Analyzing population structure for forensic STR markers in next generation sequencing data

Sanne E. Aalbers[1] and Bruce S. Weir[1]

[1]Department of Biostatistics, University of Washington, University Tower, 15th Floor, 4333 Brooklyn Ave., Box 35946, Seattle, WA, USA

Corresponding author: saalbers@uw.edu

Introduction

Forensic DNA interpretation has been centered on the analysis of short tandem repeats (STRs), traditionally relying on capillary electrophoresis (CE) to gain access to the allele numbers contained in a DNA sample. To evaluate such DNA evidence profiles, match probabilities can be calculated and these depend on appropriate estimation of the population structure quantity F_{ST}, or theta, values. It is common in forensic DNA evidence evaluations to use values of 1% - 5% [1].

With the introduction of next generation sequencing (NGS) a new dimension has been added to the field of forensic genetics, providing distinct advantages over CE systems in terms of captured information. Traditional STR analysis has been well established in the forensic community so backward compatibility with CE-based STR profiles is needed to allow the use of existing DNA databases [2]. As long as this is the case, it is expected that NGS methods will continue to be implemented, stressing the need to facilitate NGS-based population genetics analysis.

In recent years, studies have reported population statistics demonstrating the increase in discrimination power by differentiating the nucleotide sequences of alleles with identical size [3-5]. If NGS data are to be used for match probabilities there needs to be a way to accommodate population structure, which requires values for F_{ST} for NGS data.

Methods

Most published estimates of F_{ST} are produced using the Weir and Cockerham estimator [6]. A different estimator is recommended nowadays, as detailed in Weir and Goudet [7] and applied to CE-based STR data in [1]. This updated framework can be used to obtain locus-specific F_{ST} estimates as well as estimates per geographic group and a global measure. Matching proportions within and between individuals or populations are used in order to characterize identity by descent (IBD) for an individual or population relative to a reference set of IBD values.

Within-population matching between individuals j and j' in population i is defined as $\widetilde{M}^i_{jj'} = \sum_u X^i_{ju} X^i_{j'u}/4$, where X_{ju} denotes the dosage of allele u for individual j. The average between-individual matching in population i with N_i individuals $\widetilde{M}^i_S = \sum_{j \neq j'} \widetilde{M}^i_{jj'}/[N_i(N_i-1)]$ can then be averaged over populations to get $\widetilde{M}^S = \sum_i \widetilde{M}^i_S/r$. Similarly, matching between individual j from population i and individual j' from population i' $\widetilde{M}^{ii'}_{jj'} = \sum_u X^i_{ju} X^{i'}_{j'u}/4$ leads to average between-population matching $\widetilde{M}^{ii'}_B = \sum_{j \neq j'} \widetilde{M}^{ii'}_{jj'}/(N_i N_{i'})$. Averaging over pairs of populations yields $\widetilde{M}^B = \sum_{i \neq i'} \widetilde{M}^{ii'}_B/[r(r-1)]$.

Population-specific F_{ST} values for genotypic data can then be estimated relative to genotype matching between populations as $\hat{\beta}^i_{ST} = (\widetilde{M}^i_S - \widetilde{M}^B)/(1-\widetilde{M}^B)$, with an overall estimate of $\hat{\beta}_{ST} = (\widetilde{M}^S - \widetilde{M}^B)/(1-\widetilde{M}^B)$. Note that these are locus-specific estimates, which are expected to vary among loci. The average β estimates over loci are calculated as the ratio of averages rather than the average of ratios, with the former leading to smaller variances. The reader is referred to [7] and [1] for a more detailed discussion on this approach. The overall estimate $\hat{\beta}_{ST}$ may be used as the population structure quantity theta in forensic match probability calculations.

Results

DNA samples from 350 individuals, over 5 geographic groups, as part of the 1000 Genomes Project Phase 3 (http://www.1000genomes.org) were obtained from the Coriell Institute for Medical Research (Camden, New Jersey, USA) and sequenced using Illumina's MiSeq FGx™ and ForenSeq™ DNA Signature Prep Kit. Genotype calls were obtained through their Universal Analysis Software (UAS) over 27 autosomal loci for both the length-based (LB) allele callings as well as the sequence-based (SB) allele callings.

Estimated matching proportions based on length-based genotype matching yields an average within-population matching, averaged over populations and loci, of $\widetilde{M}^S = 0.2165$, while the average between-population matching is $\widetilde{M}^B = 0.1968$, yielding an overall estimate of $\hat{\beta}_{ST} = 0.0245$. Population-specific estimates range from 0.0035 for the African group to 0.0347 for the American group. These results are concordant with the worldwide survey by Buckleton et al. [1], in the sense that the smallest values are observed for the African group as compared to the rest of the world, as expected from our understanding of higher genetic diversity within those older populations from the migration of modern humans.

Matching proportions based on sequence-based genotypes show somewhat lower values of $\widetilde{M}^S = 0.1878$ and $\widetilde{M}^B = 0.1664$ due to the increase in the number of observed types as a result of the additional variation. The global estimate is in this case $\hat{\beta}_{ST} = 0.0257$, which is an increase from our previous estimate, albeit small. 95% confidence intervals can be obtained based on bootstrapping over loci. Intervals for length-based results (0.0179, 0.0315) show a great deal of overlap with sequence-based intervals (0.0191, 0.0329), suggesting that the usual recommended value of around 3% is appropriate for DNA evaluations based on NGS data.

As sequencing data are subject to population structure, the impact on theta estimates should be checked when transitioning to NGS-based methods. So far, our results show similar effects of sequencing data on theta estimates as what we've seen for CE-based results.

Acknowledgments

This work was funded in part by grants 70NANB15H323 from the US National Institute of Standards and Technology, grant 2017-DN-BX-K541 from the US National Institute of Justice, and grant GM 075091 from the US National Institutes of Health. The authors thank John Buckleton and Scott Kennedy for their support.

References

[1] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, B.S. Weir. Population-specific F_{ST} values for forensic STR markers: A worldwide survey. Forensic Sci. Int.: Genetics, 23:91-100, 2016.

[2] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P. De Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Philips. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society of Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Sci. Int.: Genetics, 22:54-63, 2016.

[3] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Sci. Int.: Genetics, 21:15-21, 2016.

[4] N.M.M Novroski, J.L King, J.D. Churchill, Lay Hong Seah, B. Budowle. Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Sci. Int.: Genetics, 25:214-226, 2016.

[5] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone. Sequence-based U.S. population data for 27 autosomal STR loci. Forensic Sci. Int.: Genetics, 37:106-115, 2018.

[6] B.S. Weir, C.C. Cockerham. Estimating F-statistics for the analysis of population-structure. Evolution, 38(6):1358-1370, 1984.

[7] B.S. Weir, J. Goudet. A Unified Characterization of Population Structure and Relatedness. Genetics, 206:2085-2103, 2017.