Biological Complexity and Clinical Oncology in the Detection of Variants

David Chien*

Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy

Introduction

The ability to interpret the potential effects of genomic changes on a patient's health and how this information can be used to fine-tune personalised therapies represents the most significant of these challenges. To address the biological complexity of cancer, we will discuss various significant technologies, computational techniques, and models that may be used to NGS data ranging from Whole Genome to Targeted Sequencing. In our final section, we'll look at the opportunities and difficulties that bioinformatics for precision medicine will face in the future, both at the molecular and clinical levels, with a particular emphasis on homologous recombination deficiency as a newly discovered biomarker in clinical practise. Numerous potential applications in research and diagnostic settings include sequence variation detection, epigenetic and transcriptional regulation, chromatin conformation, its 3D architecture, and the way these phenomena interact with one another. This process, applied to a large number of loci per subject, generates an enormous amount of sequence data. Variant identification and its preprocessing phases will be the main subjects of this review. Small DNA or RNA fragments are aligned using either a reference genome sequence that has already been assembled or by concatenating input reads using "*de novo*" techniques, which employ no reference sequence at all.

For genomic study at various scales, from a few genes to the entire genome, various methodologies are optimized. Whole genome sequencing, which involves more time and money but covers the entire genome, is used to look at previously undiscovered genetic changes. Since protein-associated mutations frequently have a negative effect on genome regulation, whole genome sequences may only cover protein-coding genes, which make up 3% of the entire genome, but at lower cost. However, the utility of WES in clinical research and practise is considerably constrained by the difficulty of data interpretation. Targeted sequencing was invented as a result to examine specific mutational hotspots for a given genome. Utilizing this method, disease-causing genomic changes with known or suspected pathogenicity are found. Sample pretreatment, library preparation, sequencing, and bioinformatics analysis are a some of the stages that make up a typical NGS workflow. Each stage is important and may conceal sources of inaccuracy that could affect the outcome. The following stages are used to carry out the well-established techniques for variation detection and annotation at the moment.

Description

Biomedical machine learning techniques became widely used as high throughput sequencing technology developed. This led to the development of novel patient stratification techniques, diagnostic tools, and medication discovery methodologies [1]. In addition to the substantial amount of publicly available NGS data from consortiums. The accessibility of inexpensive sequencing technologies led to a global increase in in-house data and, as a result, to an increase in demand for ML-based software that is computationally more efficient, accurate, and reusable [2]. Modeling how genetic variations and their interactions affect cell growth and fate, leading to cancer transformation, is one of the most difficult tasks faced by ML in clinical genomics. Despite the fact that traditional inference techniques can be very adaptable and interpretable in terms of causation, they are frequently limited by linearity or model-based assumptions, offering aggregate. Conventional techniques, such as GATK, which heavily rely on different statistical models and heuristics based on calling accuracy, allelic, and sequencing coverage, are used to predict the likelihood of variation for each genomic locus [3]. This work is significantly hampered by the presence of sequencing artefacts that are only partially controllable, such as those brought on by low-complexity and repeated genomic sequences, DNA synthesis dephasing and efficiency, and polymerase chain reaction errors.

Furthermore, sequencing data is inherently high-dimensional; several combinations of genomic changes may result in the same phenotype, and usually, only a tiny percentage of persons possess the variant that is associated with the observed disease. Deep neural networks have been widely utilised for variant finding because of their capacity to represent a very high number of characteristics and parameters. Convolutional neural networks (CNNs)' fundamental principle is to transform collections of aligned data into patterns of an image, producing clusters of interconnected variants that could have a pathogenic effect. The general-purpose programme Deep Variant, the specialist Clairvoyant, NeuSomatic is examples of CNN-based algorithms [4]. These algorithms were created for single-molecule technologies, somatic variations, and structural variants, respectively. Using ensemble approaches, where the learning process makes use of various integrated models, gains in predicted accuracy have frequently been made. To find SNV and indels from short-read sequencing data, CNNScoreVariants, for instance, uses pre-trained models in GATK. Possible information biases in their training sets are one of the key limitations of many deep learning techniques. Finding genomic regions that are causally linked to the breakdown of one or more biological activities and pathways is the aim of variant discovery [5]. Pathogenic changes to coding DNA are considerably simpler to link to a disease phenotype because they are more likely to change the structure and function of the encoded protein. As a result, the majority of clinical and cancer genomics diagnostic techniques are either based on panels of a small number of exons or WES. However, noncoding variations that are likely to be found at regulatory elements are responsible for many of the harmful characteristics of disease.

Conclusion

In conclusion, we describe some important technologies, computational algorithms and models that can be applied to NGS data from Whole Genome to Targeted Sequencing, to address the problem of finding complex cancer-associated biomarkers. In addition, we explore the future perspectives and challenges faced by bioinformatics for precision medicine both at a molecular and clinical level, with a focus on an emerging complex biomarker such as homologous recombination deficiency.

Conflicts of Interest

None.

Address for Correspondence: *David Chien, Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy, E-mail: dchien78@gmail.com*

References

1. Rizzo, Alessandro, Angela Dalia Ricci and Giovanni Brandi, et al. "PD-L1, TMB, MSI, and Other Predictors of Response to Immune Checkpoint Inhibitors in Biliary Tract Cancer." *Cancers* 13 (2021): 558.

2. Vestergaard, Lau K., Douglas N.P. Oliveira, Tim S. Poulsen and Claus K. Høgdall, et al. "Oncomine comprehensive assay v3 *vs.* oncomine comprehensive assay plus." *Cancers* 13 (2021): 5230.

3. Dehghani, Mehdi, Kevin P. Rosenblatt, Lei Li and Mrudula Rakhade, et al. "Validation and clinical applications of a comprehensive next generation sequencing system for molecular characterization of solid cancer tissues." *Front Mol Biosci* 6 (2019): 82.

4. Park, Jay J.H., Grace Hsu, Ellie G. Siden and Kristian Thorlund, et al. "An overview of precision oncology basket and umbrella trials for clinicians." *CA Cancer J Clin* 70 (2020): 125-137.

5. Prelaj, Arsela, Rebecca Tay, Roberto Ferrara and Nathalie Chaput, et al. "Predictive biomarkers of response for immune checkpoint inhibitors in non-small-cell lung cancer." *Eur J Cancer* 106 (2019): 144-159.

How to cite this article: Chien, David. "Biological Complexity and Clinical Oncology in the Detection of Variants." J Oncol Med & Pract 7 (2022): 161.