BACON: Baited Abrogation of CONtamination

Fernando Mendez[1], Ruomu Jiang[1], Simon White[1], William Lee[1]

[1] Helix, San Carlos, California

Whole exome and whole genome sequencing (WES and WGS, respectively) have proven valuable tools to hasten diagnosis of genetic disorders and enable discovery of heritable traits and disease. These approaches rely on next generation sequencing (NGS) technologies and have been scaling up due in part to the ease of sample collection with saliva (used millions of times), both in the clinic and at home. NGS methods are very sensitive, but also susceptible to the effects of contamination. Saliva-based collection methods require collection protocols (i.e. waiting 30 minutes) that attempt to minimize the impact of exogenous factors such as food on data quality. Nonetheless analytical tools are necessary to combat possible sample contamination that could lead to incorrect genotype calls, which impact ancestry estimates, estimates of relatedness, and clinical diagnosis. A number of statistical methods (like Freemix) enable the efficient detection of human cross-contamination; however, these methods were not designed for cross-species contamination. The effects of cross-species contamination on human genotype calls have not been carefully explored, possibly due to the lack of large saliva based collections that are subsequently sequenced on NGS technologies. Potential sources of contamination in saliva include oral microbiome and ingested food. Bacterial sequences are very divergent from those of humans, but some animal sequences may be mapped and used in variant calling.

We have studied the effect of mapping sequences from nine farm and pet animals to the human reference genome and found that a considerable number of human genes in a contaminated human sample could be flagged as having pathogenic mutations. This includes more than a third of the genes included in the ACMG 59 list of genes with actionable mutations. We tested various levels of animal contamination and observed that amounts of foreign DNA that are not detected as contaminating by Freemix can still result in genotyping errors. We implemented a filtering method that greatly alleviates this problem, removing most incorrect calls without significantly affecting the coverage by reads of human origin. We then applied this method to human samples that were collected from saliva and exome sequenced. We show that our method significantly reduces the risk of incorrect genotype calls with potential healthcare consequences, thereby improving consumer confidence in genomic testing results.