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]Division of Bioinformatics, Molecular Pathology, Pathology and Clinical Laboratory Medicine, King Fahad Medical City, Riyadh 11525, Saudi Arabia
[2]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 onl .

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 22, 2019; **Published** April 29, 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

Limitations not only exist with respect to the production of the Variant Call Format (VCF) file for each patient as indicated above but also the subsequent downstream clinical interpretation of the variants themselves using software packages [8,9]. This phase of the service is conducted by clinical reporting teams who exploit both public [10,11] and proprietary [12] knowledgebases in order to identify known (or candidate) pathogenic variants using ACMG classification rules. Deleterious changes in genetic loci result in negative impact upon downstream biological pathways to which they are critically associated. It is through such associations that the phenotype (using standard ontologies [13]) of the patient can be linked directly to a specific genetic cause and thus elicit a clinical diagnosis [14]. The clinical reporter must therefore bring a diverse array of information together in an effective and co-ordinated manner to identify causal pathogenic variants for inherited genetic disorders. The transfer of medical records into a digitised format is greatly facilitating the construction of large-scale datasets enabling genomic (and other laboratory test) data to be associated with particular clinical phenotypes. This is an essential step for personalised medicine (i.e., directing clinical management predicted upon defined genetic signatures) although it does not come without challenges regarding the management, curation and mining of information repositories. Automated algorithms capable of routine feature extraction are being developed to produce insights to assist in making pharmacogenomic recommendations for patients [15].

While exome analysis is making significant enhancements to the field of genomic medicine, it is also producing concomitant challenges for clinical laboratories with respect to service provision. For example, a patient VCF record can be reviewed several times if reported negative by the clinical laboratory. Negative reporting is multifactorial and can include for example insufficient evidence at the time of assessment regarding a variant's pathogenic nature or even the fact that a laboratory has not updated its reference databases at the time of reporting [16]. For these reasons re-analysis of the data is routinely requested by ordering physicians in the anticipation that a positive causal variant can be identified. Recent publications suggest a period of 6–12 months might be suitable for a re-evaluation of historical data [16,17]. Several reports have indicated that re-analysis is producing dividends for the identification of pathogenic variants and thereby enabling a positive diagnosis to be made for the patient [18-20]. Re-analysis service requests by physicians do however have a potentially negative impact for diagnostic laboratories with respect to: a) increased workload for personnel having to deal with these cases and b) reduced time spent by the reporting team on each new sample entering the laboratory. As the costs of DNA sequencing continue to reduce, it is inevitable that the volume of diagnostic requests for clinical exome will rise commensurately. This will lead to stress points regarding both sample throughput and even the possibility that quality of reporting could be negatively impacted.

Given the bottlenecks outlined above (current and impending), clinical laboratories are now actively seeking solutions using computational algorithms in order to facilitate both variant screening and patient reporting. Such approaches will be at the core of this necessitated paradigm change from manual clinical reporting to more automated methodologies. Reports are now emerging from laboratories who have developed and implemented computational algorithms to support their clinical operations [21].

The results described within this manuscript report the development of computational algorithms capable of screening a cohort of VCF files derived from 961 clinical cases. The code (written in Perl 5) was designed to identify specific pathogenic variants of interest and to have the findings (whether negative or positive) exported into a fully annotated clinical report. Such methodologies will provide a key step towards enhancing the throughput of both primary patient exome reporting, meeting increased volume requirements due to cost reductions in genome sequencing and facilitate the reanalysis of historical negative cases.

Methodology

The Department of Pathology and Clinical Laboratory Medicine at King Fahad Medical City (KFMC) provides a full clinical exome sequencing and reporting service for patients with suspected inherited disorders. The laboratory is College of American Pathologists (CAP) accredited (7538102 AU-ID1607501). Institutional Review Board approval to conduct this retrospective study was obtained (IRB: 16-085 and IRB: 16-247b).

Genomic DNA from each patient was extracted and quantitated using a Qubit Fluorimeter (Agilent). 1 ug of DNA was used for library preparation (Agilent SureSelect v5 AllExon 50 MB kit) and the constructs were sequenced on an Illumina 4000 at an average coverage of 150x. Quality control was conducted using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) with alignment and variant calling being conducted using commercially available software (DNASTAR, USA) predicated upon the MAQ alignment and variant calling algorithm [22]. 961 VCFv4.2 (https://samtools.github.io/hts-specs/VCFv4.2.pdf) files were taken for algorithm development and testing.

137 variants observed within KFMC patients and/or the wider Saudi patient population were used as a target panel for this proof-of-concept study. Variants were only considered when their quality score, $Q=-10\ log_{10}\ P$ was 20 or higher and this was built into the coding during variant selection and extraction as part of the re-formatting of the original VCFv4.2 file.

Development of the source code was carried out using Geany IDE 1.34.1 (https://www.geany.org/). Scripts were written in Perl 5 using Strawberry Perl (64-bit) 5.28.1.1-64 bit (http://strawberryperl.com/) with modules being downloaded from the CPAN (Comprehensive Perl Archive Network https://www.cpan.org/). The operating system was Windows 10 Pro. The coding was modularised as follows: STDIN requests from the user for the pathlength to: a) the directory containing patient VCF files and b) the csv file encompassing clinical information. The VCF file was formatted so that variant information only contained the following data values:

(i) Chromosome, (ii) Position, (iii) Reference base, (iv) Alternate base, (v) Zygosity.

True (input scalar values=non-zero) and false (input scalar values=zero) were used to annotate the clinical report as to whether a pathological variant had/had not been identified respectively. Regular expressions (regex) linking patient medical record number (MRN) from the variant file to the identical MRN within the clinical data were utilised in order to combine variant results to medical information. Annotation of variants (e.g., NM_code; amino acid change; clinical disorder) was achieved using standard key: value (hash) combinations. PDF clinical reports were produced from the resulting data files containing both medical information and any (annotated) variant identified within a patient VCF file.

Hardware specification: Quad core (Intel i7-6700k CPU @4.0 GHz);

64 GB RAM; 4 TB SSD (Samsung EVO 850); EVOC High Performance Systems chassis (HIDevolution, USA).

Discussion

Clinical exome sequencing was performed on a consecutive series of patients/families whom presented to King Fahad Medical City (KFMC) for diagnosis over a two-year period. The cohort of 961 cases included 25 families (80 individuals consisting of 25 probands and 55 parents/affected siblings). Summary statistics regarding age, gender and reason for referral (for a subset of conditions but does not represent an exhaustive list) are presented in Tables 1a-1c. The sample set was considered a representative test-bed in which to undertake computational analysis given the breadth of clinical indications presented to KFMC for genetic analysis during the course of two years.

Each individual had their exome sequenced to an average coverage of 150x. Raw fastq files were quality controlled using FastQC, aligned against GRCh37.p13 and genotyped using MAQ [22] according the laboratory's standard operating procedure. VCFv4.2 files derived from each exome sequencing analysis had an average of 98,177 variants (range 89,216–114,476) per patient.

To test the ability of computational algorithms to search for variants of interest within a series of VCFv4.2 files, a Perl 5 script was constructed (see below) to identify a group of 137 variants. These targets were chosen as a proof-of-principle panel given that many variants have been routinely reported as pathogenic both at KFMC and/or external organisations nationally and internationally. Variants included both SNV (single nucleotide variants n=106) and indel (insertion/deletion n=31) producing a total of 137 targets in which to screen for. One or more of the 137 variants was positioned in each of the 22 human autosomes (i.e., chromosomes 1-22) in order to reduce the risk of potential systematic bias of variant discovery being chromosome dependent by the search algorithms.

Performance characteristics of the algorithm were initiated by sensitivity testing i.e., confirming its ability to detect all variants derived from the panel of 137 targets (true positives). One patient VCF was selected at random (negative for all 137 target panel variants) and "spiked" with the full list of 137 variant target co-ordinate search strings (chromosome number, position, reference base, alternative base) for each target variant. The variants were added as a single group within the patient VCF and the algorithm applied to search the VCF. All 137 targets were successfully detected across all 22 chromosomes (data not shown). To confirm that the algorithm could detect all target variants no matter where their location within a patient VCF file, the panel of 137 targets were sort-ordered based upon their chromosome number followed by base position within each respective chromosome. All 137 variants were correctly identified suggesting that detection of a variant was not biased towards the order in which a target is presented to the algorithm nor which chromosome location the target variant was positioned (data not shown). Regarding specificity of detection (i.e., true negative VCF files), 10 patient VCF files (reported as negative by the search algorithm during screening) were searched manually for all 137 target variants. No variants were detected within these 10 patient VCF files suggesting that the algorithm is specific for the target variant search strings.

Given an average of 98,177 variants per individual, the number of comparative analyses needed to be completed by the search algorithm

	Min (Years)	Max (Years)	Average (Years)	Standard Deviation (Years)	Number of Individuals
Age	0	66	8	11	961

Table 1a: Age distribution of individuals submitted for clinical exome analysis. The average is 8 years old due to the fact that the predominating cases KFMC treats are paediatric patients.

Gender	Total	Percent (%) of Total
Female	443	46
Male	497	52
Unknown	21	2

Table 1b: Gender distribution of patients used within the analysis. Unknown cases are patients whom died before confirmation of gender could be established

Clinical indications of patients submitted for sequencing
Epilepsy
Mitochondrial disorders
Developmental delay and intellectual disability
Metabolic disorders
Muscle disorders
Hydrops fetalis
Skeletal dysmorphic features
Retinitis pigmentosa
Ataxia

Table 1c: A representative sample of phenotypes listed by physicians at the time of patient clinical presentation.

for 961 VCF files is equivalent to approximately 12.9 billion (98,177 patient variants × 137 target variants × 961 total individuals). To explore the feasibility of automating variant identification and reporting, a script was constructed in Perl 5 (Strawberry Perl 5.28.1.1-64 bit environment/package manager). The proof-of-concept script was designed to address the following limitations:

i) Variant identification from using a pre-defined panel,

ii) Consistency for detection,

iii) Annotation (e.g., NM code; amino acid change; genetic disorder),

iv) Production of clinical reports.

In order to achieve these goals, a 10-stage automated process preceded by two user input requests was implemented (Table 2):

(i) User request 1: The user is requested to input the pathlength of the folder containing the VCF files that will be searched against using the variant panel.

(ii) User request 2: The second user is requested is to input the pathlength of the file (.csv) containing the clinical information that will be used to populate the clinical report (e.g., ordering physician; patient medical record number; patient name; gender etc.). This file is formatted in a manner to allow specific column data to be extracted by the coding and to populate specific fields of the final pdf report for each individual patient.

Once the user has input the required pathlengths, the algorithm proceeds in a fully automated manner without any further input from the clinical reporting team. This consequently minimises the amount of direct hands-on-time required by laboratory staff. As each phase of the pipeline is completed (Stage 1-Stage 10) a notification is sent to the user via the STDOUT of the terminal keeping the operator informed of how the analysis is proceeding.

Table 2 highlights the key aspects of the methodological workflow and the time taken to complete each phase with respect to variant identification, annotation and production of the final clinical report. The run takes approximately 11.38 hours (real-time) to complete the process of searching 961 VCFv4.2 files for 137 target variants which equates to the production of 1.4 clinical reports per minute. The physical hardware resources used during processing were considered to be minimal with approximately 15%-20% CPU demand and less than 10 GB RAM (data not shown). The physical file size for 961 VCFv4.2 files was 7.2 GB hard disc space and a total output file size (temporary files and final pdf reports) of 3.6 GB hard disc space. As a result, these authors would contend that implementing scripts such as the one described in this manuscript could be easily done within any clinical laboratory using minimal computing resources.

Once the computational pipeline was completed, a total 961 pdf reports had been produced containing full clinical information pertaining to the patient in question and any variant findings (Figure 1 presents the first two-pages of an example output report). A built-in script simultaneously exports to the STDOUT and a logfile text report for any patient where more than one variant (either in the heterozygous or homozygous state) is identified.

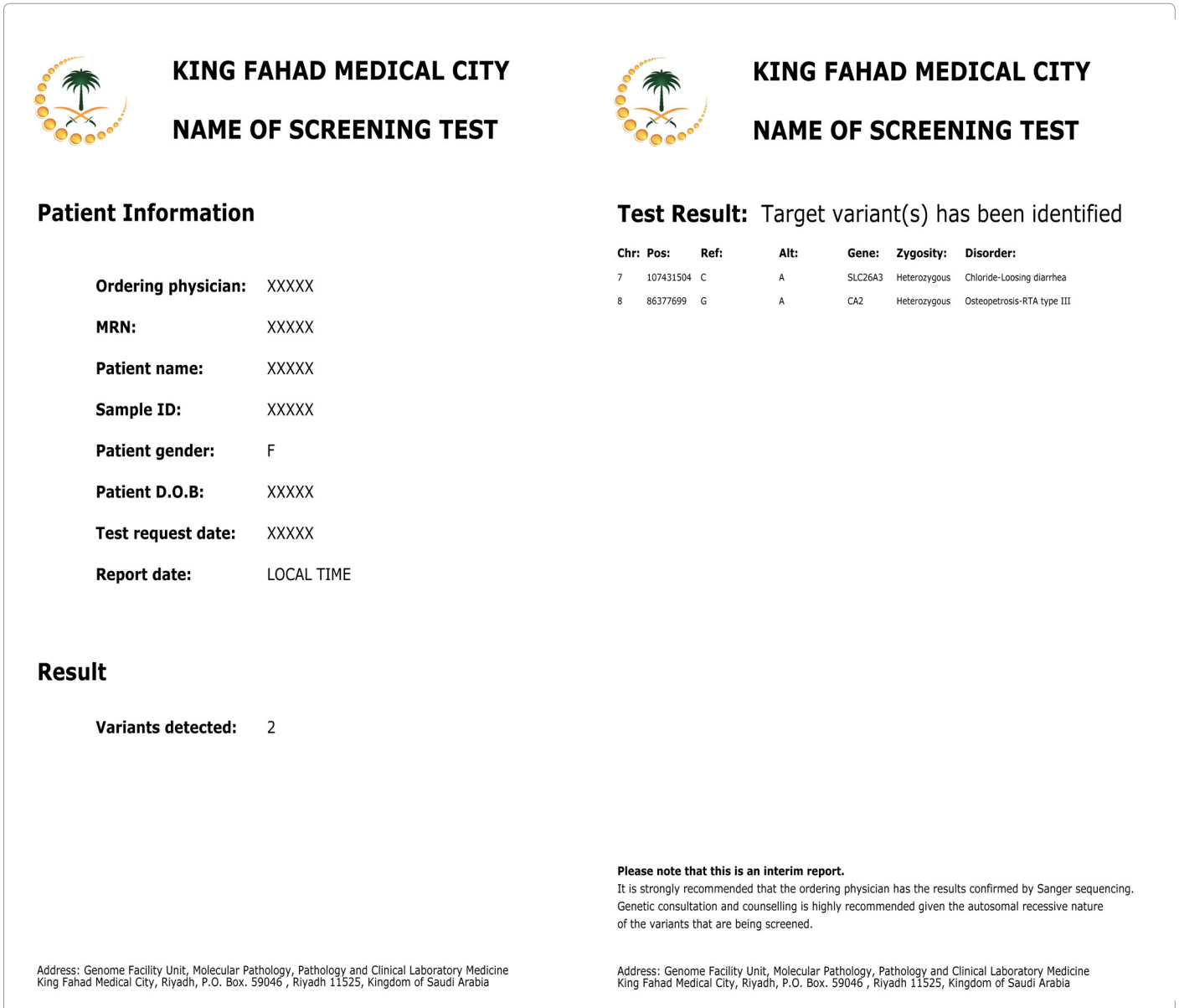
Table 3 shows the results derived from the variant identification and annotation process for 177 patients (18.4% of the total population) where a positive detection occurred for at least one variant. 15 patients were identified with two variants (1.6% of the total population). In order to confirm accuracy with respect to the specific variant identified by the coding and its zygosity (i.e., heterozygous or homozygous), all 177 VCFv4.2 files for each patient were manually checked with respect to correct chromosome number, position, reference base, alternate base and zygosity. All 177 results were concordant with the output report of the algorithm even for patients where more than one variant was detected.

From the total panel of 137 target variants, 55 (40%) were detected within the VCFv4.2 files of the 961 patients while 82 variants from the panel were not detected. Three patients were noted to be homozygous for two pathogenic variants (i.e., the patient was diagnosed with two inherited disorders). Dual and higher order diagnosis in a single patient are not uncommon in highly consanguineous populations such as Saudi Arabia [23].

In summary this proof-of-concept study indicates that computational algorithms can be used to good effect to assist clinical genomic laboratories in the identification, annotation and production of patient clinical reports for pre-defined panels of variants. This study would also suggest that sensitivity and specificity of detection are high. Due to minimal hardware requirements these algorithms can be easily deployed in hospital/commercial laboratory environments with minimum specification to analyse the VCFv4.2 file of a single or batch of patients. This study has presented data to show that a single individual VCFv4.2 file containing on average 98,000 variants can easily be screened against a set of 137 pre-defined targets and where a positive finding(s) is made, annotated (through hash arrays) and reported at a

Automation pipeline			
Stage	**Procedure**	**Manual/Automated**	**Time (hours)**
1	User enters path length for the directory containing VCF Files	Manual	N/A
2	User enters path length for the file containing patient clinical dat	Manual	N/A
3	Read directory containing VCF file	Automated	0.02
4	Trimming of VCF fil	Automated	0.12
5	Read directory containing trimmed VCF file	Automated	0.12
6	Identify target variants in patient VCF file	Automated	11.00
7	Read directory containing list of files with found target variant	Automated	0.02
8	Convert files to csv forma	Automated	0.02
9	Read directory containing list of csv file	Automated	0.02
10	Annotate variants	Automated	0.02
11	Merge text file containing patient data with variant resul	Automated	0.02
12	Produce clinical reports	Automated	0.05
		Total time	**11.38**
Average number of reports produced per minute=1.4			

Table 2: An overview of the automated pipeline to search patient VCF files (n=961) for 137 target variants and output the findings in a physician ready clinical repor



KING FAHAD MEDICAL CITY

NAME OF SCREENING TEST

Patient Information

Ordering physician:	XXXXX
MRN:	XXXXX
Patient name:	XXXXX
Sample ID:	XXXXX
Patient gender:	F
Patient D.O.B:	XXXXX
Test request date:	XXXXX
Report date:	LOCAL TIME

Result

Variants detected: 2

Address: Genome Facility Unit, Molecular Pathology, Pathology and Clinical Laboratory Medicine
King Fahad Medical City, Riyadh, P.O. Box. 59046 , Riyadh 11525, Kingdom of Saudi Arabia

(A)

KING FAHAD MEDICAL CITY

NAME OF SCREENING TEST

Test Result: Target variant(s) has been identified

Chr:	Pos:	Ref:	Alt:	Gene:	Zygosity:	Disorder:
7	107431504	C	A	SLC26A3	Heterozygous	Chloride-Loosing diarrhea
8	86377699	G	A	CA2	Heterozygous	Osteopetrosis-RTA type III

Please note that this is an interim report.
It is strongly recommended that the ordering physician has the results confirmed by Sanger sequencing. Genetic consultation and counselling is highly recommended given the autosomal recessive nature of the variants that are being screened.

Address: Genome Facility Unit, Molecular Pathology, Pathology and Clinical Laboratory Medicine
King Fahad Medical City, Riyadh, P.O. Box. 59046 , Riyadh 11525, Kingdom of Saudi Arabia

(B)

Figure 1: An example output of an automated pdf report. Part A contains the clinical information relating to each patient (populated from the user specified csv file) and includes the number of variants that have been identified during the screening process. Part B contains information relating the variants (if any) that have been identified during the screening process including gene co-ordinates, ref/alt bases, the zygosity in which the variant occurs and a finally a brief description of the clinical disorder that occurs if the patient is negatively impacted by a deleterious pathogenic variant.

rate of 1.4 clinical reports per minute. Such approaches will provide potential solutions capable of meeting the following requirements of most accredited laboratories offering genomic services. These include:

a) Throughput (e.g., screening for defined panels of variants with report production),

b) Accurate detection (high sensitivity and specificity),

c) Consistent quality of reporting by using hash arrays to annotate findings,

d) Expand the repertoire of targets screened by adding new variants to the existing list.

As noted from point d) above, it has not escaped the attention of these authors that increased numbers of variants can be easily added to the target panel that are of clinical relevance to the laboratory and also physicians using the laboratory's services. With this goal in mind studies have already been initiated by this team to explore the use of creating more complex coding/modelling involving successive rounds of logic operators (analogous to decision tree algorithms) in order to discover pathogenic variants that are not "hard coded" but meet specific criteria (for example loss-of-function mutations through the introduction frameshift, premature stop-sites or loss of start-sites). In this manner it will be possible to exploit the use of publicly available databases (e.g., OMIM https://www.omim.org/; ClinVar https://www.ncbi.nlm.nih.gov/clinvar/); Mendelian Clinically Applicable Pathogenicity (M-CAP) Score http://bejerano.stanford.edu/mcap/ [24]) to create decision tree/artificial neural network models capable of making accurate variant classifications concordant with ACMG requirements.

[a]Total number of patients	177
[b]Total number of variants detected	192
[c]Number of patients with variants concordant to VCF	177
[d]Patient zygosity con irmed in VCF iles	177
[e]Number of patients with one variant detected	162
[f]Zygosity: Heterozygous Homozygous	112 50
[g]Number of patients with two variants detected	15
[h]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 identifie

[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 identifie

[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 frequenc . 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: 495-50 .
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. 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.