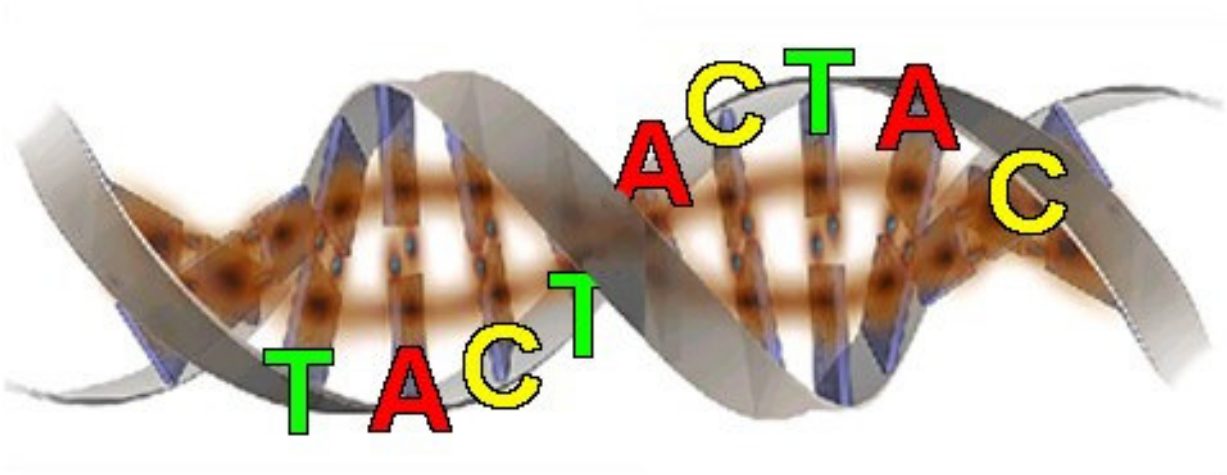
Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Biologia Geral



Simulações evolutivas com microssatélites ligados para estudos de tempos de divergência populacional

Leandro Martins de Freitas
Belo Horizonte
2007

Simulações evolutivas com microssatélites ligados para estudos de tempos de divergência populacional

Dissertação apresentada ao Programa de Pós-Graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Mestre em Genética.

Orientador: Prof Dr Fabrício Rodrigues dos Santos

Índice

Agradecimentos

Agradeço a minha família que me ajudou e apoiou neste caminho. Muito obrigado, mãe, por acreditar e torcer por mim. Agradeço aos meus irmãos (Patrícia, Leonardo e Haroldo) por me ajudarem quando precisei e por serem meus amigos.

Gostaria de agradecer aos meus amigos do departamento e do LBEM (Anderson, Camila, Claudia, Dani, Débora, Eloísa - que me orientou na parte estatística, Fabiano – que me deu uma força no MySQL, Felipe, Josimar, Letícia, Marilza, Sarah, Sibelle, Ricardo e Rodrigo) e ao Francisco Prosdocimi - que deu início a este projeto e confiou o seu desenvolvimento a mim.

Agradeço também aos professores desta universidade, principalmente os de genética com quem aprendi tanto.

Á CAPES que financiou a compra dos computadores e ao CNPq que financiou a minha bolsa.

Ao professor Doutor Fabrício Rodrigues dos Santos que me forneceu espaço no seu laboratório para o desenvolvimento deste trabalho e deu exemplo de um profissional dedicado e responsável.

Ana Raquel, minha gatinha, que passou todo o tempo durante esse trabalho comigo e me apoiou e procurou me dar novas idéias. Muito obrigado mesmo.

Muito Obrigado a todos vocês!

Resumo

Os microssatélites têm sido muito usados para o estudo de populações devido a sua grande variabilidade. Esta alta variabilidade é devida à alta taxa de mutação. Os microssatélites têm ajudado muito na reconstrução da historia evolutiva humana e estudos de migração através das Américas. Estudos usando microssatélites no cromossomo Y ajudam a entender como e quando ocorreu esta migração. A migração apresenta uma estrutura geográfica considerável, podendo ser observada ao longo de várias partes do mundo. Junto com estudos moleculares, as simulações da evolução também são muito úteis para entender como podem ter ocorrido tais eventos. Simulamos uma molécula não recombinante e com vários *loci* de microssatélites associados, usando um script (MGSim) Perl com o objetivo de avaliar os efeitos de reduções populacionais nas datações do ancestral feitas usando este tipo de marcador. Simulamos vários grupos com redução da população em pontos específicos em cada grupo. Todas as simulações têm os mesmos parâmetros diferindo somente se ocorre redução da população ou não. Os grupos com redução de 90% e de 99% da população não apresentaram diferença significativa em relação ao grupo controle quando comparado o valor de Average Squared Difference (ASD), mas os grupos com redução de 99.9% e de 99% mais 20 gerações com tamanho constante apresentaram diferença significativa. Fizemos também uma reconstrução da proporção de alelo e haplótipo ancestral recuperado corretamente ao longo das gerações e verificamos que o grupo com redução de 90% não apresenta curvas diferentes do grupo controle, mas nos outros grupos essa queda é muito rápida mostrando que podem existir erros nas datações das populações ate seu ancestral.

Abstract

The microsatellites (STR) have been used in the study of populations due to their high variability. This high variability occurs because of the high mutation rate. The microsatellites have helped significantly in the reconstruction of the human evolutionary history and in migration studies through the Americas. Studies using microsatellites in the Y chromosome help to understand how and when the migration occurred. The migration presented a considerable geographic structure that was observed in several regions of the world. Together with molecular studies, the simulations of evolution are also very important and useful to understand how such events might have happened. With the aim to evaluate a non recombinant molecule and with several microsatellite loci, we developed a script (MGSim) using Perl language, that evaluate the effects of reduction of populations in the dating of the ancestral, that are done using this kind of marker. We submitted several groups to reduction of the population in a specific point in each group. All the simulations have the same parameters differing only in the reduction of the population or not. The group with 90% and 99% of reduction of the population did not show a significant difference from the control group when comparing the Average Squared Difference (ASD) variation, but the groups with reduction of 99.9% and 99% plus 20 generations with constant size showed significant differences. We made also a reconstruction of the proportion of allele and ancestral haplotype correctly retrieved through the generations and observed that in the group with 90% of reduction there were no difference from the control group, but in the other groups this decrease is too fast, showing that mistakes may occur in the populations dating until its ancestral.

1. Introdução

Microssatélites

Nos últimos anos, microssatélites tornaram-se importantes marcadores genéticos para estudar as relações filogenéticas/genealógicas entre indivíduos relacionados ou entre populações (Pritchard e Feldman,1996) e são descritos como marcadores ideais para avaliar a dinâmica intra-populacional (Bowcock *et al*., 1996) porque existe, freqüentemente, um grande número de alelos, há codominância, são abundantes dentro do genoma e é relativamente fácil caracterizar o número de repetições (Lehmann *et al.*, 1996).

Microssatélites são *loci* com repetições em tandem em que um curto segmento de DNA (normalmente dois a seis pares de bases - pb) é repetido até aproximadamente 100 vezes (Tautz, 1993). Os alelos variáveis são caracterizados pelos diferentes números de repetições.

Os microssatélites são classificados em relação ao tamanho da unidade repetitiva, sendo denominados mono, di, tri, tetra, penta e hexanucleotídeos, e quanto à estrutura de sua repetição, sendo denominados perfeitos, quando existe um único tipo de repetição sem interrupções; imperfeitos, quando existem diferentes bases intercalando a repetição; ou compostos, quando mais de duas repetições perfeitas ou imperfeitas são separadas por no máximo 3 pb (Weber, 1993).

O alto nível de diversidade dos microssatélites, é devido à alta taxa de mutação (entre 10^{-3} a 10^{-4} mutações/*locus*/meiose, Weber e Wong, 1993, Heyer *et al*. 1997), o que os tornam mais informativos para estudar genética e evolução de populações (Goldstein e Feldmam, 1995b).

Outros interesses nos microssatélites se devem a sua importância no mapeamento do genoma de varias espécies, incluindo humanos. Alguns microssatélites têm sido também associados a doenças genéticas. Instabilidades em certas repetições de microssatélites foram identificadas em pelo menos três doenças humanas e podem ser utilizadas para compreender o mecanismo da doença e seu diagnóstico.

Os microssatélites são amplamente distribuídos ao longo dos genomas de mamíferos. Esta característica distingue os microssatélites dos minissatélites, também

chamados de VNTR (*variable number of tandem repeats*) que consistem de blocos em tandem de 20 pb ou mais e são encontrados predominantemente nas regiões sub-teloméricas (Jeffreys *et al*., 1988; Valdes e Freimer, 1993).

A identificação dos microssatélites pode ser através de busca em bancos de dados (GenBank, EMBL) ou por varredura genômica em bibliotecas de DNA por hibridização com oligonucleotídeos, compostos de seqüências especificas. Seqüências únicas que flanqueiam o microssatélite são identificadas e um par de iniciadores para estas seqüências específicas é usado para amplificar a repetição através da PCR que permite a identificação do polimorfismo através de gel em eletroforese ou em um seqüenciador automático.

O principal processo mutacional para a produção dos microssatélites parece envolver uma derrapagem da fita de DNA durante a replicação. Durante a síntese de DNA, as duas fitas podem derrapar uma sobre a outra, formando bolhas de DNA não-pareado. Muitas destas bolhas são corrigidas pelo sistema de reparo, mas a pequena proporção de bolhas não reparadas resulta no ganho ou na perda de uma unidade repetitiva. Estas mutações ao longo das gerações produzem os diferentes alelos. Em microssatélites ligados, estes *loci* se associam como haplótipos e o aparecimento de um novo alelo em qualquer *locus* gera um novo haplótipo (figura 1). Normalmente a mutação envolve uma única repetição produzindo uma inserção ou deleção (Schlotterer e Tautz, 1992), cujo modelo é conhecido como mutacional passo a passo (Stepwise Mutation Model - SMM)

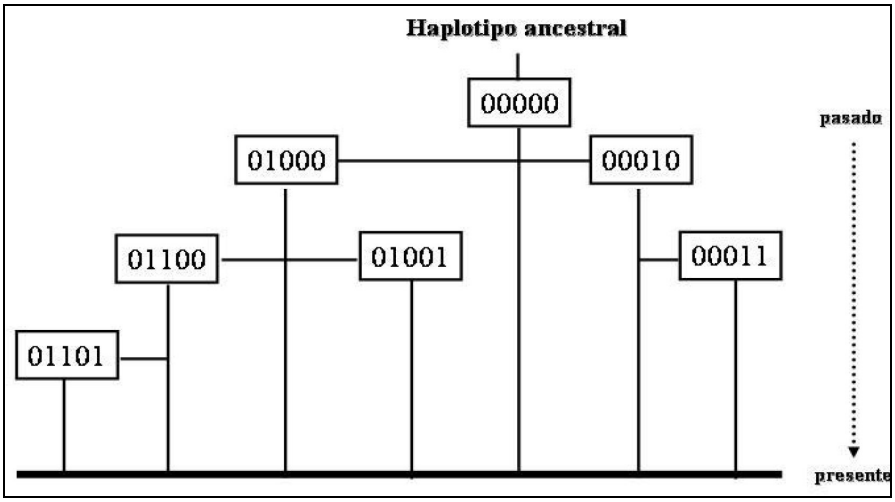


Figura 1 Estabelecimento do processo de diferenciação de haplótipos de microssatélites ao longo do tempo (Villalobos 2006).

Segundo o SMM uma seqüência de alelo pode ser expressa por inteiros (..., A-1, A0, A+1,...), e que, se um alelo muda seu estado por mutação, ele move um passo na direção positiva ou um passo na direção negativa. Deixe *v* ser a taxa mutacional por *locus* por geração, e assumir que a mudança mutacional em direção ao negativo e ao positivo ocorra com igual freqüência como mostra a figura 2

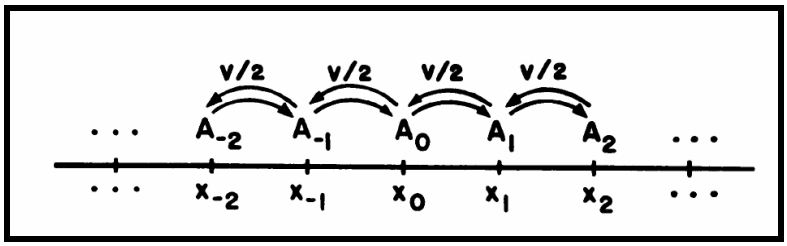


Figura 2 – Diagrama ilustrando o modelo SMM (Ohta e Kimura, 1973).

Simulações Diretas (*Forward*)

Existem muitas situações nas quais as simulações genéticas poderiam ser muito úteis. Simulações permitem a exploração e o entendimento de situações muito complexas e também permitem a validação de inferências feitas a partir de estudos empíricos. Elas podem ter diferentes parâmetros para comparações posteriores com grande conjunto de dados (Balloux, 2001).

Simulações usando o método de coalescência (Kingman, 1982) têm sido muito usadas devido às suas eficiência e flexibilidade. O método de coalescência permite simular diversos tipos de cenários evolutivos, diferente tipos de marcadores genéticos e possui uma velocidade muito grande para produção de dados. Ao contrario, Simulações "diretas" (*forward*), embora sejam simples, são computacionalmente ineficientes e foram usadas principalmente com o objetivo educacional. Entretanto recentemente, devido ao crescimento exponencial da capacidade de processamento dos computadores, as simulações *forward* passaram a serem usadas em estudos genéticos evolutivos (Balloux, 2001; Hey, 2004, http://lifesci.rutgers.edu/heylab/HeylabSoftware.htm; Balloux e Goudet, 2002). O significado de "direta" neste contexto é simples, a simulação move do ancestral para os descendentes como acontece nas populações reais ao contrário das simulações que usam o método de coalescência que vão do presente para o passado.

A simulação *forward* é mais flexível no sentido que outros efeitos podem ser aplicados na população em uma geração específica, enquanto que na simulação usando coalescência (e.g SIMCOAL, http://cmpg.unibe.ch/software/simcoal/) a adição destes fatores ainda é complicada mesmo quando queremos implementar uma simples força evolutiva como seleção (Fearnhead, 2003).

Ao contrário do método de coalescência, simulação *forward* guarda todas as informações sobre os ancestrais e gerações subsequentes. Isto permite às simulações *forward* uma exploração mais realísta do efeito de diferentes processos evolutivos. (Calafell *et al*., 2001; Balloux e Goudet, 2002).

Nos últimos anos apareceram programas que usam a metodologia de simulações *forward* como *Forward Population Genetic simulation* (Hey - http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm), SimuPOP (Peng e Kimmel 2005) e Easypop (Balloux 2001).

O uso de simulações para os estudos evolutivos é muito importante, pois é um meio de se verificar o acúmulo de diversidade genética com o tempo e a possibilidade de se usar microssatélites para datar eventos demográficos. Estes dados de simulações podem ser comparados com populações reais e podemos ver se estão dentro da variação encontrada nas simulações, indicando que os parâmetros das simulações explicariam a história desta população.

2 Objetivos

- Este trabalho tem como objetivos: desenvolver um programa que simule a evolução de microssatélites ligados e comparar grupos de simulações com e sem redução da população para inferir possíveis efeitos de fatores estocásticos na historia da população.
- Comparar os estimadores modal e médio para inferir qual seria o melhor estimador para a reconstrução do ancestral comum.
- Analisar a proporção de alelo e haplótipo ancestral recuperado corretamente ao longo da genealogia
- Construção de redes de haplótipos das genealogias simuladas em gerações especificas usando o programa Network 4.2 (Bandelt et al, 1999).

3 Resultados

Manuscript prepared for Journal of Molecular Evolution

Linked microsatellite evolution approached by forward simulations

Leandro M. Freitas [1], Francisco Prosdocimi [2], Fabrício R. Santos [1]

(1) Laboratório de Biodiversidade e Evolução Molecular, Department of General Biology and (2) Laboratório de Biodados, Department of Biochemistry and Immunology, Instituto de Ciências Biológicas , Belo Horizonte, Minas Gerais, Brazil

Keywords: Ancestral, forward evolution, simulation, microsatellite

Running header: Linked microsatellites evolutionary simulation

Corresponding author

Fabrício R. Santos
Departamento de Biologia Geral, ICB, UFMG
Av. Antônio Carlos, 6627 C.P. 486
31.270-010 Belo Horizonte, MG, Brazil
Phone: +55 31 3499-2581
Fax: +55 31 3499-2570
fsantos@icb.ufmg.br

Abstract

Human Y chromosome microsatellites have been extensively used to study human history, particularly concerning recent events as the Peopling of the Americas. These studies include genetic dating methods such as the ones that identify the ancestral microsatellite haplotype in a founding haplogroup defined by SNPs and calculate the divergence time from this ancestral haplotype to the present day population characterized by a diverse set of derived haplotyes. Our main goal was to develop a software (Microsatellite Genealogy Simulator - MGSim) that simulates forward evolution of linked microsatellites from a single haplotype ancestor. The software was built in PERL language that stores all simulated data in a MySQL database. We simulated several population sets for 2,000 generations with different evolutionary scenarios, particularly bottlenecks, each one repeated 50 times, which were compared by distinct population genetic analyses. The analysis of variance indicated that only the groups with bottleneck lasting 20 generations were significantly different from the control group, without bottleneck. After bottleneck, population diversity was recovered after around 400 to 800 generations, when reduction occurred in generation 100 and 500, respectively. Another analysis measured the proportion of ancestral alleles or haplotypes recovered after the bottleneck along the genealogy. The results showed us that when bottleneck occurs in the beginning of the genealogy (generation 100) the recovery of ancestral alleles was about 60% less efficient than in the control batch and when the reduction happened later in generation 500, the recovery was 30% less than in the control group.

Introduction

Microsatellites are tandem repeats, where a short monomer, usually two to six base pairs, is repeated until approximately 100 times (Tautz 1993). Different alleles are characterized by differences in the number of repeats. Nowadays, microsatellites are considered important genetic markers used to study phylogenetic relationships between individuals or populations (Pritchard and Feldman 1996). Besides, they are considered ideal markers to estimate population structure. The main reasons for this popularity are: they display a large number of alleles in each locus; they are codominant; they are abundant and widespread in the genome; and it is easy to characterize the repeat number of each allele (Lehmann et al. 1996).

The high level of diversity, typical of microsatellites, is due to the high mutation rate (between 10^{-3} a 10^{-4} mutation/*locus*/meiosis, Weber and Wong 1993, Heyer et al. 1997), which makes them highly informative to study population genetics and evolution (Goldstein and Feldman 1995b). Besides the use of allele frequency in population genetics, the number of repeats can be also used as a quantitative trait (Goldstein 1995a).

The main mutational process involved in microsatellite evolution is due to slippage in the DNA strands during replication, generating a DNA loop. DNA hairpins that are not corrected by the repair system can result in a gain or loses of usually one repeat, named the Stepwise Mutation Model (SMM) (Schlotterer and Tautz 1992).

The SMM was originally developed by Ohta and Kimura (1973) to characterize electrophoretic variants in a protein study, but recently it has been successfully applied to microsatellites as phylogenetic makers.

In the SMM, alleles can be expressed by integer numbers (..., X_{-1}, X_0, X_{+1},...). If an allele changes its state by mutation, it moves one step in the positive direction or in the negative direction. In the SMM, *v* is the mutation rate and assumes that the mutational changes towards the positive and negative directions occur with the same frequencies.

Linked microsatellites have been used in large scale to study human population relationships (Stumpf and Goldstein 2001). Linked *loci* in the human Y chromosome are important evolutionary tools (Seielstad et al 1999; Peter de Knijff 2000; Rootsi et al 2004), because they are paternally inherited, therefore carry the information about the paternal evolutionary past, complementing the information provided by the mitochondrial DNA (mtDNA) that is inherited maternally. These genetic data plus linguistic, archaeological and paleoanthropological data provide the main evidences to reconstruct human species history (Stumpf and Goldstein 2001).

When investigating human populations' history, Y chromosome microsatellites can be applied, for example, to date demographic and migratory events. They are usually combined with single nucleotide variations that behave as unique event polymorphisms (UEPs). An unique set of alleles at some UEPs can define a haplogroup, a Y chromosome type or lineage that is common to many populations, usually restricted to a particular geographic region (Thomas et al 1998). The phylogeographic analysis of multiple Y microsatellite *loci* in males bearing a particular UEP defined haplogroup can also reveal past migratory events, as well as temporal and geographic relationships among populations (Zerjal et al 1997).

The fast evolution of Y chromosome microsatellites, and other genomic regions with no significant recombination, such as autosomal haplotypes, allow to access the

ancestry information of hundreds to thousands generations ago (Deka et al 1996; Skorecki et al 1997; Torroni et al 1994; Foster et al 1996; de Knijff 2000).

However, microsatellite dynamics in populations is scarcely known, particularly those linked *loci* that evolve as haplotypes. Thus, microsatellite simulation studies can be an important approach to understand the intra-specific evolutionary process. Linked microsatellite simulations can be performed in a realistic way, allowing to investigate dating methods and to evaluate the influence of specific parameters in these estimates. The use of simulations can also help in the interpretation of population divergence from a common antecessor, such as linked microsatellite haplotypes that evolve from a single ancestor haplotype after many generations, since the "birth" of the haplogroup (Santos and Tyler-Smith, 1996). The comprehension of this molecular process can be important to understand, for example, the first pre-columbian migration to America that happened between 24 and 14 thousand years ago (Santos and Tarazona-Santos, 2002; Bortolini et al 2003).

Population movements are some of the main issues in human evolutionary studies where the Y chromosome has been used as a powerful tool to trace human history (Jobling and Tyler-Smith 1995; Hammer et al 1997; Underhill et al 1996; Underhill et al 2000). Because the Y chromosome has an effective population size smaller than that of the autosomes, it evolves faster by drift than all diploid markers. Besides, it tends to show restricted geographic distribution, which allows correlating molecular and population processes (Bing-Su et al 2000). Therefore the distribution of multiple Y chromosome microsatellite haplotypes associated to haplogroups specific of some human populations, can be highly informative to infer past population migrations and demographic events (Bing-Su et al 2000).

We developed the algorithm MGSim (Microsatellite Genealogy Simulator) to simulate evolution of linked microsatellites from a single ancestor, which saves the genealogical record of haplotype populations in a reference database. The MGSim software allowed simulating many genealogies with different number of linked *loci* and mutation rates. Besides, we have also simulated bottlenecks to evaluate the effect of population size reduction in the evolutionary dynamics of microsatellite haplotypes. Analyses of several simulated genealogies reveal interesting aspects of microsatellite evolutionary dynamics that can be used to diagnose possible problems in methods used for dating purposes.

Methodology

Description of the simulator

In this study, we have developed a software (PERL script) to simulate microsatellite haplotype evolution in a non-recombinant context such like the human Y chromosome. All data generated by the software were stored in a MySQL database for further analyses.

The software is executed using some command line parameters to simulate the genealogy. The parameters allowed are: (1) the number of microsatellites analyzed; (2) the current allelic status (number of unit repeats) for each microsatellite in the ancestor; (3) the mutation (4) the number of generations to be simulated and (5) the maximum effective number (N_e) populations will reach.

The algorithm creates a table in the MySQL database containing five columns that keep the information from each individual simulated. The columns keep information about the current record, such like (1) who is its father on paternal lineage, (2) which generation it

belongs, (3) its ID number, (4) its microsatellite haplotype and (5) how many mutations each locus has. The fifth column is relevant to be compared to the forth and identify recurrence of alleles (homoplasy). So, individuals presenting the same microsatellite haplotype may have different number of mutations since their common ancestor and it can be retrieved from database stored information.

The first individual inserted in the database is considered to be the ancestor of all genealogy and receives the ID number 1, having the number zero associated to its father and generation.

In silico genealogy evolution

The algorithm works producing an increasing number of generations as supported by the user in the command line. When the population is smaller than N_e, all the offspring created will go to next generation. However, when the number of offspring population becomes bigger than N_e, the software randomly chooses which individuals will be selected for the next generation.

So, in order to select next generation population, MGSim software makes sorting for each individual associating them with a random fitness number (since their fitness is dissociated to its neutral microsatellite haplotype). After choosing the high fitness individuals that will reproduce, the script draws a number based on the Poisson distribution to be the offspring number of each haplotype. The Poisson distribution permits the simulation to be more real since it doesn't determine how many and which individual will have descendents, leading the process to be more random. Using the Poisson distribution

with mean 2 a great amount of individuals will have descendents between 0 and 3, and the probability of having more than 6 individuals is less than 1%.

A new individual is created copying the haplotype of its father and modifying it if some mutation has occurred (a number is sorted to check if the mutation rate defined by the user has happened). If a mutation occurs, the haplotype of this individual is modified, following the stepwise mutational model (SMM), and stored in the database. The mutations follow the SMM and they can rise or shrink by one the number of repetitions of a given individual allele.

Bottleneck

A similar simulator algorithm was developed to take bottleneck events on account. In this modified algorithm the user is able to enter with the generation he wants the bottleneck to occur, the number of generations the bottleneck will be kept and the population size reduction desired. Other algorithm parameters may be modified in the same ways as the standard algorithm.

Genealogy analysis

In order to analyze the genealogies stored at MySQL database, another PERL script was written. This script analyzes each locus at every hundred generations until the end of the simulation (representing the living population). This script calculates many parameters, such as: (1) the mean allele, (2) modal allele, (3) average squared difference (ASD), (4) tau (τ – generation number), (5) the number of mutations occurred and (6) how many years have been passed since the common ancestral (using the repeat number of each locus). A

powerful point of our analysis is the storage of the real number of mutations that occurred and not only the allelic status of the microsatellite loci (the only data present on real population studies). So, we have used this information to recalculate indices such as: (1) ASD_R, (2) tau_R, (3) putative number of mutations and (4) how many years have been passed from the common ancestral.

An extra PERL script retrieves the descriptive statistic from the analyzed data stored in the database by the former script, producing a text file with the information of the name of the table, effective number used, minimum, maximum, mean, standard deviation, variance, modal and average from the value choice in table.

Evolutionary Scenarios

We have used MGSim to simulate seventeen evolutionary scenarios (Table 1). All the scenarios start with one individual (the putative ancestor) and the population grows until the N_e (effective population size) determined. One of these scenarios keeps a constant N_e size until the last generation (control simulation) and the other sixteen alternative scenarios suffer bottlenecks, represented by reduction in the N_e size of 90 % or 99.9 % in a single generation, or by the 99 % N_e size reduction for 20 generations.

Sample

We have simulated 17 groups and 50 genealogies were built for each one. Each genealogy was simulated with the following parameters: (1) haplotype containing 6

microsatellite loci, (2) mutation rate of 0.0021 per generation (Heyer et al. 1997), (3) 2000 generations to be simulated from ancestor and (4) effective number (N_e) of 5000 individuals (Underhill et al. 1997). In 16 groups, we choose 10 groups to present a bottleneck of 90%, 3 groups were chosen to present a bottleneck of 99% and keep this smaller population size for the 20 following generations and the last 3 groups have been simulated with a single generation bottleneck of 99.9%. Each one of the groups has a bottleneck in some specific generation. The first group was not submitted to any bottleneck effect, working as the control group for comparisons.

The probability that an individual do not present any offspring is 0.135, following the Poisson distribution with mean 2. Considering that our genealogy begins with a single ancestor individual, genealogies where the first individual did not have offspring were dropped and recreated.

Network Constructions

We simulated four new genealogies until the generation 120 to evaluate the effect of bottlenecks on the reconstruction of networks. As we stored only the data at every hundred generations until the end of the simulation we had to simulate these new genealogies to construct the network in all groups in the generation 120. We used the generation 120 since it was the last generation with reduction size in the $90_{+20\,100}$ group and because we couldn't select any ancestor in the generation 200 testing the sample effect.

In the network construction we selected randomly 50 individuals from one specific simulation in the generation 120 and created a file with the selected individuals and the haplotype frequency.

We used the software Network 4.2 (Bandelt H-J *et al.*, 1999) to construct Median Joining networks from simulations.

Testing the Modal and Mean Alleles

We used estimators based on mean and modal alleles to infer the ancestral alleles in the control batch and check which one could be the best estimator of the ancestral microsatellite allele. We tested 300 loci analyzed in each generation (six loci multiplied by 50 simulations).

Results

Comparison between batches

BoxPlot

The comparisons between batches with bottleneck and the control batch were done using the ASD in each generation analyzed through the boxplot with notches. We used the ASD to detect likely problems appearing in dating methods using an inferred ancestral microsatellite haplotype, particularly for populations experiencing size fluctuations and bottlenecks.

The comparisons between batches with bottleneck of 90% in single generation and the control batch showed no significant change in diversity estimates (figure 1). However, the boxplot graph shows an increase in variance in the ASD just after the bottleneck event, indicated by the increase of the boxplot size (figure 1).

The genealogy comparisons between control and bottleneck of 99% indicated that there is a significant difference in variation after the bottleneck event (figure 1), as there is no overlap between the notches of two plots, a 'strong evidence' that the two medians differ significantly (Chambers et al., 1983, p. 62).

The significant difference presented in the 99.9_{100} group disappeared 300 generations later, getting closer to the control batch (data not shown). The same significant difference appeared in the 99.9_{500} (figure 1), but this difference disappeared approximately 600 generations after the bottleneck. The batch 99.9_{1000} showed also a significant difference that persisted until the end of the simulation (figure 1).

The comparison between the control batch and the bottleneck of 99% lasting 20 generations shows a significant difference (figure 1). This difference is bigger than in the batch with bottleneck of 99.9%, as we can see in the boxplots graphs. As expected the difference in diversity disappeared faster in the batch with bottleneck of 99% and constant size in 20 generations. This effect is likely due to the size reduction of the population and genetic drift increasing variance of the ASD.

The ASD in the last simulated generation after the ancestor for 90_{2000}, 99_{2000} and $90_{(+20)\,2000}$ groups has shown highly variation (increasing, decreasing or keeping ASD stable), but the mean in all the batches became almost the same as in the control group. The increase or the decrease of the ASD presented no correlation with the generation where the bottleneck occurred since there are genealogies with increase and decrease of the ASD in

all batches disregarding the generation where bottleneck occurred. The ASD showed great variety; it began very low and became higher along the generations. In the groups with bottleneck the variety became bigger than the control group after each bottleneck event, as shown in the boxplot graph and the calculus of the variation (Table 2).

Analysis of variance between batches – Kruskal-Wallis test

To know whether ASD variation presents a normal distribution or not, we did a normality test (Shapiro Wilk). All groups and all generations the ASD do not follow a normal distribution (data not shown). Thus, to compare the batches of simulated genealogies we used nonparametric tests.

We used the non-parametric Kruskal-Wallis test to do the analysis of variance between batches comparing the ASD.

Comparing the ASD variation we observed that all batches with 90% (N_e=500) bottleneck in any generation did not show significant difference at the 5% level when compared with the control batch, with the exception of one batch with bottleneck in the generation 800. We think that this difference can be due to the stochastic choice of the number of descendents per individual, as well as the choice of remaining individuals after bottleneck that can rarely choose haplotypes very different from the control batch that make this batch more distant from the expected. After these generations the difference between ASD variation became not significant until the end of the simulation. We simulated again the 90_{800} group with 50 new genealogies and we didn't see any significant diference between the 90_{800} and the control group.

In the batches with 99.9% bottleneck (N_e=5), the Kruskal-Wallis test showed significant difference in the ASD mean at the 1% level when compared with the control batch as the boxplot graph indicates. The batches 99.9_{100} showed a significant difference in the ASD compared with control but this difference disappeared after 300 generations. The batch 99.9_{500} showed significant difference at the 1% level compared with the control batch and this difference disappeared after approximately 600 generations. The batch 99.9_{1000} showed a significant difference and this difference continued until the end of the simulation.

Comparing the control batch and batches with bottleneck of 99% and a constant size reduction (N_e=50) in the next 20 generations we found a significant difference. This difference was probably caused by founder effect and genetic drift that produces a differentiation among the genealogies causing an increase in the ASD. The difference appeared in all the batches and disappeared after 300 and 400 generations for the batch with bottleneck in the generation 100 the batches with bottleneck in the generations 500 and 1,000, respectively.

Testing the mean and modal allele estimators

Testing the mean and modal allele to check which one could be the best estimator we got the results showed in figure 2.

Analyzing 300 loci in the control batch, we got 222 simulated loci for the mean estimative against 173 loci for the modal right estimated in the generation 1,000 (figure 2 a), and 141 against 102, respectively, in the generation 2,000 (figure 2 b). We observed that

the use of the mean allele is the estimator that can get closer to the real repeat number present in the ancestral microsatellite locus. Using the mean we could retrieve the correct ancestral allele more frequently, as well as it presented lower variance of estimates. For example, the mean estimative has never shown alleles -5, -4, -3 and 4 that were found in the modal estimates, obtained after generation 2,000.

Reconstruction of the ancestral alleles after the bottleneck

We reconstructed the ancestral alleles and haplotypes in the simulated batches with bottleneck and compared with the reconstruction in the control batch to verify the possible problems with this methodology after population size reductions. Our hypothesis is that size reduction could lead to a wrong dating based on divergence time, calculated from an inferred ancestral haplotype (Zerjal et al. 1997).

The bottleneck effect causes a reduction in the proportion of correct reconstructions of ancestral alleles and haplotype (figures 3). When the bottleneck happens in the beginning ($99_{(+20)\,100}$), the retrieval of the correct ancestral allele is reduced by about 10% in the generation 200. However when the reduction happens in $99_{(+20)\,500}$ group we had a loss of 5% in the generation 500 and 37% in the generation 600 compared with the control group. The correct allele reconstruction when the reduction happens in the $99_{(+20)\,1000}$ is very low, approximately 30%, not very different from the control group (40%) (figure 3 a).

Comparing the correct reconstruction of the haplotype when the bottleneck happens in the beginning ($99_{(+20)\,100}$) we have a loss of 60% in the generation 200. When the reduction happens in the $99_{(+20)\,500}$ group there is a loss of 30%. Besides this in the

generation 600 after the reduction we have 0% of correct reconstruction of the ancestral haplotype. In the batch 99 $_{(+20)\ 1000}$ we saw that there is no difference comparing it with the control batch, and the correct reconstruction of the ancestral haplotype is impossible (0%) (figure 3 b).

Construction of the networks

In the control batch network it is possible to observe the ancestral haplotype as the most frequent in the generation 120. The network presents a star-like structure where derived haplotypes pop up from the ancestral haplotype, but we can also observe some points of reticulations or homoplasy (figure 4 a).

The simulation with bottleneck of 90 $_{100}$ also displays the ancestral haplotype as the most frequent one and the other haplotypes coming from it. However the number of homoplasies is higher than in the control simulation. Almost all the haplotypes are linked by one or more homoplasies. Despite the homoplasies, it is still possible to see the haplotypes deriving from the ancestral one by few steps, as expected in an expanding population (figure 4 b).

In the network with bottleneck of 99.9% we also observed the ancestral haplotype, although it is not the most frequent one. Besides, the number of haplotypes decreases very much when compared with the control network. The frequency of the haplotypes changed drastically, allowing us to infer a different ancestral haplotype or multiple ones. As we can see on the network, the haplotype H_8 would be considered the ancestral one together with H_5, because of the frequency and the star structure suggesting that they gave origin to other haplotypes (figure 4 c).

Observing the network from the simulation with bottleneck of 99 $_{(+20)\ 100}$ we can see a stronger effect in the haplotype diversity due to stochastic events, as population did not grow until generation 120. The long term reduction in population size (20 generations at N_e=50) changes drastically the haplotype network, with no homoplasy observed and the real ancestral haplotype presenting a very low frequency (figure 4d).

Discussion

The analyses of the ASD showed that bottlenecks increase the ASD variance among simulations. This variance was expected since the bottleneck can select randomly any individual to continue in the simulation (individuals that will have descendents). When the bottleneck selected a good representation of the population, the values were kept almost invariable; however when the bottleneck selected individuals with atypical haplotypes, i.e. carrying more or less mutations than the average, the ASD increases or decreases respectively. This selection of the individuals that are quite different from the mean, observed in the population without bottleneck, causes the increase of the variance in ASD.

The ASD showed a significant difference between the control batch and the batches with bottleneck of 99.9% and 99% of size reduction for 20 generations. However this difference does not cause a bias in the ASD, i.e. it may increase or decrease the values as shown in the boxplot graphs and Kruskal-Wallis analyses. However, after some generations the significant difference between bottleneck and control batches disappeared. The batch 99.9 $_{100}$ recovers faster (taking about 300 generations) than later bottleneck batches. For the batch 99.9 $_{500}$ approximately 600 generationsare needed to recover this distribution and for

the batch 99.9 $_{1000}$ the difference is maintained until the end of the simulation, but it would be likely to be recovered later in this simulation. This is the lag time after a bottleneck event that a population takes to recover diversity due to the appearance of new haplotypes, where lost alleles reappear by new mutations making the significant difference disappear. Probably, the batch with bottleneck in the generation 100 was faster to recover because in the beginning there were few alleles to be lost, but in the generations 500 and 1,000, many alleles were lost, taking longer to recover diversity again. The batch with bottleneck of 99% reduction in the population size for 20 generations suffered the effect of drift for a longer time, causing a large loss of haplotypes and increasing the variance of the ASD as seen between the simulated genealogies (data not shown).

The batch with bottleneck of 90% in the generation 800 was the only batch to show a significant difference after the bottleneck. We believe that this difference was a rare event derived from the random process of simulations. We repeated the simulation of bottleneck in another batch, and the significant difference disappeared.

Although the Kruskal-Wallis test did not show any significant difference between the control batch and groups with bottleneck of 90%, the observed differences can represent dramatic time estimates. After the bottleneck event the difference between the highest values of the control batch and in the batch with bottleneck can be of more than 1 point in the ASD and this represents 476 generations using the mutational rate of 0.0021. This number of generations for the human species represents more than 10,000 years, causing many troubles particularly in the reconstruction of the recent human history.

The mean and modal allele methods previously used to infer ancestral alleles and haplotypes (Zerjal et al. 1997) were analyzed in several simulations. The results indicated that the use of the mean was the best to retrieve the correct ancestral allele in 300

simulations. The use of mean allele was more efficient, particularly to recover the ancestral alleles after 1000 generations, when compared to the modal allele method.

The inference of ancestral alleles and haplotypes can be drastically affected by a bottleneck event. However, as expected, the reconstruction of the ancestral haplotype is more difficult than the reconstruction of ancestral alleles. For the haplotype reconstruction the correct inference is obtained in about 60% of genealogies in the generation 200 and only 30% in generation 500. These results show that bottlenecks can easily lead to a wrong dating when based on the inference of the ancestral microsatellite haplotype. Our simulations included a single bottleneck; however for real populations this problem of inference of the ancestral alleles and haplotypes can be even bigger because they may be subjected to population size fluctuations and strong bottlenecks.

The construction of networks from simulated haplotype populations was an important tool to verify the random effects of bottleneck and genetic drift, as well as to observe changes in the topology of the network and ancestral haplotype frequency changes. The topology of the network can go from a state full of homoplasies to a network with reduced homoplasies and haplotypes. Besides, the frequency of the ancestral haplotype may change (decrease) drastically, and the ancestral haplotype inference based on frequency or the phylogeographic structure can be full of flaws.

These problems should be considered in the studies including dating of human migrations with Y chromosome microsatellites. For example, previous studies on the Peopling of Americas (Bianchi et al. 1997; Bortolini et al. 2003) used Y microsatellites to investigate the entry time of the first populations between 500-1,000 generations ago (around 12,500 and 25,000 years ago). According to several independent data (Pena et al. 1995, Santos et al. 1995, Tarazona-Santos and Santos, 2002; Bonatto and Salzano, 1997)

the first peoples to occupy the Americas suffered a bottleneck followed by a population expansion, whose subpopulations were divided into small groups occupying all over the Americas. The use of Y linked microsatellites assuming inferred ancestral haplotypes can be misleading giving us a wrong idea about the time when the migration wave really occurred. Because the ancestral haplotype was usually inferred based on its frequency (Bianchi et al. 1997, Bortolini et al. 2003), we would expect that the inferred time would be usually underestimated. Although independent methods of assuming an ancestral haplotype can be used, and have been also used in the Peopling of Americas, our simulation indicates that diversity reaches saturation soon in genealogy, and it is recovered again soon after population size reductions.

The development of the software MGSim and its use in the simulation of evolving populations with and without bottleneck was important to help us to understand some details of the evolution of linked microsatellites, extensively used to reconstruct the human male past. Furthermore, MGSim can make genealogy simulations with several different parameters as mutation rates, effective population size, number of loci and generations. Thus, new evolutionary simulations of linked microsatellites can be run to test not only dating methods but also the effects of homoplasies and diversity saturation. Other changes in the software can be further introduced such as inclusion of population subdivision to test effects in phylogeographic reconstructions, or the addition of heterogeneity of mutation rates and allele size constraints among loci, a main problem related to the use of a restricted set of Y chromosome microsatellites (Carvalho-Silva et al. 1999).

References

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48. http://fluxus-engineering.com

Bianchi, N. O.; Catanesi, C. I.; Bailliet, G.; Martinez-Marignac, V. L.; Bravi, C. M.; Vidal-Rioja, B. L.; Herrera, R. J.; and Lopez-Camelo, J. S. (1998). Characterization of Ancestral and Derived Y-Chromosome Haplotypes of New World Native Populations. Am. J. Hum. Genet. 63:1862–1871,

Bing Su; Chunjie Xiao; Ranjan Deka •Mark T. Seielstad; Daoroong Kangwanpong • Junhua Xiao; Daru Lu; Peter Underhill •Luca Cavalli-Sforza; Ranajit Chakraborty; Li Jin (2000).Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet 107:582–590

Bortolini M C, Salzano F M, Thomas M G, Stuart S, Nasanen S P, Bau C H, Hutz M H, Layrisse Z, Petzl-Erler M L, Tsuneto L T, Hill K, Hurtado A M, Castro-de-Guerra D, Torres M M, Groot H, Michalski R, Nymadawa P, Bedoya G, Bradman N, Labuda D, Ruiz-Linares A. (2003). Ychromosome evidence for differing ancient demographic histories in the Americas. Am J Hum Genet 73:524-539

Bonatto SL and Salzano FM (1997). A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci U S A 94:1866-1871

Carvalho-Silva DR, Santos FR, Hutz MH, Salzano FM and Pena S D (1999). Divergent human Y-chromosome microsatellite evolution rates J Mol Evol 49:204-214

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). Graphical Methods for Data Analysis. Wadsworth & Brooks/Cole. pp 62.

Deka, R., L. Jin, M. D. Shriver, L. Mei Yu, And N. Saha et al. (1996).Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. Genome Res. 6:1177-1184.

Forster, P., R. Harding, A. Torroni, And H.-J. Bandelt (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. 59:935-945.

Goldstein, D. B., Ruiz-Linares, A., Cavalli-Sforzaf, L. L. E Feldman, M. W (1995a). An Evaluation of Genetic Distances for Use With Microsatellite Loci. Genetics 139:463–471.

Goldstein, D. B., Ruiz-Linares, A., Cavalli-Sforzaf, L. L. E Feldman, M. W (1995b). Genetic Absolute Dating Based on microsatellite and the origin of modern humans. Proceedings of the National Academy of Sciences 92:6723-6727.

Heyer, E, Puymirat, J, Dieltjes, P, Bakker, E, De Knijff P (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799-803.

Hammer, M. F., A. B. Spurdle, T. Karafet et al. (11 coauthors) (1997). The geographic distribution of human Y chromosome variation. Genetics 145:787–805.

Jobling, M. A. and Tyler-Smith, C. (1995). Fathers and sons:the Y chromosome and human evolution. TIG November 1995 11:449-456.

Lehmann, T., Hawley, W. A., Collins, F. H. (1996.) An Evaluation of Evolutionary Constrains on Microsatellite Loci using Null Alleles. Genetics. 144:1155-1163.

Ohta, T., E Kimura, M (1973). The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. 22:201–204..

Pena SDJ, Santos FR, Bianchi NO, Bravi CM, Carnese FR, Rothhammer F, Gerelsaikhan T, Munkhtuja B, Oyunsuren T (1995). A major founder Y-chromosome haplotype in Amerindians. Nature Genetics 11:15-16

Peter De Knijff (2000). Messages through Bottlenecks: On the Combined Use of Slow and Fast Evolving Polymorphic Markers on the Human Y Chromosome. Am. J. Hum. Genet. 67:1055–1061.

Pritchard, J. K. E Feldman, M. W (1996). Statistics for microsatellite variation based on coalescence. Theor. Pop. Biol. 50:325–344.

Rootsi S, Semino O (42 coauthors) (2004). Phylogoegraphy of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe

Ruiz-Linares A, David B (42 coauthors) (1999). Goldstein. Microsatellites Provide Evidence For Y Chromosome Diversity Among The Founders Of The New World. Proc. Natl. Acad. Sci. 96:6312–6317

Santos F R, Hutz M H, Coimbra-Jr C E A, Santos R V, Salzano F M, Pena S D J (1995) Further evidence for the existence of major founder Y chromosome haplotype in Amerindians. Braz J Genet 18:669-672

Santos FR and Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz J Genet 19:665-670

Schlotterer, C. E Tautz, D. (1992). Slippage synthesis of simple sequence DNA. Nucleic Acids Research 20:2211-215.

Seielstad, M. T., Bekele, E., Ibrahim, M., Touré, A., E Traoré, M. (1999). A view of modern human origins from Y chromosome microsatellite

variation. Genome Research, 9:558-567.

Skorecki, K., S. Selig, S. Blazer, R. Bradman, And N. Bradman (1997). Y chromosomes of Jewish priests. Nature 385:32.

Stumpf, M. P. H. E Goldstein, D. B. (2001) Genealogical and Evolutionary Inference with the Human Y Chromosome. Science 291:1738-1742.

Tarazona-Santos and Santos FR (2002) The peopling of the Americas: a second major migration? Am J Hum Genet 70:1377-1380

Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In DNA fingerprinting: State of science (eds. S.D.J. Pena, R. Chakraborty, J.T. Epplen, and A.J. Jeffreys). 21-28. Birkhäuser Verlag, Basel, Switzerland.

Thomas, M. G., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N., Goldstein, D. B. (1998). Origins of Old Testament priests. Nature 394:138-140.

Torroni, A., J. V. Neel, R. Barrantes, T. G. Schurr, And D. C. Wallace (1994). Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc. Natl. Acad. Sci. USA 91:1158-1162.

Weber, J. L. E Wong, C. (1993) Mutation Of Human Short Tandem Repeats. Human Molecular Genetics. 2:1123-1128.

Underhill PA, Jin L, Zemans R, Oefner PJ & and Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA 93:196-200

Underhill, P.A. et al. (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high performance liquid chromatography (DHPLC). Genome Res. 7:996–1005

Peter A. Underhill, Peidong Shen, Alice A. Lin, Li Jin, Giuseppe Passarino, Wei H. Yang, Erin Kauffman, Batsheva Bonné-Tamir, Jaume Bertranpetit, Paolo Francalacci, Muntaser Ibrahim, Trefor Jenkins, Judith R. Kidd, S. Qasim Mehdi0, Mark T. Seielstad, R. Spencer Wells, Alberto Piazza, Ronald W. Davis, Marcus W. Feldman, L. Luca Cavalli-Sforza &

Peter. J. Oefner (2000). Y chromosome sequence variation and the history of human populations. Nature Genetics 26:358-361.

Zerjal, T; Dashnyam, B; Pandya, A; Kayser, M; Roewer, L; Santos, F R; Schiefenhovel, W; Fretwell, N; Jobling, M A; Harihara, S; Shimizu, K; Semjidmaa, D; Sajantila, A; Salo, P; Crawford, MH; Ginter, E K; Evgrafov, O V; Tyler-Smith, C (1997). Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183

Tables

Table 01 – Evolutionary Scenarios

Number	Bottleneck Simulation	Bottleneck %	Bottleneck Generation	Generation Span	Name/Code
1	No	0	0	0	Control
2	Yes	90	100	1	90_{100}
3	Yes	90	200	1	90_{200}
4	Yes	90	300	1	90_{300}
5	Yes	90	400	1	90_{400}
6	Yes	90	500	1	90_{500}
7	Yes	90	600	1	90_{600}
8	Yes	90	700	1	90_{700}
9	Yes	90	800	1	90_{800}
10	Yes	90	900	1	90_{900}
11	Yes	90	1000	1	90_{1000}
12	Yes	99.9	100	1	99.9_{100}
13	Yes	99.9	500	1	99.9_{500}
14	Yes	99.9	1000	1	99.9_{1000}
15	Yes	99	100	20	$99_{100\ (+20)}$
16	Yes	99	500	20	$99_{500\ (+20)}$
17	Yes	99	600	20	$99_{1000\ (+20)}$

Table 02 – Variance in the ASD along the genealogy.

Generation	Control	90_{100}	90_{500}	90_{1000}
100	0.01	0.01	0.00	0.00
200	0.01	0.03	0.01	0.02
300	0.04	0.04	0.03	0.03
400	0.08	0.07	0.07	0.06
500	0.11	0.12	0.20	0.13
600	0.20	0.20	0.36	0.21
700	0.32	0.37	0.39	0.32
800	0.41	0.42	0.45	0.45
900	0.56	0.59	0.53	0.61
1000	0.72	0.65	0.66	0.83
1100	0.92	1.04	0.77	1.15
1200	1.16	1.19	1.07	1.42
1300	1.27	1.54	1.29	1.59
1400	1.62	1.83	1.76	2.08
1500	1.92	1.86	2.25	2.21
1600	2.33	2.60	2.62	2.40
1700	2.78	2.81	3.16	2.46
1800	2.81	3.58	3.79	2.91
1900	3.36	4.14	4.75	3.67
2000	3.83	5.29	4.95	4.19

Generation	Control	99.9 $_{100}$	99.9 $_{500}$	99.9 $_{1000}$
100	0.01	0.04	0.01	0.00
200	0.01	0.12	0.02	0.01
300	0.04	0.15	0.04	0.04
400	0.08	0.20	0.08	0.07
500	0.11	0.23	0.66	0.13
600	0.20	0.36	1.06	0.22
700	0.32	0.48	1.14	0.32
800	0.41	0.53	1.16	0.50
900	0.56	0.87	1.14	0.64
1000	0.72	1.33	1.47	2.59
1100	0.92	1.45	1.65	4.39
1200	1.16	1.83	1.88	4.52
1300	1.27	2.60	2.19	4.43
1400	1.62	3.10	2.58	4.80
1500	1.92	4.24	2.92	5.00
1600	2.33	5.02	3.50	4.72
1700	2.78	5.60	3.34	4.97
1800	2.81	5.17	3.39	5.04
1900	3.36	5.48	3.46	5.16
2000	3.83	5.91	4.44	5.02

Generation	Control	99% $_{100\ (+\ 20)}$	99% $_{500\ (+\ 20)}$	99% $_{1000\ (+\ 20)}$
100	0.01	0.01	0.01	0.00
200	0.01	0.13	0.02	0.01
300	0.04	0.17	0.03	0.04
400	0.08	0.19	0.05	0.06
500	0.11	0.26	0.17	0.11
600	0.20	0.30	1.65	0.18
700	0.32	0.38	1.85	0.25
800	0.41	0.47	1.81	0.37
900	0.56	0.52	1.93	0.52
1000	0.72	0.71	2.46	0.93
1100	0.92	0.99	2.38	4.61
1200	1.16	1.15	2.54	4.94
1300	1.27	1.49	2.86	4.71
1400	1.62	2.14	3.14	5.35
1500	1.92	2.88	3.79	6.22
1600	2.33	3.79	4.24	7.00
1700	2.78	3.95	4.92	7.30
1800	2.81	4.31	5.94	8.16
1900	3.36	5.33	6.82	8.64
2000	3.83	5.68	8.08	9.55

Figure legends

Figure 1. Boxplot of ASD distribution. 1 - control $_{100}$; 2 - control $_{200}$; 3 - control $_{300}$; 4 - control $_{400}$; 5 - control $_{500}$; 6 - control $_{600}$; 7 - control $_{700}$; 8 - control $_{800}$; 9 - control $_{900}$; 10 - control $_{1000}$; 11 - control $_{1100}$; 12 - 90 $_{100}$; 13 - 90 $_{200}$; 14 - 90 $_{300}$; 15 - 90 $_{400}$; 16 - 90 $_{500}$; 17 - 90 $_{600}$; 18 - 90 $_{700}$; 19 - 90 $_{800}$; 20 - 90 $_{900}$; 21 - 90 $_{1000}$; 22 - 99 $_{(+20)\ 200}$; 23 - 99 $_{(+20)\ 600}$; 24 - 99 $_{(+20)\ 1100}$; 25 - 99.9 $_{100}$; 26 - 99.9 $_{500}$; 27 - 99.9 $_{1000}$. The boxplot always shows a increase in the variation of the ASD after a bottleneck.

Figure 2. (a) Distribution of the estimated ancestral alleles calculated with the modal and mean estimator in generation 1,000. The mean estimator found correctly 222 loci against 173 loci for the modal estimator using of 300 loci. (b) Distribution of the estimated ancestral alleles calculated with the modal and mean estimator in generation 2,000. The mean estimator found correctly 141 loci against 102 loci for the modal estimator using of 300 loci. Theses results indicate that the mean estimator is the best.

Figure 3. (a) Reconstruction of the ancestral allele (mean allele). Black line - control batch; red line - batch 99 $1_{00\ (+20),}$ green line - batch 99 $_{500\ (+20)}$; blue line - batch 99 $_{1000\ (+20)}$. (b) Reconstruction of the ancestral haplotype (mean allele). Black line - control batch; red line - batch 99 $1_{00\ (+20),}$ green line - batch 99 $_{500\ (+20)}$; blue line - batch 99 $_{1000\ (+20)}$. After a bottleneck we noticed a reduction in the correct estimative of the ancestral allele and haplotype.

Figure 4. (a) Network using the control simulation, (b) Network using the simulation 90_{100}, (c) Network using the simulation with 99.9_{100}, (d) Network using the simulation $99_{100\,(+20)}$. In all simulations we chose 50 random individuals from the generation 120. The ancestral haplotype is the blue circle. When a bottleneck occurs there is a reduction of the ancestral frequency and this effect can change the ancestral inference.

Figure 2 A

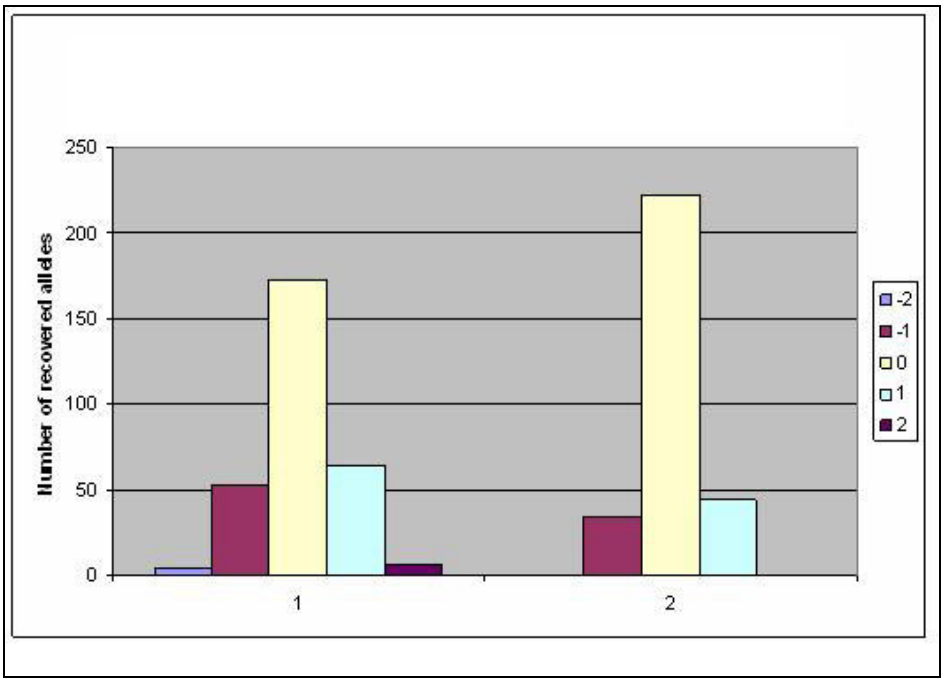


Figure 2 B

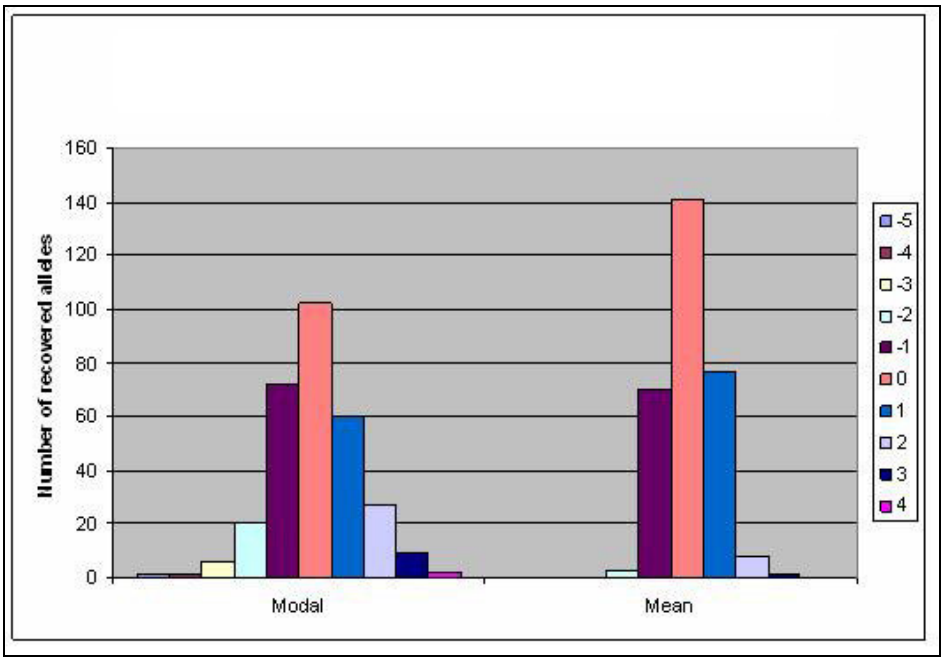


Figure 3A

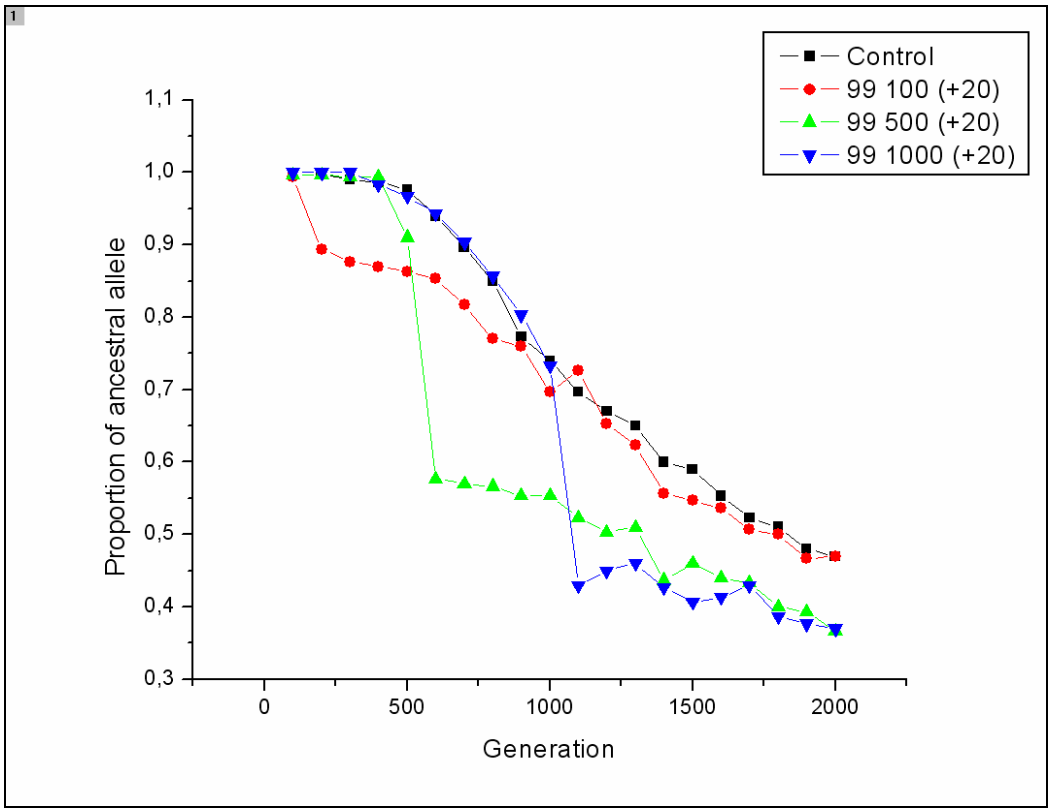
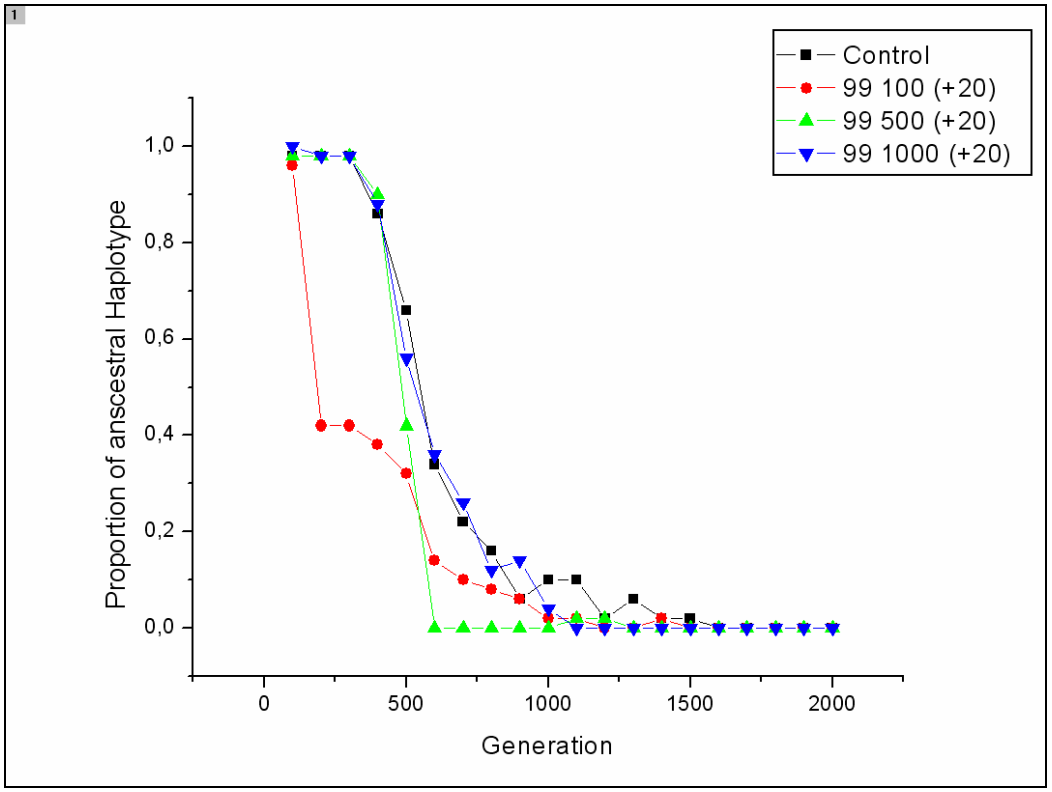


Figure 3 B



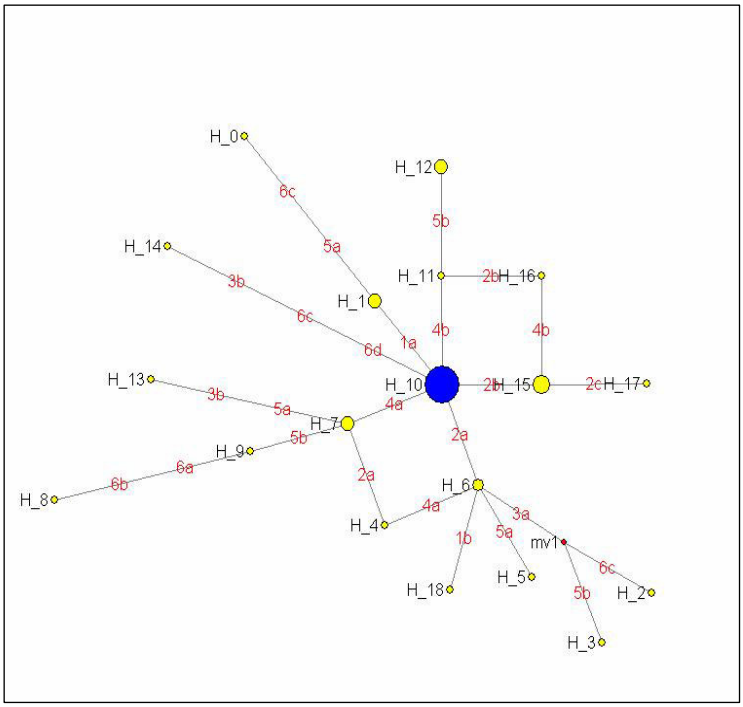


Figure 4 A

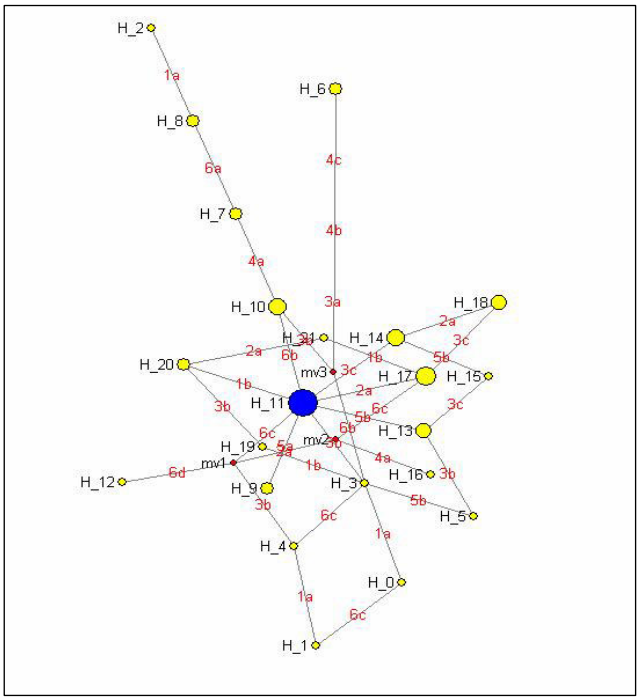


Figure 4 B

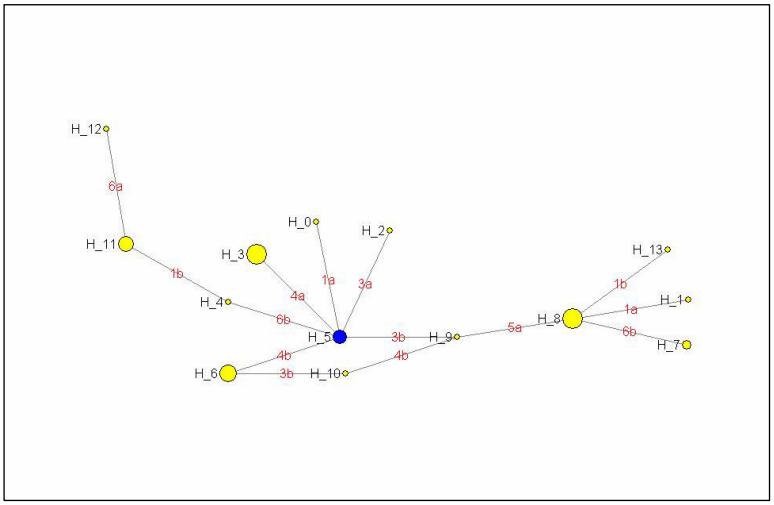


Figure 4 C

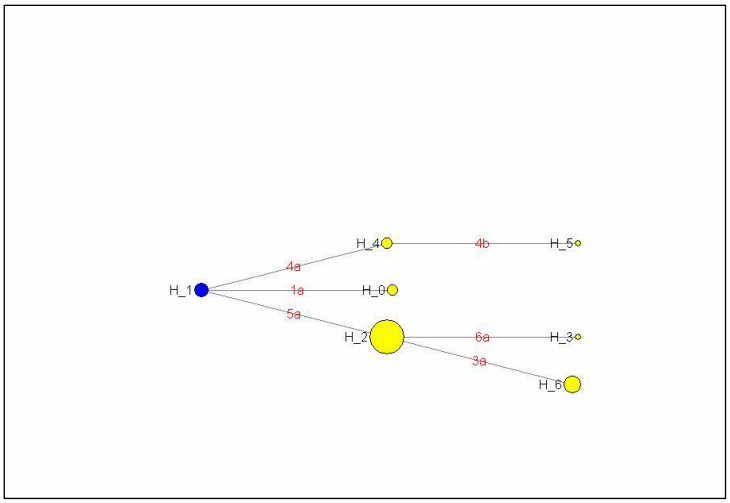


Figure 4 D

5. Referências

BOWCOCK, A. M., RUIZ-LINARES, A., TOMFOHRDE, J., MINCH, F., KIDD, J. R. *et al.* High resolution of human evolutionary trees with polimorphic microsatellites. *Nature* **368**:455-457. 1994.

BALLOUX, F. EASYPOP (Version 1.7): A Computer Program for Population Genetics Simulations. The american Genetic Association. 301-302. 2001

BALLOUX, F. e GOUDET, J.. Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Molecular Ecology 11, 771–783. 2002

CALAFELL, F. *et al*. Haplotype evolution and linkage disequilibrium: a simulation study. Hum. Hered., 51, 85–96.2001

DEKA, R., L. JIN, M. D. SHRIVER, L. MEI YU, AND N. SAHA *et al.* Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. *Genome Res*. **6**:1177-1184. 1996

FEARNHEAD, P. Ancestral processes for non-neutral models of complex diseases. Theoretical Population Biology 63 115–130. 2003

FORSTER, P., R. HARDING, A. TORRONI, and H.-J. BANDELT. Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59**:935-945. 1996.

GOLDSTEIN, D. B., RUIZ-LINARES, A., CAVALLI-SFORZAF, L. L. e FELDMAN, M. W.. An Evaluation of Genetic Distances for Use With Microsatellite Loci. *Genetics* **139**: 46.51171. 1995a.

GOLDSTEIN, D. B., RUIZ-LINARES, A., CAVALLI-SFORZAF, L. L. e FELDMAN, M. W. Genetic Absolute Dating Based on microsatellite and the origin of modern humans. *Proceedings of the National Academy of Sciences* **92**, 6723-6727. 1995b.

HAMMER, M. F. A recent common ancestry for human Y chromosomes. *Nature*. **378**:376-378. 1995

Hey, 2004, http://lifesci.rutgers.edu/heylab/HeylabSoftware.htm;

HEYER, E, PUYMIRAT, J, DIELTJES, P, BAKKER, E, DE KNIJFF P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799-803. 1997.

KINGMAN, J. F. C. The coalescent Stochastic Process Appl. **13**:235-248. 1982b

LEHMANN, T., HAWLEY, W. A., COLLINS, F. H. An Evaluation of Evolutionary Constrains on Microsatellite Loci using Null Alleles. *Genetics*. **144**:1155-1163. 1996.

OHTA, T., e KIMURA, M. The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. *Genet. Res.* **22**:201–204. 1973.

PENG, B. e KIMMEL, M. simuPOP: a forward-time population genetics simulation environment. Bioinformatics. Vol. 21 no. 18 3686–3687. 2005

PETER DE KNIJFF. Messages through Bottlenecks: On the Combined Use of Slow and Fast Evolving Polymorphic Markers on the Human Y Chromosome. *Am. J. Hum. Genet.* **67**:1055–1061. 2000.

PRITCHARD, J. K. e FELDMAN, M. W. Statistics for microsatellite variation based on coalescence. *Theor. Pop. Biol.*, **50**:325–344. 1996.

ROOTSI, S. *et al.* Phylogoegraphy of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe. Am J Hum Genet. 75(1): 128–137. 2004.

SCHLOTTERER, C. e TAUTZ, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* 20:2211-215. 1992.

SEIELSTAD, M. T., BEKELE, E., IBRAHIM, M., TOURÉ, A., e TRAORÉ, M. A. view of modern human origins from Y chromosome microsatellite variation. *Genome Research,* **9**:558-567. 1999.

SKORECKI, K., S. SELIG, S. BLAZER, R. BRADMAN, and N. BRADMAN *et al.* Y chromosomes of Jewish priests. *Nature* **385**:32. 1997.

STUMPF, M. P. H. e GOLDSTEIN, D. B. Genealogical and Evolutionary Inference with the Human Y Chromosome. *Science* 291: 1738-1742. 2001.

TARAZONA-SANTOS e SANTOS F.R. The peopling of the Americas: a second major migration? Am J Hum Genet 70:1377-1380. 2002

TAUTZ, D. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In *DNA fingerprinting: State of science* (eds. S.D.J. Pena, R. Chakraborty, J.T. Epplen, and A.J. Jeffreys). 21-28. Birkhäuser Verlag, Basel, Switzerland. 1993.

THOMAS, M. G., SKORECKI, K., BEN-AMI, H., PARFITT, T., BRADMAN, N., GOLDSTEIN, D. B. Origins of Old Testament priests. *Nature* **394**:138-140. 1998.

TORRONI, A., J. V. NEEL, R. BARRANTES, T. G. SCHURR, and D. C. WALLACE. Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. *Proc. Natl. Acad. Sci.* USA **91**:1158-1162. 1994.

VILLALOBOS, H.R. La Historia Del Hombre Descrita Por El Cromosoma “Y”. Universidad Autónoma del Estado de Hidalgo, Hidalgo, México. 2006. pp. 29-76. 2006

WEBER, J. L. e WONG, C. MUTATION OF HUMAN SHORT TANDEM REPEATS. *Human Molecular Genetics.* **2**: 1123-1128v. 1993.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo