1. Program

Program : Estimate the false positive rate based on Mendelian Errors in Trio data

Program Version : 0.1a

Author : Ryan

Date : 19 July 2011

Development Language : python 2.7 (it should be compatible to all 2.Xs)

The package consists of 2 files

1. mergeTrio.py

This script takes trio variant list files and merge per loci.

1. testTrioInheritance.py

This script tasks the output of mergeTrio.py to report the statistics.

2. Introduction

These scripts estimate the false positive rate of SNP calls based on the principles of Mendelian inheritance. SNP locations observed in the child are compared to those from both parents. When the child SNP call is not attributable to inheritance, it is assumed to represent a false positive error. SNP calls can be made with any available tool (i.e., SAMtools, GATK, etc), but the input must be formatted as a two-column text file, where the first column lists the SNP location, and the second column lists the SNP calls(convert scripts will be provided for GATK).

3. How to run

1. mergeTrio.py

python mergeTrio.py child\_snp\_list parent1\_snp\_list parent2\_snp\_list > trio\_snp\_list

1. testTrioInheritance.py

python testTrioInheritance.py trio\_snp\_list > falsePostive\_trio.txt

4. Wendy's trio samples

In Wendy's sample, overall low coverage(~23X) as well as too small number of SNPs due to much strict filtered SNP calls which lead to unable to identify the true inheritance in both parents samples.

|  |  |
| --- | --- |
| # of SNPs in Child | 9928 |
| # of SNPs attributable to inheritance | 1379 |
| # of SNPs Mendelian Errors | 8123 |
| # of SNPs excluded due to incomplete data | 0 |
| # of tri-allelic SNPs in Child | 0 |
| False Positive Error Rate | 0.854873 |

5. Selecting Filter Thresholds

Mendelian Errors in Trio data could be good measurement of false positive rate and we could use this measure to determine optimal Filter Thresholds.

We calculated the total number of Mendelian errors in the trio data from Dr. Wendy Chung's data.

One of filter parametesr that I chose is the coverage which is the number of the reads that cover the current position. The mendelian errors are significantly lower at 49X above filter setting.

The other filter parameter I investigated is Reference Allele Count which is the number of reads of the reference allele at the current position. Roughly 50% of mapped reads contributed the reference allele with the binomial distribution. Around above 25 of Reference Allele Counts could be the optimal filter to reduce false positives caused by Mendelian Errors.

As expected, the frequency of Mendelian errors diminishes as a function of increased sequence depth and Reference allele count.