Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Sequence Analysis

Skyhawk: An Artificial Neural Network-based discriminator for reviewing clinically significant genomic variants

Ruibang Luo[1,2,*], Tak-Wah Lam[1], Michael C. Schatz[2]

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, [2]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Many rare diseases and cancers are fundamentally diseases of the genome. In the past several years, genome sequencing has become one of the most important tools in clinical practice for rare disease diagnosis and targeted cancer therapy. However, variant interpretation remains the bottleneck as is not yet automated and may take a specialist several hours of work per patient. On average, one-fifth of this time is spent on visually confirming the authenticity of the candidate variants.

Results: We developed Skyhawk, an artificial neural network-based discriminator that mimics the process of expert review on clinically significant genomics variants. Skyhawk runs in less than one minute to review ten thousand variants, and among the false positive singletons identified by GATK HaplotypeCaller, UnifiedGenotyper and 16GT in the HG005 GIAB sample, 79.7% were rejected by Skyhawk.

Availability: Skyhawk is easy to use and freely available at https://github.com/aquaskyline/Skyhawk

Contact: rbluo@cs.hku.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The dramatic reduction in the cost of whole genome, exome and amplicon sequencing has allowed these technologies to be increasingly accessible for genetic testing, opening the door to broad applications in Mendelian disorders, cancer diagnosis and personalized medicine (Katsanis and Katsanis, 2013). However, sequencing data include both systematic and random errors that hinder any of the current variant identification algorithms from working perfectly. Even using state-of-the-art approaches, typically 1-3% of the candidate variants are false positives with Illumina sequencing (Zook, et al., 2016). With the help of a genome browser such as IGV (Robinson, et al., 2017), or web applications such as VIPER (Woste and Dugas, 2018), a specialist can visually inspect a graphical layout of the read alignments to assess supporting and contradicting evidence to make an arbitration. Though necessary, this is a tedious and fallible procedure because of three major drawbacks. 1) It is time-consuming and empirical studies report it requires about one minute per variant, sometimes summing up to a few hours per patient (Uzilov, et al., 2016). 2) It is tedious, not infallible, and even experienced genetic-specialists might draw different conclusions for a candidate variant with marginal or contradicting evidence. 3) There is no agreed standard between genetic-specialists to judge various types of variants, including SNPs (Single Nucleotide Polymorphisms) and Indels. A specialist might be more stringent on SNPs because there are more clinical assertions and fewer candidate SNPs will be less likely to get contradicting medical conclusions, whereas another specialist might be more demanding on indels because they are rarer and harder to be identified.

An efficient, accurate and consistent computational method is strongly needed that automates assessing the candidate variants as they would be visually validated. Importantly, the new validation method needs to be orthogonal, i.e., independent of the algorithms used to identify the candidate variants. The new validation method also needs to capture the complex non-linear relationship between read alignments and the authenticity of a variant from a limited amount of labeled training data. As a validation method, the precision must be maximized, and false positives must be minimized. Consequently, instead of using hand-coded models or rule-

based learning, a more powerful and agnostic machine learning approach such as an Artificial Neural Network (ANN) is needed.

2 Implementation

We implemented Skyhawk, a computational discriminator that is fast and accurate for validating candidate variants in clinical practice. Skyhawk mimics how a human visually identifies genomic features comprising a variant and decides whether the evidence supports or contradicts the sequencing read alignments. To reach this goal, we repurposed the network architecture we developed in a previous study named Clairvoyante (Luo, et al., 2018). The multi-task ANN was designed for variant calling in Single Molecule Sequencing and the method is orthogonal to traditional variant callers using algorithms such as Bayesian or local-assembly. In Skyhawk, we used a repurposed network to generate a probability of each possible option for multiple categories including 1) variant type, 2) alternative allele, 3) zygosity, and 4) indel-length. We then compare a candidate variant to Skyhawk's prediction on each category. Skyhawk will agree with a variant if all categories are matched but will reject and provide possible corrections if any category is unmatched. We have provided pre-trained models for Skyhawk on GitHub trained using the known variants and Illumina data of multiple human genomes, including sequencing libraries prepared by either the PCR or the PCR-free protocol. With a trained model, Skyhawk accepts a VCF input with candidate SNPs and Indels, and a BAM input with read alignments. Skyhawk outputs a judgment and a quality score on how confident the judgment was made for each candidate variant. Skyhawk was implemented in Python and Tensorflow and has been carefully tuned to maximize its speed.

3 Results

Using four deeply Illumina sequenced genomes (HG001, HG002, HG003 and HG004) with 13.5M known truth variants from the Genome In A Bottle (GIAB) project (Zook, et al., 2016), we trained Skyhawk to recognize how the truth variants are different from another 20M non-variants we randomly sampled from the four genomes. The sample details and the commands used are in the **Supplementary Note**. For benchmarking and identifying the false positive variant calls, we used the known truth variants in HG005, which was not included in the model training. Although the known truth variants were intensively sequenced and meticulously curated, there are still some true variants that have not yet indexed by GIAB (Luo, et al., 2018). Thus, we called variants using three different variant callers including GATK HaplotypeCaller (HC) (Van der Auwera, et al., 2013), GATK UnifiedGenotyper (UG) (Van der Auwera, et al., 2013) and 16GT (Luo, et al., 2017). We expect any "false positive" variants called by two or all three callers are less likely to be genuine false positives and should be accepted by Skyhawk. Conversely singletons called by only one caller are more likely to be genuine false positive and should be rejected by Skyhawk. The results are shown in **Figure 1**. Only 18.64% of the "false positive" variants called by all three callers were rejected while 79.70% of the singletons were rejected by Skyhawk. Those called by two callers have an intermediate 45.11% rejected by Skyhawk. In the true positive variants, only 1,879/3,232,539 (0.058%) in HC, 43/2,902,052 (0.0014%) in UG, and 124/3,228,537 (0.0038%) in 16GT were rejected. By deducting the rejected variants from both the number of true positives and true negatives, the precision increased from 99.77% to 99.92% for HC, 99.50% to 99.58% for UG and 99.51% to 99.84% for 16GT.

Skyhawk aims to relieve users from a heavy manual review workload without compromising the accuracy. Instead of taking over the review of all variants, Skyhawk was configured to review only 1) SNPs with a single alternative allele, and 2) Indels ≤4bp. Skyhawk also outputs a quality score

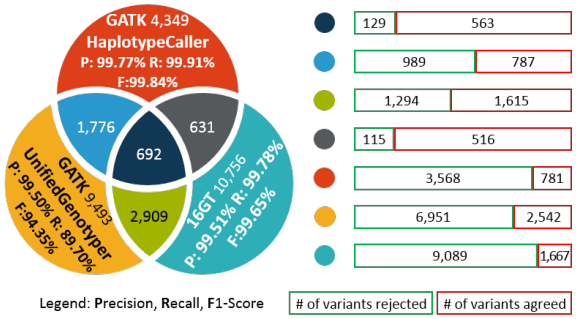

Figure 1. The variant calling results of GATK HaplotypeCaller, GATK UnifiedGenotyper, and 16GT.The Venn diagram on the left shows 1) the precision rate, recall rate and F1-score of each variant caller on the HG005 genome, and 2) the total number of variants fall on each shade. The bars on the right shows the number of variants rejected or agreed by Skyhawk. The bar length is proportionate to the total number of variants in that shade.

ranging from 0 to 999 to indicate how confident a judgment is. Among the false positive singletons, 27.46% of the judgments were with a quality score lower than 150. Reviewing these variants manually shows that these variants were often located in genome regions with homopolymer runs or very low depth. We suggest users to rely on Skyhawk only when the quality score of judgment is high and to manually review when the quality score falls below 150, or higher if the workload allows. Skyhawk requires less than a gigabyte of memory and less than a minute on one CPU core to review ten thousand variants, thus can be easily integrated into existing manual review workflows, such as VIPER (Woste and Dugas, 2018) with minimal computational burden. Overall, Skyhawk greatly reduces the workload on reviewing variants, and we believe Skyhawk will immediately increase the productivity of genetic-specialists in clinical practice.

Funding

R. L. and T. L. were partially supported by Innovative and Technology Fund ITS/331/17FP from the Innovation and Technology Commission, HKSAR. This work was also supported, in part, by awards from the National Science Foundation (DBI-1350041) and the National Institutes of Health (R01-HG006677).

Conflict of Interest: none declared.

References

Katsanis, S.H. and Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet* 2013;14(6):415-426.

Luo, R., Schatz, M.C. and Salzberg, S.L. 16GT: a fast and sensitive variant caller using a 16-genotype probabilistic model. *GigaScience* 2017.

Luo, R., *et al.* Clairvoyante: a multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing. *bioRxiv* 2018.

Robinson, J.T., *et al.* Variant Review with the Integrative Genomics Viewer. *Cancer Res* 2017;77(21):e31-e34.

Uzilov, A.V., *et al.* Development and clinical application of an integrative genomic approach to personalized cancer therapy. *Genome medicine* 2016;8(1):62.

Van der Auwera, G.A., *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;43:11 10 11-33.

Woste, M. and Dugas, M. VIPER: a web application for rapid expert review of variant calls. *Bioinformatics* 2018.

Zook, J.M., *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 2016;3:160025.