Mining mapped reads in 1000 Genomes data for accurately predicting polymorphic human endogenous retroviruses

Weiling Li¹ *, Raunaq Malhotra¹, Mary Poss², and Raj Acharya¹

¹ Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802
² Department of Biology, The Pennsylvania State University, University Park, PA, 16802

Human endogenous retroviruses (HERVs) present in all individuals have been implicated with some diseases [5]. However, the role of polymorphic HERVs in diseases has not been investigated due to their unknown prevalence in the generic population and due to the lack of computational methods for their discovery. We propose a novel method to accurately detect polymorphic HERVs in an individual by mining reads that mapped to the human reference genome (hg19 build). We leverage the fact that reads from a polymorphic HERV, if present in an individual, map at the location of a closely related fixed HERV in the genome. Thus, the proportion of distinct segments (kmers) of the polymorphic HERVs detected in an individual can be used as an indicator for their presence. We determine the threshold for identification of a polymorphic HERV using a simulation on a population of individuals containing polymorphic HERVs. The inflection point on a ranked plot of decreasing proportions of the polymorphic HERV in the population simultaneously identifies the prevalence of polymorphisms and the threshold for polymorphic HERV's detection [1]. We demonstrate our method on three known polymorphic HERV-Ks (K103, K113 and K115) [4], and determine their prevalence on the 1000 Genomes data [2]. Our results are concordant with known experimental results for polymorphic HERV-Ks: K103 with 99%, K113 with 26%, and K115 with 18% prevalence respectively, and even among individuals from different geographical locations [3]. Among 14 individuals who do not have polymorphic HERV-K103, 12 have an African ancestry, while other two have Mexican and Puerto Rican ancestries respectively.

References


* wul135@cse.psu.edu
