Bio-Medical Materials and Engineering 26 (2015) S1797–S1803
DOI 10.3233/BME-151481
IOS Press

An automatic and efficient pipeline for disease gene identification through utilizing family-based sequencing data

Dandan Song, Ning Li and Lejian Liao*
Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Lab of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

Abstract. Due to the generation of enormous amounts of data at both lower costs as well as in shorter times, whole-exome sequencing technologies provide dramatic opportunities for identifying disease genes implicated in Mendelian disorders. Since upwards of thousands genomic variants can be sequenced in each exome, it is challenging to filter pathogenic variants in protein coding regions and reduce the number of missing true variants. Therefore, an automatic and efficient pipeline for finding disease variants in Mendelian disorders is designed by exploiting a combination of variants filtering steps to analyze the family-based exome sequencing approach. Recent studies on the Freeman-Sheldon disease are revisited and show that the proposed method outperforms other existing candidate gene identification methods.

Keywords: Mendelian disorders, disease gene identification, whole-exome sequencing

1. Introduction

Recently, several different identification strategies have been developed for analyzing and interpreting whole-exome sequencing data in order to improve detection capabilities for disease-associated genes [1-3]. Typically, most published studies focus on exploiting the traditional candidate strategy and performing a filtering approach, which is based on the variant's predicted impact on protein function and structure, which are mutations in the exonic and splice-site region [4]. Moreover, these methods are always designed for analyzing a specific gene, and there is not a pipeline that can be commonly used for applying various filtering steps to the identification of a disease gene from exome resequencing experiments. In addition, such an approach is not optimal under the condition of further availability of family members or other affected individuals.

In this study, an automatic and efficient pipeline is proposed for finding disease variants in Mendelian disorders. Thus pipeline consists of a combination of features: an improved candidate strategy and

* Address for correspondence: Lejian Liao, Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Lab of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. Tel.: +86-1068913536; Fax: +86-1068913536; E-mail: liaolj@bit.edu.cn.

a family-based exome sequencing approach. As one appropriate open-source tool, ANNOVAR [5] and many in-house developed programs are integrated into the pipeline. The pipeline takes standard VCF (Variant Call Format) files and a highly configurable file as input and then automatically prioritizes variants for inherited Mendelian diseases. Specially, it can process multiple files simultaneously on per-chromosome mode. Additionally, twelve main criteria are used for predicting the effects of coding non-synonymous variants on protein function. In addition, it fully makes use of the family-bases sequencing information as well as takes non-affected family members as the control data set; thus, the amount of interesting variants can be kept to a minimum. The software is written in Perl and driven by command-line. It has been implemented as a standalone application on UNIX high performance computing system scheduled by SGE (Sun Grid Engine) as well as on a single personal computer equipped with the UNIX operating system. The proposed method revisits the Freeman-Sheldon disease and shows that it outperforms other existing candidate gene identification methods.

2. Materials and methods

2.1. An overview of the analysis pipeline

An automatic and efficient pipeline is proposed and developed for analyzing family-based exome sequencing data. The constructed steps are shown in Figure 1. The software takes VCF (Variant Call Format) files as the input of calling variant positions in DNA sequence data, which is commonly generated by methods such as Samtools (sequence alignment/map tools) [6], SOAP2 (short oligonucleotide analysis package 2) [7], and GATK (genome analysis toolkit) [8] by using sequencing data.

Due to the memory restrictions of most computing platforms, a parallel program (Algorithm 1) is designed for the analysis process by running the program on a separate chromosome-by-chromosome mode. This feature can be executed in a manner in which multiple files are performed simultaneously on an UNIX operating system automatically, and the real execution time is appreciably reduced. The results are merged into a single output file that sorts by the chromosome order.

2.2. Disease Gene Identification Process

In order to search for disease genes from amongst thousands to millions of genomic variants, several in-house developed scripts, as well as a source analysis tool, have been integrated into a single pipeline. It can be seen from Figure 1 that the proposed approach includes two phases: in the first phase, an interface to the ANNOVAR [7] software is provided to functionally annotate genetic variants from

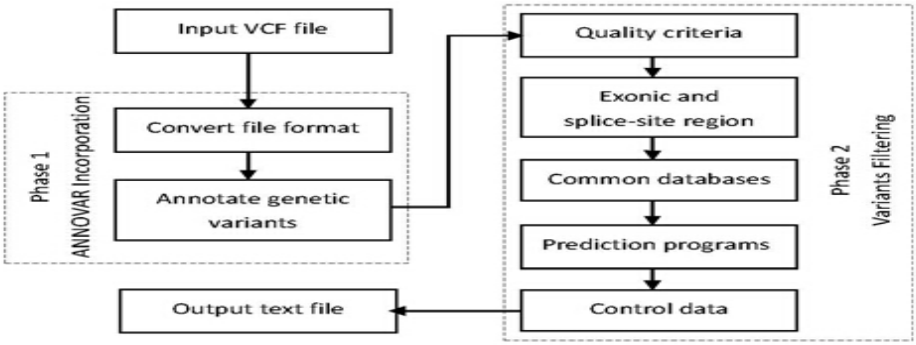

Fig. 1. Identification of disease gene from high-throughput family-based sequencing data.

Algorithm 1: Analysis procedure on per-chromosome basis

```
INPUT: A VCF file X
OUTPUT: Filtered results on per-chromosome
for i {chr1,chr2,chr3,chr4,chr5,chr6,chr7,chr8,chr9,chr10,chr11,chr12,chr13, chr14, chr15, chr16,
        chr17,chr18,chr19,chr20,chr21,chr22,chrX,chrY } do
  Yi    Xi  //VCF format converted to ANNOVAR standard format
  for j = 1,…,Ni do  //Ni is the total line number of file Yi, where each line corresponds to one variant
    Y'ij    Yij  //Gene-based annotation of genetic variants
  end for
  system "qsub  -l vf=6G -q group_ name -P project_name shell_script"  //Submit the job to the workstation
  job_num ++;    //The job number adds one
  if(job_num > max_job_num){
    sleep;      //The job is temporally turned into sleeping state.
  }
end for
```

high-throughput family-based sequencing data. In the second phase, a filtering method is performed to identify a candidate gene list for Mendelian diseases from the annotated genetic variants. Then, the details of each step that is contained in the two phases are described in the following:

2.2.1. ANNOVAR Annotation

The pipeline is supported to take VCF files as input, adopted by the 1000 Genomes Project [9] for encoding structural variations. ANNOVAR [5] is an efficient software that utilizes update-to-date information in order to functionally annotate genetic variants that are detected from high-throughput sequencing data. After annotating the list of variants with gene definition, it is convenient for the proposed filtering method to identify the causal disease genes.

2.2.2. Variants Filtering

Millions of genomic variants are identified per genome, but only a few may actually explain the Mendelian disease. Therefore, variants must be both automatically and efficiently prioritized to optimizing the disease gene identification process. The proposed filtering methods combine the improved candidate strategy with the family-based sequencing approach for selecting candidate variants. As far as is currently known, this pipeline presents the first effort to identify disease genes on this scale.

(1) Quality criteria

Firstly, the variants are filtered based on quality criteria; doing so will reduce the number of false-positive calls; the quality scores marked with PASS will be conserved.

(2) Exonic and splice-site region

Then variants outside the exonic and splice-site region will be excluded, assuming that they can not affect the mutated gene's encoded protein sequence's function and structure.

(3) Common databases

The 1000 Genomes Project and the International HapMap Project contribute to developing a public resource, including SNPs and structural variants and their haplotype contexts. VAF (Variant Allele Frequency) is the number of copies of a particular allele divided by the number of copies of all alleles at the genetic place (locus) in a population; it is usually expressed as a percentage. A VAF threshold is applied to the filtering variants from the public resources. Only variants whose allele frequency is higher or equal to the threshold are printed filtered out. Moreover, users can change the default threshold in the configurable file.

(4) Prediction program

Subsequently, twelve main criteria obtained from ANNOVAR, are used for predicting the effects of coding non-synonymous variants on protein function (see Table 1). A variant with higher scores is more deleterious and can be represented by any of the following: D (Deleterious), A (disease_causing_automatic), P (possibly damaging), or H (high). A filtering function is provided in the proposed method so as to connect a final deleterious decision to every prediction criteria, therefore indicating whether the variants are related to Mendelian diseases.

Because each prediction software has its own evaluation standard in predicting whether a mutation is harmful, the results can be variable between prediction softwares. However, in our research, a variation is considered a disease gene if two or more softwares predict said variation to be harmful. Of course, users can easily customize the number of software with positive prediction results that is appropriate for their needs so as to set their own criteria for a gene to be annotated as deleterious.

Accordingly, the deleterious function is defined as:

$$D=\sum_{i=1}^{12} Xi - N, \quad (1)$$

where Xi represents variants that are predicted as deleterious or tolerated based on twelve different scores, Xi , and N represents the minimum number of prediction software which annotates the variant as deleterious, N . Finally, if the Deleterious score (D) is greater than or equal to zero— that is, more than N software prediction results indicate that the candidate is deleterious— then the variant will be conserved.

Table 1

Detailed information for the twelve prediction programs

Score	**Categorial Prediction**	**X^i**	**Categorial Prediction**	**X^i**
SIFT	D: Deleterious	1	T: tolerated	0
PolyPhen 2 HDIV	D: Probably damaging; P: possibly damaging	1	B: benign	0
PolyPhen 2 HVar	D: Probably damaging; P: possibly damaging	1	B: benign	0
LRT	D: Deleterious	1	N: Neutral; U: Unknown	0
MutationTaster	A:disease_causing_automatic; D: disease_causing	1	N:polymorphism; P: polymorphism_automatic	0
MutationAssessor	H: high; M: medium	1	L: low; N: neutral	0
FATHMM	D: Deleterious	1	T: tolerated	0
MetaSVM	D: Deleterious	1	T: tolerated	0
MetaLR	D: Deleterious	1	T: tolerated	0
GERP++	higher scores (>=2)	1	lower scores (<2)	0
Phylop	higher scores (>=0.95)	1	lower scores (<0.95)	0
Siphy	higher scores (>=0.95)	1	lower scores (<0.95)	0

(5) Control data

Finally, in contrast with traditional candidate gene analysis approaches, the pipeline for identifying these specific genetic risk factors is improved because there is a search for an association between a specific variant and a disease via a comparison between a group of affected individuals with a group of unaffected controls. By using healthy individuals as the control data set, the list of candidate genes can be substantially trimmed down to a human-manageable number.

3. Experimental results

In order to clearly demonstrate the performance and efficiency of the proposed pipeline in identifying causal genes for Mendelian diseases, the Freeman-Sheldon disease was selected as the tested Mendelian disease in the experiments. Because its causal mutations in the genes are already known to be a rare, dominantly-inherited disorder. Since exome data for the Freeman-Sheldon cases were unavailable, we downloaded the exome data of three subjects who were reported in [10] in the HapMap Project, including one Yoruba subject (NA18507), one European American (NA12878), and one East Asian subject (NA18956). Additionally, we use the healthy European individual (SRR309293) as control data. The statistics of the testing data is summarized in Table 2.

Figure 2 lays out an overview of the variants reduction and the running time of each step. At first, ~0.67 million variants were initially inputted with gene annotation, amino acid change annotation, prediction scores, variant allele frequencies, and other information. Then, after filtering the variants based on quality criteria, ~0.59 million variants marked with PASS were conserved. Next, the variants were focused on exonic protein-changing only, and from which, a subset of 38,146 variants were identified as falling into highly-conserved genomic regions. Assuming that variants that are observed in the public common database with high variant allele frequency are less likely to be causal variants for

Table 2

Summary of the Freeman-Sheldon syndrome study data

Statistics	**NA12878**	**NA18507**	**NA18956**	**NA309293**
Raw data yield (Mb)	4,216	4,474	3,842	10,159
Data mapped to target region (Mb)	1,379	1,346	2,990	8,975
Reads mapped to target region (%)	40.03	37.42	52.65	95.35
Mean depth of target region	15.2	17.6	20.2	56.8
Coverage of target region (%)	85.39	89.29	89.87	97.81
Average read length (bp)	76	76	76	101
GC rate	42.03	41.44	41.19	53.29
Gender test result	F	M	F	M
Heterozygosity (%)	67.4	69.4	63.3	65.8
raw SNPs in total	606,506			
raw INDELs in total	63,715			

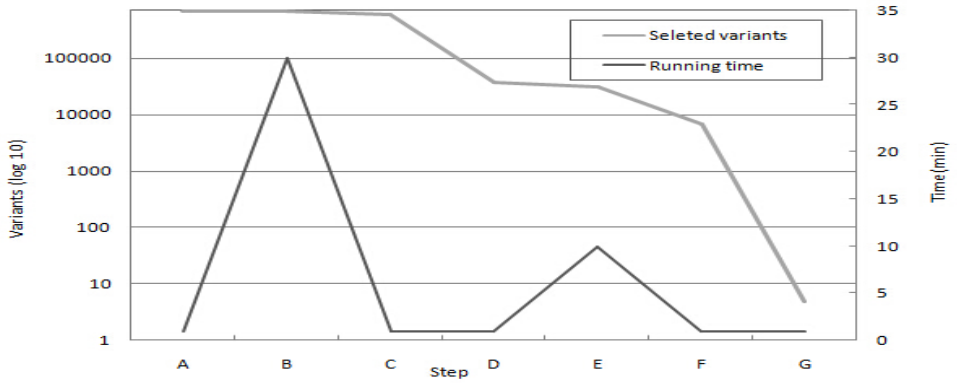

Fig. 2. Number of selected variants and the running time on each step of candidate strategy. Step A: A VCF file is converted to the ANNOVAR input format; Step B: Genetic variants are added with annotation information; Step C: Variants with low quality scores are filtered; Step D: Variants outside of the exonic and splice-site region are removed; Step E: The variants with high allele frequency in the public database are filtered; Step F: The deleterious variants are predicted; Step G: Healthy family members are set as control data so as to exclude private benign variation.

Freeman-Sheldon syndrome, the variants were filtered from the 1000 Genome Project and the HapMap project. This procedure left 31,120 variants. In the reduced variants set, 6801 mutations were predicted to be deleterious by applying Eq. (1). Finally, the four subjects were examined so as to identify variation shared among the three affected individuals; in leaving out the control case, the list of candidate genes was trimmed down to only three (HEATR1, TCF25, and MYH3), including the causal gene MYH3 as the major cause of this syndrome.

The following experiments were designed to further evaluate the performance of the proposed pipeline. The proposed pipeline was compared to the variants reduction method implemented in the ANNOVAR as the only comparative tool. The experimental results are shown in Figure 3. The causal gene MYH3 is the only major cause of this syndrome. The pipeline selected only 3 genes with 66.6% false identification, in stark contrast to the 142 candidate genes that were left by the ANNOVAR variants reduction procedure with 99.3% false identification. In addition, two variants were located in the MYH3 gene from the total number of five variants that were generated by the proposed method; therefore, the accuracy is 40%, whereas the accuracy of ANNOVAR is 2% because four useful variants were generated from the number of 204 variants. Moreover, the proposed method's superiority is also

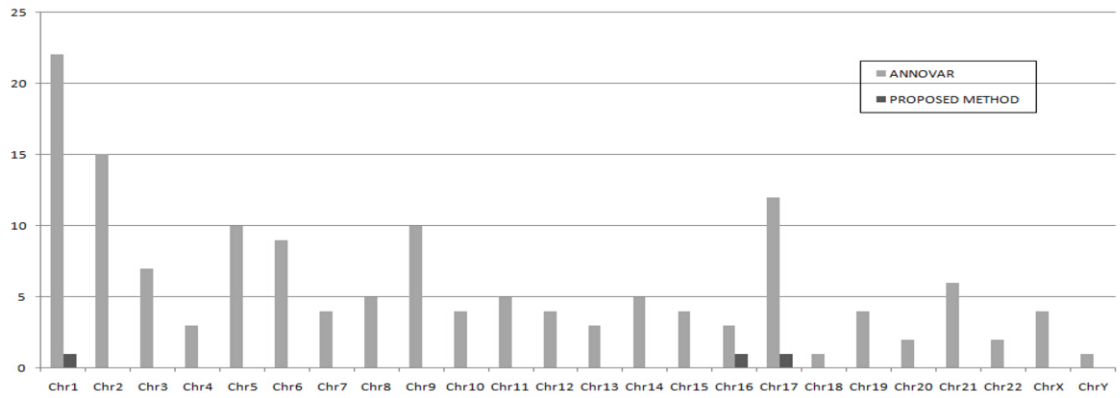

Fig. 3. Comparison between the proposed method and ANNOVAR on selected genes.

exemplified in the running time of the method; the proposed method requires four hours, which is far less than the eight hours that are necessary for ANNOVAR. These data demonstrate that the results achieved by the proposed method are more practical for further analysis in real applications.

4. Conclusion

In this paper, an automatic and efficient pipeline has been developed for identifying disease variants from family-based sequencing data. The distinguishing features of the proposed method are: (1) multiple files can be performed simultaneously on a per-chromosome basis equipped with a UNIX operating system automatically, (2) incorporation of twelve main criteria into the filtering procedure for prediction of gene function, and (3) using non-affected family members as the control data set. These features have been tested via analyzing real data sets, and experimental results show that this method achieves superior performance over another existing variants reduction pipeline. Thus, the proposed pipeline can be used in combination with other bioinformatic filters so as to streamline gene discovery in future exome sequencing projects.

Acknowledgments

This work is funded by the National Program on Key Basic Research Project (973 Program, Grant No. 2013CB329600), National Key Technology R & D Program of China (2012BAK11B01), National Natural Science Foundation of China (NSFC, Grant Nos. 61472040, 60873237), and Beijing Higher Education Young Elite Teacher Project (Grant No. YETP1198).

References

[1] J. Xu, L. Sun, Y. Gao and T. Xu, An ensemble feature selection technique for cancer recognition, Bio-Medical Materials and Engineering **24** (2014), 1001–1008.
[2] L. Sun and J. Xu, A granular computing approach to gene selection, Bio-Medical Materials and Engineering **24** (2014), 1307–1314.
[3] C. Gilissen, A. Hoischen, H.G. Brunner and J.A. Veltman, Disease gene identification strategies for exome sequencing, European Journal of Human Genetics **20** (2012), 490–497.
[4] J.L. Wang, X. Yang, K. Xia, Z.M. Hu, L.Weng and X. Jin, et al., TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing, Brain **133** (2010), 3510–3518.
[5] K. Wang, M. Li and H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Research **38** (2010), 164.
[6] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan and N. Homer, et al., The sequence alignment/map format and SAMtools, Bioinformatics **25** (2009), 2078–2079.
[7] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen and J. Wang, SOAP2: An improved ultrafast tool for short read alignment, Bioinformatics **25** (2009),1966–1967.
[8] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis and A. Kernytsky, et al., The genome analysis toolkit: A map reduce framework for analyzing next-generation DNA sequencing data, Genome Research **20** (2010), 1297–1303.
[9] N. Siva, 1000 Genomes project, Nature Biotechnology **26** (2008), 256–256.
[10] S.B. Ng, E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham and C. Lee, et al., Targeted capture and massively parallel sequencing of 12 human exomes, Nature **461** (2009), 272–276.