Genome Medicine

RESEARCH HIGHLIGHT

Improving bioinformatic pipelines for exome variant calling

Hanlee P Ji[1,2]

See research article http://www.biomedcentral.com/1471-2105/13/8

Abstract

Exome sequencing analysis is a cost-effective approach for identifying variants in coding regions. However, recognizing the relevant single nucleotide variants, small insertions and deletions remains a challenge for many researchers and diagnostic laboratories typically do not have access to the bioinformatic analysis pipelines necessary for clinical application. The Atlas2 suite, recently released by Baylor Genome Center, is designed to be widely accessible, runs on desktop computers but is scalable to computational clusters, and performs comparably with other popular variant callers. Atlas2 may be an accessible alternative for data processing when a rapid solution for variant calling is required.

Keywords Next-generation sequencing, exomes, variant calling, single nucleotide variation, insertion, deletions

Next-generation DNA sequencing (NGS) has revolutionized genetics by enabling researchers to routinely sequence genomes, either in their entirety or specific subsets [1-3]. For example, exome resequencing, in which researchers enrich for all annotated and putative exons and then sequence the genomic targets, has been widely adopted. Exome sequencing has become a popular approach owing to the availability of commercial exome enrichment assays, the generally lower cost than whole-genome sequencing and the focus on coding regions and associated variants that have a direct impact on coding sequence and thus gene function. As a result, a large number of studies are using human exome resequencing to study the genetic diversity of human populations. Furthermore, exome resequencing is frequently used in the study of human diseases, including Mendelian disorders and cancer. Given the accessibility of the technology, many groups are working towards potential clinical diagnostic applications in personalized medicine.

Challenges of variant calling from exome sequencing

Analysis has become one of the primary challenges for NGS users, as a direct result of the sheer volume of sequencing data currently being generated. Exome sequence analysis can be generally summarized as a two step process with alignment of the data to a human genome reference followed by subsequent genetic variant calling from the post-alignment data, or, more simply, the identification of specific sequence alterations that are polymorphisms, rare variants or mutations. Exome-targeted resequencing analysis is particularly useful for the discovery of single nucleotide variants (SNVs) and insertion or deletions (indels). Although a variety of robust and now widely adopted sequence alignment tools are available, the challenge of variant calling from aligned data remains. Although alignment algorithms can be used to accurately determine the location of any sequence, it is more problematic to determine whether a variation that is identified in an aligned sequence is a true genetic variation. Numerous academic and commercial groups have developed a variety of bioinformatic tools or cloud-based solutions to facilitate variant calling. For example, SAMtools [4] and the Genome Analysis Tool Kit (GATK) [5] provide SNV and indel calling and are widely used. However, experienced bioinformaticians and information technology specialists are required to implement these and other popular tools, limiting accessibility of variant calling for the research community and in clinical diagnostic laboratories. In addition, the accuracy of variant calling bioinformatic tools is highly variable. Improving the speed, accuracy and user-friendliness of sophisticated variant calling pipelines is an important step towards personalized medicine.

*Correspondence: genomics_ji@stanford.edu
[1]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[2]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA

Seeking to address this challenge, Fuli Yu and colleagues from the Baylor Genome Center recently published in *BMC Bioinformatics* a variant-calling software package, the Atlas2 suite [6], that can analyze aligned data generated from a variety of NGS platforms, including Life Sciences' SOLiD, Roche's 454 and Illumina's Genome Analyzer and HiSeq systems. I focus on application of the Atlas2 suite and its optimization for exome sequence analysis. Atlas2 relies on the standard Binary Sequence Alignment/Mapping format (BAM) sequence format, which is a widely adopted and generally supported format among NGS users [4]. For example, this format is used for cancer exome data from the National Institutes of Health's repertoire of sequencing production studies such as the Cancer Genome Atlas. Variant calls are produced using the Variant Call Format (VCF), which has been adopted by the 1000 Genomes Project. Atlas2 runs on many different computing platforms, from the standard desktop computer to highly scaled implementation with multimode computational clusters. The group also implemented Atlas2 on the Genboree Workbench [7], a genomic web resource that enables the user to view NGS data and carry out analysis. Web-driven analysis tools are increasingly popular, as seen with Galaxy [8] and DNANexus [9]. Given that most clinical diagnostic centers do not have a dedicated bioinformatics staff, the desktop computer and web implementation aspects have particular appeal for a diagnostic laboratory setting.

Developing and testing a variant calling algorithm

Challis *et al.* [6] used a trained logistic regression model to assess the quality of each potential variant as a true variant rather than a sequencing, mapping, or alignment error. Critical factors that they assessed include the following: ratio of the variant base to the reference sequence, the overall sequencing depth, the quality scores of the base calls, and the position in the read and strand direction based on whether it is the forward or reverse sequence. These factors are well known to be important in determining the accuracy of variant calls and, not surprisingly, were the most significant in the authors' statistical analysis [6]. Ultimately, these variables were an important component of the statistical model driving the Atlas2 variant calling algorithm.

One of the challenges faced by any NGS variant caller is the introduction of false positive and false negative SNVs and indel calls. All variant callers and their related publications typically report the overall sensitivity, specificity and false discovery rates of variants called from a control genome. Atlas2 was compared with two other popular variant callers, GATK and SAMtools, using sequence data from the 1000 Genomes Project. For single nucleotide polymorphisms (SNPs), the group discovered that Atlas2 had a generally high concordance with reported SNP discovery. Overall, Atlas2 showed a significantly lower number of indels by nearly an order of magnitude than either GATK or SAMtools, potentially because of higher false positive rates in GATK and SAMtools. For the analysis of Illumina sequencer data, Atlas2 indel calls were comparable to those of Dindel [10], an indel caller considered to be state-of-the-art, with a greater than 85% concordance rate.

The authors [6] also demonstrated that Atlas2 could be run on a desktop computer and that the analysis of a 28 GB whole-exome BAM file takes only 2 hours. Interestingly, they used a single core processor, whereas many current desktop computers have significantly more processors, suggesting that even faster processing times could be achieved on newer and more powerful desktop computers. The implementation of Atlas2 on a computational cluster enabled rapid processing and they cite a performance of running 92 exomes from 64 processors in 4 hours. In my own experience, GATK requires a significant number of compute nodes to process exome sequence in days, rather than hours. The speed of analysis is a major strength of the Atlas2 suite and makes such analyses accessible to research groups analyzing exome data, even those who do not have routine access to large multiprocessor computational clusters.

Overall, Challis *et al.* [6] have introduced a highly flexible and rapid SNV, SNP and indel calling program suite. It uses standard data formats that are widely adopted. Most importantly, the flexibility of Atlas2 to run on standard desktop computers and workstations provides an opportunity for many laboratories to use the software. Also, in comparison studies with GATK, SAMTools and Dindel, Atlas2 demonstrated comparable variant calling accuracy, showing that accuracy is not compromised by faster processing times. Given that this software is generally available and is open source, this represents a solution that could be readily adopted by many groups regardless of their size or resources.

Abbreviations

BAM, binary sequence alignment/mapping format; GATK, Genome Analysis Tool Kit; NGS, next-generation sequencing; SNP, single nucleotide polymorphism; SNV, single nucleotide variant; VCF, Variant Call Format.

Competing interests

The author declares that they have no competing interests.

Acknowledgements

This work was supported by the following grants from the NIH: 2P01HG000205 and RC2 HG005570-01. In addition, HPJ received support from the Doris Duke Clinical Foundation and the Howard Hughes Medical Foundation. The funding bodies had no role in writing of the manuscript or the decision to submit the manuscript.

Published: 30 January 2012

References

1. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135-1145.

2. Natsoulis G, Bell JM, Xu H, Buenrostro JD, Ordonez H, Grimes S, Newburger D, Jensen M, Zahn JM, Zhang N, Ji HP: **A flexible approach for highly multiplexed candidate gene targeted resequencing.** *PLoS One* 2011, **6:**e21088.
3. Myllykangas S, Buenrostro JD, Natsoulis G, Bell JM, Ji HP: **Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing.** *Nat Biotechnol* 2011, **29:**1024-1027.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
6. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs A, Yu F: **An integrative variant analysis suite for whole exome next-generation sequencing data.** *BMC Bioinf* 2012, **13:**8.
7. **Genboree** [http://www.genboree.org]
8. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A: **Manipulation of FASTQ data with Galaxy.** *Bioinformatics* 2010, **26:**1783-1785.
9. **DNAnexus** [http://www.dnanexus.com]
10. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R:

doi:10.1186/gm306
Cite this article as: Ji HP: **Improving bioinformatic pipelines for exome variant calling.** *Genome Medicine* 2012, **4:**7.

Dindel: accurate indel calls from short-read data. *Genome Res* 2011, **21**:961-973.