Unified detection, classification, and phasing of germline, somatic, and de novo variation from high throughput sequencing data

Daniel Cooke and Gerton Lunter

August 5, 2016

Abstract

High throughput sequencing technology has seen the rise of powerful methods to detect genetic variation directly from raw data. Most notably, haplotype-aware variant calling has become de facto for identifying germline variation.

Motivation

Methods

0.0.1 Read preprocessing

0.0.2 Candidate allele generation

In order to explicitly store allele phase information the algorithm first generates a unique set of candidate alleles. This step is intended to maximise sensitivity and therefore considers proposals from independent allele 'generators'. There are two primary methods of variant generation; directly from raw alignment (CIGAR strings) and via local reassembly. Both methods have strengths and weaknesses, in particular reassembly can resolve larger structural variation and is mapper independent, but is slow and is highly dependent on the k-mer size used to make the underlying de Bruijn graph which can lead to real variant 'bubbles' being missed or pruned. On the other hand directly using the alignments is fast and predictable, but is mapper biased and unable to resolve complex structural variation. Combing both approached achieves the superior sensitivity to one or the other.

0.0.3 Candidate haplotype generation

Once the candidate variant set is finalised the algorithm 'walks' sequentially through the set and builds a tree of non-overlapping alleles, each branch of the tree is a haplotype. Because alleles are explicitly represented the algorithm is able to dynamically remove and extend the tree whilst retaining previous phase information. The algorithm therefore selectively extends and prunes until all candidate alleles are exhausted. The pruning stage is based on a probability criteria, in practise the vast majority of haplotypes will carry little probability weight and can therefore be safety removed. This dynamic approach to haplotype generation allows the algorithm to consider allele sets far exceeding the cardinality of other approaches.

0.0.4 Model fitting

0.0.5 Calling and refinement

Results

Discussion