QUALITY SCORES IN MPS DATA: WHAT ARE THEY GOOD FOR?

August E. Woerner[1], Jonathan King[1], and Bruce Budowle[1, 2]
[1]Center for Human Identification, University of North Texas Health Science Center
[2]Center of Excellence in Genomic Medicine (CEGMR), King Abdulaziz University

While many recent studies have highlighted the rich sequence diversity found in many of the common forensically relevant STRs, making the best use of this information requires a thorough understanding of the rates of error. The simplest unit of error is the Phred-scaled quality score, which is an estimate of the per nucleotide error rate. However, the quality scores from massively parallel sequencing (MPS) data are only weakly related to the true rates of error in commonly used forensic MPS kits. Moreover, the rates and types of error vary substantially across loci and between/among sequencing runs. A machine learning framework was developed that recalibrates quality scores, thus correcting many of these biases, and allowing quality to be interpreted more effectively in downstream applications. While many tools for quality score recalibration exist, ours is the first that does not require a reference sequence, and thus it does not suffer from issues of reference bias. Further, it is specifically tailored to operate on both insertion/deletion as well as single nucleotide polymorphism variants, making the tool especially relevant to forensic applications.