DETAILED ANALYSIS OF SUBSTITUTION ERROR IN mtDNA MPS DATA FROM VARIOUS SAMPLE SOURCES AND LIBRARY PREPARATION METHODS

Jennifer A. McElhoe and Mitchell M. Holland
The Pennsylvania State University

Massively parallel sequencing (MPS) has become an important tool in medical, biological, and forensic studies. The improved resolution that can be achieved with MPS has presented the opportunity to characterize mitochondrial (mt) DNA heteroplasmy to new levels that were previously impossible with traditional sequencing platforms. One important aspect for ensuring the correct characterization of minor allele frequencies (MAFs) is the ability to generate deep sequencing data with low background noise. For example, a coverage rate of 500 reads would not be acceptable with a reporting threshold of 0.2%, as a single observation of a minor variant could be the result of instrument error. Instead, an assessment of the analytical noise or error is required to identify the point at which reliable data can be reported. Inherent with this increased level of resolution, and accompanying deep sequencing, is the challenge of discriminating between error and true observations of heteroplasmy.

This poster evaluates substitution and sequence specific error for a variety of sample types and library preparations. Sample sets included purposefully damaged samples, hair samples, and pristine buccal and blood samples that were prepared for sequencing on an Illumina MiSeq using either the Nextera XT library preparation kit (Illumina) or the PowerSeq CRM Nested System library preparation kit (Promega). Assessments included substitution error in secondary data, sequence specific error when identifying sites with the greatest frequency of error in both the forward and reverse read directions, and correlation of error to adjacent upstream sequence.

Overall, assumed substitution error rates ranged from 0.18-0.49 errors per 100 nucleotides with C nucleotides generally having the highest rate of misincorporation. Comparison of error rates across samples indicated a significant difference ($p<0.05$) between damaged and non-damaged samples. In most cases, the positions of error were varied across samples with pair-wise concordance ranging from 0-68%, while the concordance for motifs was greater with a range of 50-90%. The most commonly observed motif preceding error in the forward reads was CCC, while GGG was most common in reverse reads.