UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

MARIA LUISA DE BARROS RODRIGUES

Seleção de microhaplótipos em larga escala para inferência de ancestralidade na população brasileira

Ribeirão Preto – SP
2024

RESUMO

RODRIGUES, M. L. B. **Seleção de microhaplótipos em larga escala para inferência de ancestralidade na população brasileira.** 2024. Tese (Doutorado em Ciências – Área de Concentração: Genética) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brasil, 2024.

Os microhaplótipos (MHs) são blocos de 2 ou mais *SNPs* presentes em um segmento de *DNA* de tamanho entre 200 e 300 pb. O interesse crescente no uso de MHs é devido à presença de alelos múltiplos, que resulta em maior informatividade que os *SNPs* individualmente, e menor taxa de mutação que os *STRs*. Portanto, MHs tornam as estimativas da genética de populações, forense e clínica mais precisas. Visando estimar a ancestralidade da população brasileira pela primeira vez a partir de MHs, elaboramos um *pipeline* e desenvolvemos um *script* para seleção de MHs altamente informativos em larga escala, a partir de dados genômicos. Partimos de um *dataset* incluindo 522 indivíduos do Sudeste do Brasil, mesclados aos dados dos bancos públicos (*SGDP*, *HGDP* e *1000 Genome Project*), totalizando 4081 indivíduos genotipados em quase 1 milhão de *SNPs* a partir dos quais selecionamos um conjunto de mais de 120 mil MHs, amplamente distribuídos entre os 22 cromossomos autossômicos. Os marcadores, tanto MHs quanto *SNPs*, tiveram sua informatividade estimada e foram separados em subconjuntos de marcadores mais informativos para serem utilizados nas estimativas de ancestralidade. Os resultados foram comparados entre si e às estimativas referentes ao conjunto completo de marcadores demonstrando maior eficiência dos MHs para essa finalidade e maior proximidade de resultados dos subconjuntos de MHs em relação ao conjunto completo. Desenvolvemos também uma abordagem para estimar o que chamamos de informatividade *cluster* específica, no caso informatividade nativa americana, demonstrando maior acurácia na estimativa da proporção de ancestralidade desse grupo sub-representado em bancos de dados públicos.

Palavras-chave: microhaplótipos, ancestralidade, informatividade, população brasileira, nativo americanos, *microarray*.

ABSTRACT

RODRIGUES, M. L. B. **Large-scale selection of microhaplotype for ancestry inference in the Brazilian population.** 2024. Tese (Doutorado em Ciências – Área de Concentração: Genética) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brasil, 2024.

Microhaplotypes (MHs) are blocks of 2 or more SNPs present in a DNA segment of up to 300 bp. The growing interest in the use of MHs is due to the presence of multiple alleles, which results in higher informativeness than individual SNPs, and lower mutation rate than STRs. Therefore, MHs make estimates of population, forensic, and clinical genetics more accurate. To estimate the ancestry of the Brazilian population for the first time from MHs, we developed a pipeline and developed a script for the selection of highly informative MHs on a large scale, based on genomic data. We started from a dataset including 522 individuals from the Southeast of Brazil, merged with data from public databases (SGDP*,* HGDP and 1000 Genome Project*),* totaling 4081 individuals genotyped in almost 1 million SNPs from which we selected a set of more than 120 thousand MHs, widely distributed among the 22 autosomal chromosomes. The markers, both MHs and SNPs, had their informativeness estimated and were separated into subsets of the most informative markers to be used in ancestry estimates. The results were compared with each other and with the estimates for the complete set of markers, demonstrating greater efficiency of MHs for this purpose and greater proximity of results of MH subsets in relation to the complete set. We also developed an approach to estimate what we call specific cluster informativity, in this case Native American informativeness, demonstrating greater accuracy in estimating the proportion of ancestry of this underrepresented group in public databases.

Keywords: microhaplotypes, ancestry, informativeness, Brazilian population, Native Americans, microarray