Towards a novel method for small indels detection using Illumina/Solexa data

Chiara M[1], Horner DS[1]

Motivation

While it was long assumed that most of the genomic variation within species is due to single nucleotide polymorphisms (SNPs), the importance of genomic rearrangements such as insertions, deletions, inversions and duplications has recently become clear. Traditional sequencing based methods for the discovery of genomic structural variants (SV) are based on the mapping of paired end reads (PEM). In this approach paired end sequence reads are generated from a library of genomic DNA with a narrow range of insert sizes. The reads are mapped to a reference genome, and pairs mapping at a distance that is substantially different from the expected length, or with anomalous orientation, suggest structural variants. While earlier PEM-based methods used low-coverage Sanger sequencing, in the last few years the advent of Next Generation Sequencing has accelerated the characterization of genomic structural variation, but has also required the development of new bioinformatics approaches. In this abstract we present a novel method for the detection of small indels using Illumina/Solexa data.

Methods

First we map paired end reads from the donor genome on the reference genome, and estimate the general distributions of distances (GDD) between the mapped mate pairs. We consider only mate pairs mapping on the reference genome at approximately expected distances (80-300 bp) with unique mapping solutions. Then for every position in the genome we calculate the average value and the variance of the distance between every read spanning that position and its mate pair, in a strand specific fashion. At this point we use a 2 tailed Welsch T test to assess whether the population of distances relative to a particular genomic position and strand has a mean which significantly differs from the average of the GDD. We then cluster neighbouring anomalous genomic positions into anomalous windows (AW). The presence of indels is also expected to result in a peak of mapping of unpaired tags upstream and downstream of the indel, at a distance which will be more or less equal the average of the GDD, as reads covering junctions of rearrangement events will not map to the reference genome. This means that, given sufficient depth of sequencing, to be a genuine hallmark of an indel every AW should be linked to a peak of unpaired mapped tags. We use a single tailed Welsch T test to verify that the mean coverage from unpaired reads is significantly

[1] Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano

higher than the genomic mean at the expected distance upstream or downstream of any given AW. Finally, we join the validated AW from both strands to reconstruct the indel events.

Results

To asses the performance of our approach we generated an artificial dataset consisting of 2000 indels ranging in size from 10 to 250 bp (10,20,30,40,60,80,120,150,180,250 bp) on the mitochondrial genome of V. vinifera strain PN40040 at different coverage levels. Preliminary results suggest that our approach performs better than other methods for the detection of short insertions and deletions, with an average recovery rate of 96.4% (combined insertions + deletions) at high coverge (>40X) for indels 30 to 80 bp long with a false positive rate of 0.1%. Remarkably our method performs well in detecting this category of indels even at moderate to low sequencing coverage (recovery rate of 83% at 10X coverage for deletions of 40 bp with a false positive rate of 1.5%, recovery rate of 93% with no false positive for deletions of 10 to 20bp when the coverage is >= 30X). However, detection of very small insetions remains less tractable, probably due to the asymmetric distribution of distances between paired end read. High recovery and low false positive rates are also observed in the detection of larger deletions while larger insertions are inherently difficult to detect when the mean of the GDD is small.

Contact e-mail

matteo.chiara@unimi.it