Assessing single-nucleotide polymorphism and genotype calling using the KPGP-38 Human Genomes next-generation sequencing data from CAMDA

Wenqian Zhang, Valerii Soika, Jie Shen, Joe Meehan, Zhenqiang Su, Weigong Ge, Hong Fang, Roger Perkins, Vahan Simonyan, Weida Tong, and Huixiao Hong

Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA

Next generation sequencing (NGS) has become the preferred technology in current genetic studies largely because of the potential to identify not only common genetic variants, but also novel and rare variants as well as structural variants. However, reliability of genetic findings based on NGS data relies crucially on accurately calling single-nucleotide polymorphisms (SNPs) and their genotypes. False SNP calls and incorrect genotypes are caused in different ways, including sequencing errors, incorrect base calling, mapping errors, and sampling bias due to insufficient coverage. Analysis methods and quality control metrics to distinguish true SNPs from false variants are urgently needed to realize the full discovery potential from NGS data. The Critical Assessment of Massive Data Analysis (CAMDA) consortium hosts the KPGP-38 Human Genomes NGS data obtained from the Illumina HiSeq 2000 platform with 30x to 40x coverage. Importantly, this dataset's high coverage and the inclusion of two different sets of twins and a Caucasian female provides a suitable opportunity to explore quality control metrics for improving accuracy in SNP and genotype calling and to investigating aspects of population genetics. We first used different pipelines such as SOAP2-SOAPsnp and Hexagon to call SNPs and genotypes as well as structural variants for the KPGP-38 data. Then, twin pairs KPGP88/KPGP89 and KPGP90/KPGP91 were determined to be monozygotic based on SNP and genotype call concordance that was somewhat higher than we've previously observed for technical replicates from the same DNA. Thereafter, discordant SNPs and calls for the pair of twins were presumed to be errors for which causes could be postulated and related to the numerical procedures used. The procedure was then used to define criteria to gauge quality and identify likelihood for a calling error for application in profiling other samples. Interestingly, applying the calling criteria, we identified some SNPs in KPGP-38 that were not identified in the 1000 Genomes Project, suggesting that such metrics could refine population genetics. We will share our methods and results and will further discuss implications of our findings.