Bioinformatics, 31(22), 2015, 3715–3717
doi: 10.1093/bioinformatics/btv424
Advance Access Publication Date: 21 July 2015
Applications Note

OXFORD

Data and text mining

GeneCOST: a novel scoring-based prioritization framework for identifying disease causing genes

Bugra Ozer[1,2,*], Mahmut Sağıroğlu[1] and Hüseyin Demirci[1]

[1]Advanced Genomics and Bioinformatics Research Center, The Scientific and Technological Research Council of Turkey (TUBITAK), Gebze, Kocaeli, Turkey and [2]Department of Biological Sciences and Bioengineering, Sabanci University, 34956, Istanbul, Turkey

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 8, 2015; revised on July 15, 2015; accepted on July 16, 2015

Abstract

Summary: Due to the big data produced by next-generation sequencing studies, there is an evident need for methods to extract the valuable information gathered from these experiments. In this work, we propose GeneCOST, a novel scoring-based method to evaluate every gene for their disease association. Without any prior filtering and any prior knowledge, we assign a disease likelihood score to each gene in correspondence with their variations. Then, we rank all genes based on frequency, conservation, pedigree and detailed variation information to find out the causative reason of the disease state. We demonstrate the usage of GeneCOST with public and real life Mendelian disease cases including recessive, dominant, compound heterozygous and sporadic models. As a result, we were able to identify causative reason behind the disease state in top rankings of our list, proving that this novel prioritization framework provides a powerful environment for the analysis in genetic disease studies alternative to filtering-based approaches.

Availability and implementation: GeneCOST software is freely available at www.igbam.bilgem.tubitak.gov.tr/en/softwares/genecost-en/index.html.

Contact: buozer@gmail.com

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

1 Introduction

The progress in next-generation sequencing (NGS) technologies have led to an increased usage of whole exome and whole genome sequencing in medicine. The practice of high-throughput methods in identifying disease causing genes is becoming more prevalent with the aim of exploring the unresolved Mendelian diseases. Hence, there is still a need for effective tools to analyze the big data coming out of sequencing experiments.

An important step in extracting information from huge amount of data produced by high-throughput technologies is 'variant prioritization'. Variety of tools have been proposed in the literature for identifying the causative reason behind Mendelian diseases such as BierApp (Aleman *et al.*, 2014), Privar (Zhang *et al.*, 2013), Mendelscan (Koboldt *et al.*, 2014), KggSeq (Li *et al.*, 2012) and eXtasy (Sifrim *et al.*, 2013).

Most of these methods actually rely on tight filtering approach to reduce the number of variants. However, NGS technology is prone to errors at many steps such as optical reading and alignment. Therefore, filtering-based approaches have the possibility of missing the real causative reason behind the disease state. Additionally, some of the programs are web based and require the uploading of the variant file to the program owners' server which may compromise the privacy of the individuals. Furthermore, most of the programs can only operate on a single individual's data hence they are not able to use the whole information coming from the pedigree.

In this study, we propose a novel scoring-based approach for evaluating disease causing genes. Unlike any other method, GeneCOST focuses on genes instead of single variations. Hence, if several different mutations in a single gene are causing the same disease which is observed at sporadic cases, GeneCOST would

perfectly identify that gene as well. By adopting a scoring-based approach we aim to overcome the problems coming with the quality of the data. Thus, we show in our work that GeneCOST is superior to other prioritization tools no matter the data is faulty or not. Moreover, GeneCOST has a wider range of applicability than any other existing methods since it can make use of the pedigree information. It can accept multiple affected and healthy individuals' information, which is crucial for sporadic cases. Our method works offline on personal computers and users are not obliged to upload the vcf file to our servers which carries sensitive information. Additionally, GeneCOST does not require any prior knowledge of the disease such as phenotype information and can be easily applied using the command line.

2 Methods

When disease model is defined for a certain disease, GeneCOST evaluates each gene and ranks the genes according to their costs (likelihood scores) with the aim of finding disease causing gene among the first few candidates. For this purpose, we have validated GeneCOST for all disease models; sporadic, recessive, dominant and compound heterozygous.

2.1 Annotation

Our environment involves an annotation tool which takes the result of GATK genotyping as input and outputs an annotated vcf file which makes the vcf file suitable for GeneCOST execution. Details are provided in the Supplementary File.

2.2 Cost function

The disease likelihood score (cost function) in GeneCOST is defined to have a robust structure so that it can be applied to any kind of data.

The Cost function considers following parameters:

1. Minor Allele Frequency: frequency value taken from 1000 Genome and ESP6500, takes a value between 0 and 1,
2. Segmental Duplication Information: takes a value between 0 and 1,
3. Mutation Rankscore: 11 prediction algorithms are retrieved from dbNSFNP, a value between 0 and 3,
4. Mutation Type: a value between 2 and 3 depending on mutation severity considering type of the mutation (single nucleotide polymorphism, frameshift, splice, etc . . .),
5. PhastCons: considers the Conservation Score of the region of the variant: a value between 0 and 1,
6. Pedigree Information,
7. Sequencing Quality of a Variant.

The details of the cost function and general workflow of GeneCOST are explained in more detail at the Supplementary Section.

GeneCOST is implemented in C# and performs well with 64-bit structure. The program can easily be executed using personal Windows PCs. It requires a memory of 1.5 times of the vcf file and has also the ability to work with large vcf files where some of the previously proposed tools fail to accomplish.

3 Data and results

As current programs either only support single patient information or do not have the capability to work with our vcf files due to various reasons explained in Supplementary Section, we have compared our method with other available methods by applying sample files of eXtasy (Supplementary Table S3). Furthermore, we have proved applicability of our program for eleven different cases including both public and in-house (real-life) data.

1. Disease causing mutations for Miller Syndrome embedded to publicly available Corpas family vcf file for all disease models (Biesecker, 2010).
2. Publicly available Pigo.vcf file is used for the compound heterozygous case of hyperphosphatasia with mental retardation (Krawitz *et al.*, 2012).
3. Family with Klippel–Feil Syndrome (Bayrakli *et al.*, 2013).
4. Family with X-linked Renpenning syndrome.
5. Three family cases with autosomal recessive cerebrofaciothoracic dysplasia (Alanay *et al.*, 2014).
6. Single individual with non-ketotic hyperglycinemia.

Related pedigree information for these studies and evaluation of our program with respect to identifying disease causing gene is presented in the Supplementary Section (Supplementary Table S4).

As a result, for the cases with high sequencing quality, GeneCOST perfectly identified the causative reason behind the disease state at the first ranking of the candidate list. Even though when the read quality is low as in the case of Inhouse6 dataset and the expected variation was not observed in one of the samples, GeneCOST has successfully identified the causative reason at the 10th ranking, which would not be possible with any kind of filtering method.

GeneCOST has major advantages over existing programs. It supports pedigree information and multiple patient cases. The program does not require any prior knowledge of the disease, hence it can be used to identify novel genes. The disease likelihood score is calculated on a gene level approach, therefore, GeneCOST provides a valuable tool for sporadic cases where the patients do not have any kinship relation. Finally, the proposed method does not rely on any filtering procedure, hence, it is more robust to NGS-related errors which has not been observed at any other programs. Moreover, to the best of our knowledge, GeneCOST has been tested against a widest range of Mendelian diseases compared with other prioritization tools. We have demonstrated the usage of GeneCOST with artificial and real data over a number of Mendelian diseases from different inheritance models. Overall, GeneCOST is a powerful prioritization tool offering a strong alternative to filtering-based tools.

Funding

The project is supported by the Republic of Turkey Ministry of Development Infrastructure (Grant number 2011K120020) and BILGEM—TUBITAK (The Scientific and Technological Research Council of Turkey) (Grant number 100132). We are very grateful to Dr Nurten Akarsu, Dr Ali Dursun, Dr Fatih Bayraklı and Dr Ersan Kalay for sharing their datasets and for their invaluable suggestions.

Conflict of Interest: none declared.

References

Alanay,Y. *et al.* (2014) TMCO1 deficiency causes autosomal recessive cerebrofaciothoracic dysplasia. *Am. J. Med. Genet. Part A*, **164A**, 291–304.

Aleman,A. *et al.* (2014) A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Res.*, **42**, W88–W93.

Bayrakli,F. *et al*. (2013) Mutation in MEOX1 gene causes a recessive Klippel-Feil syndrome subtype. *BMC Genet.*, **14**, 95.

Biesecker,L.G. (2010) Exome sequencing makes medical genomics a reality. *Nat. Genet.*, **42**, 13–14.

Koboldt,D.C. *et al*. (2014) Exome-based mapping and variant prioritization for inherited Mendelian disorders. *Am. J. Hum. Genet.*, **94**, 373–384.

Krawitz,P.M. *et al*. (2012) Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. *Am. J. Hum. Genet.*, **91**, 146–151.

Li,M.X. *et al*. (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.*, **40**, e53.

Sifrim,A. *et al*. (2013) eXtasy: variant prioritization by genomic data fusion. *Nat. Methods*, **10**, 1083–1084.

Zhang,L. *et al*. (2013) PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. *Bioinformatics*, **29**, 124–125.