Informatics in Medicine Unlocked (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A new method for analysis of whole exome sequencing data (SELIM) depending on variant prioritization

Mehmet Ali Ergun[a,*], Abdullah Unal[b], Sezen Guntekin Ergun[a], E.Ferda Percin[a]

[a] Gazi University Faculty of Medicine, Department of Medical Genetics, Besevler, Ankara, Turkey
[b] ICterra, METU Technopolis Bachelor of Science Industrial Engineering Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Keywords:
Whole exome sequencing
Variant prioritization
Workflow

ABSTRACT

Background: After the first genome had been sequenced in 2003 with an international project, Human Genome Project, the 1000 Genomes Project also revealed the analysis of 1092 and 2504 genomes respectively. Whole exome sequencing of human samples was reported to detect approximately 20,000–30,000 SNV and indel calls on average. It is very important to choose the best tool that suits the related study.
Methods: In this study, it is aimed to demonstrate the results of an in-house method (SELIM) for variant prioritization of WES data without using in-silico methods.
Results: By this method, the annotated data have been decreased by 7.4–13.8 times (mean=10.9).
Conclusion: By the initiation of 1.000.000 genome project, powerful databases are needed. In this respect, SELIM is an in-house workflow that can easily be used for simplifying the annotated data without using any in-silico methods.

1. Introduction

After the first genome had been sequenced in 2003 with an international project, Human Genome Project, the 1000 Genomes Project also revealed the analysis of 1092 and 2504 genomes respectively [1–3]. Recently, the Precision Medicine Initiative Cohort Program will enroll 1 million or more volunteers in this research program. This will enable research for a wide range of human diseases by using next generation sequencing (NGS) [4,5]. Compared with Sanger sequencing, NGS had been reported to have many advantages, including high speed, low cost, less time consuming, high sensitivity, need of less amount of sample, and sequencing multiple genes at a higher coverage [6].

Whole Exome sequencing (WES), involves exome capture, which limits sequencing to the protein-coding regions of the genome, composed of about 20,000 genes, 180,000 exons, and constituting approximately 1% of the whole genome [7].

A typical workflow of WES analysis consists of the following steps: raw data quality assessment, pre-processing, alignment, post-processing, variant calling, annotation, and prioritization [8].

Regarding variant filtration and prioritization, the number of candidate variants has been reported to be reduced using a three-step filtration and prioritization steps; removing of reliable variant calls, choosing the low frequent variants and finding the variants related to the disease. As there are many tools available, it is very important to choose the best tool that suits the related study [8]. Additionally, public databases containing information on putative disease-causing mutations are incomplete and may have high error rates requiring manual curation; associations for some mutations in the database may not be causal [9].

In this study, it is aimed to demonstrate the results of an in-house method, SELIM for variant prioritization of WES data.

2. SELIM workflow

2.1. Design of the algorithm

SELIM is composed of eight steps to filter and prioritize candidate variants across individual patients and healthy controls that have been subjected to WES. SELIM was constructed using Microsoft Excel. This method is based on to filter the variants with respect to an algorithm without using in-silico tools.

In the first step after annotating the vcf data with the web interface ANNOVAR software (wANNOVAR) (http://wannovar.wglab.org/), the annotated data have been transferred to MS Excel file [10]. Secondly, each of the data has been filtered including the "exonic, exonic-splicing and splicing" parameters, excluding the “synonymous” SNV mutations. Then at the third step, the heterozygous (0/1) and homozygous (1/1)

* Corresponding author.
E-mail address: aliergun@gazi.edu.tr (M.A. Ergun).

http://dx.doi.org/10.1016/j.imu.2017.02.002
Received 21 January 2017; Received in revised form 31 January 2017; Accepted 4 February 2017
2352-9148/ © 2017 Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

alleles have been separated into two groups. At the fourth step, each of the groups has been classified according to "aa changes", "unknown" and “gene detail” parameters by removing the duplicate values using the filtering option of excel. At the fifth step, each of the data from each allele groups has been joined. The main reason for each step of the workflow is based on removing the duplicate values resulting only in the unique data.

At step 6, there are two ways to analyze the data. The step 6a refers to the removal of variants at a frequency of greater than 1% based on the 1000 Genomes Project (http://www.1000genomes.org/). Then at step 7a using online tools including Sorting Intolerant From Tolerant (SIFT) [11] and Polymorphism Phenotyping v2 (PolyPhen-2) [12] will help researchers to identify the candidate gene(s) for the associated disorder(s).

Regarding step 6b, we recommend separating the data using the “dbSNP” parameter. Then dbSNP data can be analyzed on the https://www.ncbi.nlm.nih.gov/snp/ and https://www.ncbi.nlm.nih.gov/clinvar/ websites using OR gates such as; rs0001 OR rs0002 OR rs0003… so on. At the step 7b, the data that do not confer with “dbSNP or ClinVar” databases have to be analyzed individually in order to identify the causative mutation.

Finally, at step 8, it is up to for the end-users to analyze the variants in order to find the candidate gene(s) related to the disorder. With this method, the annotated data have been decreased by 7.4–13.8 times (mean=10.9) (Fig. 1a and b).

Regarding a sample dataset composed of 14 samples, in the first place the vcf data has been annotated, and at the second step this data has been filtered including the "exonic, exonic-splicing and splicing" parameters, excluding the “synonymous” SNV mutations revealing 124.084 of the original dataset (Supplement 1). Then at the third step, the heterozygous (0/1) (70.996 data) and homozygous (1/1) (53,088 data) alleles have been separated into two groups (Supplemet 2a and b). At the fourth step, each of the groups has been classified according to "aa changes", "unknown" and “gene detail” parameters by removing the duplicate values using the filtering option of excel revealing 16.083 data for the heterozygous (0/1) and 5700 data for the homozygous (1/1) groups (Supplement 3a and b). At the fifth step, each of the data from each allele groups has been joined revealing 21.783 data (Supplement 4).

Also, as an example case, we would like to demonstrate how we used to identify the pathogenic variant in an Osteogenesis Imperfecta patient filtering from 11.814 to 2564 data (Supplement 5a and b). The filtered results led us to identify the pathogenic mutation in the COL1A1 gene (Gly560Cys) [13].

3. Discussion

After WES, a search for the disease-causing mutation is performed by comparing the sequencing data with a human genome reference, resulting in a list of all non-reference “variants.” Typically, 20–30,000 variants result for each exome sequence [14,15]. By using this method, the variants could be decreased by an average of 10 times, reaching about a thousand data per sample.

In case of the prioritization of candidate variants, a widely used approach has been reported to reduce the candidate list is to exclude known variants which are present in public SNP databases, published studies or in-house databases [16]. This method permits “apples-to-apples” comparisons of variants, enabling users to reach the unique data, without using in silico databases.

It has also been reported that the prioritization methods might have a risk in removing the pathogenic variant. So, it is advised to use these prediction tools with caution as they may not be reliable enough to indicate a definitive diagnosis. The use of different prioritization approaches and the combination of prediction results with phenotypic and pedigree data have been recommended [17]. This workflow is also useful for obtaining the unique data for the end-user without eliminating the pathogenic variants.

So, new bioinformatic tools are needed for NGS analysis. Especially, regarding WES analysis no pathogenic data have to be deleted or excluded. With this method the aim is to reduce the data by 10 times or more in order to delineate only the unique files.

4. Conclusion

As a conclusion, by the initiation of 1.000.000 genome project, powerful databases are needed. In this respect, SELIM, an in-house workflow is able to demonstrate the unique files as well as indicating the frequencies of the repeated variants without using any in-silico methods.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.imu.2017.02.002.

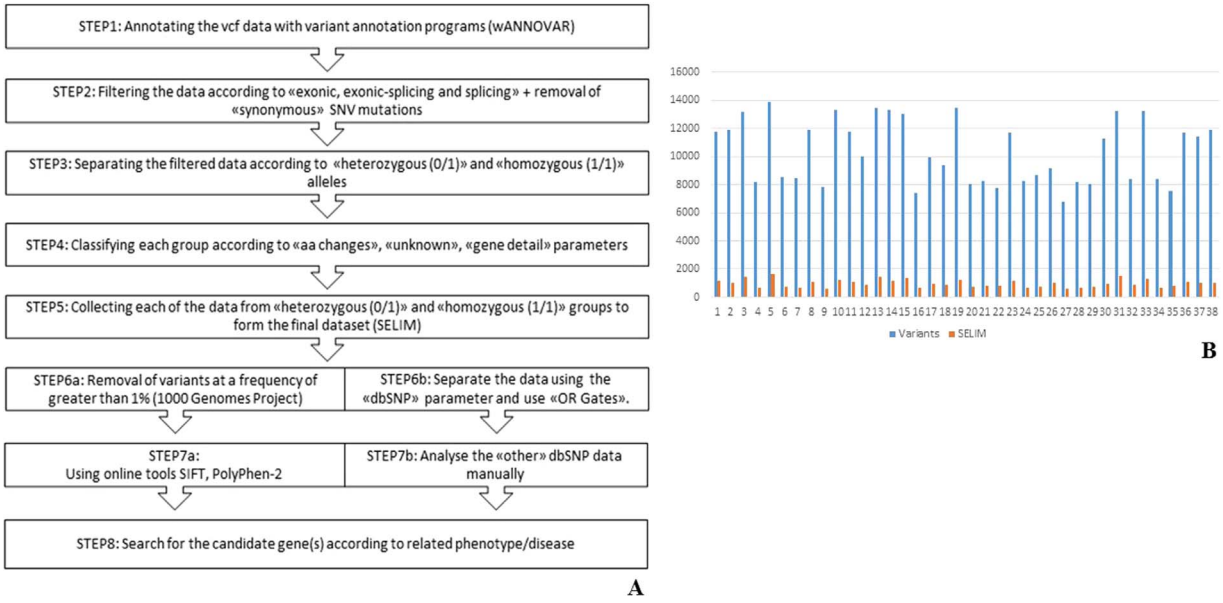


Fig. 1. (a) SELIM is composed of eight steps to filter and prioritize candidate variants. (b) The original annotated data with SELIM indicating the reduction in the big data.

References

[1] Schmutz J1, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salazar A, Tsai M, Myers RM. Quality assessment of the human genome sequence. Nature 2004;429:365–8.

[2] 1000 Genomes Project Consortium , et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.

[3] Sudmant PH, et al. An integrated map of structural variation in 2504 human genomes. Nature 2015;526:75–81.

[4] Collins FS, Varmus HA. New initiative on precision medicine. N Engl J Med 2015;26:793–5.

[5] Dong L, et al. Clinical next generation sequencing for precision medicine in cancer. Curr Genom 2015;16:253–63.

[6] Garraway LA. Genomics-driven oncology: framework for an emerging paradigm. J Clin Oncol 2013;31:1806–14.

[7] Choi M, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 2009;106:19096–101.

[8] Bao R, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inf 2014;13:67–82.

[9] Bell CJ, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 2011;3:65ra4.

[10] Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 2015;10:1556–66.

[11] Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–74.

[12] Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.

[13] Ergun MA, et al. Whole exome sequencing reveals a mutation in an osteogenesis imperfecta patient. Meta Gene 2017;11:137–40.

[14] Robinson PN, et al. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet 2011;80:127–32.

[15] O'Rawe J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 2013;5:28.

[16] Gilissen C, et al. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 2012;20:490–7.

[17] Pabinger S, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 2014;15:256–78.