Journal of Proteomics and Bioinformatics

OPEN ACCESS Freely available online

Mini Review

Three Essential Factors in Whole-Genome Sequencing for Clinical Applications

Abhishek Narain Singh[*], **Krishan Pal**

Department of Biotechnology, SunRise University, Alwar, Rajasthan, India

ABSTRACT

The advent of next-generation sequencing has now become main stream in clinical practice recently after spending close to a decade and a half in developing the technology and data science software applications for making sense of the technology impact. Various algorithms have been deployed with a mixed reaction in terms of usability as there are benefits and drawbacks in each one of the methods be it related to computational complexity, computational resources to be deployed, or the precision of results in terms of false-positive and false-negative association. This article essentially talks about some of the three essential best practices in the use of next-generation sequencing technology particularly for whole-genome sequencing to clinical use for rare disease diagnostics.

Keywords: Next-generation; Mitochondrial DNA; Whole Genome Sequencing (WGS)

INTRODUCTION

DNA is the blueprint of life and has codified information of chemical reactions that are responsible for our emotions, susceptibility to a disease, predisposition to a disease, or character traits such as height, voice tone with pitch, the color of skin and hair, and eyes and so on. At the same time, these features can vary depending on external factors and the variation would be dependent on the feature itself and the external factor. These physical attributes are termed as 'phenotypes' and in a biomedical relevant context; they would be the 'disease phenotypes'. By and large, the information in the genotype phase does carry a substantial amount of weightage in terms of dictating the disease phenotype of an individual, and cannot be undermined. Article essentially demonstrates how a customized approach can be applied for genome analysis using which the author was able to bring some novel insight to the scientific world such as the possibility of mitochondrial DNA to be paternally inherited as well in humans and the fact that SVs (Structural Variations) can contribute more to the phenotype, such as rare disease, than the Single Nucleotide Polymorphism (SNPs) [1].

LITERATURE REVIEW

The various workflows have limitations in content and constrain overall efficacy. Article recently has tried to address some of the best practices using whole-genome sequencing intended for clinical purposes [2]. Whole Genome Sequencing (WGS) has the advantage over Whole Exome Sequencing (WES), given that now we know that about 98% of the genome does not contain the gene region but regions that have significance in the regulation of expression of genes-in other words, how the gene products are recruited. WGS is thus well-positioned to replace WES, targeted NGS, and microarray techniques. Besides this, the analysis can be periodically repeated if and when needed to update the annotation. There is currently no consensus so far between various laboratories to identify a common set of guidelines for the purpose.

Essential 1st practice for NGS genome coverage: Although a general practice of use of 30 x genome coverage is recommended on average which covers most of the SNPs and SVs to be determined as was also advocated in article [3] suggests that for exhaustive detection of SNPs using the short NGS sequencing and mapping approach we must aim for coverage of 100 x and for larger structural variations a 300 x is recommended. This could be considered as one of the essential elements in translating WGS for clinical practices such as for rare diseases. A lot of this has got to do with the short NGS sequencing signal quality as well as the quality of base call decay gradually and steadily as the read length increases.

The technical and analytical elements of clinical WGS can be separated into three stages: sample preparation, including

Corresponding author: Abhishek Narain Singh, Department of Biotechnology, SunRise University, Alwar, Rajasthan, India, Email: abhishek.narain@iitdalumni.com

Received: 18-May-2022, Manuscript No. JPB-22-17558; **Editor assigned:** 20-May -2022, PreQC No JPB-22-17558 (PQ); **Reviewed:** 03-Jun-2022, QC No. JPB-22-17558; **Revised:**22-Feb-2023, Manuscript No. JPB-22-17558 (R); **Published:** 01-Mar-2023, DOI: 10.35248/0974-276X.23.16.628.

Citation: Singh AN, Pal K (2023) Three Essential Factors in Whole-Genome Sequencing for Clinical Applications. J Proteomics Bioinform. 16.628.

Copyright: © 2023 Singh AN, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

extraction and library preparation followed by sequence generation (1st step); read alignment and variant detection (2nd step); and annotation, filtering, prioritization, variant classification, and case interpretation (3rd step) followed by variant confirmation (final step).

Essential 2nd practice for NGS positive control: A high-quality reference standards materials and a truth dataset are necessary for laboratories or lab companies offering WGS for clinical applications. They should include the publicly available reference standards in addition to the commercial ones and privately held positive control for every variant detected. The National Institute of Standards and Technology (NIST) NA12878 genome and Platinum Genomes are routinely utilized by NGS laboratories seeking to establish WGS analytical validity [4]. These genomes have the advantage of many a thousand variants that have been curated and confirmed across many NGS technologies. Classes of clinically relevant genetic variation detectable by clinical WGS are summarized include Single-Nucleotide Polymorphism (SNPs), Small Deletions and Insertions (InDels), Structural Variation (SV) bases greater than 50, including Copy Number Variation (CNV) and balanced rearrangements, Mitochondrial (MT) variants, and Repeat Expansions (REs). Laboratories may not be able to validate all classes of variation and a phased approach to validation and subsequent test offering may be necessary. The laboratory companies must provide clear test definitions and identify factors affecting reportable variant types to inform the physicians. Many groups have used NA12878 for validation, and many groups also utilize the Ashkenazi Jewish and Chinese ancestry trios from the Personal Genome Project that are available, as reference materials with variant benchmarks. Thus the choice of reference and positive control will assure high confidence in the result and the results can be interpreted by others in the industry as well.

Essential 3rd practice for short NGS for quality and performance metrics: Quality metrics should be applied at every stage wherever possible, and that includes the Quality Control (QC) tools. Right from the point when the raw data is generated, to the point when mapping is done of the reads to the reference and so on, all need to be quality checked. Quality metrics are calculated for every run of the instrument and cluster computing, and after alignment and variant calling. However, it can be challenging for every laboratory to adhere to a single universal number to be assigned as a threshold and there is where individual laboratory companies can have their creativity. Important metrics for passing samples include the total gigabases (Gb;>Q30) produced per sample, the alignment rate of Purity-Filtered Bases (PF reads aligned %), the predicted usable coverage of the genome, proportion of reads that are duplicates, the % call ability, and any evidence of sample contamination. Use of precision is a more recommended metric than specificity, owing to the large number of true negatives expected by clinical WGS. Going as per the Food and Drug Authority (FDA) suggestion is a similar, but slightly different metrics, for validation of NGS assays, including positive percent agreement (PPA which is synonymous with sensitivity), Negative Percent Agreement (NPA) similar to specificity, and Technical Positive Predictive Value (TPPV) equivalent to precision, as well as reporting the 95% confidence interval. Thus, these quality and performance metrics at each step of the workflow can make clinical biomarkers and results more profound and attractive for approval by FDA and acceptable for the patient and their well-wishers. Particularly for SNPs and indels, gold-standard data are available so that repeatability and reproducibility of the workflow can be checked. There is continual updating of software versions and data sources for annotation (e.g., OMIM, ClinVar, etc.). Thus the development, validation, and deployment cycles can be challenging for laboratories. Thus a track of changes such as through Git and version control can be very useful.

DISCUSSION

Workflows such as that offered by nf-core Sarek may or may not adhere to these critical three essential points, and it would be the task of the bioinformatics to look into the vital details of the pipelines and fit in the components that might be needed. Many automated workflows exist that can yield reasonable results, however, for an FDA acceptable result; one must try to meet all these three essential factors as deemed necessary. As an example, a promising next-generation sequencing technology that is now upcoming is Long Read Sequencing (LRS) and this technology comes with its own set of new tools for mapping and variant calling. This technology seems to be more promising for SV detection, particularly long insertions, than the traditional short-read sequencing method for which several benchmarks have been set up. Now, as a clinical bioinformatics, one might have to change the pipeline in case one would like to make use of LRS to advantage. Software tools such as minimal [5,6], and SV detector using sniffles [7]. Can very well be also plugged into a combination of short and long NGS upstream pipelines and for which there might not be an already existing pipeline for the purpose. LRS read generation differs from Short Read Sequencing (SRS) in the sense that the reads are long (several kilobases) and the errors are uniformly distributed. This uniform distribution of error in LRS can be taken to advantage in the sense that we do not have to worry about trimming the ends as in the case of SRS (and so we curtail those sequences to be short). At the moment, even though we might see a shift in the usage of NGS from SRS to LRS, the three essential elements that this paper discusses in the deployment of WGS in clinical practice seem to be staying for good [8].

CONCLUSION

In this paper, we discussed the three essential aspects that should be kept in mind while conducting clinical bioinformatics work using next generation sequencing technology deployment. Good and calculated genome coverage can have a great impact on having a desirable downstream output. Particularly for short NGS sequencing, it would be good to know what coverage on average and its distribution is a good representation to capture most or almost all of the desired class of variant. While this number would greatly decrease if LRS is used, experimental figures exist of what is optimal for various variant classes for SRS, and those numbers have been stated. The use of positive control and right reference sequences, not just GRCh NCBI reference genomes but also the other ones can have a great impact such as for the population one is investigating the clinical impact. Rare diseases by rare variants can be investigated by the NCBI GRCh genome, but even better by comparing it with the reference genome of the population of concern for which some sample genomes have been stated in this paper. Also, the fact that positive control would exist differently for every population of interest can have a profound impact on the clinical outcome as this information is ordered to the physician. At every step of the workflow pipeline, there exist opportunities to assess the quality and various performance metrics. If care is taken to ensure quality checks are taken at every step of the workflow, that can ensure higher confidence in the end consumer and also have possibly higher credibility in the eyes of the FDA for approval of a test or diagnostic, particularly for rare diseases that can be population specific.

ACKNOWLEDGEMENT

Pardeep Singh, Sultan Singh, Ekta Sheoran, Pooja Lamboria, Pankaj Gupta, Jitendra Yadav, Rachana Swami helped in academic and operational support.

AUTHOR'S CONTRIBUTION

Idea was conceived, implemented and the paper was written by the first author.

REFERENCES

1. Singh AN. Customized biomedical informatics. Big Data Anal. 2018;3(1):1-12.
2. Marshall CR, Chowdhury S, Taft RJ. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. Npj Genom Med. 2020;5(1):1-12.
3. AliotoT, Buchhalter I, Derdak S. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6(1):1-13.
4. Rehm HL. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15(9):733–747.
5. Beyter D, Ingimundardottir H, Oddsson A. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021; 53(6): 779–786.
6. Heng Li. Minimap 2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:18 3094–3100.
7. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestadn M. Accurate detection of complex structural variations using single-molecule sequencing. Nat Meth. 2018;15(6):461–468.
8. Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, et al. A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F 1000 Research. 2020;9(63).