SNPPET: A Fast and sensitive algorithm for variant detection and confirmation from targeted sequencing data

Ashutosh, Devendra Joshi, Christian Le Coq, Margherita Corioni, Eric Lin and Jayati Ghosh

Agilent Technologies Genomics, Santa Clara, California

Introduction

The adoption of targeted sequencing gene panels as preferred methods of variant identification has been rapid in both clinical research and investigative studies. Due to this rapid growth, fast and sensitive methods of analysis have become of paramount importance. For cancers and other mosaic situations, it is critical to be able to detect or confirm the presence of important pathogenic variants, even if they occur at low allele fractions or in complex genotypes. First generation variant calling methods and pipelines have difficulty calling variants reliably in such situations. The SNPPET algorithm, developed as part of Agilent SureCall Analysis platform, overcomes these shortcomings.

The algorithm can be configured to run in various modes (Figure 1). Within the workflow of Agilent SureCall, each mode results in variant calls, assessment and categorization of variant.

Algorithm

The algorithm performs a statistical test between two models at each locus: A null model that assumes that all non-reference bases seen at the locus are only due to sequencing error and an alternative model that assumes each non-reference base seen at the locus is a true variant. Only the reads satisfying a specified mapping quality criterion and uniquely mapped to a location are considered and only bases covering the locus that have base qualities greater than the quality filters. Final called variants are then assessed for their biological impact and categorized by subsequent steps.

To assess the likelihood and compute each term in the product, we assume that the sequencing errors at the locus are independent and then assign a significance to the variant call by comparing the two models. The fraction of the true alternate allele is then estimated by a maximum likelihood approach by maximizing the odds ratio between the two models. If the odds in favor of a true variant exceed the null model by more than a specified threshold then the variant is considered a potential true variant.

The algorithm then examines the candidates found in each capture target region and subjects the candidate variant stretches to a series of filters to control false positives. It also assembles the variant calls into called indels and variant relationships (e.g. between TP53 and myeloid neoplasms) mentioned in literature across 80 samples. The confirmation of variants was done by Sanger sequencing.

Experimental Methods

Using results from four experiments, benchmarking and validation of SNPPET is demonstrated below. Experiment 1 and 2 validate the ability to detect variants even at low fractions of samples. Experiment 3 is used to benchmark sensitivity and specificity of the algorithm against other known tools. Finally, experiment 4 is performed on ClearSeq AML panels developed in collaboration with Dr. Robert Ohgami and Dr. Daniel Arber at Stanford University, to validate the calls made using independent assays. In experiment 1, SureSelect XT low input pre-capture libraries[1] from 200ng input generated from different dilutions of gDNA molecular reference standards (BRAF V600E, Horizon Diagnostics allelic frequency of 100%, 1% and 6%) is used and captured with the SureSelect Focused Clinical Research Exome (3Gb of data). In experiment 2, SureSelect QXT libraries[2] generated from 50ng input for the same standards is used. In both experiments, the dilutions tested are 5%, 10%, 25%, 50%, 75% and 100% series. The data from confirmed and suspected AML patients is captured and analyzed. The variants called by the algorithm were validated with independent methods (Sanger sequencing and SnapShot). SNPPET was then run again in the confirmation mode to re-verify status of validated variants. The whole genome samples are analyzed with the GATK pipeline using binned Genotypy as the variant caller using best practices with GATK v1.1 and used as gold standards.

Results and Discussion

For experiments 1 and 2, expected variants are detected down to 5% (Figures 2 and 3). For experiment 3 the results are compared against calls from GATK. All samples are analyzed both with the GATK pipeline as well as SNPPET (Table 1). SNPPET identifies all NA1083 variants at the 0% dilution. At increased fractions HCT116 variants are detected as expected. SNPPET performs better at lower allele fraction in terms of both false positives and false negative rates. For Exp. 4, SureCall is able to detect or confirm over 95% of the validated variants and the known relationships (e.g. between TP53 and myeloid neoplasms) mentioned in literature across 80 samples. The confirmation of variants was done by Sanger sequencing.

Implementation and Performance

SNPPET is implemented completely in Java 7 and uses the fork-join framework to parallelize and fully utilize multicore architectures. As part of SureCall it can run on a computer with 8GB RAM (16GB RAM recommended). Each target region is mapped for variants in parallel and finally the candidate variants are reduced per region. This speeds up the variant detection and makes the algorithm useful for large batch analyses in a clinical research setting. The typical performance metrics on various sizes of capture are given in table 2. SureCall is available for free to Agilent customers for either Windows PC or Mac OS.

Conclusions

SNPPET is a fast variant caller which makes less stringent assumptions about allelic models and hence is able to detect lower allele fractions and complex variants reliably. This ability, coupled with its speed and ease of use makes it a very useful tool for variant detection and confirmation.

References and Acknowledgements

1. SureSelect XT Target Enrichment System for Illumina Pared-End Sequencing Library Illumina HSieq and MiSeq Multiplexed Sequencing Platforms. Protocol Version 1.6, October 2013*

We acknowledge Dr. Robert Ohgami and his colleagues at Stanford University for invaluable input in design and implementation experiments.