DATING POPULATION EXPANSION BASED ON STR VARIATION WITHIN Y-CHROMOSOME SNP-HAPLOGROUPS

Zhivotovsky L.A.

N.I. Vavilov Institute of General Genetics RAS, Moscow, Russia; e-mail: levzh@hotmail.com

Keywords: human evolution, DNA markers, STRs, SNPs, dating, mathematical models, statistical methods

Summary

Motivation: Dating divergence within- and between DNA-lineages is of great importance for revealing historic information from observed DNA variation in human populations. The increasing amount of data on Y-chromosome SNPs and STRs require adequate methods for estimation of population expansion time. Such methods should necessarily involve mathematical models that describe evolution of DNA variation, and statistical methods to estimate the dates of population events.

Results: Dynamic models of population divergence under mutation, genetic drift and migration are analyzed. Estimates of the dynamic parameters are introduced and applied to several data sets on worldwide human populations.

Introduction

Increasing attention has recently been paid to microsatellite variation within Y-chromosome haplogroups defined by binary polymorphisms, such as SNPs (single nucleotide polymorphisms), or, as a general term, UEPs – unique event polymorphisms (Underhill et al., 1996; Zerjal et al., 1997; Kayser et al., 2000a; de Knijff, 2000), many of which are specific to populations related through their recent or past history (Underhill et al., 2000; Hammer et al., 2001; Y Chromosome Consortium, 2002). Microsatellites, or short tandem repeat (STR) polymorphisms, are abundant in the human genome, can be easily genotyped and scored, and thus have become a useful tool for the elucidation of human population history and for forensic purposes. The mutation rate at Y-chromosome STRs is important for calibration of the molecular clock in evolutionary studies and has been estimated (Heyer et al., 1997; Forster et al., 2000; Zhivotovsky et al., 2004). Similar linked SNP and STR haplotypes are also available in autosomes (Mountain et al., 2002). How this information can be used for dating historic population events is illustrated in this paper.

Dynamic models

Variation at STR loci is amenable to the use of population statistics that treat allele repeat scores (the number of repeats) as a quantitative trait (Goldstein et al., 1995a, b; Slatkin, 1995; Zhivotovsky, Feldman, 1995). Among them is a specific measure of genetic distance, the squared difference in the number of repeats, which is linear with time since divergence (Goldstein et al., 1995b; Zhivotovsky, Feldman, 1995). However, this property is fulfilled only if populations under divergence are at genetic equilibrium, have constant equal sizes and are not subject to gene flows; therefore a different, though related genetic distance has been suggested (Zhivotovsky, 2001) that is relatively robust to these demographic processes. We model these processes and a modified distance to apply it for Y chromosome STRs within lineages defined by SNPs.

Statistical method of dating

The age of STR variation within each haplogroup can estimated as the average squared difference ($ASD_0$) in the number of repeats between all current chromosomes of a sample and the founder
(assumed to be modal) haplotype divided by $w = 6.9 \times 10^{-4}$ per 25 years (Zhivotovsky et al., 2004). The upper bound for expansion time (the time of divergence of populations) is suggested to be calculated using $T_o'$ (Zhivotovsky, 2001) and assuming an STR-variance in repeat scores at the beginning of population separation ($V_o'$) equal to zero. The lower bound is calculated as $T_o$, with $V_0$ taken as a predicted value of the within-population STR-variance prior to population split; the latter was computed as a linear approximation of the within-population variance in repeat scores as a function of time. We investigate how these estimates evolve using the model just described above, and evaluate how informative they are under uncertainties in demographic parameters.

**Population data, Results and Discussion**

For illustration of our approach we use data on populations with various histories. Among those are Polynesian populations (Maoris, Cook Islanders, and Samoans) whose Y-chromosome lineages reflect Polynesian origin in Melanesia and eastern/southeastern Asia, in particular lineage C2 characterized by mutations at RPS4Y711 and M38 marked additionally with the mutation M208 (Su et al., 2000; Kayser et al., 2000a; Underhill et al., 2001). The Gypsy populations from Bulgaria were analyzed with a Y chromosome lineage defined by mutation M82, which is derived from the Indian subcontinent and is exceedingly rare in Europe (Semino et al., 2000; Underhill et al., 2000; Gresham et al., 2001). The Bantu expansion was investigated using the E3a7-M191 haplogroup, which occurs at high frequency in the Bantu populations, with traces in other, non-Bantu-speaking groups from sub-Saharan Africa (Cruciani et al., 2002). Haplogroups E and J was investigated in samples from Europe and the Mediterranean and also from Africa and Asia, which are distributed differentially within the Near East, North Africa and Europe probably associating to the diffusion of Arab people or reflecting the spread of Anatolian farmers, or tracing the subsequent diffusion of people from the southern Balkans to the West (Semino et al., 2004). Around 20 Native American, 28 Asian, and 5 European populations (including 342 Amerind speakers, 186 Na-Dene speakers, and 60 Aleut-Eskimo speakers) were used to investigate the origins of Native American paternal lineages based on SNP analysis of three major haplogroups C, Q, and R, accounted for almost all Native American Y chromosomes (Zegura et al., 2004).

The data illustrate how the statistical procedure works and how the estimates correspond to available historic and archaeological records.

**Acknowledgements**

This work was partially supported by RAS programmes “Genetic aspects of the Biosphere evolution” and “Genofond dynamics” and grants RFBR 04-04-48639 and NIH R03 TW005540.

**References**


