LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets

Dreycey Albin (da39)

March 7, 2019

Scientific question

The scientific question and goal for this paper focuses on *creating a quality-aware method for calling single nucleotide variants (SNVs) in genomic sequencing data.* The method Wilm et al. created focuses on cell-population heterogeneity using statistical methods and a pruned-dynamic-programming approach, thereby enabling a fast and robust way to find low-frequency single nucleotide changes within a sequencing sample.

Hypothesis tested

The hypothesis tested in this research is that *there method will be able to call low frequency variants, and in particular, better than the currently (at the time) existing methods.* This hypothesis was tested by comparing their method to Breseq, SNVer, Goto et al., and Wright et al on multiple test datasets (in silico, DENV2 cell-culture isolates). In addition, LoFreq's ability to call SNVs on low coverage genomic data was compared to samtools using gastric adenocarcinoma samples with 30x coverage. However, the true novelty in proving their hypothesis came with the experimental validation, where they tested their results using micro-fluidic digital PCR systems (using the "Fluidigm Digital Array").

Methods/Results

The methods and results for this paper focus on (1) The statistical methods employed (i.e. SNV calling); (2) detection limit testing; (3) Testing on gastric cancer genomes and mitochondrial genomes; (4) robustness and false positive rates; and (5) experimental validation using a Fluidigm Digital Array and time-of-flight mass spectrometry. These methods efficiently yield valid comparisons between LoFreq and other methods for calling SNVs.

While the results may be segmented into multiple categories, the in silico testing with the simulated E. Coli genomes and cultured dengue viral samples may be grouped together. A distinct gain using LoFreq is the ability for the method to maintain a perfect positive predictive value (PPV), while also delivering a high sensitivity at multiple coverages (simulated). A big takeaway is that the SNV calls from Breseq and SNVer represented a subset of the total call made by LoFreq on the simulated genomes. As for validation, the frequencies for the cultured samples called LoFreq consistently give similar results to the Fluidigm Digital Array. An important note to be made is that all methods have trouble detecting low frequency SNVs in samples with lower coverage, and the reproduciblity and robustness for all model-based approaches.

LoFreq was also tested against samtools for being able to measure frequencies on low coverage samples. Data from gastric adenocarcinoma samples with 30x coverage were used to test the differential ability for detecting low frequency variants. This is difficult because the tumor cells are masked by normal tissue samples, causing SNV frequencies to become lower. This is evident in the evaluation performed by samtools, as the frequency distribution has a stark cut off at lower frequencies. In contrast, LoFreq is able to reestablish a symmetric frequency distribution for the samples (these were compared to germline cells).

Another important result is the ability for LoFreq to identify "hot" and "cold" regions in a genome. The "cold" regions correspond to areas of the genome that are less likely to have single nucleotide polymorphisms (SNPs), and "hot" being the opposite definition. The researchers applied this to three Dengue virus serotypes that in a drug-trial study for the nucleoside-analog *Balapiravir.* The resultant data showed that proteins needed for viral viability contained "cold" spots in regions needed for proper function.

Key implications of the results

The key implication from these data is that LoFreq increases sensitivity, maintains reproducible, minimizes false positives, all the while establishing a new low in runtimes for model-based SNV methods (O(NK), N=number

of rows in alignment, K=number of variants in the column). Their method expands upon the methods that were currently available by also incorporating a quality-aware approach, and has a exponentially (literally) decreased runtime by making statistically relevant assumptions and using pruned-dynamic-programming.