BMC Bioinformatics

POSTER PRESENTATION Open Access

Improving variant calling by incorporating known genetic variants into read alignment

Nam S Vo[*], Vinhthuy Phan

From 14th Annual UT-KBRIN Bioinformatics Summit 2015
Buchanan, TN, USA. 20-22 March 2015

Background

The identification of genetic variants has great significance in genetic research. To call variants using next-generation sequencing data, current methods rely primarily on mapped reads produced by a separate read aligner without taking into account existing genetic variants [1]. Thus, these methods usually require a large number of reads (high coverage) to be able to detect variants accurately [2]. Moreover, the separation of read alignment and variant calling results in a workflow is complex and involves many separate steps and different tools [3].

Materials and methods

We introduce a novel method that leverages existing information about genetic variants to improve performance of variant calling. Incorporating known variants allows reads to be aligned more accurately and variants to be detected accurately with low coverage. This method further integrates two separate processes of read alignment and variant calling into one unified workflow, which results in a much more automated, simplified and faster process. A Bayesian method is used to calculate quality of variant calls.

Results

We showed that this method significantly improved the accuracy of variants on simulated data on human chromosomes, especially with low-coverage data, compared to popular methods such as GATK. At low coverage (<= 5x), this method achieved recall rates that were 2-19% higher while maintaining competitive precision compared to GATK. In particular, the method showed a significant improvement on identifying INDELs with recall rates 33-42% higher than GATK, and precision rates 9-34% higher than GATK. Our method also simplifies the workflow greatly, requiring 2 steps to call variants while GATK requires 6-8 steps and 2 external tools including Picard and SAMtools to preprocess the data.

Conclusions

As genetic variants are being collected for more and more people, the integration of existing information into the calling of variants is realistic. We demonstrated that by incorporating existing variant information, accurate detection of variants could be achieved even with low coverage. Thus, the method is promising in helping to reduce experimental cost.

Availability

Source code and testing data are available at https://github.com/namsyvo/IVC.

Acknowledgements
This work is partly supported by NSF CCF-1320297. We thank Quang Tran for his help in initial stage of the project.

Published: 23 October 2015

References

1. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and snp calling from next-generation sequencing data.** *Nature Reviews Genetics* 2011, **12**:443-451.
2. Yu X, Sun S: **Comparing a few snp calling algorithms using low-coverage sequencing data.** *BMC Bioinformatics* 2013, **14**:274.
3. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant analysis of next-generation genome sequencing data.** *Briefings in Bioinformatics* 2014, **15**:256-278.

doi:10.1186/1471-2105-16-S15-P18
Cite this article as: Vo and Phan: **Improving variant calling by incorporating known genetic variants into read alignment.** *BMC Bioinformatics* 2015 **16**(Suppl 15):P18.

* Correspondence: nsvo1@memphis.edu
Department of Computer Science, University of Memphis, Memphis, TN 38152, USA