Impact of library-specific artifacts on single nucleotide variant analysis in whole-genome sequencing data

Tiziana Sanavia, Minseok Kwon, Maxwell A. Sherman, Alison Barton, Rachel Rodin, Christopher Walsh, Peter J. Park

Background: Whole-genome sequencing (WGS) offers the ability to detect genomic alterations in both coding and non-coding regions that might be involved in regulatory and epigenetic mechanisms. Common WGS experiments are currently performed at 30-40X of coverage, using a single technical replicate per biological sample. Aim of this study is the quantification of library-specific artifacts in single nucleotide variant (SNV) calls from WGS data using multiple technical replicates for the same biological sample and the identification of specific features that allow their detection.

Methods: SNV analysis was performed in PCR-free Illumina WGS data of human brain tissues from 7 individuals using GATK HaplotypeCaller with default settings. For each biological sample, 7 technical replicates at 30X of coverage were provided and analyzed separately. To detect those SNVs that are more likely to be possible artifacts, library-specific calls were defined as those non-polymorphic SNVs (i.e. population frequency <1%) called in only one library per sample.

Results: 4.5% of the non-polymorphic SNVs (116,132 on average per sample) are library-specific, and more than 12% are missed in at least one library. Compared to the SNVs found in all the 7 libraries, library-specific calls are characterized by lower allelic fractions (0.16 vs 0.5 on average) and genotype quality (62% vs 99% on average) and they are mainly described by a specific mutational signature showing higher fractions of ACA>AAA and TCT>TTT mutations at homopolymeric regions. These results show a surprisingly high number of library-specific SNVs which are identifiable with mutational properties at specific genomic contexts. Thus our findings can be used as guidelines to build filters able to remove potential false positives in analyses like somatic or de novo mutation detection.