Identification and Characterization of Transposable Elements in Genetic Reference Populations

Anwica Kashfeen[1], Fernando Pardo Manuel de Villena[2], Leonard McMillan[1]
Department of Computer Science[1], Department of Genetics[2]
University of North Carolina at Chapel Hill, NC, USA

A large fraction of eukaryotic genomes, including those of mammals, consists of transposable elements (TEs). Spontaneous TE insertions also cause deleterious mutations, and drive chromosome evolution. It is difficult to identify, map, characterize, and determine the zygosity of TEs using current high-throughput short-read sequencing data. Existing approaches search for TEs by aligning billions of mostly irrelevant short reads to either a reference genome or a TE sequence library. These methods are computationally slow, have high false negative rates, and are unable to determine the TE's genomic context and/or zygosity status. Here we present a new msBWT-based TE identification and characterization pipeline that significantly outperforms previous methods in each one of these areas. We apply this method to two different laboratory mouse populations, the Collaborative Cross and a well defined set of commercially available substrains. In each population, we are able to detect fixed, shared and private TEs. We consider private TEs as those whose presence differs between individual samples with identical haplotypes, and these TEs tend to segregate in the relevant population. Thus we conclude that most private TEs represent de novo insertion events. We have identified hundreds of private TEs using our approach, we provide preliminary evidence that the number of private TEs depends on genetic background and that private TEs are more deleterious than either shared or fixed TEs. We will discuss the implications of these findings in classical genetic analyses and their impact on rigor and reproducibility.