Mian et al., J Comput Sci Syst Biol 2019, 12:2
DOI: 10.4172/0974-7230.1000299

Journal of
Computer Science & Systems Biology

Research Article Open Access

Automating the Computational Analysis of Exome Sequencing Data: A Prototype Methodology to Overcome Bottlenecks Observed with Operator Driven Clinical Interpretation for Known Pathogenic Mutations

Shahid Mian[1,2*], Wafaa Al-Turaif[1], Abdullah Al-Nawfal[1], Mohammed Mudhish[1], Eissa Faqeih[3] and Manar Samman[2]

[1]*Molecular Pathology, Pathology and Clinical Laboratory Medicine, Division of Bioinformatics, King Fahad Medical City, Riyadh 11525, Saudi Arabia*
[2]*Division of Bioinformatics, Molecular Pathology, Pathology and Clinical Laboratory Medicine, King Fahad Medical City, Riyadh 11525, Saudi Arabia*
[3]*Clinical Research Department, Children's Hospital, King Fahad Medical City, Riyadh 11525, Saudi Arabia*

Abstract

Objective: Clinical exome sequencing produces between 90,000-100,00 variants per individual. Bottlenecks are manifested due to manual (operator based) interpretation of data. Given an increasing demand for genomic screening, automated computational methodologies are urgently required to meet both throughput and interpretation. Objective: determine if algorithms can be developed to identify and report specific pathogenic variants.

Methods: Clinical exome sequencing was performed on 961 individuals presented for diagnostic analysis to King Fahad Medical City (KFMC). Variant Call Format (VCF 4.2) files from each patient were used for algorithm development. Perl (v5.28.1) was used as the construct language. 137 known pathogenic variants were used as a search test bed. A 10-step procedural workflow was implemented to automate the process of searching for targets. Where a positive identification was elicited, variants were annotated, merged with clinical data and output as a pdf report. Negative findings were output as a pdf report with clinical data only.

Results: 961 VCF files were screened for 137 pathogenic variants of interest to KFMC. Target variants were compared against each variant within a patient's VCF using logic operators. A total processing time including report production for 961 individuals was completed in 11.38 hours. 177 patients (18.4%) were positive for at least one variant and 15 patients had two variants (1.6%). All positive cases were verified manually in the originating VCF. The 137-target list of variants were "spiked" into a negative control patient VCF to act as a positive control (sensitivity). All variants were detected by the algorithm. 10 negative finding patients were chosen at random and manually checked for the absence (specificity) of the 137 variants. No variants were detected.

Conclusion: Automated searching and production of reports for specific pathogenic variants using computational searching is feasible for diagnostic laboratories undertaking clinical exome sequencing.

Keywords: Diagnostics; Bioinformatics; Algorithm; Exome; Screening; Clinical; Automation

Introduction

The advent of high throughput Next Generation Sequencing technologies (NGS) has facilitated a paradigm change for clinical diagnostic laboratories and translational genomics. The ability to rapidly survey the genomes of patients with suspected inherited germline disorders, identify causal pathogenic variants and report them to the treating physician has become standard practice. The American College of Medical Genetics and Genomics routinely publishes guidelines on how to standardise the classification of variants into five groups based upon the likely severity of the mutation upon protein function [1]. The organisation also provides an evidenced based rationale for clinical laboratories to screen specific genes for deleterious pathogenic variants not directly linked to the patient's phenotype (under the "secondary findings" screening initiative) given that early medical intervention can lead to improved healthcare outcomes for these genetic classes [2]. Continued developments in sensitivity (quality scores and depth of sequence coverage) in addition to breadth of genomic interrogation (whole exome sequencing (WES); whole genome sequencing (WGS); targeted panels) serve to enhance the diagnostic rate for inherited conditions [3,4] including those disorders targeted within pre-natal screening programmes [5].

While the potential of DNA sequencing remains significant for clinical laboratories, there are still technological challenges that must be addressed. For example, NGS elicits tens of thousands of genetic variations during comparison to a standardised reference genome even though an exceptionally small number are likely to be causal for a patient's clinical disorder [6]. Limitations still exist in bioinformatic pipelines used to identify pathogenic variants elicited from WES/WGS analysis. Standard practice is the use of "truth sets" (i.e., high confidence variants) that can be used for statistical measures (e.g., sensitivity, specificity, positive predictive value, negative predictive value). Algorithms therefore benchmark against truth sets in order to measure their performance levels. Challenges are still pervasive however as exemplified by Krushe et al. who have shown that single-nucleotide variant concordance using two different computational pipelines is 99.7% within high-confidence regions compared to 76.5% outside of these regions [7]. This may suggest that even the use of "truth sets" could lead to systematic bias during the establishment of threshold settings in the core alignment/variant calling algorithms.

***Corresponding author:** Mian S, Division of Bioinformatics, Molecular Pathology, Pathology and Clinical Laboratory Medicine, King Fahad Medical City, Riyadh 11525, Saudi Arabia, Tel: 00966539417360; E-mail: smian@kfmc.med.sa

Received April 08, 2019; **Accepted** April 20, 2019; **Published** April 27, 2019

Citation: Mian S, Al-Turaif W, Al-Nawfal A, Mudhish M, Faqeih E, et al. (2019) Automating the Computational Analysis of Exome Sequencing Data: A Prototype Methodology to Overcome Bottlenecks Observed with Operator Driven Clinical Interpretation for Known Pathogenic Mutations. J Comput Sci Syst Biol 12: 47-52. doi:10.4172/0974-7230.1000299

Total number of patients	177
Total number of variants detected	192
Number of patients with variants concordant to VCF	177
Patient zygosity confirmed in VCF files	177
Number of patients with one variant detected	162
Zygosity: Heterozygous **Homozygous**	112 50
Number of patients with two variants detected	15
Dual variants: Dual heterozygous **Dual homozygous** **Single heterozygous and single homozygous**	7 3 5

Table 3: Results of variant identification for all 177 patients positively identified with a target variant

[a] Total number of patients. This value indicates the number of patients from the cohort of 961 patients in which a target variant was identified

[b] Total number of variant detections. This value indicates the total number of times any of the 137 target variants were detected in all patient VCF files. The value is higher than the total number of patients given that a single patient could have more than one variant identified

[c] Number of patients with variants concordant to VCF. Variants detected by the automated coding were manually checked in the patient VCF file. The value shown relates to the number of patients where concordance occurred between the VCF and output result

[d] Patient zygosity confirmed in VCF files. The VCF file records the zygosity (homozygous or heterozygous) for each variant. The output zygosity for each variant detected by the automated coding was manually checked against the original VCF file for each patient. The reported value indicates the number of patients where variant zygosity was concordant even for those patients where more than one variant detected

[e] Number of patients with one variant detected. This value shows the number of patients in which only one variant (homozygous or heterozygous) was detected

[f] Zygosity. From the 177 patients where a variant was detected the zygosity (homozygous or heterozygous) is presented

[g] Number of patients with two variants detected. As indicated, this is the number of patients (from the total of 177 where a variant was identified) in which two variants were detected

[h] Dual variants. These values indicate the zygosity for patients in which two variants were detected. Three patients were homozygous for both variants and thus represent a "dual diagnosis" condition

As clinical genomics becomes more widespread, personalising healthcare through genomically guided therapeutics and management protocols will be at the cornerstone of this change. Computational algorithms will be requisite at every stage of this rapidly developing field.

Conclusions

Automated computational algorithms can expedite the discovery process for known targeted variants of clinical significance and automate reporting without user intervention. This translates to reduced turn-around-time for diagnostic laboratories without impacting the quality of service delivery.

Author Contributions

SM developed the clinical bioinformatics pipeline for NGS, produced patient VCF files used within this study, developed and tested the Perl source code and constructed the manuscript. WT, AN and MM tested the NGS bioinformatics pipeline, produced patient VCF files, developed patient databases containing clinical information and confirmed accuracy of the algorithm's output reports. EF provided the team with a list of pathogenic variants from the Saudi population as a test bed and reviewed the manuscript. MS collaborated with EF to refine the variant list used in this study and reviewed the manuscript.

Acknowledgements

The authors would like to acknowledge the continued support from Dr Musa Faqeih (Director of Pathology and Clinical Laboratory Medicine Administration at KFMC) without whom this work would not be possible.

References

1. Richards S, Aziz N, Bale S, Bick D, Das S, et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17: 405-424.
2. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, et al. (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med 19: 249-255.
3. Al-Dewik N, Mohd H, Al-Mureikhi M, Ali R, Al-Mesaifri F, et al. (2019) Clinical exome sequencing in 509 Middle Eastern families with suspected Mendelian diseases: The Qatari experience. American Journal of Medical Genetics Part A.
4. Calpena E, Hervieu A, Kaserer T, Swagemakers SMA, Goos JAC, et al. (2019) De Novo Missense Substitutions in the Gene Encoding CDK8, a Regulator of the Mediator Complex, Cause a Syndromic Developmental Disorder. American Journal of Human Genetics.
5. de Koning MA, Haak MC, Adama van Scheltema PN, Peeters-Scholte C, Koopmann TT, et al. (2019) From diagnostic yield to clinical impact: a pilot study on the implementation of prenatal exome sequencing in routine care. Genetics in Medicine: Official Journal of the American College of Medical Genetics.
6. Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, et al. (2016) VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics 17: 444.
7. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, et al. (2019) Best practices for benchmarking germline small-variant calls in human genomes. Nature Biotechnology.
8. Hansen AT, Bernth Jensen JM, Hvas AM, Christiansen M (2018) The genetic component of preeclampsia: A whole-exome sequencing study. PloS ONE 13: e0197217.
9. Flygare S, Hernandez EJ, Phan L, Moore B, Li M, et al. (2018) The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool. BMC Bioinformatics 19: 57.
10. Landrum MJ, Kattman BL (2018) ClinVar at five years: Delivering on the promise. Human Mutation 39: 1623-1630.
11. de Andrade KC, Frone MN, Wegman-Ostrosky T, Khincha PP, Kim J, et al. (2019) Variable population prevalence estimates of germline TP53 variants: A gnomAD-based analysis. Human Mutation 40: 97-105.
12. Rim JH, Lee JS, Jung J, Lee JH, Lee ST, et al. (2019) Systematic evaluation of gene variants linked to hearing loss based on allele frequency threshold and filtering allele frequency. Scientific Reports 9: 4583.
13. Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, et al. (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research 47: D1018-D1027.
14. Salmon LB, Orenstein N, Markus-Bustani K, Ruhrman-Shahar N, Kilim Y, et al. (2018) Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. Genetics in Medicine: Official Journal of the American College of Medical Genetics.
15. Reisberg S, Krebs K, Lepamets M, Kals M, Magi R, et al. (2018) Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. Genetics in Medicine: Official Journal of the American College of Medical Genetics.
16. Li J, Gao K, Yan H, Xiangwei W, Liu N, et al. (2019) Reanalysis of whole exome sequencing data in patients with epilepsy and intellectual disability/mental retardation. Gene.
17. Ewans LJ, Schofield D, Shrestha R, Zhu Y, Gayevskiy V, et al. (2018) Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genetics in Medicine: Official Journal of the American College of Medical Genetics 20: 1564-1574.
18. Al-Nabhani M, Al-Rashdi S, Al-Murshedi F, Al-Kindi A, Al-Thihli K, et al. (2018) Reanalysis of exome sequencing data of intellectual disability samples: Yields and benefits. Clinical Genetics 94: 495-501.
19. Shashi V, Schoch K, Spillmann R, Cope H, Tan QK, et al. (2019) A comprehensive iterative approach is highly effective in diagnosing individuals who are exome negative. Genetics in Medicine: Official Journal of the American College of Medical Genetics 21: 161-172.

20. Al-Murshedi F, Meftah D, Scott P (2019) Underdiagnoses resulting from variant misinterpretation: Time for systematic reanalysis of whole exome data? European Journal of Medical Genetics 62: 39-43.

21. Baker SW, Murrell JR, Nesbitt AI, Pechter KB, Balciuniene J, et al. (2019) Automated Clinical Exome Reanalysis Reveals Novel Diagnoses. The Journal of Molecular Diagnostics 21: 38-48.

22. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851-1858.

23. Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, et al. (2017) The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. Human Genetics 136: 921-939.

24. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, et al. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics 48: 1581-1586.