ELSEVIER

Theory and Methodology for Utilizing Genes as Biomarkers to Determine Potential Biological Mixtures

SADEEP SHRESTHA, PHD, MICHAEL W. SMITH, PHD, TERRI H. BEATY, PHD, AND STEFFANIE A. STRATHDEE, PHD

PURPOSE: Genetically determined mixture information can be used as a surrogate for physical or behavioral characteristics in epidemiological studies examining research questions related to socially stigmatized behaviors and horizontally transmitted infections. A new measure, the probability of mixture discrimination (P_{MD}), was developed to aid mixture analysis that estimates the ability to differentiate single from multiple genomes in biological mixtures.
METHODS: Four autosomal short tandem repeats (STRs) were identified, genotyped and evaluated in African American, European American, Hispanic, and Chinese individuals to estimate P_{MD}. Theoretical P_{MD} frameworks were also developed for autosomal and sex-linked (X and Y) STR markers in potential male/male, male/female and female/female mixtures.
RESULTS: Autosomal STRs genetically determine the presence of multiple genomes in mixture samples of unknown genders with more power than the apparently simpler X and Y chromosome STRs. Evaluation of four autosomal STR loci enables the detection of mixtures of DNA from multiple sources with above 99% probability in all four racial/ethnic populations.
CONCLUSIONS: The genetic-based approach has applications in epidemiology that provide viable alternatives to survey-based study designs. The analysis of genes as biomarkers can be used as a gold standard for validating measurements from self-reported behaviors that tend to be sensitive or socially stigmatizing, such as those involving sex and drugs.
Ann Epidemiol 2005;15:29–38.

KEY WORDS: Probability of Mixture Determination, Short Tandem Repeats, Biological Mixtures, Genetic Biomarkers.

INTRODUCTION

Advances in genetic studies using molecular biology techniques have accelerated dramatically in recent decades and extensions of these methods to epidemiological studies are proving fruitful (1–3). In genetic studies, innovative molecular technologies are routinely used as means to translate genetic information from individual laboratory specimens to build inferences about the human genome and its influence on the risk of disease (4). These molecular techniques can also be integrated into conventional epidemiological studies, expanding the application of a genetic approach to novel problems.

In the era of a known human genome sequence (5, 6), human molecular genetics is broadening its focus beyond identifying the contribution of genes, the environment and their interactions to providing a better understanding of disease processes. Applying these powerful laboratory techniques to examine DNA and protein biomarkers has led to new approaches for addressing epidemiological questions (7, 8). Biomarkers have been applied to epidemiological studies to assess individual exposures such as the molecular measure of infectious microbes or the titration of chemical compounds in the environment, pre-clinical effects such as prognosis and diagnosis of cancer and other diseases, and indicators of susceptibility to infectious and chronic diseases such as the determination of DNA damage or genetic polymorphisms (2, 9–11); but their usefulness extends to many additional aspects of epidemiology.

Analyzing polymorphisms in the human genome provides a valuable approach in genetic and molecular biology

From the Laboratory of Genomic Diversity and Basic Research Program, Science Applications International Corporation, National Cancer Institute at Frederick, Frederick, MD (S.S., M.W.S.); Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD (S.S., T.H.B., S.A.S.); and Division of International Health and Cross Cultural Medicine, University of California, San Diego, La Jolla, CA (S.A.S.).

Address correspondence to: Dr. Michael W. Smith, Laboratory of Genomic Diversity, and Basic Research Program, SAIC-Frederick, National Cancer Institute at Frederick, Building 560, Rm. 21–74, Frederick, MD 21702. Tel.: (301) 846-1913; Fax: (301) 846-6327. E-mail: smithm@ncifcrf.gov

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. The project included in this manuscript has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400 and the National Institute on Drug Abuse (DA09225 and DA8009).

Received December 9, 2002; accepted May 25, 2004.

360 Park Avenue South, New York, NY 10010

1047-2797/05/$–see front matter
doi:10.1016/j.annepidem.2004.05.006

Selected Abbreviations and Acronyms

AA = African Americans
AS = Asians
CAE = capillary array electrophoresis
DNA = deoxyribose nucleic acid
EA = European Americans
ELSI = ethical, legal, and social implications
H = heterozygosity
HBV = hepatitis B virus
HCV = hepatitis C virus
HIV = human immunodeficiency virus
HS = Hispanics
IRB = institutional review board
PCR = polymerase chain reaction
P_m = probability of matching
P_{MD} = probability of mixture discrimination
RFLP = restriction fragment length polymorphism
SNP = single nucleotide polymorphism
STR = short tandem repeat
VNTR = variable number of tandem repeat

studies (12–14). Many types of polymorphic genetic markers have been identified (15), such as restriction fragment length polymorphism (RFLP), variable number of tandem repeat (VNTR), single nucleotide polymorphism (SNP), and short tandem repeat (STR) (Table 1). However, STRs offer enormous potential for mixture analyses that could be instrumental in various epidemiological studies. STRs are tandemly repeated simple DNA sequence motifs of 2 to 7 bases in length (16–18) that are arranged head-to-tail and are well distributed throughout the human genome, primarily in the intragenic regions. They are abundant in essentially all ethnically and geographically defined populations and are characterized by simple Mendelian inheritance (19). STR polymorphisms originate due to mutations caused by slipped-strand mispairing during DNA replication that results from either the gain or loss of repeat units. Mutation rates typically range from 10^{-3} to 10^{-5} events per gamete per generation (20–22), compared with single nucleotide rates of mutation of 10^{-7} to 10^{-9}(23). Generally larger STR alleles have higher mutation rates (21, 24). In humans, STR markers are routinely used in gene mapping (25, 26), paternity testing and forensic analysis (27, 28), linkage (29, 30) and association studies (31), along with evolutionary studies (32, 33). Herein, we present and review the usefulness of STR biomarkers to differentiate between mixed and unmixed biological samples in epidemiological studies.

METHODS

Detecting Biological Mixtures with Highly Polymorphic Genetic Markers

The diploid (2N) nature of human beings results from inheriting a haploid (1N) genome from each parent. Genotypes consist of two alleles at any autosomal STR marker, which are either homozygous or heterozygous. When biological materials or fluids of two or more people are mixed (e.g., blood or saliva), one can distinguish the alleles of the two individuals at an STR when the alleles differ in size. Detecting three or more distinct alleles at any autosomal STR locus in a single biological sample provides essentially definitive evidence that the sample is a biological mixture from at least two individuals (true positives, Table 2).

However, several considerations complicate interpretation of STRs in relation to determining mixtures. First, in extremely rare cases of abnormalities (such as chimerism or trisomy) or somatic mosaicism, individuals with three distinct alleles have been reported (34, 35). The frequency of such three-allele genotypes in standard forensics markers is 0.04% (36) and the highest level reported, to our knowledge, is for TPOX at 0.18% (37). Markers with minimal false positive rates from three-allele genotypes can be identified by simple genotype determination and screening. Second, false positives (Table 2) due to the failure in amplifying the STR can be a concern. The occurrence of slippage (38, 39) from the *Taq* polymerase enzyme extending the labeled strand across a slipped repeat with a bulge (40) produces minor fragments that are

TABLE 1. Characterization of polymorphic biomarkers used in biomolecular and genetic epidemiology studies

Genetic Biomarkers	Abbreviation	Genetic Basis	Relevant References
Single nucleotide polymorphisms	SNP	Variable nucleotide site in the sequence with two alternate alleles (e.g., G or A)	(101, 102)
Restriction fragment length polymorphism	RFLP	Individual RFLP alleles are usally an SNP or an insertion/deletion variation. Alleles are differentiated by analysis of restriction enzyme digestions and subsequent size resolution of the resulting DNA fragments	(103, 104)
Variable number of tandem repeats	VNTRs	Variation of nucleotide repeat units of 12 to 60 or more and total length ranging from 500 to 3000 bp usually found at telomeres (also known as minisatellites)	(105, 106)
Short tandem repeats	STRs	Tandemly repeated simple DNA sequence motifs of 2 to 7 bases in length and are usually distributed in the intragenic region of the genome (also known as microsatellites)	(16, 107)