594	13	123	-808.6	37.9	8793	83	840	0.009999	0.00875	0.009958	0.99825
50	2	1	1	0.5	-152.75	40.98	1520	15	141	-849.44	42.82	8763	87	905	0.01006	0.009	0.010611	1.009875
50	2	1	1	1	-134.58	38.1	1537	17	139	-840.36	36.82	8750	92	918	0.010018	0.009375	0.010819	1.008375
50	2	1	48	0	-175.68	38.94	1475	13	140	-786.42	38.8	8664	113	850	0.009985	0.0125	0.009861	0.999875
50	2	1	48	0.5	-152.59	40.44	1495	21	134	-764.17	40.1	8622	104	837	0.009899	0.010375	0.009764	0.989125
50	2	1	48	1	-133.12	40.6	1513	23	144	-790.44	39.82	8758	107	886	0.010063	0.0105	0.010306	1.008875
50	2	5	1	0	-871.88	201.48	7248	87	728	-4065.7	215.62	42801	483	4283	0.009876	0.0099	0.009875	0.9876
50	2	5	1	0.5	-767.39	203.74	7239	76	695	-3922.5	175.56	43621	472	4259	0.010106	0.0099	0.0099	1.00855
50	2	5	1	1	-766.84	203.18	7312	66	749	-4006.3	206.64	43046	447	4383	0.009926	0.009525	0.010094	0.993725
50	2	5	48	0	-859.14	199.64	7159	97	725	-4015.1	198.34	42968	480	4263	0.009947	0.009575	0.009828	0.99325
50	2	5	48	0.5	-750.79	198.22	7201	84	694	-3939.7	203.2	43402	471	4251	0.010056	0.009675	0.009881	1.003625
50	2	5	48	1	-716.78	199.66	7385	89	737	-3849.8	197.62	43061	505	4274	0.00991	0.0104	0.009825	0.990725
50	2	10	1	0	-1667.2	424.12	14457	154	1380	-8037.3	392.08	86276	888	8480	0.009975	0.009175	0.009861	0.995663
50	2	10	1	0.5	-1551.7	411.06	14689	156	1420	-7873.9	403.68	86609	955	8500	0.009989	0.009988	0.009833	0.997488
50	2	10	1	1	-1478.6	393.28	14723	162	1471	-7887.8	396.06	86673	915	8658	0.009993	0.009413	0.009982	0.998625

50	2	10	48	0	-1707.4	398.26	14538	156	1401	-7961.7	397.14	86082	941	8450	0.009937	0.009813	0.00979	0.992225
50	2	10	48	0.5	-1618.9	396.4	14311	155	1463	-8214	398.84	86397	929	8844	0.010012	0.009675	0.010251	1.003013
50	2	10	48	1	-1501.9	397.6	14415	174	1503	-7827	390.82	86729	1024	8698	0.010044	0.010625	0.009993	1.004488
50	2	20	1	0	-3361.2	811.44	28701	331	2875	-16230	776.56	172812	1941	17355	0.010008	0.010063	0.010056	1.001256
50	2	20	1	0.5	-3159.6	750.86	29758	338	2930	-16181	779.06	173486	1935	17523	0.009981	0.009981	0.010134	0.999488
50	2	20	1	1	-2998.2	784.14	29394	283	2840	-15645	804.28	173115	1852	17070	0.009981	0.009806	0.009882	0.997
50	2	20	48	0	-3427.2	787.08	28868	312	2882	-16190	792.08	173280	1943	17301	0.010029	0.010194	0.010013	1.002888
50	2	20	48	0.5	-3226.4	795.02	28838	315	2950	-15858	793.74	172724	1972	17225	0.009992	0.010356	0.009913	0.998863
50	2	20	48	1	-3074.4	788.18	29069	297	2973	-15960	785.56	172990	1884	17465	0.009995	0.009919	0.010064	1
50	5	1	1	0	173	14.22	1929	172	0	894.26	15.34	9903	894	0	0.00443	0.0361	0	1.087
50	5	1	1	0.5	1189.76	9.5	1805	1181	0	5965.55	11.08	9114	5955	0	0.004061	0.2387	0	1.510375
50	5	1	1	1	894.32	16.98	1600	846	0	4159.34	18	8844	4119	9	0.004024	0.16365	0.00005	1.31575
50	5	1	48	0	212.14	20.04	1876	212	0	1038	18.26	9779	1038	0	0.004391	0.0413	0	1.091125
50	5	1	48	0.5	1585.38	18.54	1730	1566	0	7858.94	18.12	9115	7851	0	0.004103	0.31425	0	1.70875
50	5	1	48	1	1703.22	19.22	1601	1670	0	8562.42	18.7	8793	8520	0	0.003996	0.3425	0	1.75525
50	5	5	1	0	457.44	68.74	9280	457	0	2280.7	64.76	47787	2286	5	0.004279	0.01829	5.56E-06	1.008525
50	5	5	1	0.5	1800.91	56.36	8952	1786	10	8953.94	56.58	45959	8972	44	0.004112	0.07186	3.78E-05	1.105675
50	5	5	1	1	870.52	131.9	7719	877	92	3767.6	126.54	43941	4251	579	0.004025	0.03374	0.000541	1.002075
50	5	5	48	0	693.2	90.4	8676	688	0	3403.94	95.48	46684	3401	0	0.004223	0.02713	0	1.018025
50	5	5	48	0.5	3093.14	93.76	8332	3053	0	15416.2	93.82	45057	15377	0	0.004081	0.12324	0	1.226225
50	5	5	48	1	3295.86	92.66	8224	3178	0	16528.6	94.76	44064	16418	2	0.003982	0.1324	2.22E-06	1.22705
50	5	10	1	0	564.96	106.68	18445	577	9	2874.96	102.24	94259	2910	33	0.004212	0.011665	1.33E-05	0.977138
50	5	10	1	0.5	1927.76	108.4	17743	1966	63	9764.57	117.34	91317	10060	320	0.004087	0.04047	0.000143	1.024063
50	5	10	1	1	573.74	274.16	15201	864	370	2361.44	312.76	86210	4522	2212	0.003945	0.01829	0.001023	0.956363
50	5	10	48	0	983.54	181.32	16899	982	1	5085.24	185.9	91433	5093	10	0.004141	0.020555	0.000005	0.983175
50	5	10	48	0.5	4001.55	180.6	16719	3948	1	20222.8	186.68	89377	20165	2	0.004037	0.081085	5.56E-07	1.11095
50	5	10	48	1	4220.08	186.32	16603	4060	0	21022.5	184.76	88672	20875	11	0.004004	0.084075	6.11E-06	1.111188
50	5	20	1	0	377.58	165.66	36634	694	294	1954.28	182.16	185399	3530	1556	0.004132	0.00709	0.000351	0.955394
50	5	20	1	0.5	1941.74	199.44	35648	2276	347	9731.22	194.12	182595	11552	1833	0.004082	0.02319	0.000413	0.985681
50	5	20	1	1	-16.8	576.9	29414	1072	1048	-976.42	602.68	173540	5660	6548	0.004004	0.01147	0.001528	0.963838
50	5	20	48	0	1427.9	348.2	33794	1446	13	7393.1	350.44	180776	7448	56	0.004083	0.015005	1.19E-05	0.956419
50	5	20	48	0.5	5035.85	385.32	33514	4946	7	25441.9	361.48	179633	25372	25	0.004059	0.051065	0.000005	1.041019
50	5	20	48	1	5096.26	372.68	33211	4899	24	25375.1	371.04	177141	25312	130	0.003998	0.051033	2.94E-05	1.027806
50	10	1	1	0	191	7.66	1953	191	0	1007	10.54	9737	1007	0	0.002162	0.0204	0	1.075
50	10	1	1	0.5	1692.26	8.1	1700	1681	0	8574.76	8.34	8813	8567	0	0.001976	0.17215	0	1.749875
50	10	1	1	1	830.8	26.2	1426	739	0	3311.18	81.96	7218	3027	3	0.001609	0.0572	8.33E-06	1.010375
50	10	1	48	0	234	13.32	1835	234	0	1164	14.9	9883	1164	0	0.002236	0.02325	0	1.12225
50	10	1	48	0.5	2272.86	15.02	1742	2261	0	11290.6	13.1	9397	11282	0	0.002126	0.225525	0	2.0845
50	10	1	48	1	2437.7	16.06	1693	2385	0	12194.1	12.74	8638	12151	0	0.001929	0.24415	0	2.088875
50	10	5	1	0	556.5	42.14	9522	556	1	2755.32	47.38	48370	2758	3	0.002158	0.01101	1.11E-06	1.0263
50	10	5	1	0.5	2546.24	37.28	9111	2534	7	12684.9	36.68	46232	12707	40	0.002062	0.050865	1.83E-05	1.183175
50	10	5	1	1	841.66	145.28	7085	835	92	2961.1	418.58	36827	3329	585	0.001652	0.01247	0.000274	0.818225
50	10	5	48	0	792.92	70.04	9240	791	0	3846.16	54.2	48196	3842	0	0.002164	0.015255	0	1.050175
50	10	5	48	0.5	4279.03	75.3	8682	4234	0	21502	76.76	45187	21466	0	0.002028	0.08616	0	1.343425
50	10	5	48	1	4531.34	76.78	8304	4422	0	22761.6	71.98	44065	22652	0	0.001987	0.09115	0	1.349775
50	10	10	1	0	748.66	72.44	18741	749	3	3741.24	88.54	95149	3752	12	0.002122	0.007508	2.5E-06	0.99275
50	10	10	1	0.5	2829.16	72.34	18110	2845	44	14210.2	84.9	92340	14390	204	0.002062	0.028863	4.44E-05	1.074188
50	10	10	1	1	604.22	296.26	13773	886	335	1379.04	822.82	74686	3618	2122	0.001692	0.00683	0.000496	0.8179
50	10	10	48	0	1217.56	141.2	17808	1217	0	6021.22	132.44	93630	6018	3	0.002106	0.012003	8.33E-07	1.007825

50	10	10	48	0.5	5540.66	143.62	17217	5486	0	27933.2	152.44	91248	27875	1	0.002056	0.055973	2.78E-07	1.205263
50	10	10	48	1	5707.4	137.82	16851	5562	1	28763.2	145.32	88694	28619	6	0.001996	0.057643	1.39E-06	1.186313
50	10	20	1	0	802.02	137.2	37132	893	85	4073.38	147.78	187777	4515	436	0.002092	0.004528	4.88E-05	0.966363
50	10	20	1	0.5	2999.6	146.6	36456	3197	214	15029.2	153.76	184469	16115	1107	0.002056	0.016148	0.000124	1.0114
50	10	20	1	1	-49.86	580.12	28402	1091	1046	-2244	1289.82	156785	4872	6519	0.001783	0.004726	0.00076	0.860231
50	10	20	48	0	1705.56	248.3	34955	1706	5	8477.1	247.22	183217	8521	45	0.002059	0.008519	5.56E-06	0.969481
50	10	20	48	0.5	7084.89	299.66	34554	6998	3	35517	291.56	180660	35442	19	0.002029	0.035555	2.22E-06	1.091038
50	10	20	48	1	6973.02	280.98	33767	6796	21	35609.6	298.8	178261	35445	60	0.002007	0.035811	5.42E-06	1.082388
100	2	1	1	0	-222.94	46.66	1087	13	126	-868.59	45.54	8237	81	858	0.004965	0.00425	0.005083	0.99375
100	2	1	1	0.5	-161.95	44.82	1117	11	109	-849.3	44.59	8333	88	873	0.005011	0.004813	0.005306	1.007125
100	2	1	1	1	-116.4	42.29	1084	17	115	-716.34	47.64	8137	97	782	0.004898	0.005	0.004632	0.975
100	2	1	48	0	-183.51	46.31	1120	14	105	-790.81	45.06	8241	98	789	0.004945	0.00525	0.00475	0.986125
100	2	1	48	0.5	-148.06	45.44	1108	14	100	-783.81	45.9	8230	90	815	0.004946	0.00475	0.004965	0.989125
100	2	1	48	1	-125.01	46.29	999	13	114	-744.48	46.44	8256	97	821	0.00504	0.00525	0.00491	1.006
100	2	5	1	0	-966.99	226.43	5646	55	563	-4080.9	222.01	41775	476	4150	0.005018	0.005263	0.004982	1.003425
100	2	5	1	0.5	-832.33	226.98	5561	53	547	-4050.8	223.15	41547	427	4177	0.004998	0.004675	0.005042	0.99975
100	2	5	1	1	-673.14	231.12	5402	52	539	-3857.1	223.34	41636	468	4155	0.005033	0.0052	0.005022	1.00665
100	2	5	48	0	-972.54	226.17	5442	50	555	-4241.2	227.38	41489	430	4231	0.005007	0.00475	0.005106	1.002575
100	2	5	48	0.5	-765.14	225.26	5556	69	535	-3797.5	226.03	41461	527	4034	0.004987	0.005725	0.00486	0.99655
100	2	5	48	1	-638.64	223.96	5483	68	561	-3936	227.28	41470	469	4243	0.004998	0.005013	0.005114	1.00175
100	2	10	1	0	-1863.7	460.48	11099	122	1148	-8097.2	445.05	82478	904	8166	0.004957	0.004888	0.004874	0.989738
100	2	10	1	0.5	-1584.2	453.29	11127	113	1088	-7878.8	454.05	83117	917	8176	0.004999	0.005025	0.004922	0.998525
100	2	10	1	1	-1358.2	454.59	10706	85	1047	-7862.3	442.41	83134	915	8362	0.00503	0.005188	0.00508	1.007163
100	2	10	48	0	-1836.7	451.3	10907	113	1087	-8150.9	453.28	82894	906	8197	0.004999	0.004956	0.004938	0.998625
100	2	10	48	0.5	-1671.6	453.45	10883	111	1151	-8006.5	454.03	82291	912	8276	0.004959	0.005006	0.004948	0.991675
100	2	10	48	1	-1341	451.66	11048	122	1089	-7737.3	453.19	82449	966	8320	0.004958	0.005275	0.005022	0.99345
100	2	20	1	0	-3510.8	903.14	21858	246	2162	-16044	919.56	166088	1868	16299	0.005008	0.005069	0.004909	0.999931
100	2	20	1	0.5	-3139.5	923.62	21016	264	2056	-15835	920.37	165061	1809	16312	0.005002	0.004828	0.00495	0.999038
100	2	20	1	1	-2890.7	874.4	22431	250	2298	-15820	872.1	166290	1835	16709	0.004995	0.004953	0.005004	0.999094
100	2	20	48	0	-3518.1	890.75	22135	255	2128	-16298	893.89	166188	1926	16586	0.005002	0.005222	0.00502	1.001138
100	2	20	48	0.5	-3169.5	891.58	22055	232	2143	-16171	896.7	166077	1798	16734	0.005001	0.004894	0.005066	1.001119
100	2	20	48	1	-2764.1	892.44	22055	223	2120	-15404	893.16	166542	1856	16358	0.005017	0.005103	0.004944	1.002238
100	5	1	1	0	222.07	19.45	1763	222	0	1116.01	13.7	9686	1116	0	0.002201	0.02235	0	1.102125
100	5	1	1	0.5	1496.23	9.93	1740	1481	0	7575.14	11.37	9051	7554	0	0.002031	0.151825	0	1.673
100	5	1	1	1	1192.96	17.34	1560	1089	0	5649.09	18.22	8573	5544	5	0.001948	0.111375	1.39E-05	1.434125
100	5	1	48	0	260.64	21.59	1682	260	0	1416.03	22.11	9533	1416	0	0.002181	0.0289	0	1.125875
100	5	1	48	0.5	2215.28	18.85	1594	2195	0	10999.9	19.24	8991	10981	0	0.002055	0.21965	0	2.022875
100	5	1	48	1	2404.59	20.89	1577	2333	0	12152.2	20.5	8735	12068	0	0.001988	0.243375	0	2.111625
100	5	5	1	0	516.55	60.21	9101	515	1	2602.86	62.94	47594	2609	8	0.002139	0.01047	3.89E-06	1.01485
100	5	5	1	0.5	2208.77	48.1	9134	2185	7	11084.1	49.55	45943	11093	43	0.002045	0.04454	0.00002	1.143825
100	5	5	1	1	1304.85	120.11	6838	1146	39	6103.47	114.92	42835	6207	294	0.002	0.025305	0.000142	1.032825
100	5	5	48	0	948.97	102.2	7932	944	0	4632.61	105.24	45713	4628	0	0.002099	0.01842	0	1.036625
100	5	5	48	0.5	4297.99	98.95	7674	4210	0	21874.7	96.32	44406	21801	0	0.002041	0.087955	0	1.358075
100	5	5	48	1	4771.42	99.49	7716	4545	1	23797.4	95.15	43693	23545	2	0.001999	0.095	5.56E-07	1.37445
100	5	10	1	0	675.67	108.19	17996	682	8	3305.12	108.94	92739	3339	36	0.002076	0.006643	7.78E-06	0.96785
100	5	10	1	0.5	2467.22	103.95	17709	2471	53	12359.9	102.49	91178	12555	238	0.002041	0.02521	5.14E-05	1.046725
100	5	10	1	1	1310.23	242.51	13567	1302	172	5758.49	262.26	85074	6772	1196	0.001986	0.013675	0.000284	0.975013
100	5	10	48	0	1385.29	195.19	15822	1371	1	7183.06	198.95	89792	7169	5	0.002055	0.014495	1.11E-06	0.99715
100	5	10	48	0.5	5676.92	195.86	15645	5544	0	28840.8	196.13	88157	28699	2	0.002014	0.057888	5.56E-07	1.195863

100	5	10	48	1	6132.92	195.29	15482	5808	0	30543.5	193.09	87084	30219	1	0.001989	0.061028	2.78E-07	1.200175
100	5	20	1	0	644.95	191.92	35769	797	134	3260.98	201.25	183867	3982	698	0.002057	0.003981	7.83E-05	0.949044
100	5	20	1	0.5	2525.17	203.63	34973	2743	263	12759	202.39	181949	14043	1318	0.002041	0.014125	0.000147	0.995819
100	5	20	1	1	930.86	505.9	27161	1446	604	3789.41	516.98	170222	7900	4173	0.001987	0.008068	0.000496	0.956775
100	5	20	48	0	2147.72	379.49	30871	2133	5	11027.5	389.61	177300	11039	33	0.002034	0.011133	3.89E-06	0.971019
100	5	20	48	0.5	7297.04	388.1	31177	7073	1	37099	389.85	175937	36902	13	0.002011	0.037286	1.67E-06	1.091256
100	5	20	48	1	7728.38	391.02	30764	7238	5	38555	389.6	174851	38127	34	0.002001	0.038611	4.03E-06	1.093781
100	10	1	1	0	233.01	7.58	1910	233	0	1213	12.82	9710	1213	0	0.001083	0.01225	0	1.0975
100	10	1	1	0.5	2095.78	7.15	1794	2078	0	10452.5	6.84	8859	10444	0	0.000981	0.104575	0	1.928875
100	10	1	1	1	1184.85	18.56	1488	1055	1	4261.7	56.63	7019	3810	3	0.000768	0.034438	2.78E-06	1.036
100	10	1	48	0	281.02	16.57	1704	281	0	1458.01	12.07	9852	1458	0	0.001132	0.014713	0	1.165625
100	10	1	48	0.5	3045.92	15.75	1669	3018	0	15315.1	16.04	9057	15294	0	0.001026	0.15345	0	2.458
100	10	1	48	1	3291.5	15.49	1597	3217	0	16580.9	15.52	8659	16497	0	0.000981	0.166	0	2.54275
100	10	5	1	0	616.1	46.31	9407	615	0	3138.06	42.29	48524	3147	10	0.001087	0.00633	2.78E-06	1.041475
100	10	5	1	0.5	3057.46	38.01	9049	3034	11	15207.1	37.75	45801	15218	41	0.001021	0.03046	8.33E-06	1.22415
100	10	5	1	1	1437.31	96.24	7303	1302	58	4621.77	266.01	36201	4522	282	0.000803	0.00805	6.22E-05	0.80855
100	10	5	48	0	975.28	71.28	8835	968	0	5124.29	72.84	47481	5123	0	0.001074	0.010388	0	1.070025
100	10	5	48	0.5	5929.35	71.76	8377	5853	0	29759.2	80.17	44610	29675	0	0.001006	0.059555	0	1.501375
100	10	5	48	1	6245.36	70.61	8303	6039	0	31411.8	76.08	44229	31175	1	0.000998	0.06284	2.78E-07	1.526575
100	10	10	1	0	798.07	91.63	18573	801	7	4123.08	99.77	95090	4146	24	0.001063	0.004181	2.36E-06	0.998488
100	10	10	1	0.5	3424.28	78.26	17878	3410	28	17173.7	78.43	91448	17297	165	0.001022	0.017359	1.9E-05	1.094925
100	10	10	1	1	1184.25	250.52	12453	1155	153	4475.95	492.46	74295	5504	1233	0.000859	0.005436	0.00015	0.840888
100	10	10	48	0	1619.03	147.81	16864	1611	1	8228.42	142.23	92646	8217	3	0.001053	0.008258	2.78E-07	1.029875
100	10	10	48	0.5	7838.3	154.4	16462	7686	0	39128.1	145.11	89683	38999	2	0.001017	0.039141	2.78E-07	1.3067
100	10	10	48	1	8120.07	148.75	16199	7797	0	40304.3	141.47	88370	39978	4	0.001002	0.040226	5.56E-07	1.30445
100	10	20	1	0	977.55	152.22	37058	1021	39	4938.48	149.31	188392	5139	199	0.001051	0.002574	1.11E-05	0.972575
100	10	20	1	0.5	3643.92	161.37	35593	3776	185	18429.4	160.2	183774	19290	909	0.001029	0.009696	5.03E-05	1.027619
100	10	20	1	1	1224.56	406.2	28256	1693	582	3109.35	735.02	160475	7464	4119	0.000918	0.003607	0.000246	0.884544
100	10	20	48	0	2338.17	248	33683	2319	4	12057.9	269.07	182435	12055	16	0.001033	0.006085	8.33E-07	0.990625
100	10	20	48	0.5	9912.88	305.18	33149	9710	2	49902.3	307.81	178386	49699	11	0.001009	0.024993	6.25E-07	1.157719
100	10	20	48	1	10165.2	277.24	32278	9744	3	50766	285.64	176190	50366	13	0.000999	0.025389	6.94E-07	1.1534

Table S2.2

Input Parameters					Results after 2,000 Generations					Results after 10,000 Generations					Final Results			
N (Pop Size)	***α*** (Offspr per ind)	***μ*** (Muta per gam)	***r*** (recom per gam)	***h*** (Dom coeff)	Ave fit per ind	Ave #SNPs per ind	Number of Fixed Mutations			Ave fit per ind	Ave #SNPs per ind	Number of Fixed Mutations			Fixation Probability (π)			K/μ ratio
							S = 0	S = 1	S = -1			S = 0	S = 1	S = -1	s=0	s=1	s=-1	
24	2	1	1	0	-1480	33	185	177	1481	-7330	34	968	840	7970	0.020391	0.019184	0.020862	0.991875
24	2	1	1	0.5	-1451	36	188	146	1459	-7354	32	945	869	8101	0.019714	0.02092	0.021354	1.01525
24	2	1	1	1	-1325	37	194	154	1382	-7169	34	985	863	7949	0.020599	0.020515	0.021113	1.008375
24	2	1	48	0	-1435	33	182	171	1412	-7160	34	986	903	7860	0.020938	0.021181	0.02073	0.998
24	2	1	48	0.5	-1398	33	198	146	1419	-7089	32	973	860	7849	0.020182	0.02066	0.020673	0.989875
24	2	1	48	1	-1398	34	198	133	1468	-7060	36	979	931	7924	0.020339	0.02309	0.020756	1.004375

24	2	5	1	0	-7328	147	934	847	7431	-36047	154	4969	4589	39920	0.021016	0.021655	0.020891	1.00665
24	2	5	1	0.5	-7179	167	825	792	7339	-36204	163	4915	4419	39977	0.021302	0.02099	0.020986	1.008875
24	2	5	1	1	-6894	167	853	876	7343	-35880	167	4882	4525	39997	0.020984	0.021117	0.020997	1.0083
24	2	5	48	0	-7336	165	932	819	7301	-36073	170	4992	4331	39517	0.021146	0.020324	0.020715	0.9947
24	2	5	48	0.5	-7341	166	894	823	7489	-36056	165	4949	4369	39766	0.02112	0.020521	0.020754	0.99695
24	2	5	48	1	-6894	170	903	817	7278	-35529	166	4990	4387	39481	0.021286	0.02066	0.020707	0.9965
24	2	10	1	0	-14708	372	1855	1686	14671	-72248	344	9827	8758	79336	0.02076	0.020463	0.02079	0.996363
24	2	10	1	0.5	-14324	327	1786	1616	14622	-71980	379	9829	8907	79491	0.020945	0.021097	0.020856	1.002538
24	2	10	1	1	-14030	341	1804	1687	14631	-71685	307	9856	8914	79682	0.020969	0.020911	0.020914	1.004125
24	2	10	48	0	-14833	331	1799	1644	14884	-72170	340	9846	8864	79310	0.020956	0.020891	0.020713	0.996163
24	2	10	48	0.5	-14360	332	1780	1616	14735	-72157	339	9657	8777	79565	0.020513	0.02072	0.020843	0.99835
24	2	10	48	1	-14184	330	1749	1663	14942	-71365	330	9764	8884	79402	0.020872	0.020894	0.020724	0.9962
24	2	20	1	0	-29183	683	3598	3394	29417	-143957	676	19674	17856	158663	0.020932	0.020923	0.020776	0.99865
24	2	20	1	0.5	-28841	659	3483	3311	29475	-144469	632	19652	17732	159887	0.021053	0.020864	0.020964	1.006263
24	2	20	1	1	-28504	701	3595	3199	29275	-143526	654	19313	17618	159027	0.020466	0.020861	0.020858	0.999306
24	2	20	48	0	-29526	672	3646	3363	29803	-144898	663	19612	17699	159456	0.020789	0.020741	0.020842	0.999719
24	2	20	48	0.5	-28842	662	3717	3274	29513	-144226	665	19677	17601	159222	0.020781	0.020728	0.020851	0.999975
24	2	20	48	1	-28219	669	3659	3347	29512	-143618	660	19549	17916	159472	0.02069	0.021078	0.020891	1.002619
24	5	1	1	0	550	7	311	637	82	2697	6	1700	3045	342	0.014469	0.02787	0.000334	0.507125
24	5	1	1	0.5	1614	9	250	1727	122	8009	9	1262	8528	529	0.010542	0.078715	0.000523	1.0275
24	5	1	1	1	90	32	176	633	542	-22	31	957	3097	3108	0.008135	0.028519	0.0033	0.726375
24	5	1	48	0	843	8	264	844	1	4389	8	1376	4407	15	0.011583	0.041238	1.8E-05	0.586125
24	5	1	48	0.5	2766	12	245	2739	0	13948	11	1224	13932	5	0.010198	0.129549	6.43E-06	1.522125
24	5	1	48	1	2580	17	162	2534	10	12978	16	888	12987	64	0.007563	0.120984	6.94E-05	1.404125
24	5	5	1	0	-3481	31	1169	1331	4738	-17070	31	5957	6545	23543	0.009975	0.012069	0.004837	0.720175
24	5	5	1	0.5	-172	53	1092	2704	2831	-904	49	5662	13841	14704	0.009521	0.02578	0.003054	0.6895
24	5	5	1	1	-3776	158	858	1341	4746	-19773	145	4942	7243	26720	0.008508	0.013662	0.005652	0.799
24	5	5	48	0	-1125	43	1073	1600	2614	-5856	44	5407	7869	13601	0.009029	0.014512	0.002826	0.53975
24	5	5	48	0.5	3695	64	1031	4296	572	19116	61	5291	21991	2854	0.008875	0.040961	0.000587	0.605925
24	5	5	48	1	2168	80	942	3689	1562	10794	72	4897	18903	8130	0.00824	0.035218	0.001689	0.643425
24	5	10	1	0	-9942	74	2223	2270	11994	-49542	65	11174	11497	60883	0.009324	0.010679	0.006287	0.838338
24	5	10	1	0.5	-4322	103	2313	3849	8003	-21339	90	11285	19955	41132	0.009346	0.018641	0.00426	0.727588
24	5	10	1	1	-9575	331	1751	2240	10626	-48068	325	10006	12583	59551	0.008599	0.011971	0.006292	0.844038
24	5	10	48	0	-6922	86	2080	2464	9145	-34793	98	10670	12710	47226	0.008948	0.011859	0.004897	0.711463
24	5	10	48	0.5	2240	135	2030	5520	3112	12078	156	10539	28473	16212	0.008864	0.026566	0.001685	0.557025
24	5	10	48	1	-644	161	1952	4753	5271	-3010	161	10133	24733	27647	0.008522	0.023125	0.002878	0.631713
24	5	20	1	0	-23437	122	4288	4172	27302	-116244	160	21517	21222	137032	0.008973	0.009867	0.007056	0.900056
24	5	20	1	0.5	-14298	211	4298	6269	20248	-71280	208	21660	31537	102414	0.009043	0.014623	0.005283	0.779975
24	5	20	1	1	-21179	550	3652	4116	23405	-109155	647	19954	22569	129452	0.008491	0.010679	0.006819	0.880013
24	5	20	48	0	-19631	170	4101	4538	23630	-95555	153	20821	23056	118121	0.008708	0.010716	0.006076	0.810806
24	5	20	48	0.5	-4073	306	3905	7839	11296	-19524	285	20390	40882	59804	0.008586	0.019122	0.003119	0.612725
24	5	20	48	1	-8387	314	3890	7197	15148	-42563	328	20135	36556	78624	0.008461	0.01699	0.004082	0.68175
24	10	1	1	0	869	5	409	914	43	4463	4	2145	4667	201	0.009042	0.021719	0.000102	0.705875
24	10	1	1	0.5	2460	7	269	2525	77	12267	7	1421	12622	373	0.006	0.058432	0.00019	1.443125
24	10	1	1	1	143	34	171	649	506	-681	143	795	2428	2846	0.00325	0.010295	0.001505	0.592875
24	10	1	48	0	1188	7	377	1191	3	6081	6	1777	6095	15	0.007292	0.02838	7.72E-06	0.7895
24	10	1	48	0.5	4125	9	267	4095	0	20661	9	1276	20642	2	0.005255	0.095758	1.29E-06	2.19475
24	10	1	48	1	3807	14	196	3734	1	19059	14	978	19001	10	0.004073	0.088351	5.79E-06	2.00725
24	10	5	1	0	-2984	21	1203	1552	4488	-14756	23	6615	7661	22370	0.005638	0.007071	0.0023	0.735075

24	10	5	1	0.5	1330	34	1166	3665	2319	6679	39	6041	18539	11832	0.005078	0.017215	0.001223	0.73155
24	10	5	1	1	-3529	147	887	1438	4617	-20943	561	4514	6447	24529	0.003778	0.005797	0.002561	0.7137
24	10	5	48	0	-1102	31	1213	1721	2773	-5095	26	6029	8500	13539	0.005017	0.007846	0.001385	0.559025
24	10	5	48	0.5	6119	56	1057	6324	243	30838	55	5584	32032	1224	0.004716	0.029755	0.000126	0.7804
24	10	5	48	1	4633	59	1051	5379	827	23042	70	4984	27399	4426	0.004097	0.025486	0.000463	0.7388
24	10	10	1	0	-9788	41	2268	2515	12218	-47961	39	11787	12674	60566	0.004958	0.005879	0.003109	0.850325
24	10	10	1	0.5	-2236	70	2333	4912	7069	-10866	70	11706	24793	35592	0.004882	0.011505	0.001834	0.722213
24	10	10	1	1	-8926	298	1785	2409	10229	-48937	852	9251	11584	55750	0.003889	0.00531	0.002927	0.777025
24	10	10	48	0	-7314	60	2151	2609	9783	-36894	50	11055	12979	49719	0.004638	0.006001	0.002568	0.740125
24	10	10	48	0.5	5771	107	2050	7600	1791	29363	116	10609	38838	9423	0.004458	0.018078	0.000491	0.592863
24	10	10	48	1	3029	129	1955	6676	3679	14561	128	10079	33661	19088	0.004231	0.015616	0.000991	0.631475
24	10	20	1	0	-23517	92	4308	4269	27581	-117140	73	21960	21687	138659	0.004597	0.00504	0.003571	0.913425
24	10	20	1	0.5	-11372	145	4517	7540	18695	-56535	112	22614	37301	93676	0.004713	0.008611	0.002411	0.767744
24	10	20	1	1	-21176	550	3659	4239	23315	-111031	1811	17833	20430	120464	0.003691	0.004685	0.003123	0.796963
24	10	20	48	0	-20382	86	4267	4587	24729	-102114	100	21210	23076	124952	0.004412	0.00535	0.003222	0.847844
24	10	20	48	0.5	1595	229	4022	10076	8158	8008	252	20962	51124	42774	0.004411	0.011877	0.001113	0.578775
24	10	20	48	1	-3057	239	3932	8994	11809	-16309	272	20420	45267	61376	0.004294	0.010496	0.001594	0.63955
50	2	1	1	0	-1528	40	170	144	1282	-7278	42	945	840	7704	0.009688	0.009667	0.00991	0.986625
50	2	1	1	0.5	-1470	42	155	147	1312	-7286	40	984	855	7854	0.010363	0.009833	0.010096	1.009875
50	2	1	1	1	-1290	40	149	150	1281	-7127	41	927	864	7845	0.009725	0.009917	0.01013	1.007
50	2	1	48	0	-1592	40	139	148	1329	-7420	41	936	878	7884	0.009963	0.010139	0.010116	1.01025
50	2	1	48	0.5	-1480	39	158	150	1358	-7545	40	943	858	8137	0.009813	0.009833	0.010461	1.034
50	2	1	48	1	-1302	39	156	137	1287	-7104	40	996	840	7787	0.0105	0.009764	0.010031	1.005375
50	2	5	1	0	-7499	201	785	740	6429	-36437	209	4742	4370	38941	0.009893	0.010083	0.010035	1.002475
50	2	5	1	0.5	-7121	208	755	700	6314	-35658	197	4676	4406	38682	0.009803	0.010294	0.00999	0.999875
50	2	5	1	1	-6950	196	873	714	6633	-36040	210	4749	4242	39086	0.00969	0.0098	0.010016	0.996425
50	2	5	48	0	-7566	198	807	700	6504	-35879	198	4851	4290	38468	0.01011	0.009972	0.009865	0.98995
50	2	5	48	0.5	-7209	205	842	766	6571	-35904	199	4689	4371	38890	0.009618	0.010014	0.009975	0.994275
50	2	5	48	1	-6869	198	793	749	6576	-35778	201	4826	4370	39120	0.010083	0.010058	0.010044	1.00495
50	2	10	1	0	-14854	393	1625	1475	13089	-72516	387	9529	8664	78035	0.00988	0.009985	0.010023	1.000488
50	2	10	1	0.5	-14263	401	1729	1454	12864	-71366	384	9647	8779	77484	0.009898	0.010174	0.009972	0.998288
50	2	10	1	1	-13935	373	1639	1475	13407	-71940	394	9512	8700	78428	0.009841	0.010035	0.010034	1.001488
50	2	10	48	0	-15117	403	1645	1395	13139	-72322	400	9706	8647	77582	0.010076	0.010072	0.009945	0.99695
50	2	10	48	0.5	-14464	402	1649	1421	13043	-72092	393	9592	8631	77926	0.009929	0.010014	0.010013	1.00045
50	2	10	48	1	-13731	401	1588	1426	12893	-70750	397	9615	8797	77307	0.010034	0.010238	0.00994	0.99765
50	2	20	1	0	-29598	836	3166	2888	25688	-145043	786	19158	17364	156022	0.009995	0.010053	0.010057	1.005013
50	2	20	1	0.5	-28949	809	3390	2958	26370	-144179	768	19306	17216	155899	0.009948	0.009901	0.009995	0.998144
50	2	20	1	1	-28139	800	3249	2997	26210	-143412	769	19256	17687	156360	0.010004	0.010201	0.010042	1.005294
50	2	20	48	0	-29369	790	3182	2917	25944	-144355	806	19368	17360	155267	0.010116	0.01003	0.009979	0.9997
50	2	20	48	0.5	-29047	790	3247	2855	26322	-144408	797	19295	17207	155978	0.01003	0.009967	0.010004	1.00035
50	2	20	48	1	-28063	793	3226	2971	26243	-143290	797	19084	17283	155711	0.009911	0.009939	0.00999	0.997738
50	5	1	1	0	697	7	322	729	30	3642	7	1649	3749	105	0.006635	0.016778	4.63E-05	0.55275
50	5	1	1	0.5	2086	8	270	2133	67	10370	8	1295	10706	351	0.005125	0.047628	0.000175	1.23525
50	5	1	1	1	655	28	174	885	301	2684	28	986	4526	1902	0.00406	0.020228	0.000988	0.75675
50	5	1	48	0	1197	8	235	1191	0	6076	9	1349	6081	10	0.00557	0.027167	6.17E-06	0.75175
50	5	1	48	0.5	4121	13	234	4057	0	20596	12	1119	20536	1	0.004425	0.09155	6.17E-07	2.170625
50	5	1	48	1	4102	18	178	3945	1	20314	18	980	20144	3	0.00401	0.089994	1.23E-06	2.125375
50	5	5	1	0	-2491	38	1211	1348	3723	-12120	34	5950	7032	19048	0.004739	0.006316	0.001892	0.6437
50	5	5	1	0.5	756	48	1132	3162	2362	3786	49	5799	15846	12001	0.004667	0.014093	0.00119	0.67475

50	5	5	1	1	-2159	139	798	1599	3258	-12622	140	4893	8742	20954	0.004095	0.007937	0.002185	0.72335
50	5	5	48	0	1399	54	967	2282	730	7971	51	5192	11982	3860	0.004225	0.010778	0.000386	0.426375
50	5	5	48	0.5	6301	71	991	6354	149	32216	70	5253	32868	753	0.004262	0.02946	7.46E-05	0.7845
50	5	5	48	1	5734	84	969	5901	388	28758	80	4942	30532	1986	0.003973	0.027368	0.000197	0.75505
50	5	10	1	0	-8496	71	2234	2391	10656	-41591	69	11137	12163	53522	0.004452	0.005429	0.002646	0.769263
50	5	10	1	0.5	-2946	94	2121	4292	7051	-14394	96	10886	22025	36266	0.004383	0.009852	0.001803	0.696413
50	5	10	1	1	-7237	294	1696	2583	8353	-38093	279	9907	14202	50868	0.004106	0.006455	0.002624	0.779313
50	5	10	48	0	-1658	116	1975	3430	4608	-7646	107	10518	17964	25147	0.004272	0.008074	0.001268	0.5452
50	5	10	48	0.5	6192	152	1973	7585	1316	32162	155	10400	39619	7381	0.004214	0.017797	0.000374	0.581575
50	5	10	48	1	4866	167	1940	7054	2243	24724	165	10202	36678	11965	0.004131	0.016458	0.0006	0.5951
50	5	20	1	0	-20858	140	4386	4466	24881	-104658	146	21440	22242	126385	0.004264	0.004938	0.003133	0.852088
50	5	20	1	0.5	-12177	206	4179	6539	18207	-60531	178	21330	33582	93671	0.004288	0.007512	0.002329	0.747863
50	5	20	1	1	-18337	556	3325	4324	19125	-94491	535	19657	24609	115882	0.004083	0.005635	0.002986	0.833588
50	5	20	48	0	-10481	217	3942	5768	15198	-51770	222	20069	29762	80486	0.004032	0.006665	0.002015	0.658806
50	5	20	48	0.5	2127	334	3823	9898	7028	13230	321	20401	52322	38378	0.004145	0.011784	0.000968	0.5647
50	5	20	48	1	-19	357	3642	9261	8646	576	328	19938	48655	47495	0.004074	0.010943	0.001199	0.590869
50	10	1	1	0	1042	5	394	1062	17	5173	5	2097	5263	89	0.004258	0.011669	2.22E-05	0.747
50	10	1	1	0.5	3021	7	259	3065	61	15114	7	1293	15360	270	0.002585	0.034153	6.45E-05	1.69225
50	10	1	1	1	600	31	130	822	264	1450	80	812	3376	1863	0.001705	0.007094	0.000494	0.604375
50	10	1	48	0	1522	6	307	1523	4	7827	6	1714	7839	13	0.003518	0.017544	2.78E-06	0.9665
50	10	1	48	0.5	5781	10	233	5712	2	29007	11	1200	28946	2	0.002418	0.064539	0	3.025125
50	10	1	48	1	5559	15	195	5395	1	27838	15	967	27690	5	0.00193	0.061931	1.23E-06	2.883875
50	10	5	1	0	-2353	19	1285	1569	3875	-11240	21	6499	7860	19048	0.002607	0.003495	0.000937	0.66695
50	10	5	1	0.5	2360	34	1210	4182	1815	11761	35	6131	21164	9387	0.002461	0.009434	0.000467	0.736875
50	10	5	1	1	-2096	124	809	1695	3343	-12532	194	4690	8826	20255	0.001941	0.003962	0.001044	0.6981
50	10	5	48	0	-40	31	1194	1945	1905	37	27	5734	9554	9441	0.00227	0.004227	0.000465	0.492125
50	10	5	48	0.5	9115	56	1017	9055	56	46397	57	5452	46619	345	0.002218	0.020869	1.78E-05	1.0572
50	10	5	48	1	8441	64	953	8388	212	42287	63	5067	43071	1034	0.002057	0.019268	5.07E-05	0.990475
50	10	10	1	0	-8795	44	2409	2510	11174	-43754	43	11741	12706	56348	0.002333	0.002832	0.001394	0.808775
50	10	10	1	0.5	-813	77	2321	5283	5987	-3422	66	11513	27259	30611	0.002298	0.006104	0.00076	0.6974
50	10	10	1	1	-6963	247	1706	2666	8251	-40858	658	8827	12284	46600	0.00178	0.002672	0.001184	0.6886
50	10	10	48	0	-6566	49	2179	2637	9036	-33119	45	11115	13257	46209	0.002234	0.00295	0.001147	0.709113
50	10	10	48	0.5	10050	114	2038	10696	723	51117	125	10623	54918	3862	0.002146	0.012284	9.69E-05	0.699325
50	10	10	48	1	8684	129	1896	9788	1308	43513	129	10061	50416	7110	0.002041	0.011286	0.000179	0.682438
50	10	20	1	0	-22492	79	4297	4424	26725	-111605	80	21889	22398	133726	0.002199	0.002496	0.001651	0.891044
50	10	20	1	0.5	-9055	149	4537	7856	16652	-45396	127	22386	39515	84663	0.002231	0.004397	0.00105	0.734494
50	10	20	1	1	-17670	477	3395	4410	18892	-96857	991	18195	22902	109580	0.00185	0.002568	0.0014	0.774875
50	10	20	48	0	-20402	87	4249	4342	24397	-100103	99	21337	22417	122124	0.002136	0.00251	0.001508	0.830563
50	10	20	48	0.5	7890	262	3935	13024	4848	41582	247	20691	67402	25543	0.002095	0.007553	0.000319	0.573931
50	10	20	48	1	5535	248	3776	12043	6362	28380	246	20032	61930	33420	0.002032	0.006929	0.000418	0.582506
100	2	1	1	0	-1571	44	122	104	941	-7490	45	947	779	7523	0.005156	0.004688	0.005079	1.01025
100	2	1	1	0.5	-1475	47	119	102	985	-7229	44	959	779	7470	0.00525	0.004701	0.005004	1.00025
100	2	1	1	1	-1282	46	126	109	1016	-7089	46	889	806	7529	0.004769	0.00484	0.005025	0.996625
100	2	1	48	0	-1661	45	113	124	1034	-7383	45	936	812	7416	0.005144	0.004778	0.004924	0.986625
100	2	1	48	0.5	-1430	46	111	114	990	-7304	45	913	820	7573	0.005013	0.004903	0.005079	1.011375
100	2	1	48	1	-1223	45	104	104	989	-6966	45	923	813	7432	0.005119	0.004924	0.004971	0.996375
100	2	5	1	0	-7600	224	674	548	4921	-36123	235	4619	4131	36646	0.004931	0.004976	0.004896	0.981325
100	2	5	1	0.5	-7252	231	626	522	4910	-36257	228	4644	4219	37792	0.005023	0.005135	0.005074	1.014925
100	2	5	1	1	-6630	228	622	528	4934	-35426	229	4557	4176	37292	0.004919	0.005067	0.004994	0.998525

100	2	5	48	0	-7624	228	574	520	4795	-36264	229	4712	4211	37140	0.005173	0.005126	0.004992	1.00435
100	2	5	48	0.5	-7266	224	597	574	5075	-36205	231	4447	4148	37463	0.004813	0.004964	0.004998	0.9953
100	2	5	48	1	-6505	225	554	543	4835	-34908	227	4639	4108	36780	0.005106	0.004951	0.00493	0.989875
100	2	10	1	0	-15208	444	1276	1110	10068	-72623	456	9298	8221	74749	0.005014	0.004938	0.004991	0.997675
100	2	10	1	0.5	-14471	454	1157	1178	9750	-71962	446	9251	8419	74735	0.005059	0.005028	0.005014	1.004
100	2	10	1	1	-13876	447	1219	1150	10155	-70931	463	9109	8415	74275	0.004931	0.005045	0.004948	0.990938
100	2	10	48	0	-14779	453	1188	1181	9561	-72214	453	9217	8341	74022	0.005018	0.004972	0.004974	0.995625
100	2	10	48	0.5	-14301	451	1198	1134	9867	-72258	453	9306	8420	74955	0.005068	0.00506	0.005022	1.006025
100	2	10	48	1	-13622	453	1192	1078	9879	-71503	451	9173	8370	75027	0.004988	0.005064	0.005027	1.005263
100	2	20	1	0	-29714	900	2405	2145	19614	-144399	893	18119	16596	148657	0.004911	0.005018	0.004979	0.99505
100	2	20	1	0.5	-28944	892	2424	2217	19780	-144111	946	18309	16570	148506	0.004964	0.004984	0.004966	0.993525
100	2	20	1	1	-28193	890	2443	2254	20389	-143396	901	18394	16699	149744	0.004985	0.005016	0.004991	0.998444
100	2	20	48	0	-30002	899	2358	2231	19895	-145185	894	18719	16663	149532	0.005113	0.005011	0.005001	1.002688
100	2	20	48	0.5	-28915	899	2512	2237	20001	-143910	894	18711	16530	149377	0.005062	0.004963	0.004991	0.999175
100	2	20	48	1	-27731	905	2507	2186	19651	-143544	897	18218	16554	149973	0.00491	0.004989	0.005028	1.002506
100	5	1	1	0	814	7	347	826	11	4045	7	1667	4103	58	0.0033	0.009103	1.45E-05	0.5805
100	5	1	1	0.5	2559	9	235	2575	54	12857	9	1229	13089	278	0.002485	0.029206	6.91E-05	1.4665
100	5	1	1	1	1265	24	146	1216	149	5823	24	962	6639	1000	0.00204	0.015064	0.000263	0.88625
100	5	1	48	0	1647	9	240	1640	1	8409	9	1301	8405	8	0.002653	0.018792	2.16E-06	0.979125
100	5	1	48	0.5	5716	13	179	5578	0	29012	14	1089	28865	2	0.002275	0.064686	6.17E-07	3.024875
100	5	1	48	1	5909	18	169	5604	0	29445	18	1003	29117	2	0.002085	0.065314	6.17E-07	3.043625
100	5	5	1	0	-1542	38	1153	1494	2919	-7871	40	5948	7435	15180	0.002398	0.003301	0.000757	0.574925
100	5	5	1	0.5	1634	46	1094	3491	1826	8473	48	5539	18190	9678	0.002223	0.008166	0.000485	0.6749
100	5	5	1	1	-767	114	776	1952	2304	-5213	117	4954	11132	15977	0.002089	0.0051	0.000844	0.675775
100	5	5	48	0	3309	59	959	3505	144	18429	65	5304	19283	799	0.002173	0.008766	4.04E-05	0.51945
100	5	5	48	0.5	9498	75	830	9240	44	48382	75	4929	48283	243	0.00205	0.021691	1.23E-05	1.083525
100	5	5	48	1	9349	83	916	8820	69	47252	83	4831	47103	421	0.001958	0.021268	2.17E-05	1.06375
100	5	10	1	0	-7006	70	2191	2470	9193	-34580	75	11168	12773	47053	0.002244	0.002862	0.001169	0.71425
100	5	10	1	0.5	-1649	97	2055	4650	6071	-7597	100	10777	24325	31693	0.002181	0.005465	0.000791	0.675238
100	5	10	1	1	-4942	232	1583	2892	6400	-26265	245	9708	16648	41284	0.002031	0.003821	0.001077	0.709563
100	5	10	48	0	3016	128	1772	5224	1756	18121	133	10032	28361	9792	0.002065	0.006427	0.000248	0.492913
100	5	10	48	0.5	10427	159	1821	10689	498	53823	165	10118	56484	2879	0.002074	0.012721	7.35E-05	0.705913
100	5	10	48	1	10076	173	1778	10322	719	51093	172	10019	54651	4069	0.00206	0.012314	0.000103	0.699
100	5	20	1	0	-18790	148	4168	4620	22782	-93196	152	21489	23689	116308	0.002165	0.002648	0.001443	0.811975
100	5	20	1	0.5	-10313	187	4190	7042	16724	-50564	196	21288	35996	85949	0.002137	0.004021	0.001068	0.720481
100	5	20	1	1	-14803	473	3037	4631	15290	-77889	480	19215	26996	100586	0.002022	0.003106	0.001316	0.773994
100	5	20	48	0	-2054	281	3714	7725	8184	-5518	283	20146	42292	46191	0.002054	0.004801	0.000587	0.556288
100	5	20	48	0.5	8461	335	3523	12648	3703	46828	325	19840	68765	21379	0.00204	0.007794	0.000273	0.563188
100	5	20	48	1	7642	350	3600	12295	4432	40884	343	19897	66270	25170	0.002037	0.007497	0.00032	0.568813
100	10	1	1	0	1131	5	382	1144	11	5668	5	2047	5720	52	0.002081	0.006356	6.33E-06	0.78525
100	10	1	1	0.5	3557	6	232	3565	42	18003	7	1276	18164	202	0.001305	0.020276	2.47E-05	1.975375
100	10	1	1	1	1538	20	163	1480	161	5199	39	922	6044	1023	0.000949	0.006339	0.000133	0.773125
100	10	1	48	0	2032	7	289	2023	0	10364	6	1574	10358	5	0.001606	0.011576	7.72E-07	1.203125
100	10	1	48	0.5	7895	11	207	7755	0	39482	11	1124	39347	2	0.001146	0.043878	3.09E-07	4.063875
100	10	1	48	1	7811	14	196	7509	1	39141	14	983	38829	2	0.000984	0.0435	1.54E-07	4.0135
100	10	5	1	0	-1271	25	1269	1693	2897	-6143	22	6508	8479	14546	0.00131	0.001885	0.00036	0.59185
100	10	5	1	0.5	3241	34	1192	4750	1518	16400	35	5862	24045	7649	0.001168	0.00536	0.000189	0.7524
100	10	5	1	1	-744	108	776	2036	2329	-8282	237	4127	9030	14507	0.000838	0.001943	0.000376	0.563075
100	10	5	48	0	2509	32	1146	2957	387	13168	33	5776	15128	1896	0.001158	0.003381	4.66E-05	0.45775

100	10	5	48	0.5	12983	60	930	12698	32	65582	55	5283	65415	169	0.001088	0.014644	4.23E-06	1.430175
100	10	5	48	1	12761	64	903	12245	43	63664	66	5014	63351	247	0.001028	0.014196	6.3E-06	1.385525
100	10	10	1	0	-7184	48	2316	2717	9726	-36516	49	12015	13196	49582	0.001212	0.001455	0.000615	0.750425
100	10	10	1	0.5	635	70	2204	5913	5201	3353	70	11508	29953	26522	0.001163	0.003339	0.000329	0.683313
100	10	10	1	1	-4484	190	1609	3063	6315	-25171	191	9843	17340	41183	0.001029	0.001983	0.000538	0.717238
100	10	10	48	0	-2587	50	2149	3174	5533	-12980	68	10865	16207	28907	0.00109	0.00181	0.000361	0.564038
100	10	10	48	0.5	14823	123	1861	14755	307	76208	125	10420	77377	1564	0.00107	0.008698	1.94E-05	0.905475
100	10	10	48	1	14224	120	1855	14056	417	72295	121	9935	73909	2213	0.00101	0.008313	2.77E-05	0.871613
100	10	20	1	0	-20397	88	4375	4693	24763	-100969	91	21936	23353	124021	0.001098	0.001296	0.000766	0.846744
100	10	20	1	0.5	-7055	134	4326	8311	15066	-35175	140	21681	42356	77203	0.001085	0.002364	0.000479	0.709606
100	10	20	1	1	-14094	350	3298	5270	16635	-79478	649	18426	26239	96412	0.000946	0.001456	0.000616	0.724213
100	10	20	48	0	-14492	109	4131	5207	19196	-72131	109	20965	26016	97592	0.001052	0.001445	0.000605	0.725244
100	10	20	48	0.5	14782	262	3749	17173	2389	76545	252	20263	89926	13384	0.001032	0.005052	8.48E-05	0.626638
100	10	20	48	1	13793	254	3756	16472	2961	70206	243	20187	86269	16345	0.001027	0.004847	0.000103	0.622575

2.10 Other Supplementary Files

2.1 The User Guide for GEMA

(Due to the large volume of this file, it can be assessed by the URL: http://gbe.oxfordjournals.org/content/suppl/2014/04/10/evu075.DC1/GEMA_SupplementaryFileS1.docx)

2.11 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).

2.12 References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. doi: 10.1038/nature11632

Bechtel JM, Wittenschlaeger T, Dwyer T, Song J, Arunachalam S, Ramakrishnan SK, Shepard S, Fedorov A. 2008. Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. Bmc Genomics 9: 284. doi: 1471-2164-9-284 [pii] 10.1186/1471-2164-9-284

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. doi: Doi 10.1073/Pnas.0701652104

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. doi: Artn 223 Doi 10.1186/1471-2105-9-223

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

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008a. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9. doi: Artn 364 Doi 10.1186/1471-2105-9-364

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008b. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9: 364. doi: 10.1186/1471-2105-9-364

Charlsworth BaC, Deborah. 2010. Elements of Evolutionary Genetics. Greenwood Village, Colorado: Roberts and Comapany 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. doi: 10.1038/ncomms3417

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. doi: 10.1007/s00285-007-0121-7

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P. 2011. Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714. doi: 10.1038/ng.862

Consortium TIH. 2005. A haplotype map of the human genome. Nature 437: 1299-1320. doi: 10.1038/nature04226

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, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima

Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861. doi: 10.1038/nature06258

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. 2007. Principles of population genetics.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786-2787. doi: Doi 0.1093/Bioinformatics/Btn522

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. doi: 10.1038/84773

Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, van Duijn CM, Slagboom PE, van Ommen GJ, Wijmenga C, de Bakker PI, Sunyaev SR. 2013. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS genetics 9: e1003301. doi: 10.1371/journal.pgen.1003301

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. doi: 10.1038/nature10231

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. doi: 10.1146/annurev-genom-082908-150129

Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theoretical Population Biology 49: 128-142. doi: Doi 10.1006/Tpbi.1996.0007

Patwa Z, Wahl LM. 2008. The fixation probability of beneficial mutations. Journal of the Royal Society, Interface / the Royal Society 5: 1279-1289. doi: 10.1098/rsif.2008.0248

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 12: 32-42. doi: Doi 10.1038/Nrg2899

Prakash A, Shepard SS, Mileyeva-Biebesheimer O, He J, Hart B, Chen M, Amarachintha SP, Bechtel J, Fedorov A. 2009. Evolution of Genomic Sequence Inhomogeneity at Mid-range Scales. Bmc Genomics 10: 513. doi: 10.1186/1471-2164-10-513

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. doi: 10.1073/pnas.0700890104

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. doi: 10.1098/rstb.2009.0278

Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nature reviews. Genetics 9: 965-974. doi: 10.1038/nrg2473

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. doi: 10.1098/rspb.1965.0026

Zhang C, Plastow G. 2011. Genomic Diversity in Pig (Sus scrofa) and its Comparison with Human and other Livestock. Current Genomics 12: 138-146. doi: 10.2174/138920211795564386

Chapter 3

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

Shuhao Qiu[1, 2] and Alexei Fedorov[1, 2*]

Authors' Affiliations:

[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.

* To whom correspondence should be addressed. Tel +419-383-5270; Fax +419-383-3102; e-mail: Alexei.fedorov@utoledo.edu

Keywords: genomics, computational biology, allelic age, SNP, haplogroups

Accepted in the Genomics, February 16, 2015, doi: 10.1016/j.ygeno.2015.02.005

3.1 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.

3.2 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 (Kimura and Ohta 1973; Maruyama 1974a; Maruyama 1974b; Li 1975). Initially, prediction of allelic age relied upon mathematical modeling – a diffusion approximation for a branching process. In 1973 Kimura and Ohta (Kimura and Ohta 1973) 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 (Maruyama 1974a) 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 (LI 1975) 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 (Griffiths and Tavare 1999); and early 2000s (Slatkin and Rannala 2000). The allelic age has been indirectly estimated in several independent experimental studies that statistically examined the distributions of multiple mutant alleles. Slatkin and Rannala estimated the allelic age by use of intra-allelic

variability (Slatkin and Rannala 1997). Further, Rannala and Reeve applied high-resolution multipoint linkage-disequilibrium mapping (Rannala and Reeve 2001), while Genin and colleagues analyzed shared haplotypes of rare disease mutations (Genin et al. 2004). Last year Kiezun and co-authors, 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 (Kiezun et al. 2013a). 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 (Abecasis et al. 2010a). 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 (Arnheim et al. 2003).

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 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.

3.3 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 (QIU *et al.* 2014). Both v2 and v3 versions are freely available from our web site: http://bpg.utoledo.edu/~afedorov/lab/GEMA.html. 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 ($\boldsymbol{\alpha}$=5); 5) mutation rate per gamete ($\boldsymbol{u}$=20); 6) recombination rate ($\boldsymbol{r}$=1 or $\boldsymbol{r}$=48); 7) dominance coefficient ($\boldsymbol{h}$=0, $\boldsymbol{h}$=0.5, or $\boldsymbol{h}$=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 Figure 1. Experiments B and C were first described in our paper (Qiu et al. 2014) and we keep 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 ($\boldsymbol{s}$) always equal to -1. Consequently, all beneficial mutations are also equal to each other ($\boldsymbol{s}$=+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. 2014), 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. Thousand-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.

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; 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 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.

3.4 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 (***μ***=20 mutations per gamete); 3) the mating scheme was a default GEMA choice -- permanent random male-female pairs (MatingScheme =1) with 5 offspring per mating pair (***α***=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 (Qiu et al. 2014)]. 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); 3) Distribution of mutations by their selection coefficients corresponded to the "Experiment B" or "Experiment C" shown in Figure 1. We specifically used ***r***=48, because it represents the average number of pieces of paternal and maternal genomes in a human gamete (Qiu et al. 2014). The alternative ***r***=1 settings model the regions with low recombination rate frequency, which are abundant in the human genome.

Distribution of SNPs by their age for different modeled parameters is shown in Figure 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 Figure 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 (Qiu et al. 2014), 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 Figure 3. These SNPs were grouped by their selection coefficients (***s***) into being deleterious (***s***=-1), neutral (***s***=0), or beneficial (***s***=+1). Figure 3 illustrates that recombination rate (***r***), dominance coefficient (***h***), and the distribution of mutation by selection coefficients (experiments B or C) may significantly influence on the mean allelic age. In 75% of experiments no difference in the allelic ages for neutral, beneficial, or deleterious mutations was detected. Only with high recombination (***r***=48) and "Experiment B" settings (90% neutral SNPs) was the mean age of neutral mutations 1.4-2.6 times higher than for deleterious or beneficial ones, in accordance to Maruyama's predictions (Maruyama 1974a).

Finally, the distribution of SNPs by their ages is demonstrated in Figure 4. For proper comparison of different experiments with various parameters, the number of SNPs

has been normalized by division by the total number of SNPs in the experiment. Thus, Figure 4 represents SNP frequency density and provides an overall view of the age distribution of all SNPs. Statistical information about these distributions including beneficial, deleterious, and neutral groups of SNPs is presented in Table 1 (detailed information on the age of each SNP is presented in supplementary Tables S1 and S2).

3.5 Discussion

Maruyama made a non-obvious and intriguing theoretical prediction about the average age of deleterious, beneficial, and neutral mutations. Experimental verification of SNP age encounters two major problems. First is ascertaining the real age of mutations that occurred many generations ago. Second is assessing the deleterious, beneficial, or neutral effects for mutations, which is unknown for the vast majority of human SNPs. The unexpectedly young age has been deduced only for several hundred mutations located in about 50 different loci that are associated with recent strong positive selection in the human genome (Sabeti et al. 2002; Voight et al. 2006; Sabeti et al. 2007). Among the 22 strongest candidate loci for positive selection in humans presented by (Sabeti et al. 2007) in Table 1, the authors characterized 41 possibly functional SNPs and additional closely located 439 SNPs, which are in strong linkage disequilibrium and propagate with these beneficial mutations by genetic hitchhiking. Because the set of characterized beneficial mutations is tiny compared to all known human SNPs, and because the set of hitchhiking SNPs is many fold larger than the set of known beneficial SNPs, it is impossible to evaluate the Maruyama effect from these principal public datasets. One of the most comprehensive experimental evaluations of the Maruyama effect has been reported by Keizun and co-authors (Kiezun et al. 2013a). Using whole-genome computation analysis the authors examined thousands of putatively deleterious missense SNPs inside protein coding sequences and compared them with synonymous mutations. The authors concluded that deleterious alleles are, on average, younger than neutral ones. However, the analysis was qualitative and did not provide a precise quantitative estimation of the Maruyama effect. Also, the influence of important genomic parameters on the mutation

age (*e.g.*, local recombination rates, coefficient of dominance for genes under analysis) has not been examined.

Presently, even with the availability of about 3,000 completed human genomes in public databases, there is a limitation of genomic data on families that includes sequences from members of several generations. Yet this kind of information is required for evaluating allelic age. However, in a few years the technology race to develop a fast sequencing device with "$100 per genome" capacity should be accomplished. With such technology, whole-genome sequencing analysis of large pedigrees will become routine. In addition, there are several long-running selection projects with laboratory animals (like mice and rats), where frozen materials from animals across numerous generations have been preserved (Wisloff et al. 2005). The availability of cheap sequencing in the near future will provide unprecedented genomic data on extra-long pedigrees, across multiple generations of humans and other species. Such data open the possibility of a direct investigation of the fate and the age of many mutations. Hence, a precise estimation of the Maruyama predictions will soon be possible.

In this respect, mathematical modeling provides an important insight into this problem. However, existing mathematical approaches for inferring allelic age consider only one mutation at a time, while possible interactions of SNPs with one another have been ignored (Kimura and Ohta 1973; Maruyama 1974b; Li 1975). The "1000 Genomes" project recently revealed 38 million SNPs within the pool of sequenced genomes, and demonstrated that two non-related humans from the same population have over three million SNP differences between them (Abecasis et al. 2010a). Each human gene bears hundreds of SNPs, arranged in several major haplogroups having strong

linkage disequilibrium between SNPs from the same haplotype. Since mutations never exist alone, to understand their dynamics they should be modeled/analyzed in the context of haplotypes. Keeping this in mind, we implemented whole-genome computational simulations to investigate how different haplogroup structures influence the average age of SNPs. In our GEMA simulations, we used the lowest estimated value of the influx of novel mutations observed in humans (20 novel mutations per gamete) (Kondrashov and Shabalina 2002b; Conrad et al. 2011b; Li and Durbin 2011). Such an influx, even in a very small population of 100 modeled individuals, generates thousands of SNPs randomly distributed among 600 genes. Closely located mutations are linked together and form haplotypes. The length of the haplotypes depends on the recombination rate (***r***).

These haplotypes compete with one another via natural selection. Each non-neutral mutation contributes to the total fitness of the model individual, which is calculated by taking into account all beneficial and deleterious mutations and the dominance coefficients (***h***) of the genes in the modeled genomes. In our simulations we applied the ultimate selection mode, in which only the fittest offspring survive and form the next generation. Our computations demonstrated that the recombination rate, dominance coefficient, and overall distribution of the entire pool of SNPs by their selection coefficients significantly affect the mean allelic age of SNPs. The Maruyama effect (Maruyama 1974a) was detected only when the recombination rate was high (***r***=48) and the neutral mutations were overabundant (90% of SNPs are neutral in Experiment B). Under these conditions, the average age of neutral mutations was 1.4 times higher than deleterious and 2.3 times higher than beneficial ones (Figure 3 and Table 1). However under the same conditions (***r*** = 48 and ***h*** = 0.5) if the frequency of neutral mutations is

decreased to 10% and the frequencies (but not ratio) of deleterious and beneficial mutations are increased (experiment C), the average ages of mutations with different selection coefficients were practically the same (no Maruyama effect).

Our results demonstrate the fruitfulness of the whole-genome computational simulation approach for population genetics, and its benefits over mathematical modeling. All in all, GEMA programs allow investigation of the integrative effects of thousands of mutations per individual, and evaluation of the effects of grouping of mutations into haplotypes.

3.6 Table and Figure Legends

Table 3.1 Distributions of beneficial, deleterious, and neutral SNPs by their ages.

The parameters provided in the table are the following: ***s*** means the selection coefficients of SNPs; ***r*** - recombination rate; ***h*** - dominance coefficient; <age>± CI - mean allelic age and its 95% confidence interval; STDEV - standard deviation for the distribution of SNPs by their ages; #SNPs - number of SNPs; and "F Range" - the SNP population Frequency range for the analyzed SNPs ("High"= 40% - 60%; "Low"= 10% - 30%).

Figure 3-1 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.

Figure 3-2 Distribution of relative frequencies of all SNPs in population by their age (measured in generations since SNP arrival).

The extreme right block of columns marked as "1000" on the horizontal axis shows the relative frequencies of all accumulated SNPs at 110-1000 generations.

Figure 3-3 Mean allelic age of SNPs with different selection coefficients calculated for different experimental conditions.

Allelic age was measured in generations passed after the SNP arrival. All analyzed SNPs had current population frequencies in the range from 40% to 60% (Panel A) or in the range from 10% to 30% (Panel B). Each error bar reflects 95% confidence interval (**CI**) for the respective experiment. The modeled selection coefficient (***s***) for each SNP had three possible values: either +1 (for beneficial mutations), 0 (for neutral mutations), or -1 (for deleterious mutations). All computations were performed for a population size of N=100; twenty novel mutations per gamete (μ=20); and five offspring per individual (α=5). In each individual experiment, the modeled number of recombination events per gamete was either r=1 or r=48; the dominant coefficient for each gene was either h=0 (dominant genes), or h= 0.5 (co-dominant genes), or h=1 (recessive genes). The distribution of mutations by selection coefficients was either from "experiment B" (90% of starting SNPs neutral) or "experiment C" (10% of starting SNPs neutral), as described in Figure 1. Exact allelic age for each SNP in these experiments is provided in Supplementary Table S2.

Figure 3-4 Distribution of SNPs by their age.

Panel A – shows all SNPs having frequencies from 40% to 60% in the population. **Panel B** – all SNPs having frequencies of 10% to 30% in the population. Each curve

represents an experiment with specific ***h*** and ***r*** parameters. The exact numbers of SNPs in the experiments are provided in the Table 1 and Supplementary Table S2.

3.7 Tables and Figures

Table 3.1

ExperimentC							ExperimentB						
r	h	s	<age> ± CI	STDEV	# SNPs	F Range	r	h	s	<age> ± CI	STDEV	# SNPs	F Range
1	0.5	-1	51.7±1.4	34.5	2417	High	1	0.5	-1	46.3±2.6	28.5	471	High
		0	51.1±3.1	34.6	480				0	49.5±0.3	31.0	34241	
		1	51.7±2.4	32.7	732				1	47.6±1.5	30.1	1476	
48	0.5	-1	160.5±3.2	108.4	4338	High	48	0.5	-1	130.5±32.1	83.4	26	High
		0	176.4±5.5	121.6	1881				0	187.9±0.8	132.1	92944	
		1	160±3.4	108.5	3803				1	83±1.6	48.8	3417	
1	0.5	-1	29.9±0.6	26.5	8467	Low	1	0.5	-1	26.3±0.8	21.5	2985	Low
		0	30.2±1.3	26.0	1480				0	31.7±0.1	26.0	128074	
		1	32.8±1.2	28.7	2041				1	32.1±0.8	25.8	4166	
48	0.5	-1	83.7±1	80.1	26786	Low	48	0.5	-1	40.7±1	30.3	3265	Low
		0	100.2±2.5	98.5	6148				0	105.3±0.3	106.4	462951	
		1	97.8±2.1	93.8	7923				1	52.2±0.9	42.4	9415	
1	0	-1	25±0.9	12.9	878	High	1	0	-1	27.4±6.2	10.0	10	High
		0	25.1±2.3	14.0	145				0	40.7±0.8	20.9	2800	
		1	24±1.9	11.6	149				1	39.9±9.4	24.0	25	
48	0	-1	145.7±2.5	99.5	6175	High	48	0	-1	118.2±56.4	70.5	6	High
		0	152±4.7	104.3	1857				0	209.3±2.2	152.1	17704	
		1	149.2±4.1	102.7	2458				1	100.4±8.5	57.2	173	
1	0	-1	15.6±0.2	8.3	4893	Low	1	0	-1	20±1.9	12.2	161	Low
		0	15.8±0.6	8.5	832				0	27±0.3	18.4	13366	
		1	15.6±0.6	8.5	752				1	24.8±2.7	16.6	144	
48	0	-1	76.4±0.9	77.4	26512	Low	48	0	-1	54.6±16.7	87.1	105	Low
		0	85.6±2.4	87.9	5208				0	115.9±1	122.2	61609	
		1	85.8±2.2	85.9	5871				1	65.9±5.1	63.4	599	
1	1	-1	335.6±3.2	262.7	25814	High	1	1	-1	384.7±13.5	314.3	2077	High
		0	333.2±8.2	266.2	4076				0	376±3	306.0	40146	
		1	329.1±7.1	257.4	5111				1	380.4±17.2	316.6	1295	
48	1	-1	171.4±3	114.2	5655	High	48	1	-1	121.3±18.4	74.6	63	High
		0	184.5±5.3	121.9	2018				0	194.8±2	138.3	18575	
		1	165.5±3.5	110.5	3789				1	98.8±3.9	62.9	1018	
1	1	-1	184.9±1.6	222.2	76851	Low	1	1	-1	199.2±5.7	257.5	7813	Low
		0	185.4±4.2	220.3	10558				0	201.7±1.6	258.1	99795	
		1	188.4±4.2	227.7	11271				1	204.2±11.7	265.1	1960	
48	1	-1	91.3±0.9	87.8	35368	Low	48	1	-1	67±2.1	53.2	2455	Low
		0	104.1±2.5	102.5	6596				0	108.7±0.9	110.2	58373	
		1	100.2±2.2	98.4	7825				1	52.5±2.5	44.6	1216	

Figure 3-1

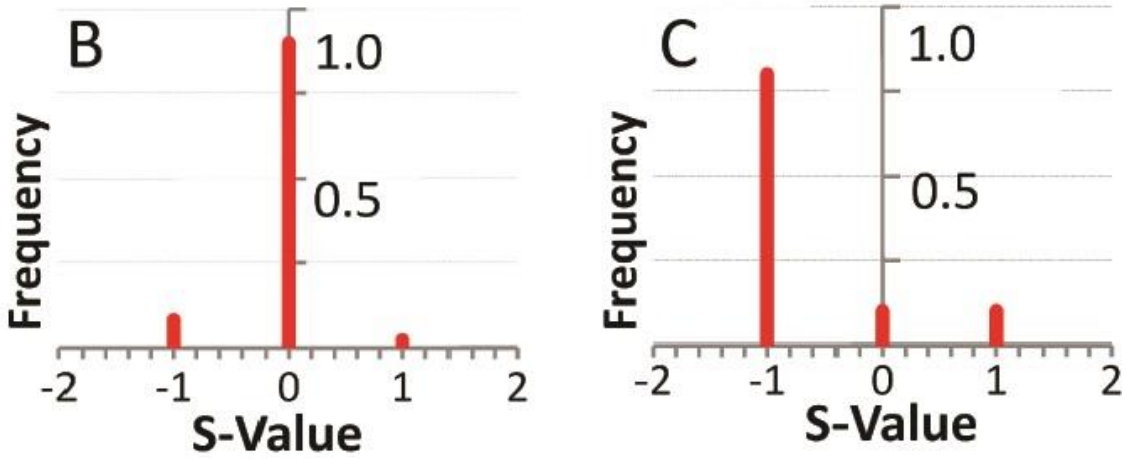

B
1.0
0.5
Frequency
-2
-1
0
1
2
S-Value
C
1.0
0.5
Frequency
-2
-1
0
1
2
S-Value

Figure 3-2

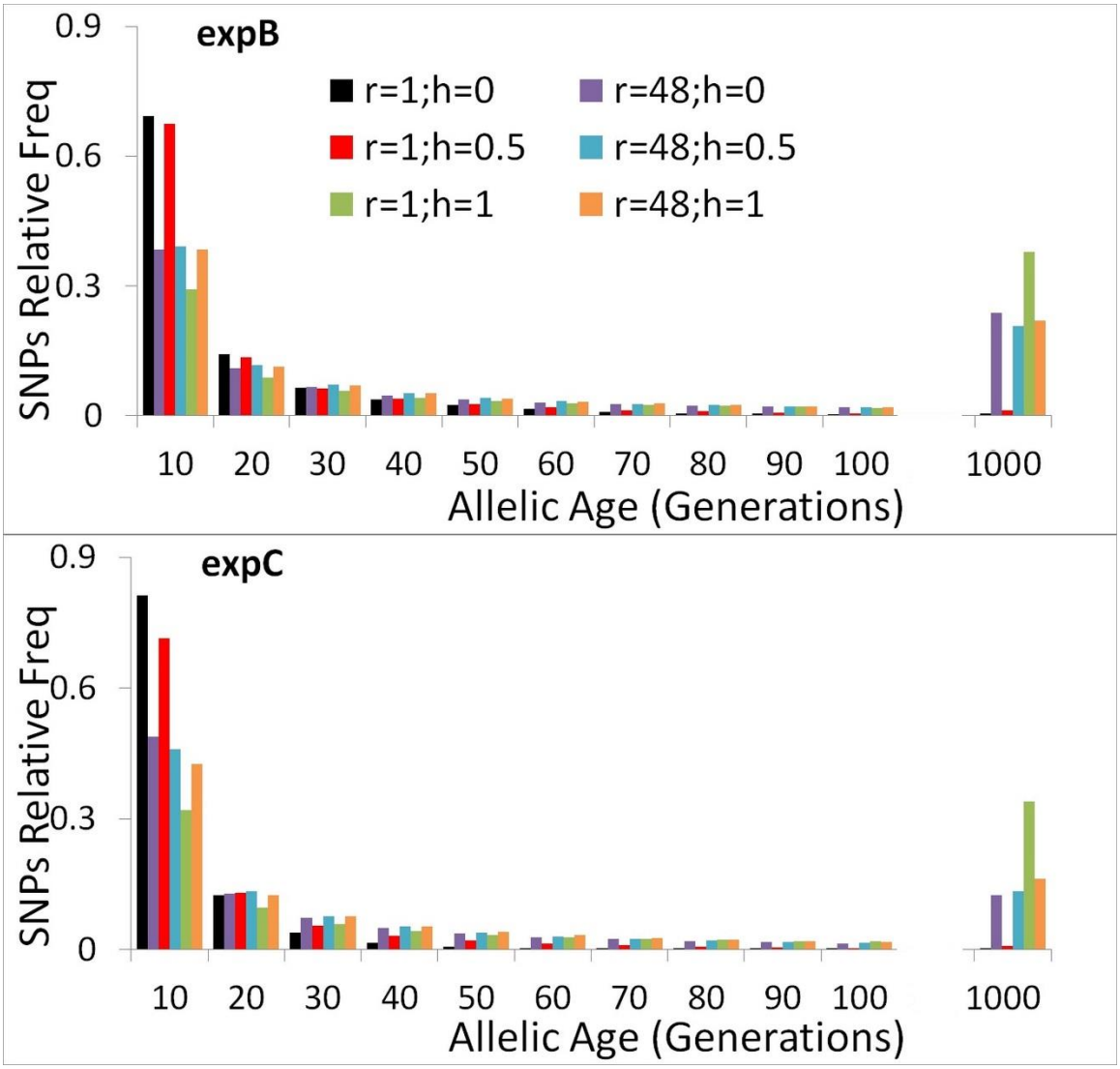
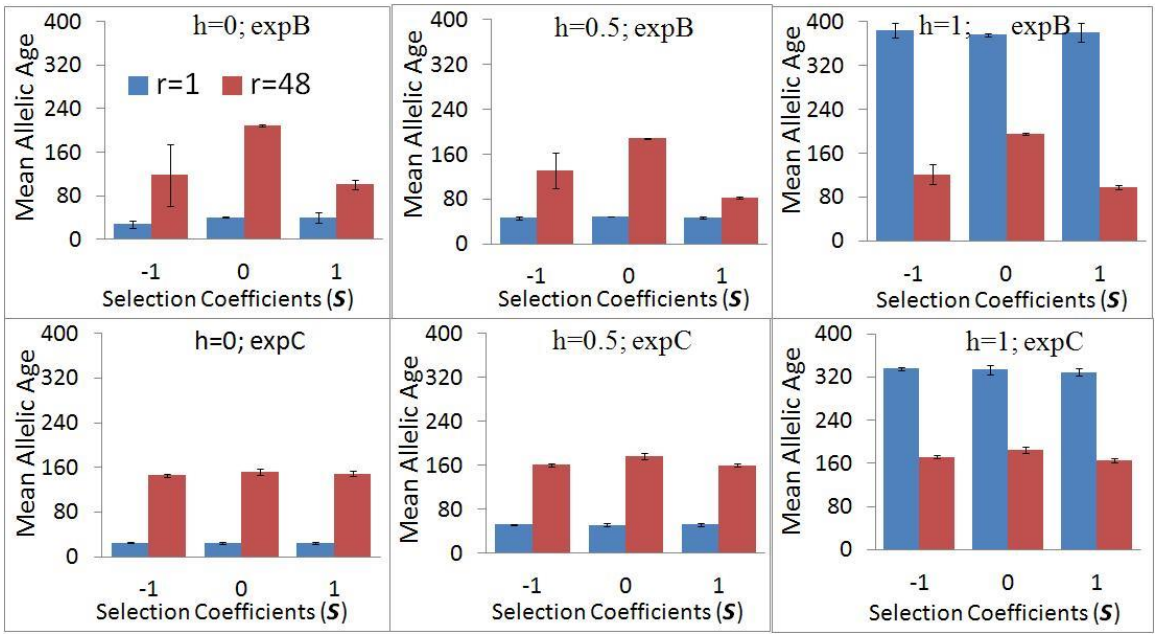

expB
r=1;h=0
r=1;h=0.5
r=1;h=1
r=48;h=0
r=48;h=0.5
r=48;h=1
SNPs Relative Freq
0.9
0.6
0.3
0
10
20
30
40
50
60
70
80
90
100
1000
Allelic Age (Generations)
expC
SNPs Relative Freq
0.9
0.6
0.3
0
10
20
30
40
50
60
70
80
90
100
1000
Allelic Age (Generations)

Figure 3-3

Panel A



Panel B

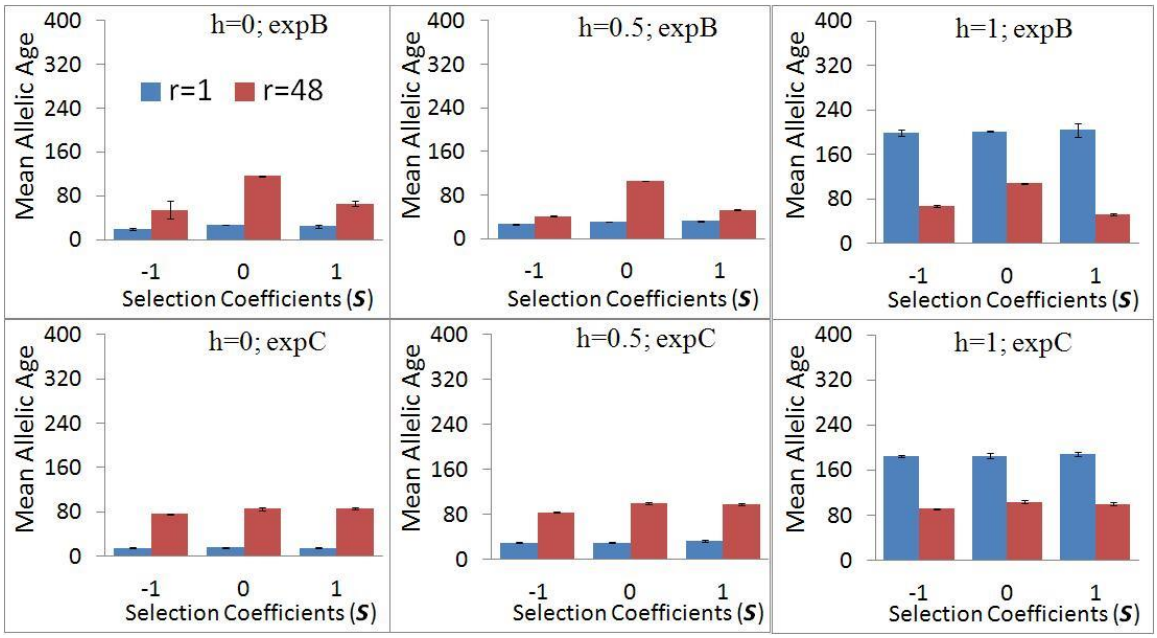
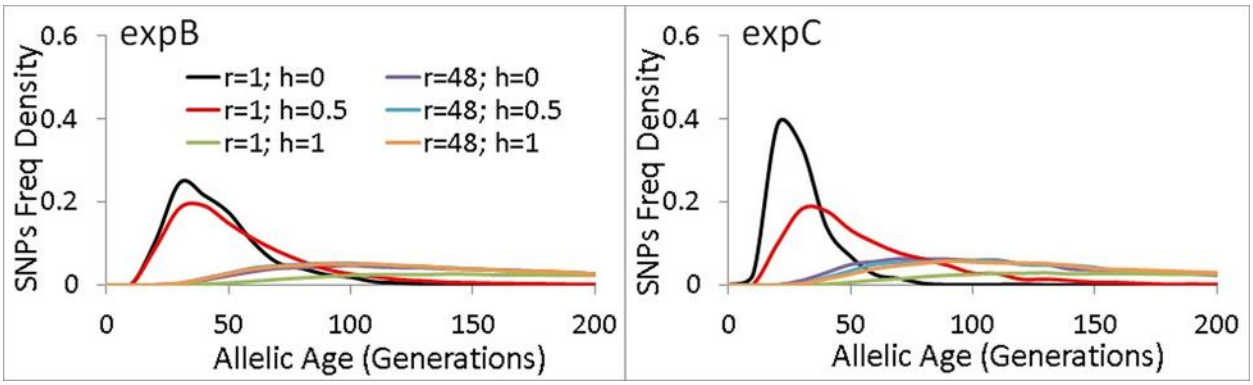
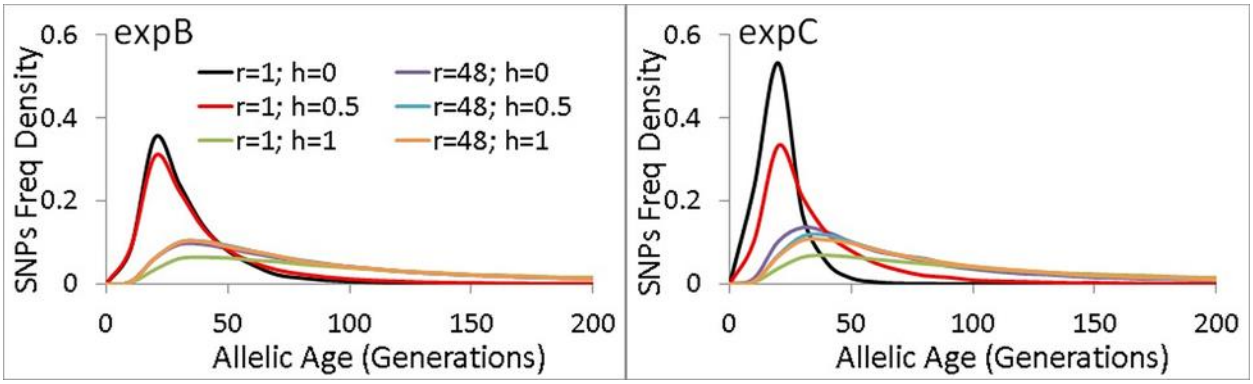


Figure 3-4

Panel A



Panel B



3.8 Supplemental Table Legends

Table S3.1 Number and Frequency density of SNPs for different allelic age ranges in Experiment C

The allelic age column means the allelic age ranges. For example, 10 means the generations from (0...10) while 20 means (10...20). Results were collected after 10,000 generations. The results of 12 experiments were collected from running the AllelicAge_10bin.pl with our raw material provided by the Backup file from GEMA.pl

Table S3.2 Number and Frequency density of SNPs for different allelic age ranges in Experiment B

The allelic age column means the allelic age ranges. For example, 10 means the generations from (0...10) while 20 means (10...20). Results were collected after 10,000 generations. The results of 12 experiments were collected from running the AllelicAge_10bin.pl with our raw material provided by the Backup file from GEMA.pl.

Table S3.3 Results of 12 experiments' raw data were collected from running the AllelicAge_csv.pl with our raw material provided by the Backup file from GEMA.pl with SNPs frequency from 40% to 60% in Experiment C

Table S3.4 Results of 12 experiments' raw data were collected from running the AllelicAge_csv.pl with our raw material provided by the Backup file from GEMA.pl with SNPs frequency from 40% to 60% in Experiment B

Table S3.5 Results of 12 experiments' raw data were collected from running the AllelicAge_csv.pl with our raw material provided by the Backup file from GEMA.pl with SNPs frequency from 10% to 30% in Experiment C

Table S3.6 Results of 12 experiments' raw data were collected from running the AllelicAge_csv.pl with our raw material provided by the Backup file from GEMA.pl with SNPs frequency from 10% to 30% in Experiment B

3.9 Supplemental Tables

Table S3.1

ExpC; N=100; μ=20; α=5;

	Number of SNPs for different allelic age ranges						Frequency density of SNPs for different allelic age ranges					
Allelic Age	r=1; h=0	r=48; h=0	r=1; h=0.5	r=48; h=0.5	r=1; h=1	r=48; h=1	r=1; h=0	r=48; h=0	r=1; h=0.5	r=48; h=0.5	r=1; h=1	r=48; h=1
10	124088	141749	135686	154408	148591	147106	0.8132545	0.4879064	0.7147839	0.4590626	0.3203656	0.4255702
20	18935	37120	24801	44834	44448	43251	0.1240972	0.1277687	0.1306499	0.1332937	0.0958309	0.125123
30	5956	21102	10454	25721	26913	26224	0.0390347	0.072634	0.0550709	0.0764698	0.058025	0.0758647
40	2305	14253	5993	17609	19503	18190	0.0151066	0.0490595	0.0315707	0.0523524	0.0420489	0.0526227
50	867	10691	3865	13073	15464	14003	0.0056822	0.0367989	0.0203605	0.0388667	0.0333407	0.04051
60	290	8128	2567	10204	12978	11258	0.0019006	0.0279769	0.0135228	0.030337	0.0279809	0.0325688
70	100	6817	1823	8147	11143	9154	0.0006554	0.0234644	0.0096034	0.0242214	0.0240246	0.0264821
80	29	5571	1317	6850	10155	7780	0.0001901	0.0191756	0.0069379	0.0203654	0.0218944	0.0225071
90	7	4739	1001	5769	9094	6665	4.588E-05	0.0163118	0.0052732	0.0171515	0.0196069	0.0192815
100	2	4086	627	5000	8300	5824	1.311E-05	0.0140642	0.003303	0.0148652	0.017895	0.0168485
110	1	3658	469	4413	7433	5298	6.554E-06	0.012591	0.0024707	0.0131201	0.0160257	0.0153268
120	2	3236	293	3867	7144	4676	1.311E-05	0.0111385	0.0015435	0.0114968	0.0154026	0.0135274
130	0	2905	233	3476	6540	4216	0	0.0099991	0.0012274	0.0103343	0.0141004	0.0121967
140	0	2538	157	3081	5962	3749	0	0.0087359	0.0008271	0.00916	0.0128542	0.0108457
150	0	2279	130	2716	5716	3303	0	0.0078444	0.0006848	0.0080748	0.0123238	0.0095554
160	0	2018	101	2494	5451	3117	0	0.006946	0.0005321	0.0074148	0.0117525	0.0090173
170	0	1882	62	2240	5242	2725	0	0.0064779	0.0003266	0.0066596	0.0113019	0.0078833
180	0	1649	56	1965	4735	2551	0	0.0056759	0.000295	0.005842	0.0102088	0.0073799
190	0	1527	57	1853	4479	2328	0	0.005256	0.0003003	0.0055091	0.0096568	0.0067348
200	0	1373	41	1605	4259	2131	0	0.0047259	0.000216	0.0047717	0.0091825	0.0061649
210	0	1265	32	1571	4070	1839	0	0.0043542	0.0001686	0.0046707	0.008775	0.0053201
220	0	1122	14	1405	4010	1732	0	0.003862	7.375E-05	0.0041771	0.0086457	0.0050106
230	0	1009	14	1242	3703	1579	0	0.003473	7.375E-05	0.0036925	0.0079838	0.004568
240	0	935	8	1116	3562	1477	0	0.0032183	4.214E-05	0.0033179	0.0076798	0.0042729
250	0	801	13	1042	3374	1338	0	0.0027571	6.848E-05	0.0030979	0.0072744	0.0038708
260	0	782	5	942	3294	1226	0	0.0026917	2.634E-05	0.0028006	0.0071019	0.0035468
270	0	731	5	806	3110	1085	0	0.0025161	2.634E-05	0.0023963	0.0067052	0.0031388
280	0	642	2	774	2908	1030	0	0.0022098	1.054E-05	0.0023011	0.0062697	0.0029797
290	0	538	0	697	2928	864	0	0.0018518	0	0.0020722	0.0063128	0.0024995
300	0	498	1	671	2838	846	0	0.0017141	5.268E-06	0.0019949	0.0061188	0.0024474
310	0	455	1	622	2654	764	0	0.0015661	5.268E-06	0.0018492	0.0057221	0.0022102
320	0	370	0	558	2655	732	0	0.0012736	0	0.001659	0.0057242	0.0021176
330	0	410	0	516	2449	605	0	0.0014112	0	0.0015341	0.0052801	0.0017502

340	0	354	0	446	2341	604	0	0.0012185	0	0.001326	0.0050472	0.0017473
350	0	317	0	416	2271	525	0	0.0010911	0	0.0012368	0.0048963	0.0015188
360	0	258	0	392	2201	466	0	0.000888	0	0.0011654	0.0047454	0.0013481
370	0	232	0	333	2085	440	0	0.0007986	0	0.00099	0.0044953	0.0012729
380	0	246	0	294	2063	417	0	0.0008467	0	0.0008741	0.0044479	0.0012064
390	0	205	0	284	1921	369	0	0.0007056	0	0.0008443	0.0041417	0.0010675
400	0	215	0	243	1816	340	0	0.00074	0	0.0007225	0.0039153	0.0009836
410	0	172	0	222	1773	339	0	0.000592	0	0.00066	0.0038226	0.0009807
420	0	149	0	209	1794	298	0	0.0005129	0	0.0006214	0.0038679	0.0008621
430	0	156	0	187	1655	278	0	0.000537	0	0.000556	0.0035682	0.0008042
440	0	125	0	182	1575	252	0	0.0004303	0	0.0005411	0.0033957	0.000729
450	0	123	0	154	1545	230	0	0.0004234	0	0.0004578	0.0033311	0.0006654
460	0	99	0	144	1531	214	0	0.0003408	0	0.0004281	0.0033009	0.0006191
470	0	106	0	135	1387	210	0	0.0003649	0	0.0004014	0.0029904	0.0006075
480	0	88	0	119	1344	159	0	0.0003029	0	0.0003538	0.0028977	0.00046
490	0	90	0	119	1326	148	0	0.0003098	0	0.0003538	0.0028589	0.0004282
500	0	62	0	99	1246	127	0	0.0002134	0	0.0002943	0.0026864	0.0003674
510	0	59	0	117	1207	144	0	0.0002031	0	0.0003478	0.0026023	0.0004166
520	0	59	0	98	1189	127	0	0.0002031	0	0.0002914	0.0025635	0.0003674
530	0	43	0	80	1233	114	0	0.000148	0	0.0002378	0.0026584	0.0003298
540	0	57	0	67	1135	95	0	0.0001962	0	0.0001992	0.0024471	0.0002748
550	0	38	0	60	1131	99	0	0.0001308	0	0.0001784	0.0024385	0.0002864
560	0	35	0	68	1014	87	0	0.0001205	0	0.0002022	0.0021862	0.0002517
570	0	33	0	55	994	73	0	0.0001136	0	0.0001635	0.0021431	0.0002112
580	0	30	0	50	995	74	0	0.0001033	0	0.0001487	0.0021452	0.0002141
590	0	34	0	45	913	69	0	0.000117	0	0.0001338	0.0019684	0.0001996
600	0	27	0	41	901	61	0	9.294E-05	0	0.0001219	0.0019426	0.0001765
610	0	22	0	39	834	51	0	7.572E-05	0	0.0001159	0.0017981	0.0001475
620	0	25	0	31	806	45	0	8.605E-05	0	9.216E-05	0.0017378	0.0001302
630	0	21	0	32	782	45	0	7.228E-05	0	9.514E-05	0.001686	0.0001302
640	0	12	0	26	752	39	0	4.13E-05	0	7.73E-05	0.0016213	0.0001128
650	0	11	0	26	691	39	0	3.786E-05	0	7.73E-05	0.0014898	0.0001128
660	0	16	0	20	707	49	0	5.507E-05	0	5.946E-05	0.0015243	0.0001418
670	0	11	0	19	676	29	0	3.786E-05	0	5.649E-05	0.0014575	8.39E-05
680	0	16	0	21	646	31	0	5.507E-05	0	6.243E-05	0.0013928	8.968E-05
690	0	11	0	17	611	33	0	3.786E-05	0	5.054E-05	0.0013173	9.547E-05
700	0	5	0	20	645	28	0	1.721E-05	0	5.946E-05	0.0013906	8.1E-05
710	0	9	0	11	612	25	0	3.098E-05	0	3.27E-05	0.0013195	7.232E-05
720	0	8	0	14	580	22	0	2.754E-05	0	4.162E-05	0.0012505	6.364E-05
730	0	9	0	14	538	17	0	3.098E-05	0	4.162E-05	0.0011599	4.918E-05
740	0	9	0	13	493	19	0	3.098E-05	0	3.865E-05	0.0010629	5.497E-05
750	0	6	0	6	489	19	0	2.065E-05	0	1.784E-05	0.0010543	5.497E-05
760	0	5	0	19	472	10	0	1.721E-05	0	5.649E-05	0.0010176	2.893E-05
770	0	4	0	4	453	16	0	1.377E-05	0	1.189E-05	0.0009767	4.629E-05

780	0	6	0	7	489	10	0	2.065E-05	0	2.081E-05	0.0010543	2.893E-05
790	0	1	0	6	431	12	0	3.442E-06	0	1.784E-05	0.0009292	3.472E-05
800	0	2	0	4	415	7	0	6.884E-06	0	1.189E-05	0.0008947	2.025E-05
810	0	2	0	7	437	14	0	6.884E-06	0	2.081E-05	0.0009422	4.05E-05
820	0	2	0	7	346	6	0	6.884E-06	0	2.081E-05	0.000746	1.736E-05
830	0	2	0	6	395	8	0	6.884E-06	0	1.784E-05	0.0008516	2.314E-05
840	0	2	0	6	348	10	0	6.884E-06	0	1.784E-05	0.0007503	2.893E-05
850	0	1	0	3	325	6	0	3.442E-06	0	8.919E-06	0.0007007	1.736E-05
860	0	1	0	5	331	6	0	3.442E-06	0	1.487E-05	0.0007136	1.736E-05
870	0	0	0	2	311	4	0	0	0	5.946E-06	0.0006705	1.157E-05
880	0	0	0	6	337	8	0	0	0	1.784E-05	0.0007266	2.314E-05
890	0	2	0	1	282	10	0	6.884E-06	0	2.973E-06	0.000608	2.893E-05
900	0	1	0	2	293	2	0	3.442E-06	0	5.946E-06	0.0006317	5.786E-06
910	0	1	0	5	275	3	0	3.442E-06	0	1.487E-05	0.0005929	8.679E-06
920	0	2	0	2	264	2	0	6.884E-06	0	5.946E-06	0.0005692	5.786E-06
930	0	3	0	0	273	3	0	1.033E-05	0	0	0.0005886	8.679E-06
940	0	0	0	0	273	2	0	0	0	0	0.0005886	5.786E-06
950	0	0	0	0	231	2	0	0	0	0	0.000498	5.786E-06
960	0	1	0	3	241	0	0	3.442E-06	0	8.919E-06	0.0005196	0
970	0	0	0	2	238	1	0	0	0	5.946E-06	0.0005131	2.893E-06
980	0	3	0	1	191	3	0	1.033E-05	0	2.973E-06	0.0004118	8.679E-06
990	0	1	0	1	205	2	0	3.442E-06	0	2.973E-06	0.000442	5.786E-06
1000	0	1	0	0	229	3	0	3.442E-06	0	0	0.0004937	8.679E-06
>1000	0	0	0	1	179	2	0	0	0	2.973E-06	0.0003859	5.786E-06
SUM	152582	290525	189828	336355	463817	345668						

Table S3.2

ExpB; N=100; μ=20; α=5;

	Number of SNPs for different allelic age ranges						Frequency density of SNPs for different allelic age ranges					
Allelic Age	r=1; h=0	r=48; h=0	r=1; h=0.5	r=48; h=0.5	r=1; h=1	r=48; h=1	r=1; h=0	r=48; h=0	r=1; h=0.5	r=48; h=0.5	r=1; h=1	r=48; h=1
10	137500	146779	136308	151113	147439	148623	0.692577	0.384401	0.674712	0.390265	0.291399	0.384125
20	28318	41818	27342	45351	44085	44020	0.142636	0.109518	0.13534	0.117124	0.08713	0.113772
30	12791	25033	12712	27677	28474	26936	0.064427	0.065559	0.062923	0.071479	0.056276	0.069618
40	7324	17899	7731	20035	20877	19593	0.03689	0.046876	0.038268	0.051742	0.041261	0.050639
50	4698	14068	5110	15484	16779	14913	0.023663	0.036843	0.025294	0.039989	0.033162	0.038544
60	2968	11676	3677	12772	13819	12524	0.01495	0.030578	0.018201	0.032985	0.027312	0.032369
70	1673	9889	2477	10435	12435	10561	0.008427	0.025898	0.012261	0.026949	0.024577	0.027296
80	1037	8669	1824	9192	11017	9121	0.005223	0.022703	0.009029	0.023739	0.021774	0.023574
90	739	7874	1359	8119	10044	8169	0.003722	0.020621	0.006727	0.020968	0.019851	0.021113
100	525	7319	974	7140	9125	7318	0.002644	0.019168	0.004821	0.01844	0.018035	0.018914
110	348	6384	799	6479	8495	6604	0.001753	0.016719	0.003955	0.016733	0.01679	0.017068
120	190	5769	484	5802	7736	6036	0.000957	0.015109	0.002396	0.014984	0.015289	0.0156
130	148	5278	363	5329	7261	5608	0.000745	0.013823	0.001797	0.013763	0.014351	0.014494
140	97	4918	254	4824	6692	5124	0.000489	0.01288	0.001257	0.012458	0.013226	0.013243
150	59	4613	162	4388	6382	4652	0.000297	0.012081	0.000802	0.011332	0.012613	0.012023
160	34	4307	142	4052	6030	4345	0.000171	0.01128	0.000703	0.010465	0.011918	0.01123
170	26	3884	98	3729	5876	3918	0.000131	0.010172	0.000485	0.009631	0.011613	0.010126
180	14	3678	43	3394	5426	3663	7.05E-05	0.009632	0.000213	0.008765	0.010724	0.009467
190	11	3423	44	3113	5241	3382	5.54E-05	0.008965	0.000218	0.00804	0.010358	0.008741
200	7	3187	25	2920	5132	3058	3.53E-05	0.008346	0.000124	0.007541	0.010143	0.007904
210	6	2921	25	2791	4830	2836	3.02E-05	0.00765	0.000124	0.007208	0.009546	0.00733
220	6	2660	24	2521	4561	2608	3.02E-05	0.006966	0.000119	0.006511	0.009014	0.006741
230	4	2527	12	2191	4472	2388	2.01E-05	0.006618	5.94E-05	0.005658	0.008838	0.006172
240	4	2441	19	2060	4252	2245	2.01E-05	0.006393	9.4E-05	0.00532	0.008404	0.005802
250	1	2210	7	1998	4124	2068	5.04E-06	0.005788	3.46E-05	0.00516	0.008151	0.005345
260	1	2154	3	1807	3940	1943	5.04E-06	0.005641	1.48E-05	0.004667	0.007787	0.005022
270	0	1944	4	1644	3750	1818	0	0.005091	1.98E-05	0.004246	0.007412	0.004699
280	1	1867	2	1563	3712	1696	5.04E-06	0.00489	9.9E-06	0.004037	0.007336	0.004383
290	4	1762	0	1405	3766	1528	2.01E-05	0.004615	0	0.003629	0.007443	0.003949
300	0	1590	0	1331	3333	1415	0	0.004164	0	0.003437	0.006587	0.003657
310	0	1537	0	1255	3409	1291	0	0.004025	0	0.003241	0.006738	0.003337
320	0	1371	0	1123	3142	1226	0	0.003591	0	0.0029	0.00621	0.003169
330	0	1277	0	1077	3069	1148	0	0.003344	0	0.002781	0.006066	0.002967
340	0	1230	0	957	2940	1032	0	0.003221	0	0.002472	0.005811	0.002667
350	0	1184	0	875	2846	967	0	0.003101	0	0.00226	0.005625	0.002499
360	0	1040	0	880	2876	908	0	0.002724	0	0.002273	0.005684	0.002347
370	0	1028	0	771	2866	817	0	0.002692	0	0.001991	0.005664	0.002112
380	0	972	0	697	2684	790	0	0.002546	0	0.0018	0.005305	0.002042

390	0	891	0	649	2597	721	0	0.002333	0	0.001676	0.005133	0.001863
400	0	885	0	650	2490	685	0	0.002318	0	0.001679	0.004921	0.00177
410	0	800	0	522	2285	598	0	0.002095	0	0.001348	0.004516	0.001546
420	0	711	0	522	2267	582	0	0.001862	0	0.001348	0.004481	0.001504
430	0	707	0	478	2176	538	0	0.001852	0	0.001234	0.004301	0.00139
440	0	697	0	452	2117	502	0	0.001825	0	0.001167	0.004184	0.001297
450	0	603	0	414	2043	428	0	0.001579	0	0.001069	0.004038	0.001106
460	0	574	0	415	1964	423	0	0.001503	0	0.001072	0.003882	0.001093
470	0	526	0	382	1819	379	0	0.001378	0	0.000987	0.003595	0.00098
480	0	526	0	326	1886	399	0	0.001378	0	0.000842	0.003727	0.001031
490	0	440	0	291	1760	319	0	0.001152	0	0.000752	0.003478	0.000824
500	0	425	0	296	1650	326	0	0.001113	0	0.000764	0.003261	0.000843
510	0	372	0	241	1629	299	0	0.000974	0	0.000622	0.00322	0.000773
520	0	355	0	233	1512	305	0	0.00093	0	0.000602	0.002988	0.000788
530	0	362	0	235	1462	245	0	0.000948	0	0.000607	0.002889	0.000633
540	0	325	0	202	1484	219	0	0.000851	0	0.000522	0.002933	0.000566
550	0	290	0	181	1339	250	0	0.000759	0	0.000467	0.002646	0.000646
560	0	281	0	192	1297	209	0	0.000736	0	0.000496	0.002563	0.00054
570	0	272	0	184	1274	193	0	0.000712	0	0.000475	0.002518	0.000499
580	0	246	0	155	1275	164	0	0.000644	0	0.0004	0.00252	0.000424
590	0	233	0	139	1172	159	0	0.00061	0	0.000359	0.002316	0.000411
600	0	186	0	147	1155	140	0	0.000487	0	0.00038	0.002283	0.000362
610	0	209	0	108	1121	134	0	0.000547	0	0.000279	0.002216	0.000346
620	0	185	0	114	1036	136	0	0.000484	0	0.000294	0.002048	0.000352
630	0	170	0	109	1038	125	0	0.000445	0	0.000282	0.002052	0.000323
640	0	182	0	98	978	118	0	0.000477	0	0.000253	0.001933	0.000305
650	0	141	0	84	962	111	0	0.000369	0	0.000217	0.001901	0.000287
660	0	163	0	84	903	95	0	0.000427	0	0.000217	0.001785	0.000246
670	0	105	0	88	920	85	0	0.000275	0	0.000227	0.001818	0.00022
680	0	130	0	72	814	84	0	0.00034	0	0.000186	0.001609	0.000217
690	0	125	0	57	817	78	0	0.000327	0	0.000147	0.001615	0.000202
700	0	119	0	58	761	70	0	0.000312	0	0.00015	0.001504	0.000181
710	0	89	0	70	772	67	0	0.000233	0	0.000181	0.001526	0.000173
720	0	134	0	56	709	57	0	0.000351	0	0.000145	0.001401	0.000147
730	0	87	0	43	690	65	0	0.000228	0	0.000111	0.001364	0.000168
740	0	87	0	44	639	55	0	0.000228	0	0.000114	0.001263	0.000142
750	0	84	0	51	671	61	0	0.00022	0	0.000132	0.001326	0.000158
760	0	73	0	42	664	57	0	0.000191	0	0.000108	0.001312	0.000147
770	0	66	0	32	598	50	0	0.000173	0	8.26E-05	0.001182	0.000129
780	0	61	0	27	617	39	0	0.00016	0	6.97E-05	0.001219	0.000101
790	0	65	0	35	592	46	0	0.00017	0	9.04E-05	0.00117	0.000119
800	0	60	0	44	557	35	0	0.000157	0	0.000114	0.001101	9.05E-05
810	0	49	0	30	556	41	0	0.000128	0	7.75E-05	0.001099	0.000106
820	0	45	0	24	501	33	0	0.000118	0	6.2E-05	0.00099	8.53E-05

	830	0	50	0	26	508	32	0	0.000131	0	6.71E-05	0.001004	8.27E-05
	840	0	38	0	25	521	16	0	9.95E-05	0	6.46E-05	0.00103	4.14E-05
	850	0	45	0	29	495	25	0	0.000118	0	7.49E-05	0.000978	6.46E-05
	860	0	39	0	21	476	15	0	0.000102	0	5.42E-05	0.000941	3.88E-05
	870	0	38	0	16	421	19	0	9.95E-05	0	4.13E-05	0.000832	4.91E-05
	880	0	26	0	9	463	17	0	6.81E-05	0	2.32E-05	0.000915	4.39E-05
	890	0	30	0	14	411	17	0	7.86E-05	0	3.62E-05	0.000812	4.39E-05
	900	0	27	0	18	419	11	0	7.07E-05	0	4.65E-05	0.000828	2.84E-05
	910	0	25	0	8	412	17	0	6.55E-05	0	2.07E-05	0.000814	4.39E-05
	920	0	32	0	5	382	12	0	8.38E-05	0	1.29E-05	0.000755	3.1E-05
	930	0	26	0	7	341	19	0	6.81E-05	0	1.81E-05	0.000674	4.91E-05
	940	0	27	0	10	368	15	0	7.07E-05	0	2.58E-05	0.000727	3.88E-05
	950	0	23	0	10	365	12	0	6.02E-05	0	2.58E-05	0.000721	3.1E-05
	960	0	20	0	4	339	12	0	5.24E-05	0	1.03E-05	0.00067	3.1E-05
	970	0	21	0	5	347	8	0	5.5E-05	0	1.29E-05	0.000686	2.07E-05
	980	0	18	0	8	303	14	0	4.71E-05	0	2.07E-05	0.000599	3.62E-05
	990	0	15	0	10	295	15	0	3.93E-05	0	2.58E-05	0.000583	3.88E-05
	1000	0	16	0	8	306	12	0	4.19E-05	0	2.07E-05	0.000605	3.1E-05
	1010	0	6	0	3	292	8	0	1.57E-05	0	7.75E-06	0.000577	2.07E-05
SUM		198534	381838	202024	387206	505970	386913						

Table S3.3, Table S3.4, Table S3.5 and Table S3.6

(Due to the large volume for each of these tables, they can be assessed online by the URL: http://www.sciencedirect.com/science/article/pii/S0888754315000397#ec0020)

3.10 Other Supplemental Files

3.1 Supplementary File 1 Perl Programming Codes for creating the SNPs based on their generations

```
#!/usr/local/perl
#this program will counts the SNPs based on their generations. It will counts about 100 bins based on our experiment and each bin represents 10 generations in the computer calculations.

$backup = $ARGV[0];
open FILEHANDLEBACKUP,$backup or die "can't open the $backup";
open (OUT, ">$ARGV[0].doc");
@Backup = <FILEHANDLEBACKUP>;
$count = 0;
$Combine = 5;


for($i = 0; $i<=$#Backup; $i++){
        $str1 = $Backup[$i];    chomp $str1;            @tempArray = split(/\t/,$str1);
        if ($i == 0){
                $Generation = $tempArray[0];
        }
        if ($tempArray[0] =~ m/pi*/ ) {
                if ($tempArray[1] <= 600 ){
                        $count ++;
                        if ($count == 1){
                                $Combine .= $tempArray[1].'_';
                                $Combine .= $tempArray[4].'_';
                        } elsif ($count == 2){
                                $Combine .= $tempArray[4].'_';
                        } elsif ($count == 3){
                                $Combine .= $tempArray[4];
                        } else {
                                $hash{$Combine}++;
                                $count = 0;
                                $GenerationHash{$Combine} += $tempArray[4];
                                $Combine = 5;
                        }
```

```
                }
        }
        if ($tempArray[0] =~ m/mi*/ ) {
         if ($tempArray[1] <= 600 ){
                $count ++;
                if ($count == 1){
                        $Combine .= $tempArray[1].'_';
                        $Combine .= $tempArray[4].'_';
                } elsif ($count == 2){
                        $Combine .= $tempArray[4].'_';
                            } elsif ($count == 3){
                        $Combine .= $tempArray[4];
                } else {
                    $hash{$Combine}++;
                    $count = 0;
                    $GenerationHash{$Combine} += $tempArray[4];
                    $Combine = 5;
                }
         }
      }
}
$start = 0;
foreach (keys %hash) {
        $Temp = $hash{$_} / 200; #frequency for the SNPs among the population
        $temp = $Temp * 200; # Real Times for the SNPs among the Population
        $tTemp = $GenerationHash{$_}; # Total Generation Time for each SNPs;
        $Average =  $tTemp /  $temp;
        $Time = $Generation - $Average;
        @myArray = split(/_/,$_);
        $s = sprintf "%0.2f",$Time;
        $h = int ($s/10);    # Every Ten number is a bin; So this will generate an integer.
        $CountHash{$h}++;
}
```

```
for ($i = 0; $i <=100; $i++){
	print OUT "$i\t$CountHash{$i}\n";
}
```

3.2 Supplementary File 2 Perl Programming Codes for creating the SNPs based on their frequencies

```
#!/usr/local/perl
#this program will create the SNPs frequencies range from 40% to 60% from the Backup file in GEMA.
#the following codes will open the Backup file from the command line;
$backup = $ARGV[0];
open FILEHANDLEBACKUP, $backup or die "can't open the $backup";
open (OUT, ">$ARGV[0].csv");
@Backup = <FILEHANDLEBACKUP>;
$count = 0;
$Combine = 5;

for($i = 0; $i<=$#Backup; $i++){
		$str1 = $Backup[$i];	chomp $str1;		@tempArray = split(/\t/,$str1);
		if ($i == 0){
			$Generation = $tempArray[0];
		}
		if ($tempArray[0] =~ m/pi*/ ) {
			if ($tempArray[1] <= 600 ){
				$count ++;
				if ($count == 1){
					$Combine .= $tempArray[1].'_';
					$Combine .= $tempArray[4].'_';
				} elsif ($count == 2){
					$Combine .= $tempArray[4].'_';
				} elsif ($count == 3){
					$Combine .= $tempArray[4];
				} else {
					$hash{$Combine}++;
					$count = 0;
					$GenerationHash{$Combine} += $tempArray[4];
					$Combine = 5;
				}
```

```
                }
        }
        if ($tempArray[0] =~ m/mi*/ ) {
                if ($tempArray[1] <= 600 ){
                        $count ++;
                        if ($count == 1){
                                $Combine .= $tempArray[1].'_';
                                $Combine .= $tempArray[4].'_';
                        } elsif ($count == 2){
                                $Combine .= $tempArray[4].'_';
                        } elsif ($count == 3){
                                $Combine .= $tempArray[4];
                } else {
                        $hash{$Combine}++;
                        $count = 0;
                        $GenerationHash{$Combine} += $tempArray[4];
                        $Combine = 5;
              }
         }
      }
}
$start = 0;
foreach (keys %hash) {
        $Temp = $hash{$_} / 200;              #frequency for the SNPs among the population
        $temp = $Temp * 200;                  # Real Times for the SNPs among the Population
        $tTemp = $GenerationHash{$_};     # Total Generation Time for each SNPs;
        $Average =  $tTemp /  $temp;
        $Time = $Generation - $Average;
        @myArray = split(/_/,$_);
        $s = sprintf "%0.2f",$Time;
        if (($Temp <= 0.6)&&($Temp > 0.4)){
                print OUT "$Generation,$myArray[3],$s \n";
        }
}
```

3.11 Acknowledgements

We are grateful to Dr. Robert Blumenthal, UT, for his valuable discussions and comments. We also appreciate the financial support from the Department of Medicine to conduct our research.

3.12 References

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin et al., 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved. American Journal of Human Genetics 73: 5-16.

Conrad, D. F., J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. J. Zhang et al., 2011 Variation in genome-wide mutation rates within and between human families. Nature Genetics 43: 712-U137.

Genin, E., A. Tullio-Pelet, F. Begeot, S. Lyonnet and L. Abel, 2004 Estimating the age of rare disease mutations: the example of Triple-A syndrome. J Med Genet 41: 445-449.

Griffiths, R. C., and S. Tavare, 1999 The ages of mutations in gene trees. Annals of Applied Probability 9: 567-590.

Kiezun, A., S. L. Pulit, L. C. Francioli, F. van Dijk, M. Swertz 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., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics 75: 199-212.

Kondrashov, A. S., and S. A. Shabalina, 2002 Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet 11: 669-674.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.

Li, W. H., 1975 The first arrival time and mean age of a deleterious mutant gene in a finite population. Am J Hum Genet 27: 274-286.

Maruyama, T., 1974a The age of a rare mutant gene in a large population. Am J Hum Genet 26: 669-673.

Maruyama, T., 1974b The age of an allele in a finite population. Genet Res 23: 137-143.

Qiu, S., A. McSweeny, S. Choulet, A. Saha-Mandal, L. Fedorova et al., 2014 Genome Evolution by Matrix Algorithms: Cellular Automata Approach to Population Genetics. Genome Biology and Evolution 6: 988-999.

Rannala, B., and J. P. Reeve, 2001 High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am J Hum Genet 69: 159-178.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter et al., 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter et al., 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

Slatkin, M., and B. Rannala, 1997 Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60: 447-458.

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu Rev Genomics Hum Genet 1: 225-249.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS biology 4: e72.

Wisloff, U., S. M. Najjar, O. Ellingsen, P. M. Haram, S. Swoap et al., 2005 Cardiovascular risk factors emerge after artificial selection for low aerobic capacity. Science 307: 418-420.

Chapter 4

Inference of Distant Genetic Relations in Humans Using "1000 Genomes"

Ahmed Al-Khudhair[1], Shuhao Qiu[2,3], Meghan Wyse[2], Shilpi Chowdhury[2], Xi Cheng[2], Dulat Bekbolsynov[2], Arnab Saha-Mandal1, Rajib Dutta[2], Larisa Fedorova[4], Alexei Fedorov[1,3*]

Authors' Affiliations:

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

[2]Program in Biomedical Sciences, University of Toledo, Health Science Campus, OH 43614, USA.

[3]Department of Medicine, University of Toledo, Health Science Campus, OH 43614, USA.

[4]GEMA-biomics, Ottawa Hills, OH 43606, USA

*To whom correspondence should be addressed. Tel +419-383-5270; Fax +419-383-3102; e-mail: Alexei.fedorov@utoledo.edu

My contributions to the work include developing a shell script to help running the program in the Ohio Super Computer center. I also help to develop Table 4.1, Table 4.2 and Figure 4.3.

Published in the Genome Biol Evol, January 7, 2015, doi: 10.1093/gbe/evv003

4.1 Abstract

Nucleotide sequence differences on the whole-genome scale have been computed for 1092 people from 14 populations publicly available by the 1000 Genomes Project. Total number of differences in genetic variants between 96,464 human pairs has been calculated. The distributions of these differences for individuals within European, Asian or African origin were characterized by narrow unimodal peaks with mean values of 3.8, 3.5, and 5.1 million respectively and standard deviations of 0.1-0.03 million. The total numbers of genomic differences between pairs of all known relatives were found to be significantly lower than their respective population means and in reverse proportion to the distance of their consanguinity. By counting the total number of genomic differences it is possible to infer familial relations for people that share down to 6% of common loci identical-by-descent. Detection of familial relations can be radically improved when only very rare genetic variants are taken into account. Counting of total number of shared very rare SNPs from whole-genome sequences allows establishing distant familial relations for persons with 8th and 9th degree of relationship. Using this analysis we predicted 271 distant familial pair-wise relations among 1092 individuals that have not been declared by 1000 Genomes Project. Particularly, among 89 British and 97 Chinese individuals we found three British-Chinese pairs with distant genetic relationships. Individuals from these pairs share identical by descent DNA fragments that represent 0.001%, 0.004%, and 0.01% of their genomes. With affordable whole-genome sequencing techniques, very rare SNPs should become important genetic markers for familial relationships and population stratification.

4.2 Introduction

Accomplishment of "1000 Genome Project" revealed immense amount of information about variation, mutation dynamics, and evolution of the human DNA sequences. The obtained critical data were originally reported by the Project Consortium (Abecasis et al. 2010b; Abecasis et al. 2012). These genomes have been already used in a number of studies, which added essential information about human populations, allele frequencies, local haplotype structures, distribution of common and rare genetic variants, and determination of human ancestry and familial relationships (see, for example, articles most relevant to this paper (Gravel et al. 2013; Harris and Nielsen 2013; Hochreiter 2013; Moore et al. 2013; Fagny et al. 2014)).

Knowledge of population stratification is important for medicine, specifically, in case-control association and cohort studies since unknown distant familial relationships could potentially compromise interpretation of collected data. Proper genetic identification of familial relationships is also critical for forensic identification, in criminal investigations, inheritance claims, and in other areas of human life.
Widely used haplotype data such as Y chromosome or mitochondrial DNA for identification of distant genetic relationships have limited applications due to the consideration of male or female lines of descent (Parson and Bandelt 2007; Willuweit et al. 2011). Estimation of genetic relatedness on autosomal genomic sequences is mainly based on genome-wide averages of the estimated number of alleles shared identically by descent (IBD) (Weir et al. 2006; Huff et al. 2011; Browning and Browning 2013). Various methods have been used to detect IBD familial relationships (Thompson 1975; Boehnke and Cox 1997; Li et al. 2014). The most commonly used GEMLINE, fastBD,

ISCA and ERSA. A most sophisticated approach, ERSA2.0, for IBD identification depends on the complicated statistical methods. Yet, only with confidence (97%) it can identify up to 5th degree relatives while deeper relations with confidence of less than 80% in simulated or mixed populations using genome-wide genotyping arrays and whole genome sequencing (Huff et al. 2011; Li et al. 2014). Recent analysis by Durand and co-authors demonstrated that GEMLINE method when applied for analysis of nearly three thousand real, non-simulated, father-mother-child trios had over 67% of false positive rate (Durand et al. 2014). The same authors introduced non-probabilistic additional computationally effective metric to score IBD fragments, HaploScore, to improve accuracy of IBD detection methods. However the efficiency and reliability of such approach to testing of familial relationship in generations deeper than first was not tested.

Aiming to advance identification of distant familial relationships, we undertook computational examination of publicly available 1092 genomes. Genomic differences across all autosomes (total number of different genetic variants) have been computationally assessed for all possible 45,747 human pairs from the same populations and also for 50,717 pairs of individuals taken from different populations, which represent 9% of all possible inter-population pairs and chosen randomly. We found that in-line with previous publications most genetic variations are found within human populations (Barbujani et al. 1997; Jobling and Gill 2004). We also observed that pairs with declared familial genetic relations have the least genomic differences compared to other non-related pairs from the same population. By simply counting the total number of genomic differences it is possible to infer familial relations for people that share down to 6% of common IBD genetic materials. Here we demonstrated that the detection of familial

relations would be drastically improved (by the order of magnitude) when only very rare genetic variants (vrGVs, with frequencies less than 0.2%) are taken into account. This paper demonstrates that simple counting of total number of shared vrGVs from whole-genome sequences allows establishing with high certainty (p-value < 0.001) distant familial relations for persons with 8^{th} and 9^{th} degree of relationship (people that have merely a fraction of a percentage of a coefficient of relationship (***r***) as defined by Sewall Wright (Wright 1922)). This is a very simple and powerful method for estimation of familial relationship based on vrGVs comparison, which requires whole-genome sequencing. With the availability of Illumina's new HiSeq X Ten device, the price of human genome sequencing this year was reduced three times to $1000 per genome. After accomplishment of the technology race to $100 per genome in the nearest future, vrGVs should become affordable important genetic markers for familial relationships and a broad range of population genetics studies.

4.3 Materials and Methods

4.3.1 Assessing the Total Number of the Genomic Variants Differences

We used data from the 1000 Genomes Project that are available through public ftp site ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/ (Abecasis et al. 2012). Specifically, Variant Call Format (VCF) files version 4.1 that contained a total of 38.2 million SNPs, 3.9 million short insertions/deletions and 14 thousand deletions for all the human chromosomes have been used. Information about genotype for each sequenced individual was extracted from the GT-field of VCF files "as is" in the 1000 Genomes dataset. The genotype likelihood information (GL field) has not been considered.

A large-scale computational analysis using a combination of Perl programs was carried out to process and assess the total genetic differences between each pair of individuals. The programs were run on the Oakley supercomputer (https://www.osc.edu/supercomputing/computing/oakley) in the Ohio Supercomputer Center or their optimized versions on our local Linux workstation. All Perl Programs utilized in this project are available at our web page "http://bpg.utoledo.edu/~afedorov/lab/prog.html". These programs include the following: 1) *Intra_PopGenomeDif.pl* and *Inter_PopGenomeDif.pl* that computes the total number of genetic variant differences between pairs of individuals from the same and different populations respectively; 2) shell script *Batch_Populations.sh* for batch-distributing the program to multiple cores in Oakley; 3) *2individualsGenomeDif_vrGVs.pl;* 4) *IDs_seperator_rareSNPs.pl*; 5) *Intra_PopGenomeDif_vrGVs.pl* and *Inter_PopGenomeDif_vrGVs.pl* that computes the total number of shared vrGVs between

individuals from the same and different populations respectively. The step-by-step description how to use these programs is presented in the Supplementary file S1.

Computer modeling of genomic differences has been performed with the program *GenomeDiffSimulation.pl.* The explicit instructions to this program are inserted as the comments into this script.

Each insertion or deletion has been counted as a single disparity not taking into account the length. Both parents' alleles have been considered. As an example, for a polymorphic site containing alleles A_1 and A_2, we counted as two differences between persons homozygous with A_1 and homozygous with A_2 and as a single difference between a heterozygous person and a homozygous one. For a population of size *N*, all possible pairs ($N^2/2$) have been computationally processed for their intra-population genomic differences.

Supplementary Table S1 shows a summary of the samples the 1000 Genomes Project has sequenced and been used in this project. Our analysis included the entire set of human autosomes while X and Y-chromosomes have been omitted to allow a proper comparison between males and females.

4.3.2 Statistics

A non-parametric statistical method (Kruskal-Wallis Test (Kruskal 1952) for testing equality of population medians among groups is used to assess for significant differences among populations and among continental ancestors.

Kruskal-Wallis Test is identical to an ANOVA (5.1.4) with the data replaced by their ranks. The data analysis is performed using R commander package.

4.3.3 Number of Very Rare Genetic Variants Shared Between Relatives

We set our frequency threshold for vrGVs as less than 0.2% based on the number of studied individuals (1092) that provide data for the 2184 haploid genomes. With this threshold, the genetic variants with less than 5 minor allele counts (in other words singletons, doubletons, tripletons, and quadratons) among 2184 studied haploid genomes were considered as vrGVs.

A subset table of the autosomal vrGVs information for the 1092 individuals is created using a Perl program (*IDs_seperator_rareSNPs.pl*). The table included solely variants (very rare genomic variants) with frequency as less as 0.2%. Using the rare variants table, a second Perl program (*Intra_PopGenomeDif_vrGVs.pl*) used to assess the number of rare variants shared between each pair of individuals within the same population. In order to assess the rare variants shared between individuals from different population, a Perl program named (*Inter_PopGenomeDif_vrGVs.pl*) was developed. We referred to familial relations following Sewall Wright (Wright 1922) in degree of relationship and coefficient of relationship (***r***). However, 1000 Genomes Project uses another term – first, second, and third order of relations, which is not well defined. Since we examined 1000 Genomes datasets we also used “order of relations” referring to the 1000 Genomes Project data.

4.4 Results

4.4.1 Genomic Differences among Humans

We have computed the total number of genomic differences between pairs of individuals whose DNA sequences are available from the "1000 Genomes" project. Our analysis included the entire set of human autosomes while X and Y-chromosomes have been omitted to allow a proper comparison between males and females. Figure 1 illustrates the intra-population results for 14 populations from Africa, America, Asia, and Europe. All pairs of individuals with declared family relationships are marked by stars in Figure 1B. These pairs have significantly fewer genomic differences than the remaining non-related pairs from the same population. Statistical examination of the intra-population distributions using Kruskal-Wallis test showed that, with the 0.05 significance level, the distributions are different from each other except for CHB and JPT populations (see statistical details in Supplementary file S2). The inter-population genomic differences are presented in the Supplementary Figure S1.

4.4.2 Computer Modeling Of Genomic Differences

Intriguingly, the number of genomic differences within Asian, European and African populations are shaped as narrow peaks with mean values of 3.5, 3.8, and 5.1 million respectively, and standard deviations in the range of 0.03-0.1 million (Figure 1A). Since a majority of human genes have several major mutually-exclusive haplotypes, comprising dozens to hundreds of frequent SNPs (Consortium 2003), the number of genomic differences for a particular gene between pairs of human individuals should range from 0 (when compared individuals carry the same gene haplotypes) to dozens or hundreds of differences (when compared individuals carry different haplotypes of the gene under

analysis). In order to understand the reason why the genomic differences for African, Asian, and European populations on the Figure 1 are distributed as single narrow peaks, a computer program *GenomeDiffSimulation.pl* has been created. This program models the genomes of virtual individuals that, on an average, contain 3,800,000 different SNPs between them. In addition, these SNPs are grouped into several (four by default) mutually exclusive haplotypes for each genomic locus of the virtual individuals. The variable parameter for this program is the total number of loci that are in linkage equilibrium with each other.

The computational results for the distribution of the total differences in SNPs between pairs of virtual individuals are shown in Figure 2. The width of the peaks in the Figure 2A essentially depends on the number of genomic loci, in which SNPs are in linkage equilibrium with each other. In the model where the number of loci with linkage equilibrium is 5,000, the peak for the total genomic differences between virtual individuals (shown in blue) closely matches the shape of the peak computed for the actual Great Britain population (which, for comparison, is also present in Figure 2A and shown as a red bold line). This number (5,000) of chromosomal loci with linkage equilibrium with each other roughly corresponds to that in the human genome. There is an ambiguity in the estimation of the exact number of such loci in humans because of the fact that linkage disequilibrium between SNPs in humans decays continuously with increasing physical distance between SNPs, and also depends on the local recombination rate, which is highly variable along chromosomes (Arnheim et al. 2003). If the human genome consisted of 5,000 loci with mutual linkage equilibrium, the average size of the locus would have to be 600 Kb. This nucleotide length in the human genome corresponds to

0.6 centimorgan for genetic distance, which seems reasonable for modeling of the locus size. Hence, 5,000 loci with mutual linkage equilibrium give a rough approximation of the human genome. This estimation is congruent to common view in Hartl and Clark textbook (page 543) (Hartl 2007). However, for more precise estimation, the population history and demography should be taken into account. All in all, we attribute the narrow width of the peaks for the genomic differences in long-established African, Asian, and European populations to the presence of several thousand chromosomal loci in mutual linkage equilibrium. In each of these relatively old populations, the haplotypes of the loci have been well shuffled and all individuals have equal chances of carrying a particular haplotype. Figure 1, also reveals much wider peaks for the American populations. We attribute this increased width to the recent admixture in populations of the New World, where European, African, and Native American genomic ancestry may be observed in various proportions in different people.

Our *GenomeDiffSimulation.pl* program has an option to mimic close genetic relations for several pairs of virtual individuals. A user may assign specific genetic relations for these pairs such as siblings (which share 50% of common genetic material IBD), second order of genetic relations (for example aunt/niece with 25% of common genetic material IBD), third order of relations (cousins with 12.5% of common IBD loci), or other more distant relatives with any user-defined percentage for common genetic loci. The genetically related pairs of virtual individuals have been simulated and five of these computational experiments are presented on the Figure 2B, where positions of pairs with genetic relations are marked by stars. Positions of virtual individual pairs with first and second order of genetic relationships (50% and 25% of common IBD loci respectively)

correspond well to the positions of the actual human pairs having declared family relationships from the 1000 Genomes. For example, genetically-related pairs of virtual individuals are compared with pairs from Great Britain populations in Figure 2B. We observed that positions of siblings and parent/child pairs are always located in the extreme left of their corresponding population peak, followed by pairs with the second order of relations, which are closer to the corresponding peaks, and so on.

In the Figure 1B, the positions of several pairs within Luhya in Webuye, Kenya (LWK), Southern Han Chinese, China (CHS), British in England and Scotland (GBR) populations that are located close to the left slopes of their respective population peaks should correspond to the fourth or fifth order of genetic relations (6.2%-3.1% of shared IBD genetic materials). The genetic relations for these pairs have not been declared, yet with this analysis we can infer their putative genetic relations (which also has been confirmed by the distributions of very rare SNPs, see next paragraph). However, according to our computer simulations, the pairs with the fifth and higher order of relations (3.1% and less percentage of common genetic materials) may frequently be located within the left slopes of the corresponding peaks together with genetically non-related pairs (see Figure 2B). Thus, prediction of fifth and higher orders of genetic relations based on the total number of genomic differences appears to be unreliable. This limitation in identifying genetic relationships exists because a majority of genomic differences between pairs of individuals is contributed by frequent SNPs that form several (usually from two to five) major haplotypes in each loci (Consortium 2003). These major haplotypes have a high probability of being the same between genetically non-related individuals. This obstacle can be overcome if we consider only the very rare

SNPs, for which probabilities of being shared by chance in non-related individuals drop dramatically (in the direct reverse proportion to the frequency of the considered SNPs).

4.4.3 Distributions of Shared Very Rare Genetic Variants in Humans

In order to explore this possible method for predicting distant genetic relations in humans, we computationally filtered a complete subset of very rare genetic variants (vrGVs) from the "1000 Genomes" database having frequencies of less than 0.2% in the 2184 chromosomes from 1092 sequenced individuals. The distributions of positions of vrGVs along chromosomes are uniform and cover a vast majority of genomic regions, as exemplified in Figure 3 and detailed in the Supplementary Table S2. About 99% of these vrGVs are inside introns or intergenic regions. The number of shared vrGVs between each pair of individuals from the same population has been calculated (Figure 4). The graph reveals that a vast majority of examined pairs from American, Asian, and European populations shared from 50 to 300 vrGVs and form unimodal peaks for each population (Figure 4A). A majority of pairs from three African populations (African Ancestry in Southwest US (ASW), Luhya in Webuye, Kenya (LWK), and Yoruba in Ibadan, Nigeria (YRI)) share from 200 to 800 vrGVs, and also form unimodal peaks for each population. However, among all 14 populations, 311 pairs shared much higher number of vrGVs, (more than a thousand per pair) with the highest number of shared vrGVs being 46,745. Such extra-long tails in the distributions of shared vrGVs were even problematic to illustrate in the same figure together with the main peaks. Therefore, we presented these tails separately in Figure 4B, which has a 50 fold different scale compared to the peaks in Figure 4A. All 40 pairs with declared genetic relationships from 1000 Genomes are marked by stars in Figure 4B. These declared relatives share 6,252 to 46,745 vrGVs and

represent the right-most points in the tails of distributions in Figure 4B. Besides these 40 pairs of known relatives, there are 271 pairs on Figure 4B that shared more than a thousand vrGVs (see Supplementary Table S3) and also dozens of pairs in Figure 4A that share several hundreds of vrGVs, which are on the right side of corresponding peaks and clearly separated from the peaks.

Interestingly, these right tails of distributions of vrGVs have population-specific patterns. For example, one of the African populations, LWK, has the highest number of pairs (260), each with more than a thousand of shared vrGVs. At the same time another African population (YRI) has only two of such pairs that share 1193 and 1841 vrGVs. Since the information about the individuals and strategies of their sampling for 1000 Genome Project is publicly unavailable, it is impossible to investigate this issue further. We hypothesize that pairs of individuals that share more than a thousand of vrGVs should have family relationships. Even those pairs, that share hundreds of vrGVs and are clearly separated from the main peaks, are likely formed by distant relatives.

This hypothesis is strongly supported by the calculations of the number of shared vrGVs between populations, shown in Supplementary Table S4. All studied 44,278 pairs formed by individuals from two different continents have less than 118 shared vrGVs. (For example, the highest number of shared vrGVs between LWK and JPT is 37; LWK-FIN is 80; and GBR-CHB is 117.) The number of shared vrGVs between populations from the same Asian or European continent is also low (for instance, maximal number between GBR and FIN is 159 and between CHB and JPT is 78). This means that a pair of European and/or Asian individuals that shares more than 300 vrGVs very likely has a familial relationship. The distributions of shared vrGVs between African populations

(LWK vs. YRI and LWK vs. ASW) are demonstrated in Figure 5. With three exceptions, all studied 14,453 pairs formed by individuals from two different African populations have less than 623 shared vrGVs (these three exception pairs are discussed in the next section). Detailed examination of the inter-population distribution of shared vrGVs was performed on the entire set of 8633 British-Chinese pairs formed by one individual from GBR and another individual from CHB population (see Table 1). This table demonstrates that a vast majority (8547) of these pairs have only single digit numbers of shared vrGVs. Only 3 out of 8633 pairs have 30 or more shared vrGVs. The distribution of shared vrGVs along chromosomes for these three pairs has been analyzed with a Perl program – *2individualsGenomeDif_vrGVs.pl*. The results for the HG00255-NA18614 pair, which has 30 shared vrGVs, are shown in the Table 2, while the data for other two pairs with 59 and 117 shared vrGVs are shown in the Supplementary Table S5. Table 2 demonstrates that 27 out of 30 shared vrGVs are located inside a 71 Kb genomic segment (positions from 90,787,654 to 90,858,949 nts) within chromosome 11. All clustered vrGVs do not show correlations with structural variants in this region. In addition, supplementary Table S6 demonstrates that shared vrGVs for HG00255 and NA18614 individuals are present on the same haplotype background. Similar clustering of shared vrGVs was observed for two other British-Chinese pairs (see Table S5). The pair HG01334-NA18627 has all 59 shared vrGVs located within 284 Kb locus on chromosome 1, while another HG00263-NA18541 pair has 115 shared vrGVs within a 806 Kb region inside chromosome 6. Given the enormous size of the human genome (3,300 Mb), the probability (*P*) of occurrence by chance for the case presented in the

Table 2 that corresponds to 27 out of 30 shared vrGVs located inside 71 Kb region is less than 10^{-117}, according to the formula (1).

$$P = C_{30}^{3} * (71000/33000000000)^{26} \quad (1)$$

This formula (1) assumes that all 30 vrGVs are independent and in equilibrium with each other. Therefore, undoubtedly, 27 out of 30 independent vrGVs cannot be located within the same short locus by chance. This means that these three British-Chinese pairs represent very distant genetic relatives and their shared vrGVs located in the same locus are identical by descent and are in linkage disequilibrium with each other. Our observations of the chromosomal distributions of shared vrGVs are in a complete accordance with the population genetics theory that genetic inheritance occurs through chromosomal IBD segments, which are likely to become smaller and smaller with generations due to meiotic recombination events (Browning and Browning 2010; Huff et al. 2011). In agreement with this theory, these three British-Chinese pairs with very distant genetic relations should likely have only one short IBD per pair. The percentage of common genetic materials (C%) identical by descent for the British-Chinese pairs under consideration may be calculated by the formula:

$$C\% = (\Delta l/2L)*100\% \quad (2)$$

Where Δl is the size of the IBD segment and L is the size of haploid genome. According to (2), these three pairs with 30, 59, and 117 shared vrGVs should have 0.0011%, 0.0043% and 0.012% of common genetic materials respectively.

If we consider relatively old population that existed for many hundreds of years (like GBR, FIN, or CHS), a majority of its individuals are likely to be in extremely distant genetic relations to each other (let's say 20 generations apart). Hence, they should

share multiple and very short IBD chromosomal segments (a few thousands of nucleotides) because these IBD segments have been divided by recombinations in multiple generations. All these short IBD segments should contain only a few vrGVs due to their small size. In this respect, let's consider for example the Chinese (CHS) population in which the distribution of shared vrGVs has a peak of 94 (see Fig 4A). A NA18548-NA18567 pair from this population has 303 shared vrGVs and is clearly separated from the corresponding peak on the Fig 4A. The distribution of shared vrGVs for this pair is demonstrated in the Table S5. This pair also has a single 36.9 Mb IBD segment on chromosome 2 that contains 199 shared vrGVs. The rest 104 vrGVs have a relatively random distribution across all chromosomes. Several of these 104 shared vrGVs may occasionally be grouped within a short chromosomal region (see Table S5). On the contrary, if we consider a pair from CHS that has a number of shared vrGVs around the peak value of 94 (for instance pair HG00557-HG00610 with 80 shared vrGVs) the distribution of shared vrGVs along chromosomes for this pair does not have any prominent IBD that contains more than 9 shared vrGVs (see Table S5). Supplementary Table S5 also contains examples of two intra-population pairs for GBR individuals (HG00109-HG00117 and HG00101-HG00099) containing 276 and 324 shared vrGVs respectively. The number of shared vrGVs corresponding to the two pairs are significantly higher than the peak value of 42 for this population. These pairs have several IBD segments on different chromosomes each containing dozens of shared vrGVs, so these individuals should be in distant genetic relation to each other.

Finally, we examined the inter-population distribution of shared vrGVs for three populations with African origin (see Figure 5). There are three pairs that have the highest numbers of shared vrGVs and they are clearly separated from the rest of the pairs illustrated in Figure 5. They are the following: (NA19443– NA18508) pair for LWK-YRI populations with 1121 shared vrGVs and two pairs for LWK-ASW populations, NA19350 - NA20348 and NA19397- NA20348 with 903 and 939 shared vrGVs respectively. Distributions of shared vrGVs along chromosomes for these three pairs are also presented in the Table S5. The LWK-YRI pair has a prominent 8.5 Mb IBD region on chromosome 8 that contains a vast majority (1037) of all shared vrGVs. Therefore this pair has 0.13% of common genetic material according to formula (2). The other two pairs from LWK-ASW share the same person NA20348 from the ASW population. These two pairs also have a single prominent IBD spanning 14 Mb genomic segment on chromosome 11, which contains more than half (709) of all shared vrGVs for these two pairs. Therefore these individuals share 0.21% of common IBD genetic materials and should be distantly related to one another.

We did not perform the exhaustive inter-population comparisons of shared vrGVs because of the enormous amount of computational space required for computation of 549,842 pairs in total, which is beyond the scope of our resources. However, we expect that many more cases for inter-and intra-population distant genetic relationships will be revealed for the 1092 sequenced individuals. All in all, our approach is able to detect distant genetic relations that may share as small as 0.001% of genomic DNA.

4.5 Discussion

We demonstrated that human populations are distinct from one another by distribution of genomic differences among their individuals (see Figure 1) and also distribution of shared vrGVs (see Figure 4). Those populations that were formed thousands years ago -- African (LWK and YRI), Asian (CHS, CHB, and JPT), and European (GBR, FIN, TSI, CEU) have sharp and narrow peaks in the corresponding distributions of genomic differences, while populations from America that experienced admixture a few hundred years ago, via inclusion of people from different continents, have much wider distributions of genomic differences (see Figure 1A).

Some human populations differ from others by distribution of shared vrGVs. For example, in the LWK population we observed the largest number of pairs (156) that shared more than 800 vrGVs. However, another African population, YRI, has only 7 of such pairs shared >800 vrGVs (see Fig 4). LWK population has the widest peak of the distribution of shared vrGVs with the mean-to-SD ratio of 1.2, whereas this ratio in European populations is about 0.3. One of the possible interpretations of this observation is that LWK might have experienced a high level of inbreeding, or it has a distinct subpopulation structure and the sample of LWK individuals were collected disproportionately from a few subpopulations.

Here we showed that genetic relationships can be effectively determined by the analysis of distribution of shared vrGVs between individuals. This analysis should take into account population structure. For example, number of vrGVs per individual varies among different geographic regions, being the highest in Africa (average vrGVs per individual in LWK is 67,200 and standard deviation, σ, is 7,500) and dropping to 16,200

in Europe (GBR population; σ=2,650) and 24,100 in Asia (CHB population; σ=4,100). In these calculations the threshold (0.2%) for vrGVs determination has been established based on the entire set of 1092 people from 14 populations. It makes sense to put such a threshold for each population discretely. This has not been done in this paper since we have not got enough statistics (the number of people in each population is less than 100). Due to the differences in population structures, we observed significant variations in the number of shared vrGVs between the first and the second order relatives in different populations (see Figure 4 and Supplementary Table S3). First order relatives (shared 50% common genetic materials) have 28,000-46,000 shared vrGVs in Africa and only about 14,000-20,000 in Asia. This number is proportionally decreased for the second order relatives and further on.

There is a constant and intense influx of novel mutations in humans and other species. On average, each person has from 40 to 100 novel mutations that are absent in the genome his/her parents (Conrad et al. 2011; Kondrashov and Shabalina 2002; Li and Durbin 2011). A majority of these novel mutations are eliminated soon after their arrival by genetic drift and selection. Yet the remaining portion of novel mutations is an important endless source for vrGVs, which pool continuously renovates and maintains at a very high level (14-40 thousand vrGVs per individual in European and Asian populations). Recent computational analysis of the 1000 Genome database by Moore and coauthors (Moore et al. 2013) also demonstrated the highest abundance of rare GVs, yet they used slightly higher threshold (0.3%) for their frequencies. In the review by Keinan and Clark the authors summarized the common viewpoint that an excess of rare genetic variants has resulted from the recent explosive growth of human population (Keinan and

Clark 2012). Whole-genome dynamics of millions of genetic variants is a very intricate issue that only recently has been touched in computer simulations (Qiu et al. 2014) and also in large-scale computations of 1000 Genomes Project data (Moore et al. 2013).

4.5.1 Impact of sequencing errors on the analysis of shared vrGVs.

As demonstrated on the Figure 3, the distribution of vrGVs along chromosomes is relatively even. A majority of vrGVs occurs inside largest genomics regions with the longest spans, namely the intergenic regions and introns. According to the publication of 1000 Genome consortium, these non-exome regions have the lowest sequencing coverage (on average x5 times), and thus they have the highest level of sequencing errors. On page 1065 of the 1000 Genome publication, the authors estimated that "in low-coverage project, the overall genotype error rate was 1-3%" (Abecasis et al. 2010). According to the same publication (page 1067), in some cases the error rate maybe ~4% (for CEU population) and ~10% for YRI depending on the sequence coverage for a genomic region. Misinterpretation of heterozygous sites with homozygous sites is the main cause of errors in interpreting genomic regions with low depth of sequencing coverage. For example, for a heterozygous person with a (G/A) SNP, when a sequence coverage is x6, there is a 1/32 chance that only G or only A nucleotides will be detected in all of the six reads (3% error). It means that, on average, 3% (and in some occasions up to 10%) of vrGVs are randomly missed in 1000 Genomes database. This effect partially explains the large intra-population variations in total number of vrGVs between individuals (see the Results section and Supplementary table S2). Another type of sequencing error is the misinterpretation of one nucleotide instead of another. The frequency of such type of errors has not been explicitly discussed in the reports of 1000 Genome. However, such

errors should occur pretty randomly across the genome and in a majority of cases should be interpreted as an arrival of a novel mutation – a singleton. Such singletons should be sparsely distributed across the genome and should increase the number of vrGVs in individuals. Since the length of the human genome is huge (3 billion nucleotides), one vrGV occurs, on an average, per 100 Kb region. Hence the probability that non-related individuals share the same vrGV is very low (less than one shared vrGVs) per pair. Taking into account that mutations did not occur randomly, but rather at particular hot-spots, this estimation may be raised to a handful of randomly shared vrGVs between non-related individuals. Indeed, when we compared number of shared vrGVs between continents (see Supplementary Table S4) the median number of shared vrGVs was 2 (for CHB-GBR populations), 6 (LWK-FIN), and 8 (for LWK-JPT). Therefore, sequencing errors due to nucleotide misinterpretation should be at most accountable for a handful of shared vrGVs between pairs of individuals and their impact on the overall vrGV analysis should be negligibly small.

In some populations marriage between relatives is a common practice. (http://www.consang.net/index.php/Global_prevalence). For example, we detected a pair from Colombian in Medellin, Colombia (CLM) (HG01277 and HG01278) that has the highest number (2863) of shared vrGVs for this population. According to "1000 Genomes" annotation table, this pair represent a husband and wife, and we project that they share about 6% of common IBD genetic materials. Therefore, we expect that their child (HG01279, not sequenced yet) should have more than 50% of common genetic materials with each of his/her parents. Presumably, due to this reason, the observed variation of numbers of shared vrGVs among the first order relatives is very wide

compared to our modeling. For instance, in LWK population, this variation is from 31,000 up to 46,500. We conjecture that the highest numbers may correspond to the families where marriage occurred between genetic relatives. It is also worth mentioning that actual relationship between siblings or parent/offspring pairs may fluctuate noticeably from 50% (Odegard and Meuwissen 2012). Finally, even within the same population, the number of vrGVs among individuals significantly varies. For example, in Chinese population CHB the average number of vrGVs per individual is 24,100 while σ=4,100. In this population the lowest number of vrGVs (16,745) was detected in HG00403 person, while the highest 40,444 in HG00702 individual. All these facts together may explain the large variations in the numbers of shared vrGVs between the pairs of relatives with the same degree of relationship.

In summary, if two individuals share less than a dozen of vrGVs they should descend from different ethnic and geographically diverse populations. In case persons share several dozens of vrGVs located in the same chromosomal region they should have some degree of genetic relationship to each other. Finally, a pair may have dozens to hundreds of shared vrGVs that have a uniform spread over all chromosomes without a strong signal for preferential association or clustering within a particular locus. This means that some predecessors of these individuals belonged to the same population. All in all, in addition to well-established DNA fingerprinting, application of vrGVs analysis for obtaining distant genetic relations could be a valuable molecular genetic technique in criminal investigations, in civil familial searching as well as for population, clinical and association studies.

4.6 Table and Figure Legends

Table 4.1 Distribution of numbers of shared vrGVs for 8633 human pairs, where one person of a pair represents British population (GBR), while another person – Chinese population (CHB).

*NOTES: detail characterization of shared vrGVs for the pair, which has 30 shared vrGVs, is shown in the Table 2. Detail characterization of shared vrGVs for three pairs at the bottom of this table (marked by *) is shown in the Supplementary Table S5.

Table 4.2 Characterization of 30 shared vrGVs for the British-Chinese pair composed by HG00255 and NA18614 individuals.

Those vrGVs that are located in the same locus on chromosome 11 are shaded. The detailed description of shared vrGVs for this pair and also for eight other pairs described in the Results section, is provided in the Supplementary Table S5.

Figure 4-1 Distribution of number of genetic variants (GVs) between all possible pairs of individuals within the same population.

Three populations from Africa (ASW, LWK, YRI), three populations from America (CLM, MXL, PUR), three from Asia (CHB, CHS, JPT), and five from Europe (CEU, FIN, GBR, IBS, TSI) have been examined. Numbers of individuals in the populations are shown on the graph behind the population identifier (e.g. 66 people for MXL-66). The number of pairs has been calculated for bins (X; X+10,000), where number of genetic variants X is plotted on the graph and the bin size was 10,000 genetic variations. **A** – Two-dimensional view of the distribution. **B** -three dimensional view of

the distribution where all pairs with declared genetic relations are marked by stars. The color of a star reflects a specific genetic relationship: red stands for siblings, blue – parent/child pair, green – second order relations, and yellow – third order.

Figure 4-2 Distribution of number of genetic variants (GVs) between pairs of individuals from the same real and modeling populations.

A – Two-dimensional expanded view of the distribution. Real population from Great Britain (GBR) is shown as a red bold line, while the five other curves represent model populations of virtual individuals. Virtual individuals in all models have on average 3.8 million differences of genetic variants between them. Various models have different number of genomic loci that are in linkage equilibrium with each other. The model with the lowest number (50) of loci with equilibrium is shown by orange line and has the widest span. The model with the highest number of loci in linkage equilibrium, 25,000, has the narrowest peak (brown line). When the number of loci with equilibrium in the modeling genome is 5,000 (navy blue line) the modeling distribution is most similar to the real one from GBR population (red line).

B – Three-dimensional view of the distribution where pairs with known genetic relations are marked with stars. The color of a star reflects a specific genetic relationship: red stands for siblings, green – second order relations, and yellow – third order, pink - fourth order, and black - fifth order of genetic relations. The front most distribution (red) represents the real population from Great Britain (GBR). The next five curves represent distributions for five model populations of virtual individuals (M1 to M5). In each of these five models the number of loci in linkage equilibrium with each other is the same -

5,000. Three pairs of virtual individuals mimic genetic relationships in every model. In M1 these three pairs are represented by siblings (that share 50% of common genetic materials from the most recent common ancestor). M2 represents three pairs with the second order of relations that share 25% of common genetic materials (e.g. aunt/niece). M3 represents three pairs with third order of relations that share 12.5% of common genetic materials (e.g. cousins). M4 –fourth order with 6.25%; and M5 – fifth order with 3.12% of common genetic materials. All three pairs with fifth order of genetic relations from M5 model are located in the same left-most bin together with one pair of virtual individuals that does not have genetic relations.

Figure 4-3 Distribution of vrGVs along chromosome 3 for four randomly picked individuals: two from Chinese (CHS) population (HG00404 and HG00407 individuals) and two from British (GBR) population (HG00097 and HG00099).

Every vrGV is represented by a dot. The detail information about distribution of vrGVs along all chromosomes for these individuals is available from Supplementary Table S2.

Figure 4-4 Distribution of number of shared very rare genetic variants (vrGVs) between all possible pairs of individuals from the same population.

Three populations from Africa (ASW, LWK, YRI), three populations from America (CLM, MXL, PUR), three from Asia (CHB, CHS, JPT), and five from Europe (CEU, FIN, GBR, IBS, TSI) have been examined.

A – Three-dimensional view of the part of the distributions where the majority of pairs are located.

B – Two dimensional view of the tails of the distributions, where pairs are presented by circles, triangles, rectangles, and crosses specific for each population. All pairs with declared genetic relations are marked by stars. The color of a star reflects a specific genetic relationship: red stands for siblings, blue – parent/child pair, green – second order relations, and yellow – third order. Scale for the number of shared vrGVs in the graph 4A is expanded 50 fold compared to 4A.

Figure 4-5 Distribution of number of shared vrGVs between pairs of individuals from different African populations.

First distribution (shown in red) represents pairs in which one individual belongs to LWK population while another person to ASW. Second distribution (blue) represents pairs in which one individual is from LWK while the other is from YRI.

4.7 Tables and Figures

Table 4.1

# Shared vrGVs	# Human Pairs
0	903
1	1828
2	2045
3	1584
4	1009
5	605
6	298
7	149
8	68
9	58
10	22
11	20
12	13
13	3
14	3
15	5
16	4
17	2
18	2
19	0
20	0
21	3
22	1
23	1
24	0
25	0
26	0
27	0
28	1
29	1
30	1 *
……	0
59	1 *
……	0
117	1*

Table 4.2

Chromo-some	Position of vrGVs	Identifiers of vrGVs	Reference allele	Alternative allele
CHR3	163910979	rs147633047	C	T
CHR9	42323192	rs184959358	G	A
CHR11	90787654	rs138781903	A	G
CHR11	90788511	rs141690807	C	T
CHR11	90788759	rs187621230	T	C
CHR11	90798281	rs144138129	G	A
CHR11	90806962	rs183908202	A	G
CHR11	90808684	rs147862657	C	A
CHR11	90812601	rs147197102	G	A
CHR11	90815124	rs190205439	C	T
CHR11	90816996	rs147226573	A	G
CHR11	90817266	rs185201515	G	A
CHR11	90826778	rs139867381	T	A
CHR11	90828732	rs149763439	G	A
CHR11	90835123	rs140038072	G	A
CHR11	90835556	rs140255793	A	G
CHR11	90840943	rs147799849	A	G
CHR11	90842479	rs142999510	G	T
CHR11	90843070	rs141928306	T	C
CHR11	90844258	rs139273514	A	G
CHR11	90847782	rs150575842	C	T
CHR11	90848531	rs147400508	G	A
CHR11	90848728	rs149904020	G	C
CHR11	90850157	rs138217375	G	T
CHR11	90852915	rs187692214	A	G
CHR11	90856705	rs144056495	G	A
CHR11	90858178	rs189208470	C	T
CHR11	90858721	rs139417643	G	A
CHR11	90858949	rs150070179	A	G
CHR20	42290810	rs146883107	C	T

Figure 4-1

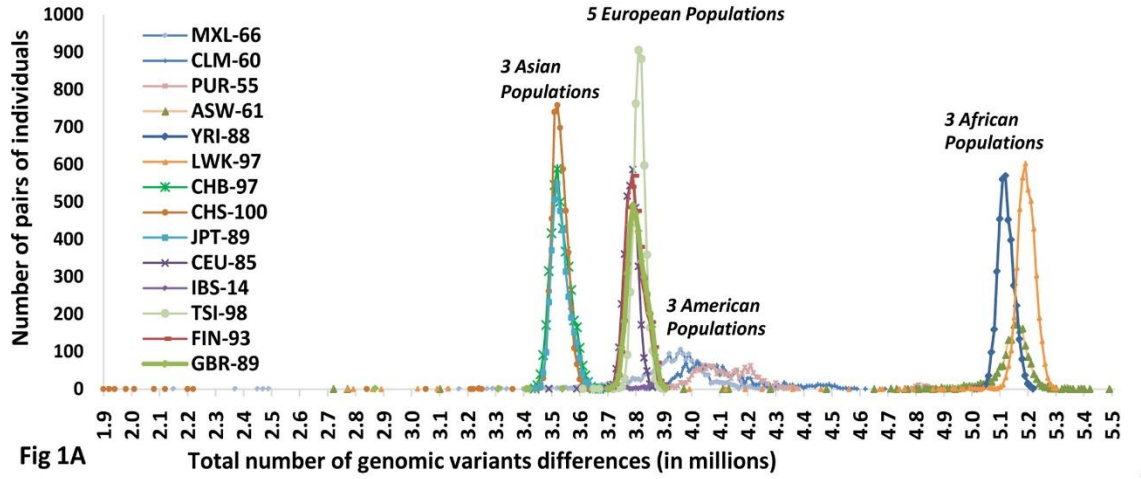


Fig 1A

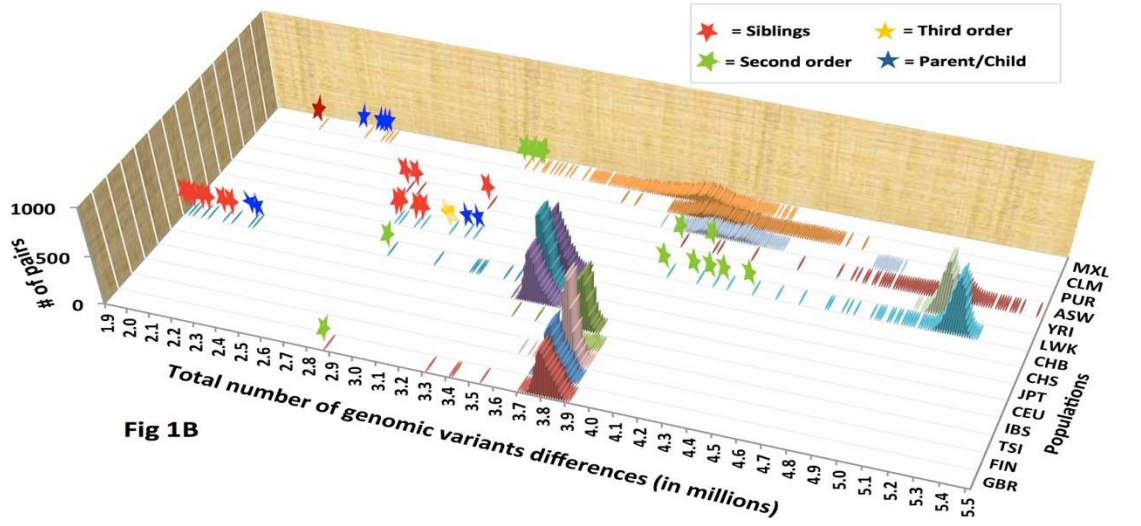


Fig 1B

Figure 4-2

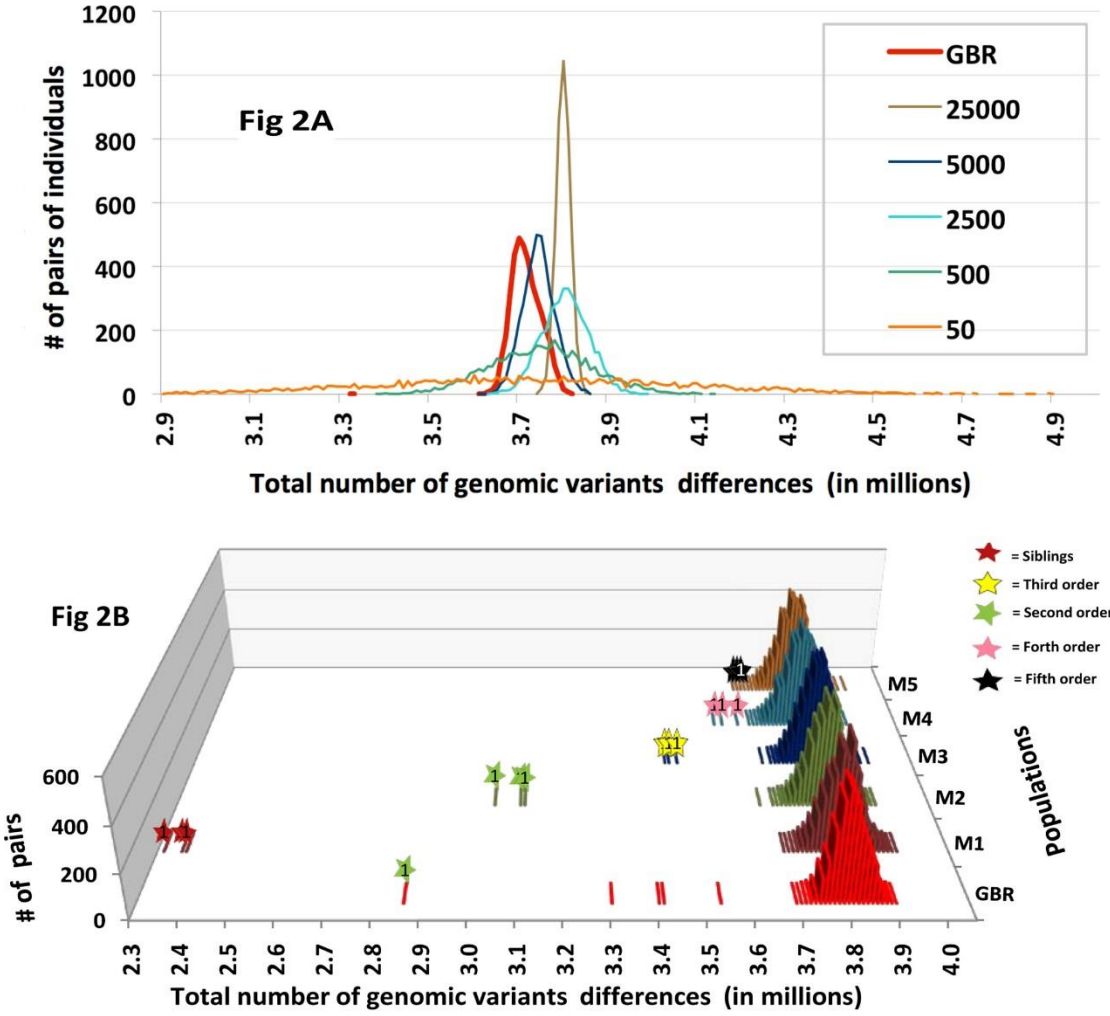


Fig 2A
# of pairs of individuals
1200
1000
800
600
400
200
0
2.9
3.1
3.3
3.5
3.7
3.9
4.1
4.3
4.5
4.7
4.9
Total number of genomic variants differences (in millions)
GBR
25000
5000
2500
500
50
Fig 2B
# of pairs
600
400
200
0
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4.0
Total number of genomic variants differences (in millions)
M5
M4
M3
M2
M1
GBR
populations
= Siblings
= Third order
= Second order
= Forth order
= Fifth order

Figure 4-3

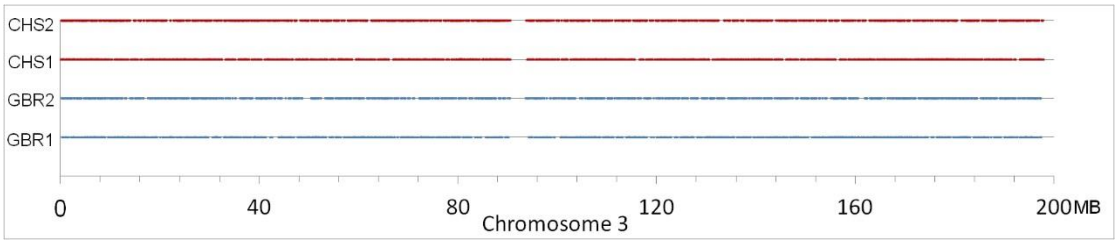


Figure 4-4

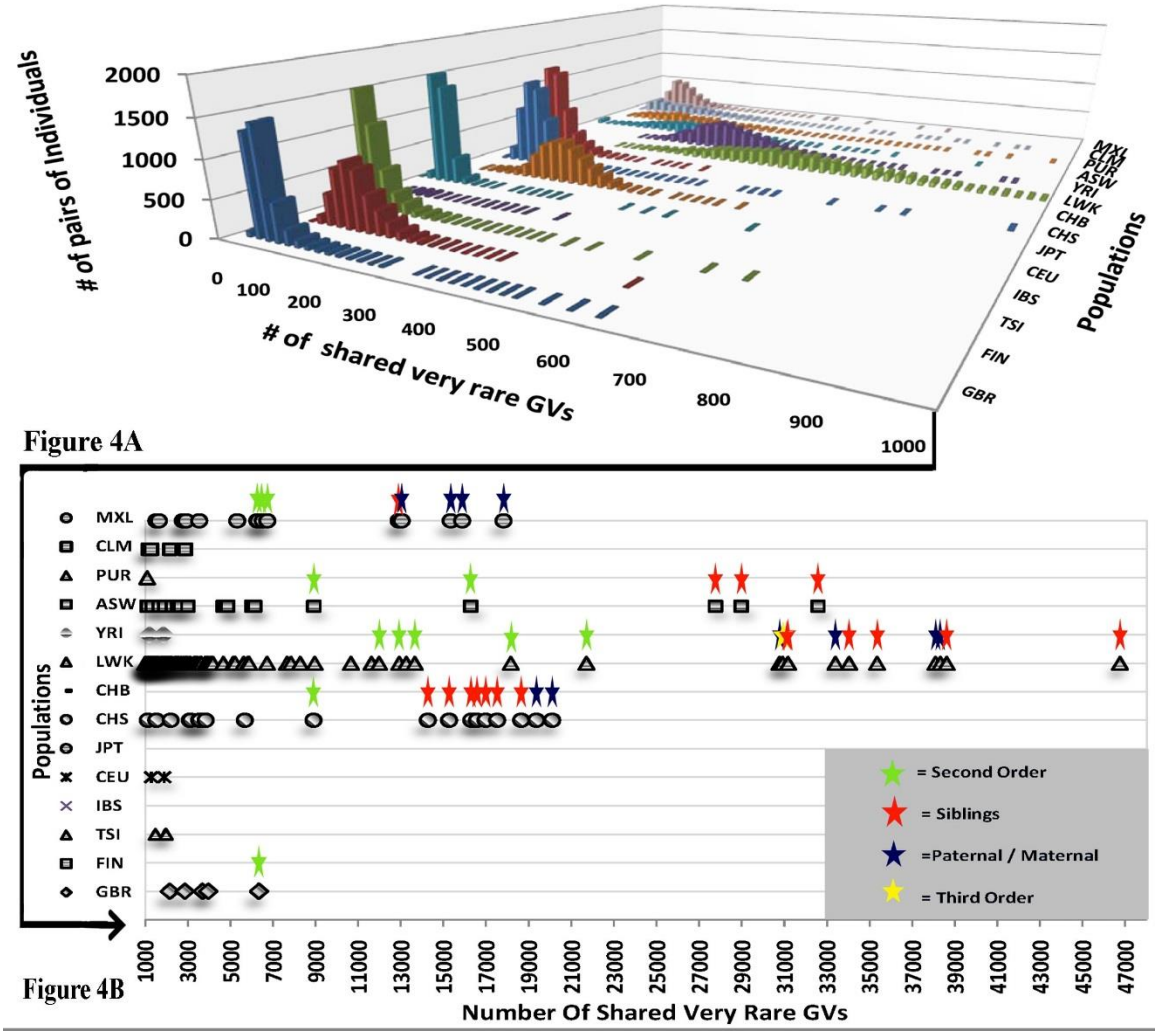
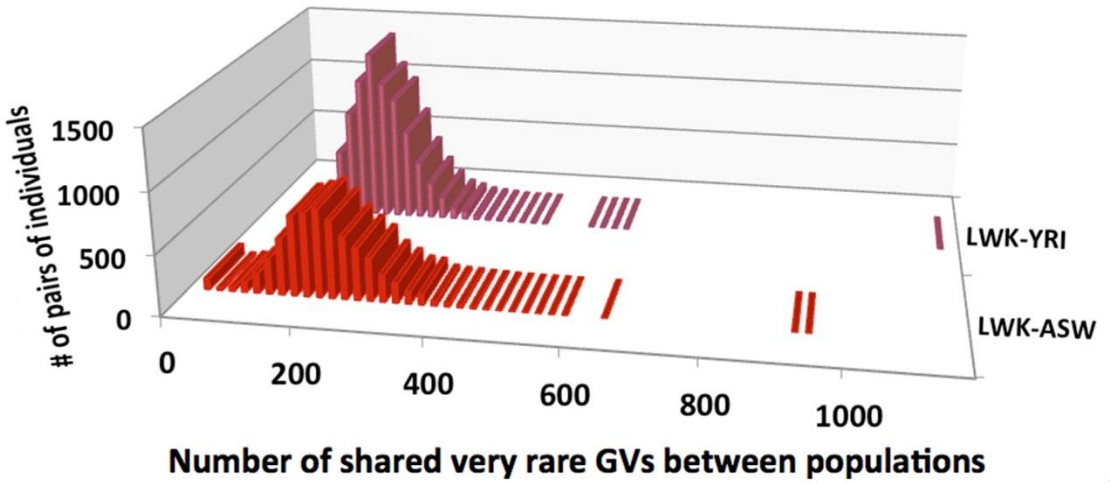

# of pairs of Individuals
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
900
1000
# of shared very rare GVs
Populations
MXL
CLM
PUR
ASW
YRI
LWK
CHB
CHS
JPT
CEU
IBS
TSI
FIN
GBR
Figure 4A
Populations
MXL
CLM
PUR
ASW
YRI
LWK
CHB
CHS
JPT
CEU
IBS
TSI
FIN
GBR
= Second Order
= Siblings
=Paternal / Maternal
= Third Order
1000
3000
5000
7000
9000
11000
13000
15000
17000
19000
21000
23000
25000
27000
29000
31000
33000
35000
37000
39000
41000
43000
45000
47000
Figure 4B
Number Of Shared Very Rare GVs

Figure 4-5


# of pairs of individuals
1500
1000
500
0
0
200
400
600
800
1000
Number of shared very rare GVs between populations
LWK-YRI
LWK-ASW

4.8 Supplementary Tables, Figures and Files

Table S4.1 Populations that have been used from the 1000 Genomes project.

1000 Genomes project Samples	
Population	**# of individuals**
Han Chinese in Beijing, China (**CHB**)	97
Japanese in Tokyo, Japan (**JPT**)	89
Southern Han Chinese, China (**CHS**)	100
Total East Asian Ancestry (ASN)	**286**
African Ancestry in Southwest US (**ASW**)	61
Luhya in Webuye, Kenya (**LWK**)	97
Yoruba in Ibadan, Nigeria (**YRI**)	88
Total African Ancestry (AFR)	**246**
British in England and Scotland (**GBR**)	89
Finnish in Finland (**FIN**)	93
Iberian populations in Spain (**IBS**)	14
Toscani in Italy (**TSI**)	98
Utah residents with Northern and Western European ancestry (**CEU**)	85
Total European Ancestry (EUR)	**379**
Colombian in Medellin, Colombia (**CLM**)	60
Mexican Ancestry in Los Angeles, California (**MXL**)	66
Puerto Rican in Puerto Rico (**PUR**)	55
Total Americas Ancestry (AMR)	**181**
Total	**1092**

Table S4.2 The entire set of vrGVs from 3 GBR Individuals and 3 CHS Individuals.

(Due to the large volume for this table, they can be assessed online by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/SupplementaryTableS2new.xlsx)

Table S4.3 Pairs of individuals sharing more than 1000 vrGVs.

(Due to the large volume for this table, they can be assessed online by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/SupplimentaryTableS3new.pdf)

Table S4.4 Numbers of shared vrGVs for pairs of individuals representing different populations (sorted).

(Due to the large volume for this table, they can be assess by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/Supplementary_TableS4new.xlsx)

Table S4.5 Characterization of 939 shared vrGVs for the (Americans of African Ancestry in SW USA)-(Luhya in Webuye, Kenya) pair composed by NA19397 and NA20348 individuals.

(Due to the large volume for this table, they can be assess by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/SupplementaryTableS5_Dec30.xlsx)

Table S4.6 Core haplotypes composed of all frequent (>20%) GVs from the chromosome 11 region: 90,787,654-90,858,949 (GRCh37 [hg19]) for two individuals HG00255 and NA18614.

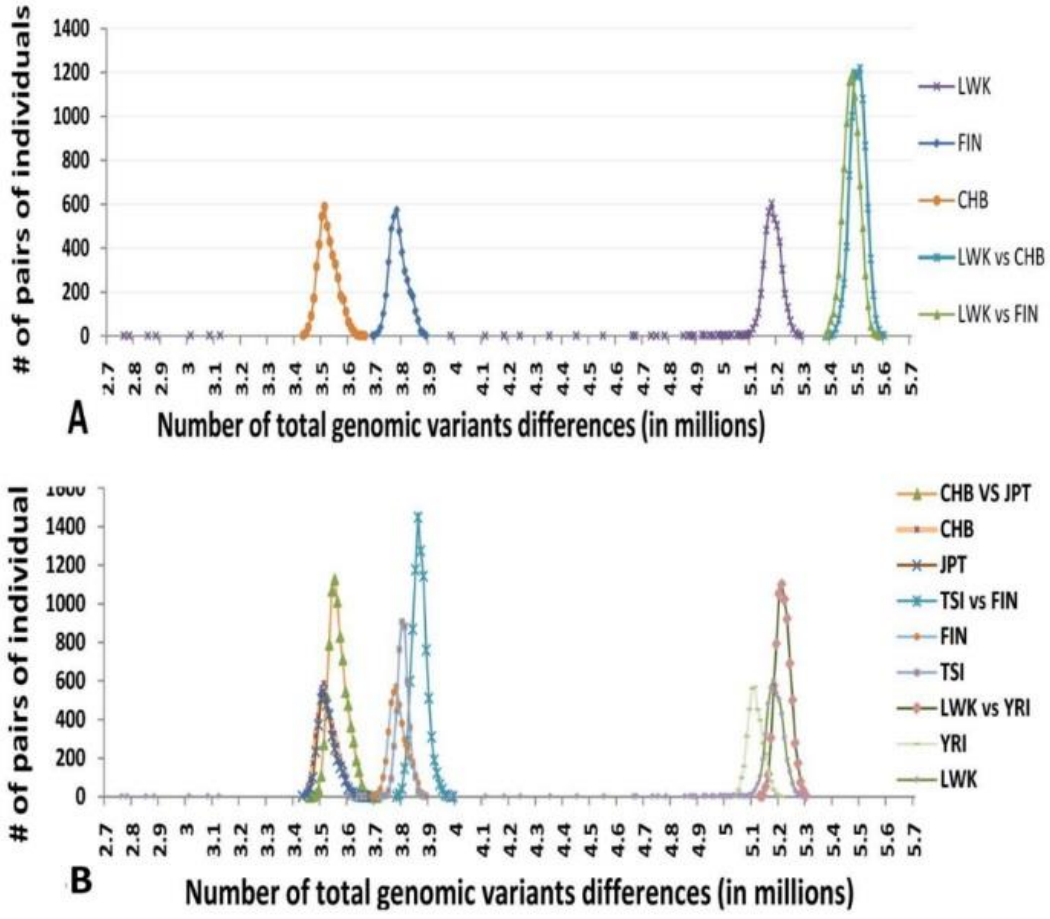
(Due to the large volume for this table, they can be assess by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/Supplementary_TableS6.xlsx)

Figure S4-1 Inter-population distributions of the total number of genomic differences in humans. The graphs reveal the fact that populations of the same continental ancestors share

more genetic differences in common than the populations from different continental ancestors.

A - The total number of genomic differences between pairs of individuals who belong to different continental ancestors e.g. (African "LWK" vs European "FIN") and (African "LWK" vs Asian "CHB"). The two distributions to the far right of the graph show the amount of genomic variability between pairs of different continental ancestors.

B - The total number of genomic differences between individuals of the same continental ancestor. e.g. (Asians "CHB" vs "JPT") , (Europeans "TSI" vs "FIN") and (Africans "LWK" vs "YRI").



File S4.1 Protocol for Assessing the Total Number of Genomic Differences in the 1000 Genomes Database.

(Due to the large volume for this table, they can be assess by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/Supplimentary_FileS1.pdf)

File S4.2 Statistical analysis.

(Due to the large volume for this table, they can be assess by the URL: http://gbe.oxfordjournals.org/content/suppl/2015/01/07/evv003.DC1/Supplementary_FileS2_Update12-30-2014.pdf)

4.9 Acknowledgements and Disclosure

We are grateful to Dr. Robert Blumenthal, University of Toledo Health Science Campus, for his insightful discussion of the project. The computations were performed in Oakley supercomputer with support from Ohio Supercomputer Center. We also appreciate the financial support from the Department of Medicine to conduct our research.

DISCLOSURE: The patent of our approach for detection of distant genetic relationships is pending.

4.10 References

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin et al., 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. Depristo, R. M. Durbin et al., 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. American journal of human genetics 73: 5-16.

Barbujani, G., A. Magagni, E. Minch and L. L. Cavalli-Sforza, 1997 An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America 94: 4516-4519.

Boehnke, M., and N. J. Cox, 1997 Accurate inference of relationships in sib-pair linkage studies. American journal of human genetics 61: 423-429.

Browning, B. L., and S. R. Browning, 2013 Detecting identity by descent and estimating genotype error rates in sequence data. American journal of human genetics 93: 840-851.

Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. American journal of human genetics 86: 526-539.

Conrad, D. F., J. E. Keebler, M. A. Depristo, S. J. Lindsay, Y. Zhanget al., 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714.

Consortium, I. H., 2003 The International HapMap Project. Nature 426: 789-796.

Durand, E. Y., N. Eeiksson and C. Y. Mclean, 2014 Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. Mol Biol Evol 31: 2212-2222.

Fagny, M., E. Patin, D. Enard, L. B. Barreiro, L. Quintana-Murci et al., 2014 Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol 31: 1850-1868.

Gravel, S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio et al., 2013 Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS genetics 9: e1004023.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS genetics 9: e1003521.

Hartl, D. L., Clark, A.G., 2007 Principles of Population Genetics. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, USA.

Hochreiter, S., 2013 HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res 41: e202.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins et al., 2011 Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research 21: 768-774.

Jobling, M. A., and P. Gill, 2004 Encoded evidence: DNA in forensic analysis. Nature reviews. Genetics 5: 739-751.

Keinan, A., and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-743.

Kondrashov, A. S., and S. A. Shabalina, 2002 Classification of common conserved sequences in mammalian intergenic regions. Human molecular genetics 11: 669-674.

Kruskal, W. H. W., W.A., 1952 Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47: 583-621.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.

Li, H., G. Glusman, H. Hu, Shankaracharya, J. Caballero et al., 2014 Relationship estimation from whole-genome sequence data. PLoS genetics 10: e1004144.

Moore, C. B., J. R. Wallace, D. J. Wolfe, A. T. Frase, S. A. Pendergrass et al., 2013 Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS genetics 9: e1003959.

Odegard, J., and T. H. Meuwissen, 2012 Estimation of heritability from limited family data using genome-wide identity-by-descent sharing. Genet Sel Evol 44: 16.

Parson, W., and H. J. Bandelt, 2007 Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 1: 13-19.

Qiu, S., A. Mcsweeny, S. Choulet, A. Saha-Mandal, L. Fedorova et al., 2014 Genome evolution by matrix algorithms: cellular automata approach to population genetics. Genome Biol Evol 6: 988-999.

Thompson, E. A., 1975 The estimation of pairwise relationships. Annals of human genetics 39: 173-188.

Weir, B. S., A. D. Anderson and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;10:771-780.

Willuweit, S., A. Caliebe, M. M. Andersen and L. Roewer, 2011 Y-STR Frequency Surveying Method: A critical reappraisal. Forensic Sci Int Genet 5: 84-90.

Wright, S., 1922 Coefficients of inbreeding and relationship. American Naturalist 56: 330-338.

Chapter 5

Conclusions

In this dissertation, 3 aspects of mutations within a population are analyzed. They are: the fate of a mutation, mutational age, and the relationship between individuals revealed by vrGVs in the population. Specifically, in order to answer the first two questions, a new computational algorithm (GEMA) was designed and employed to analyze the fate and age of a mutation under various environments by simulation. After that, by thorough analysis of next generation sequencing data from the "1000 Genome Project", a new method was designed to reveal relationships between individuals using very rare genetic variations.

Compared to previous computational simulations in the population genetics field, the software GEMA has unique features to answer specific questions. It simulates a constant and intense influx of new mutations occurring in each individual in a population with a situation close to that found in reality (Kondrashov and Shabalina 2002a; Conrad et al. 2011a; Li and Durbin 2011). Thus, it can simulate and address the fate of mutations occurring in actual populations. By using GEMA and running it with a number of different population parameters, we demonstrate that the specific combinations of different population parameters intricately and dramatically affect the probability of

fixation of a mutation and the fitness of an individual. 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 further modified the codes of GEMA and improved it to address other questions, e.g. related to the age of a mutation after its occurrence in the population. While it is very hard to verify the age of an SNP through thousands of generations experimentally, it can be simulated by computational approaches to record and trace a mutation. Since it is a simulation including thousands of mutations happening in a population, GEMA can include multiple mutations under various environments by applying different population parameters (recombination rate, number of offspring, mutation number, selection coefficient, and dominance coefficient and population size) in the analysis. Through our analysis, we demonstrated that the Maruyama effect (Maruyama 1974a) was detected only when the recombination rate was high (***r***=48) and neutral mutations were overabundant. However, when we decreased frequencies of neutral mutations and increased those of beneficial and deleterious mutations under the same conditions, the average ages of mutations were practically the same irrespective of fitness effects (no Maruyama effect).

After having simulated thousands of mutations computationally, our research focus turned to the existing real human sequencing data. With the advent of high-throughput sequencing technology and availability of more and more human sequences, it is now possible to analyze the total number of genetic variations between two individuals. Furthermore, from this analysis, it is possible to infer the relationship between

individuals. For this purpose, after computational analysis for every possible pair of individuals from the Phase I data released by "1000 Genome Project", we demonstrated that human populations are distinct from one another by distribution of genomic differences among their individuals. For example, the American population that experienced admixture a few hundred years ago, via inclusion of people from different continents, have much wider distributions of genomic differences compared to the other populations (African, Asian and European). The further analysis on the distribution of vrGVs between individuals demonstrated that vrGVs could potentially reveal the genetic relationships for these individuals. Thus, in addition to well-established DNA fingerprinting, we now contribute a new method for deriving distant genetic relationships.

All these three projects together, instead of looking one mutation at a time, we viewed and studied hundreds of mutations as an integrity. Our GEMA simulation results demonstrated that this integrity is linked and shaped by different population parameters as detailed in Chapter 2. The subsequent allelic age analysis for different mutations in Chapter 3 further illustrated the influence of these population parameters on the maintenance of such integrity. In Chapter 4, by examining real genetic variants data from 1000 Genome Project Phase I, we revealed that the distribution of number of variations between individuals is population-specific (Figure 4.1), which indirectly shows that different populations have various population parameters.

Finally, with the advancement of technology, cheap sequencing in the near future will provide unprecedented large genomic data. Such data opens the possibility of direct investigation of the fate and the age of many mutations. We could expect to compare and

validate our current simulation results from GEMA. Also, with such technology, whole-genome sequencing analysis of large pedigrees will also become routine. We can also expect more computational approaches to be designed to investigate such large datasets that can then reveal significant predictions on the fate of mutations that can be validated.

References

Chapter 1

Aminetzach, Y. T., J. M. Macpherson and D. A. Petrov, 2005 Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in Drosophila. Science 309: 764-767.

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

Boehnke, M., and N. J. Cox, 1997 Accurate inference of relationships in sib-pair linkage studies. American journal of human genetics 61: 423-429.

Browning, B. L., and S. R. Browning, 2013 Detecting identity by descent and estimating genotype error rates in sequence data. American journal of human genetics 93: 840-851.

Burrus, V., and M. K. Waldor, 2004 Shaping bacterial genomes with integrative and conjugative elements. Research in Microbiology 155: 376-386.

Carvajal-Rodriguez A. 2008. GENOMEPOP: A program to simulate genomes in populations. Bmc Bioinformatics 9. doi: Artn 223 Doi 10.1186/1471-2105-9-223

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

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008a. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9. doi: Artn 364
Doi 10.1186/1471-2105-9-364

Conrad, D. F., J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang et al., 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714.

Durand, E. Y., N. Eriksson and C. Y. Mclean, 2014 Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree analysis. Mol Biol Evol 31: 2212-2222.

Fisher RA. 1930. The Genetic Theory of Natural Selection. Dover: Oxford University Press.

Gavrilets S, Hastings A. 1994. Dynamics of genetic variability in two-locus models of stabilizing selection. Genetics 138: 519-532.

Griffiths, R. C., and S. Tavare, 1999 The ages of mutations in gene trees. Annals of Applied Probability 9: 567-590.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786-2787. doi: Doi 0.1093/Bioinformatics/Btn522

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genetical research 8: 269-294.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins et al., 2011 Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research 21: 768-774.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics 75: 199-212.

Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Human molecular genetics 11: 669-674.

Li, W. H., 1975 The first arrival time and mean age of a deleterious mutant gene in a finite population. Am J Hum Genet 27: 274-286.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475: 493-496. doi: 10.1038/nature10231

LI, H., G. GLUSMAN, H. HU, SHANKARACHARYA, J. CABALLERO et al., 2014 Relationship estimation from whole-genome sequence data. PLoS genetics 10: e1004144.

Maruyama, T., 1974a The age of a rare mutant gene in a large population. Am J Hum Genet 26: 669-673.

Maruyama, T., 1974b The age of an allele in a finite population. Genet Res 23: 137-143.

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. doi: 10.1146/annurev-genom-082908-150129

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.

Sawyer, S. A., Z. Parsch J Fau - Zhang, D. L. Zhang Z Fau - Hartl and D. L. Hartl, Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila.

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu Rev Genomics Hum Genet 1: 225-249.

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. doi: 10.1098/rstb.2009.0278

Thompson, E. A., 1975 The estimation of pairwise relationships. Annals of human genetics 39: 173-188.

Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nature reviews. Genetics 9: 965-974. doi: 10.1038/nrg2473

Weir, B. S., A. D. Anderson and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;10:771-780.

Wright S. 1965. Factor Interaction and Linkage in Evolution. Proc. R. Soc. Lond. B 162: 80-104. doi: 10.1098/rspb.1965.0026

Chapter 2

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. doi: 10.1038/nature11632

Bechtel JM, Wittenschlaeger T, Dwyer T, Song J, Arunachalam S, Ramakrishnan SK, Shepard S, Fedorov A. 2008. Genomic mid-range inhomogeneity correlates with

an abundance of RNA secondary structures. Bmc Genomics 9: 284. doi: 1471-2164-9-284 [pii] 10.1186/1471-2164-9-284

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. doi: Doi 10.1073/Pnas.0701652104

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. doi: Artn 223 Doi 10.1186/1471-2105-9-223

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

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008a. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9. doi: Artn 364 Doi 10.1186/1471-2105-9-364

Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008b. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. Bmc Bioinformatics 9: 364. doi: 10.1186/1471-2105-9-364

Charlsworth BaC, Deborah. 2010. Elements of Evolutionary Genetics. Greenwood Village, Colorado: Roberts and Comapany 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. doi: 10.1038/ncomms3417

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. doi: 10.1007/s00285-007-0121-7

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P. 2011. Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714. doi: 10.1038/ng.862

Consortium TIH. 2005. A haplotype map of the human genome. Nature 437: 1299-1320. doi: 10.1038/nature04226

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, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM,

Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD,

McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861. doi: 10.1038/nature06258

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. 2007. Principles of population genetics.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786-2787. doi: Doi 0.1093/Bioinformatics/Btn522

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. doi: 10.1038/84773

Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, van Duijn CM, Slagboom PE, van Ommen GJ, Wijmenga C, de Bakker PI, Sunyaev SR. 2013. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS genetics 9: e1003301. doi: 10.1371/journal.pgen.1003301

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. doi: 10.1038/nature10231

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. doi: 10.1146/annurev-genom-082908-150129

Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theoretical Population Biology 49: 128-142. doi: Doi 10.1006/Tpbi.1996.0007

Patwa Z, Wahl LM. 2008. The fixation probability of beneficial mutations. Journal of the Royal Society, Interface / the Royal Society 5: 1279-1289. doi: 10.1098/rsif.2008.0248

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 12: 32-42. doi: Doi 10.1038/Nrg2899

Prakash A, Shepard SS, Mileyeva-Biebesheimer O, He J, Hart B, Chen M, Amarachintha SP, Bechtel J, Fedorov A. 2009. Evolution of Genomic Sequence Inhomogeneity at Mid-range Scales. Bmc Genomics 10: 513. doi: 10.1186/1471-2164-10-513

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. doi: 10.1073/pnas.0700890104

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. doi: 10.1098/rstb.2009.0278

Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nature reviews. Genetics 9: 965-974. doi: 10.1038/nrg2473

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. doi: 10.1098/rspb.1965.0026

Zhang C, Plastow G. 2011. Genomic Diversity in Pig (Sus scrofa) and its Comparison with Human and other Livestock. Current Genomics 12: 138-146. doi: 10.2174/138920211795564386

Chapter 3

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin et al., 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved. American Journal of Human Genetics 73: 5-16.

Conrad, D. F., J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. J. Zhang et al., 2011 Variation in genome-wide mutation rates within and between human families. Nature Genetics 43: 712-U137.

Genin, E., A. Tullio-Pelet, F. Begeot, S. Lyonnet and L. Abel, 2004 Estimating the age of rare disease mutations: the example of Triple-A syndrome. J Med Genet 41: 445-449.

Griffiths, R. C., and S. Tavare, 1999 The ages of mutations in gene trees. Annals of Applied Probability 9: 567-590.

Kiezun, A., S. L. Pulit, L. C. Francioli, F. van Dijk, M. Swertz 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., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics 75: 199-212.

Kondrashov, A. S., and S. A. Shabalina, 2002 Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet 11: 669-674.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.

Li, W. H., 1975 The first arrival time and mean age of a deleterious mutant gene in a finite population. Am J Hum Genet 27: 274-286.

Maruyama, T., 1974a The age of a rare mutant gene in a large population. Am J Hum Genet 26: 669-673.

Maruyama, T., 1974b The age of an allele in a finite population. Genet Res 23: 137-143.

Qiu, S., A. McSweeny, S. Choulet, A. Saha-Mandal, L. Fedorova et al., 2014 Genome Evolution by Matrix Algorithms: Cellular Automata Approach to Population Genetics. Genome Biology and Evolution 6: 988-999.

Rannala, B., and J. P. Reeve, 2001 High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am J Hum Genet 69: 159-178.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter et al., 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter et al., 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

Slatkin, M., and B. Rannala, 1997 Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60: 447-458.

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu Rev Genomics Hum Genet 1: 225-249.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS biology 4: e72.

Wisloff, U., S. M. Najjar, O. Ellingsen, P. M. Haram, S. Swoap et al., 2005 Cardiovascular risk factors emerge after artificial selection for low aerobic capacity. Science 307: 418-420.

Chapter 4

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin et al., 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. Depristo, R. M. Durbin et al., 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. American journal of human genetics 73: 5-16.

Barbujani, G., A. Magagni, E. Minch and L. L. Cavalli-Sforza, 1997 An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America 94: 4516-4519.

Boehnke, M., and N. J. Cox, 1997 Accurate inference of relationships in sib-pair linkage studies. American journal of human genetics 61: 423-429.

Browning, B. L., and S. R. Browning, 2013 Detecting identity by descent and estimating genotype error rates in sequence data. American journal of human genetics 93: 840-851.

Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. American journal of human genetics 86: 526-539.

Conrad, D. F., J. E. Keebler, M. A. Depristo, S. J. Lindsay, Y. Zhanget al., 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714.

Consortium, I. H., 2003 The International HapMap Project. Nature 426: 789-796.

Durand, E. Y., N. Eeiksson and C. Y. Mclean, 2014 Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. Mol Biol Evol 31: 2212-2222.

Fagny, M., E. Patin, D. Enard, L. B. Barreiro, L. Quintana-Murci et al., 2014 Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol 31: 1850-1868.

Gravel, S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio et al., 2013 Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS genetics 9: e1004023.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS genetics 9: e1003521.

Hartl, D. L., Clark, A.G., 2007 Principles of Population Genetics. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, USA.

Hochreiter, S., 2013 HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res 41: e202.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins et al., 2011 Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research 21: 768-774.

Jobling, M. A., and P. Gill, 2004 Encoded evidence: DNA in forensic analysis. Nature reviews. Genetics 5: 739-751.

Keinan, A., and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-743.

Kondrashov, A. S., and S. A. Shabalina, 2002 Classification of common conserved sequences in mammalian intergenic regions. Human molecular genetics 11: 669-674.

Kruskal, W. H. W., W.A., 1952 Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47: 583-621.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.

Li, H., G. Glusman, H. Hu, Shankaracharya, J. Caballero et al., 2014 Relationship estimation from whole-genome sequence data. PLoS genetics 10: e1004144.

Moore, C. B., J. R. Wallace, D. J. Wolfe, A. T. Frase, S. A. Pendergrass et al., 2013 Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS genetics 9: e1003959.

Odegard, J., and T. H. Meuwissen, 2012 Estimation of heritability from limited family data using genome-wide identity-by-descent sharing. Genet Sel Evol 44: 16.

Parson, W., and H. J. Bandelt, 2007 Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 1: 13-19.

Qiu, S., A. Mcsweeny, S. Choulet, A. Saha-Mandal, L. Fedorova et al., 2014 Genome evolution by matrix algorithms: cellular automata approach to population genetics. Genome Biol Evol 6: 988-999.

Thompson, E. A., 1975 The estimation of pairwise relationships. Annals of human genetics 39: 173-188.

Weir, B. S., A. D. Anderson and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;10:771-780.

Willuweit, S., A. Caliebe, M. M. Andersen and L. Roewer, 2011 Y-STR Frequency Surveying Method: A critical reappraisal. Forensic Sci Int Genet 5: 84-90.

Wright, S., 1922 Coefficients of inbreeding and relationship. American Naturalist 56: 330-338.

Chapter 5

Conrad, D. F., J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang et al., 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics 43: 712-714.

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. doi: 10.1038/nature10231

Maruyama, T., 1974a. The age of a rare mutant gene in a large population. Am J Hum Genet 26: 669-673.