Aspects of Molecular Medicine 6 (2025) 100097

Contents lists available at ScienceDirect

Aspects of Molecular Medicine

journal homepage: www.journals.elsevier.com/aspects-of-molecular-medicine

A cautionary note on genomic complexity and bioinformatic analysis in cancer

Victoria A. Patten[a,1], Hocine Bendou[b], Denver T. Hendricks[a], Christopher G. Mathew[c,d], Wenlong Carl Chen[d,e,f,g], Joanna C. Fowler[h], Roshan K. Sood[h], Philip H. Jones[h], M. Iqbal Parker[a,*]

[a] *Department of Integrative Biomedical Sciences, Division of Medical Biochemistry and Structural Biology, Faculty of Health Sciences, University of Cape Town, South Africa*
[b] *Department of Integrative Biomedical Sciences, Division of Computational Biology, Faculty of Health Sciences, University of Cape Town, South Africa*
[c] *Department of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, Kings College, SE1 9RT, London, United Kingdom*
[d] *Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa*
[e] *Strengthening Oncology Services Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa*
[f] *National Cancer Registry, a Division of the National Institute for Communicable Diseases, National Health Laboratory Service, Johannesburg, South Africa*
[g] *Network for Oncology Research in Africa (NORA), Global Health Working Group, Martin-Luther-University, Halle-Wittenberg, Germany*
[h] *Wellcome Sanger Institute, Hinxton, CB10 1SA, United Kingdom*

ARTICLE INFO

Keywords:
Bioinformatics
Next generation sequencing
Gene complexity
Somatic mutations
Variant callers
Cancer

ABSTRACT

Variant calling in complex genomic regions remains a critical challenge in cancer genomics, yet systematic evaluations of false positive rates in such regions are rarely reported. This investigative study examined somatic mutations in esophageal squamous cell carcinoma (ESCC) using Whole Genome Sequencing (WGS) data, that identified a high frequency of putative mutations in *MUC3A*, a gene with an inherently complex sequence architecture. Quantitative laboratory validation attempts failed to confirm any of these computationally predicted mutations, prompting systematic re-analysis. By assessing multiple variant calling algorithms and implementing a Panel of Normals (PON) filtering strategy, we demonstrate that standard bioinformatics pipelines generated extensive false positive calls in *MUC3A*, with false positive rates approaching 100 % for this gene. While previous studies have acknowledged limitations in variant calling for repetitive or homologous regions, our work provides evidence of complete analytical failure in the *MUC3A* gene, and establishes a reproducible framework for identifying such artefacts. These findings address a critical research gap by quantifying the magnitude of false discovery in complex genomic contexts and demonstrating that multi-tool consensus approaches combined with PON filtering are insufficient without accompanied experimental validation. We recommend mandatory quantitative confirmation for variants identified in sequence-complex genes and advocate for transparent reporting of validation rates in cancer genomic studies to prevent propagation of spurious findings in literature. This paper provides a cautionary warning to future research to take into consideration the limitations of alignment and variant calling tools and to employ a combination of tools to obtain robust and reliable results.

1. Introduction

Africa faces a substantial and growing complex burden of disease characterized by the dual challenge of persistent infectious diseases and a rapidly rising incidence of non-communicable diseases that pose major public-health challenges(Goswami, 2024; de-Graft Aikins et al., 2010). While conditions such as HIV/AIDS, malaria and tuberculosis remain major public health concerns, cancer has emerged as a significant and critical health challenge with late-stage diagnoses, limited access to molecular diagnostics and targeted therapies, and inadequate oncology infrastructure contributing to poor outcomes (Umobong et al., 2025; Boutayeb, 2006). The continent's unique genetic diversity and

This article is part of a special issue entitled: Science in Africa published in Aspects of Molecular Medicine.

[*] Corresponding author. Department of Integrative Biomedical Sciences University of Cape Town, South Africa.
E-mail address: iqbal.parker@uct.ac.za (M.I. Parker).

[1] Current Affiliation: Biomedical Research and Innovations Platform, South African Medical Research Council, Tygerberg, South Africa.

https://doi.org/10.1016/j.amolm.2025.100097
Received 21 May 2025; Received in revised form 31 October 2025; Accepted 1 November 2025
Available online 4 November 2025

environmental exposures shape distinct cancer susceptibilities and molecular signatures that remain underexplored.

Recent studies published highlights the growing application of bioinformatics and mutational analysis tools to unravel the molecular underpinnings of disease in African populations. A genomic study on hypercholesterolemia conducted in children from Ghana demonstrated the potential of variant annotation pipelines in detecting pathogenic mutations (Opoku-Agyeman et al., 2025). Similarly, a different study in Ghana highlighted how genomic tools can be used to identify clinically significant variants relevant to gastric carcinogenesis through mutational profiling of antimicrobial resistance genes in *Helicobacter pylori* (Ofori et al., 2025). A further study on the use of integrative bioinformatics analyses in Alzheimer's disease demonstrated how disease pathogenesis can be understood through pathway enrichment and molecular-network mapping (Nguyen et al., 2024), while biomarker discovery and informed targeted therapies in cancer can be driven through the advances in computational approaches and mutational profiling for precision oncology (Namini et al., 2025). In this context, bioinformatics and mutational-analysis tools have an essential role in addressing the African disease burden. They enable large-scale variant calling, annotation of pathogenic versus benign variants, and the mapping of genomic mutations.

These studies illustrate that leveraging bioinformatics for mutational analysis, whether in inherited non-communicable disease, infectious resistance, or cancer genomics, offers a path to better understand the unique genomic and disease-environment interplay in African populations and make possible more effective, precision medicine strategies that account for the continent's genetic and epidemiological uniqueness.

As roughly 1 % of the human genome encodes protein-coding regions (Pertea et al., 2018), misfunctioning (loss-of-function) or dysregulation (gain-of-function) of critical proteins involved in homeostasis are often the result of mutations introduced into these genes, frequently leading to the development of cancer (Vestergaard et al., 2021), a disease characterized by the accumulation of somatic mutations in several associated genes (Futreal et al., 2004). Therefore the discovery and identification of events that contribute to tumorigenesis are critical for our ongoing understanding of cancer as a disease (Dietlein et al., 2020).

Somatic mutations are acquired throughout the lifetime of an individual and are distinguishable from germline mutations that are inherited from parents and transmitted to offspring (Stratton et al., 2009). Somatic mutations occur in healthy cells and in most cases do not cause alterations to cell behaviour (Martincorena et al., 2017). However, occasionally, key genes become altered in a manner that provides a competitive advantage to the mutated cell, promoting the formation of persistent mutant clones and initiating the process of tumour cell transformation (Jaiswal et al., 2017; Anglesio et al., 2017).

With the advent of next-generation sequencing (NGS), insights into the genome have provided meaningful knowledge into unravelling the genetic conundrums of diseases such as cancer. NGS performs massively parallel sequencing and generates vast amounts of data, posing a challenge to researchers in terms of the handling, interpretation, and analysis of the data. The subsequent development of a large number of specialized bioinformatics tools, was aimed at navigating and handling large quantities of raw data generated by NGS (Pereira et al., 2020). However, several studies have shown that the application of different tools often varies in consistency (Kumaran et al., 2019; Liu et al., 2013), suggesting cautious interpretation as outputs could lead to spurious results (Vestergaard et al., 2021). The two most prominent bioinformatics analysis processes that have the potential to influence the interpretation of the data are the tools used for alignment and variant calling (Kumaran et al., 2019; Liu et al., 2013). Both of these processes have numerous and diverse algorithms in their original design and purpose. Challenges encountered with material artefacts, library preparation sequencing technologies, and gene regions selected for sequencing all further highlight the importance of selecting appropriate tools for the downstream analysis of NGS data. NGS is a complex technology, and caution is advised when interpreting results that may be influenced by the type of biological specimen; preanalytical treatment; pseudogenes and repetitive regions; bioinformatic challenges dealing with alignment and variant calling (Vestergaard et al., 2021).

One of the main advantages of DNA sequencing is the accurate identification and annotation of mutations, including single nucleotide variants (SNV), insertions/deletions (indels), copy number variants (CNV), and structural variants (SV), with high sensitivity and specificity (true positives and true negatives) (Vestergaard et al., 2021). Whole Genome Sequencing (WGS) is often described as explorative given its broader scope and lesser read depth (30-50x) (Bewicke-Copley et al., 2019), but it is effective in identifying most germline mutations and allowing for a comprehensive large-scale detection of the relevant variants (Griffith et al., 2015). However, some reports suggest that WGS may be insufficient in detecting rare somatic mutations that could harbour cancer genomes (Vestergaard et al., 2021).

We used esophageal squamous cell carcinoma (ESCC) as a model to explore the use of bioinformatics tools for downstream WGS analysis to elucidate somatic mutations within patients. The objective of the study was to compare the number of high and moderate impact mutations detected in *MUC3A* and *TP53* in our WGS study. The selection of these two genes was made on the basis of their DNA structural differences, with the former having extensive repetitive sequences composed of tandem repeats and the latter having minimal or no repetitive sequences (Gum et al., 1997; Pratt et al., 2000), together with our preliminary results which identified multiple mutations in the *MUC3A* gene. It is well established that extensive tandem repeats pose a significant challenge to the accurate alignment of reads due to the potential for ambiguity (Treangen and Salzberg, 2011). Short sequencing reads frequently align well to multiple similar locations within the genome. This misalignment of reads has a direct impact on the number of false positive variant calls. The present study reports the number of high and moderate mutations detected using Vardict and Mutect2+PON and the number discarded by the PASS filter in Vardict and the PASS + PON filter in Mutect2.

This study identified numerous shortcomings and limitations associated with a bioinformatics-only approach without laboratory confirmation. The mucin gene *MUC3A* is discussed as an example of the complexities of genome sequencing and the caution required when analysing data, considering false positives and spurious pipeline outputs.

2. Methods and materials

2.1. Patient recruitment

Patients were recruited from Groote Schuur Hospital in Cape Town (associated with the University of Cape Town) and Charlotte Maxeke Johannesburg Academic Hospital (associated with the University of the Witwatersrand). Patients presenting with histologically confirmed esophageal squamous cell carcinoma (ESCC) were recruited through informed consent. Matched normal and tumour biopsies and blood samples were collected from each patient, processed, and stored as previously described (Ferndale et al., 2022; Matejcic et al., 2019). Since there are no early symptoms associated with ESCC, all patients presented with advanced stage 4 cancer, typically with lymph node metastases. No early-stage cancers were present in the recruited patients. Once patient biopsies and blood samples were processed, extracted DNA was subjected to Whole Genome Sequencing (WGS) at the Wellcome Sanger Institute in Cambridge in the United Kingdom. The total patient cohort comprised twenty females and fifteen males with a mean patient age of 62 years for females and 54 years for males. A breakdown according to age, gender, % tumour cells, sequencing coverage, and sequencing duplication factor is shown in Table 1. In WGS, the sequencing coverage refers to the average number of times each base in the genome is sequenced, and in this instance, it was set to >30x coverage. The sequence duplication factor refers to the proportion of

Table 1
Patient cohort age and gender, where F represents females and M represents males. Patient DNA was subjected to whole genome sequencing as described in Materials and Methods. Blinded histological testing was performed and the % tumour cells in the biopsy was determined as indicated below. Sequencing coverage for each tumour (T) and normal (N) sample is shown, as well as the duplication factor (the fraction of mapped reads where any two reads share the same 5′ and 3’ co-ordinates). Patients recruited from the University of Cape Town, and the University of the Witwatersrand are indicated as UCT and WITS respectively.

UCT Patients						WITS Patients					
Patient Number	Age	Sex	% Tumour Cells	WGS Coverage T/N	Duplication Factor T/N	Patient Number	Age	Sex	% Tumour Cells	WGS Coverage T/N	Duplication Factor T/N
PD39445	57	F	n.d.	42.64/55.65	0.09/0.08	PD44691	70	F	39	33.77/41.83	0.06/0.08
PD39446	45	M	n.d.	42.42/51.14	0.09/0.07	PD44692	54	M	47	31.48/39.89	0.06/0.08
PD39447	41	M	28	50.64/49.49	0.14/0.07	PD44693	59	M	54	42.6/38.02	0.07/0.08
PD39448	52	M	44	46.73/48.17	0.14/0.07	PD44694	54	F	21	40.01/38.1	0.07/0.08
PD39449	79	F	57	47.33/50.99	0.13/0.07	PD44695	63	F	64	41.5/42.37	0.07/0.09
PD39450	50	F	64	51.99/46.1	0.14/0.07	PD44696	54	F	69	34.86/39.44	0.06/0.08
PD39451	71	M	47	55.27/47.23	0.14/0.11	PD44697	38	M	29	36.02/37.06	0.07/0.07
PD39452	53	F	64	53.17/49.06	0.14/0.11	PD44698	45	F	70	38.59/37.61	0.07/0.08
PD39453	37	M	22	51.67/43.32	0.16/0.11	PD44699	81	F	61	34.13/42.24	0.07/0.09
PD39454	67	F	n.d.	51.68/51.82	0.16/0.11	PD44700	71	F	43	34.85/35.76	0.06/0.07
PD39455	48	F	91	47.11/53.46	0.16/0.12	PD44701	69	F	13	35.49/36.71	0.06/0.07
PD39456	41	M	51	50.71/45.18	0.17/0.11	PD44702	65	F	46	36.33/40.62	0.06/0.07
PD39457	57	M	62	48.83/45.3	0.16/0.08	PD44703	78	M	38	37.12/36.23	0.06/0.07
PD39458	60	F	30	48.58/44.17	0.17/0.08	PD44704	56	M	30	34.21/39.34	0.06/0.07
PD39459	64	F	22	62.77/51.09	0.12/0.09						
PD39460	56	M	66	48.05/47.45	0.10/0.09						
PD50649	66	F	55	34.13/31.14	0.10/0.09						
PD50650	60	F	22	37.02/37.83	0.09/0.09						
PD50651	70	M	56	34.68/40.05	0.09/0.09						
PD50653	57	F	48	29.09/33.14	0.09/0.09						
PD51372	60	M	27	36.73/32.16	0.09/0.08						

*n.d. = not determined.

reads that were found to be duplicates sharing the same 5′ and 3’ co-ordinates, often arising during library preparation or from sequencing artefacts.

Ethical approval for the study was obtained from the UCT/Groote Schuur Hospital Human Research Ethics Committee (Ethics number: 040/2005), and the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (Certificate number M170871).

2.2. DNA extraction

DNA was extracted from patient blood and biopsies using the Qiagen AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, 80224, Hilden, Germany) as per manufacturer's instructions.

Following extraction, DNA integrity was determined by gel electrophoresis using a 1 % agarose gel (SeaKem®, Lonza, 50002, Rockland, ME, USA) together with 1 μl Novel Juice (Bio-Helix, LD001-1000, Taipei, Taiwan) detection dye. A suitable gene-ladder was loaded into the gel (GeneRuler ™ 100bp Plus DNA Ladder (ThermoFisher, SM0321, Vilnius, Lithuania)). This standard protocol is essentially as previously described (Lee et al., 2012), with the amendment that Novel Juice fluorescent reagent was added to the samples to provide an environmentally safe, non-hazardous alternative to ethidium bromide for DNA detection.

2.3. Whole Genome Sequencing

DNA isolated from paired blood samples and tumour biopsies were subjected to WGS at the Wellcome Sanger Institute in Cambridge, UK. Samples were genotyped for single nucleotide polymorphisms (SNP) using a Fluidigm chip array to confirm that the tumour and normal samples were patient matched. Samples were then sequenced on an Illumina HiseqX10 using 150 bp paired-end reads to a depth of >30x coverage.

2.4. Bioinformatics analysis of WGS data

A variant-calling pipeline for small variants, including Single Nucleotide Variants (SNV) and insertions and deletions (indels), was set up utilizing the opensource software package bcbio-nextgen (Chapman et al.). This software allows for the analysis of sequences through specialized pipelines with further visualization and additional processing made possible. The variant calling analysis pipeline aligns reads to selected reference genomes, allowing for the identification of variants within the query sequences (Guimera, 2011). In this way, calls were compared against the common reference genome GRCh38 using the BWA tool for alignment. Preparation and variant calling were incorporated into the pipeline to ensure an unbiased comparison of algorithms (Chapman et al.). Once the setup was confirmed, configuration files were constructed for each patient in the sample cohort following the guidelines described in the software documentation (bcbio-nextgen 1).

All pipeline scripts used and query searches can be found in the online repository at https://github.com/VictoriaPatten/phd-scripts/tree/main/bcbio-nextgen.

In setting up the configuration files, reads were aligned to the GRCh38 human reference genome using the Burrows-Wheeler Aligner (BWA 0.7.17), which maps low-divergent sequences against large reference genomes (Li and Durbin, 2010). To select somatic and germline calls, individual variant callers were specified for each. Bcbio-nextgen carried out a single alignment for the normal sample first and then split at the variant calling stage using the normal sample as a baseline for germline and somatic calling. Freebayes v1.3.6 (Garrison and Marth, 2012), a genetic variant detector designed to locate small polymorphisms, specifically SNP's and indels, was specified for germline mutations and Vardict (Lai et al., 2016), an ultra-sensitive variant caller that simultaneously calls SNV's and indels, performing local re-alignments for more accurate allele frequency estimation was stipulated for somatic variant calling.

Variant Call Format (VCF) annotation was performed using the SnpEff tool (Cingolani et al., 2012). The effects of variants in a genome sequence are rapidly categorized and annotated based on their genomic locations and prediction of coding effects. Structural and copy number variants (CNV's) were called using Lumpy v0.3.1 (Layer et al., 2014), a probabilistic prediction framework for structural variant discovery.

A 'Panel of Normals' (PON) was incorporated into the bcbio-nextgen pipeline to eliminate false positives. Using a PON approach, a baseline level for variant calling is determined from a combined a set of normal samples typically derived from the same library preparation and sequencing workflow used for tumour samples to allow for non-sample specific system level biases to be subtracted (CreateSomaticPanelOfNormals). In this way, variant calling results are improved as recurrent technical artefacts are removed. For short variant calling, it is recommended that the PON should be created and run using the variant caller Mutect2, which is a variant detector for SNPs and indels and is part of the Genome Analysis Toolkit (GATK) (GATK). PON files were created for each matched normal sample in the cohort and combined into a single zipped VCF file. The original bcbio-nextgen configuration files described above were edited to include Mutect2 as the somatic variant caller instead of the previously used Vardict, and the *background:* parameter was set to include the VCF PON file as a background of all 35 normal samples to be run against tumour samples.

Bcbio-nextgen pipelines were re-run for all 35 tumour-normal pairs, and the resulting VCF files were filtered for HIGH impact *MUC3A* and *TP53* mutations.

2.5. Polymerase chain reaction (PCR) validation

DNA from patient biopsies was extracted in accordance with standard operating protocols previously described (Patten et al., 2023), and was stored at −20 °C until needed. DNA was available for PCR amplification from the 16 UCT patients that made up part of the patient sample cohort (Table 1). PCR primer design and optimisation is described in Patten et al. (2023). Primers for different clusters of identified *MUC3A* mutations in exon 2 of the gene were designed and used for PCR amplification of patient DNA. Post-PCR amplified products were subjected to bi-directional Sanger sequencing. Chromatograms were analysed using Chromas v2.6.6 (available at http://technelysium.com.au/wp/chromas/) a free trace viewer for simple DNA sequencing projects that is free to download.

3. Results

3.1. Analysis of whole genome sequence data using bcbio-nextgen software

Annotation of the output was performed using the SnpEff tool (Cingolani et al., 2012). GEMINI v.0.20.1 (Paila et al., 2013) was then used to create a database of the output to facilitate the query of the annotated VCF files (from Vardict). GEMINI is a genome mining tool for exploring human variations. Using the command line, the GEMINI output database files were explored and filtered to search for particular parameters of interest annotated within the variant/variant_impacts tables of the output database files.

The impact severity of the mutations can be described as the functional consequence of a given variant, ranked as either HIGH, MED (medium), or LOW) (GEMINI). A search was performed for all genes presenting HIGH impact variants across all 35 patient genomes to determine the top genes with the highest number of HIGH impact mutations within the patient cohort.

Table 2 shows the GEMINI search results of the top 20 genes in the cohort presenting with HIGH impact severity somatic mutations, indicating the number of patients with these mutations, thus providing a ranking of the genes. Fig. 1 shows the total number of variants detected in several genes across the patient cohort.

The results shown in Table 2 were unexpected. At the top of the list, with 30 out of 35 (86 %) patients presenting HIGH impact mutations, was the *MUC3A* gene. Furthermore, 258 incidences of mutations of this gene were detected across the patient cohort. These numbers far exceed those for known cancer driver genes such as *TP53, CDKN2A,* and *KMT2D* and were thus deemed 'suspicious' given the lack of literature reporting similar results. It was, therefore, imperative to reanalyse the data using an alternative approach incorporating the PON together with a different variant caller, Mutect2, in place of Vardict.

Table 2
Top 20 genes with the most HIGH impact severity variants detected using Vardict variant caller, in the 35 patient WGS cohort.

Gene	Number of patients with mutations (% of Total)
MUC3A	30 (86 %)
TP53	18 (51 %)
HEATR9	14 (40 %)
VPS52	12 (34 %)
NCOR1	10 (29 %)
AHNAK	9 (26 %)
OR4D10	8 (23 %)
CDKN2A	8 (23 %)
FIGN	7 (20 %)
OR4D11	7 (20 %)
KMT2D	7 (20 %)
CST3	6 (17 %)
GRIN2A	6 (17 %)
DEF8	6 (17 %)
FGFR1	6 (17 %)
TBL1X	6 (17 %)
FAM135A	6 (17 %)
TDG	6 (17 %)
CST5	5 (14 %)
SLC4A3	5 (14 %)

After re-running the bcbio-nextgen pipeline for all tumour-normal pairs incorporating Mutect2, an entirely new set of HIGH impact *MUC3A* mutations was identified. The Mutect2 variant caller was selected based on results reported in Bian et al. (2018) suggesting that Mutect2 identified the lowest number of false positives in their comparison of variant callers. More than 400 incidences of *MUC3A* mutations were now detected across all 35 samples in the cohort, with HIGH impact severity status. Furthermore, all the mutations identified using the Vardict variant caller were filtered out and were no longer present. This strongly suggested that the mutations in the *MUC3A* gene using the initial approach with the Vardict variant caller were false positives. Fig. 2 shows the total number of *MUC3A* mutations identified per patient using the Mutect2 and the PON approach.

From this second analysis using the Mutect2 variant caller, a much larger number of *MUC3A* mutations were identified, all falling in the second exon of the *MUC3A* gene and in 100 % of the patient cohort, which again did not seem credible. Analysis of the data using an Integrative Genome Viewer (IGV) representation of this gene (Fig. 3), shows that in both tumour and normal samples, this genomic region is extremely noisy, and it is very likely that most, if not all, of these putative mutations are not valid. All of the mutations previously identified with Vardict were filtered out, suggesting they were false positives (explaining why the presence of these mutations could not be verified experimentally). In the new set of mutations, every patient in the cohort was found to have *MUC3A* mutations, which is highly improbable. It is likely that this revised data also suffers from false positives due to the complex nature of the *MUC3A* genomic structure. To identify whether spurious sequence reads may have played a role in these results, depth of coverage for each read and each mutation was investigated and proved computationally valid with high coverage.

Tables 3 and 4 were obtained by filtering the Vardict and Mutect2 outputs with and without restricting the mutation to the PASS filter. The implementation of the filtering process involved the utilization of GEMINI for Vardict and SnpSift for Mutect2+PON, as the pipeline outputs involving the two callers are a GEMINI database and a VCF file, respectively. When considering all identified mutations in *MUC3A* as false positives, the numbers in the No PASS column for both callers reveal that Mutect2 performed better in identifying false positives of MEDIUM (MED) impact. However, Vardict demonstrated superiority in HIGH impact, which are predominantly frameshifts. The application of a PASS filter to the Vardict VCF queries led to a substantial reduction in

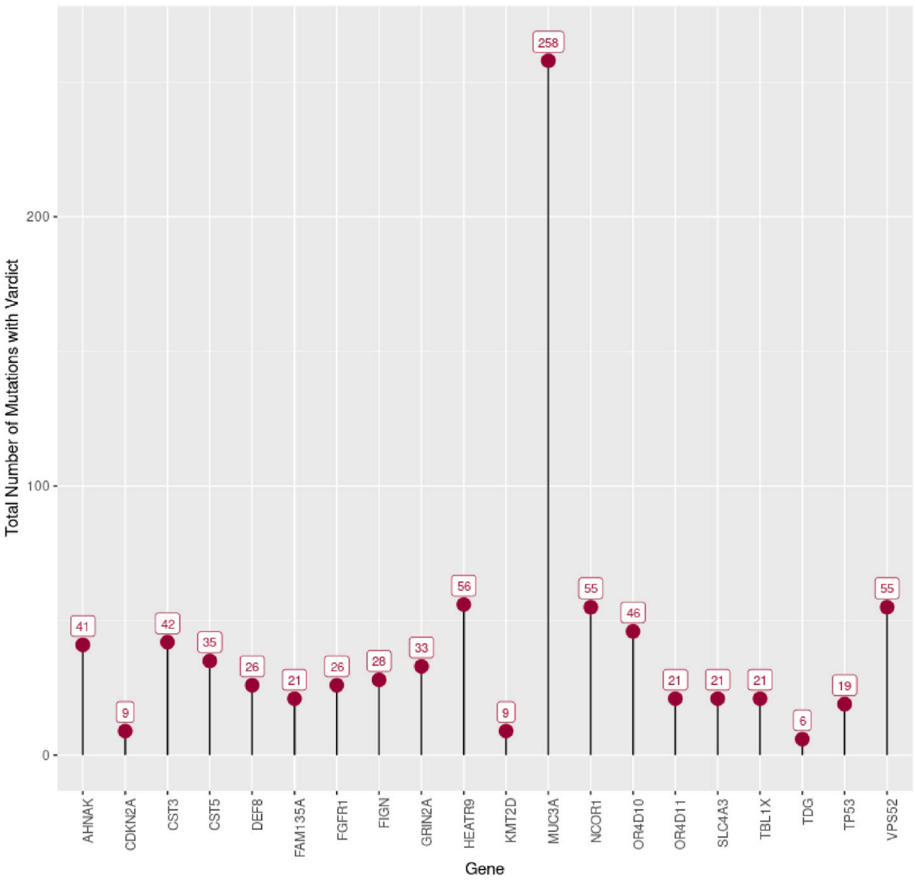

Fig. 1. Lollipop plot indicating the distribution of somatic variants across the top 20 genes with the most HIGH impact severity variants detected using Vardict variant caller with bcbio-nextgen pipeline. The *MUC3A* gene shows a vastly greater total number of mutations detected across the entire cohort compared to all other genes.

the number of mutations that passed this filter, though a number did still pass. Conversely, the PASS + PON approach employed in Mutect2 categorized all mutations as false positives, consequently deleting them from the query outputs.

The number of mutations identified in *TP53* was also investigated (Table 4), and the results showed a far lower number of mutations detected compared to those of *MUC3A*. However, similar trends can be seen when comparing Vardict and Mutect2 variant callers. In this instance, the No PASS filter was not needed for many of the patients as the same number of mutations were detected in both cases. This suggests that bioinformatic pipelines and tools should be tailored when investigating different genes with different complexities.

3.2. Laboratory confirmation of bioinformatics data

Attempts were made to validate the initial *MUC3A* mutations identified through bioinformatic analysis using the Vardict variant caller by PCR amplification of the relevant *MUC3A* regions from patient DNA. Mutations fell into five cluster areas within the large second exon of the *MUC3A* gene, and thus multiple primer sets were designed and optimised. Table 5 shows the primers designed for the first cluster of mutations, while Fig. 4 shows a representative agarose gel of amplified DNA for cluster 1 mutations using patients PD39456 and PD39457 as examples. All optimised primers and PCR data are discussed in Patten et al. (2023) and all primers and results can be found in the supplementary materials. The limited samples present in this paper are merely to show that laboratory confirmation was attempted.

Post-PCR bi-directional Sanger sequencing was performed on the amplified PCR products after excision from the agarose gel, and purification. Using sample PD39456 as an example, when analysing the resulting chromatogram (Fig. 5) in conjunction with Table 6, it can be seen that the expected *MUC3A* variants identified through bioinformatic analysis, shown in the ‘Alt’ column of Table 6 were not confirmed in the PCR product sequence. The variant positions have been indicated in the reference sequence and on the chromatogram to show where the alternate variants were expected, but the chromatogram sequences matched the reference gene sequence exactly, showing no mutations (Fig. 5). The clearly resolved peaks on the chromatogram show high confidence that the DNA sequence is accurate, and that the PCR product matched the reference sequence.

The PCR amplification and post-PCR sequencing was repeated for all initial *MUC3A* mutations identified through bioinformatic analysis using the Vardict variant caller in all relevant patient DNA samples. All results can be found in the supplementary materials, showing no validation was possible for any of the initial mutations identified.

4. Discussion

NGS downstream analyses are multifactorial technologies, and caution is crucial when analysing and interpreting the data. There are a number of factors that could influence output and results, including the type of biological specimen, pseudogenes and repetitive regions, and numerous and complex bioinformatic tools, especially pertaining to alignment and variant calling.

In the case of *MUC3A* gene, the investigations into the presence of somatic mutations in the sample cohort proved to be a challenging endeavour with conflicting results. The initial bioinformatic pipeline setup and testing was a lengthy process requiring many reconfigurations

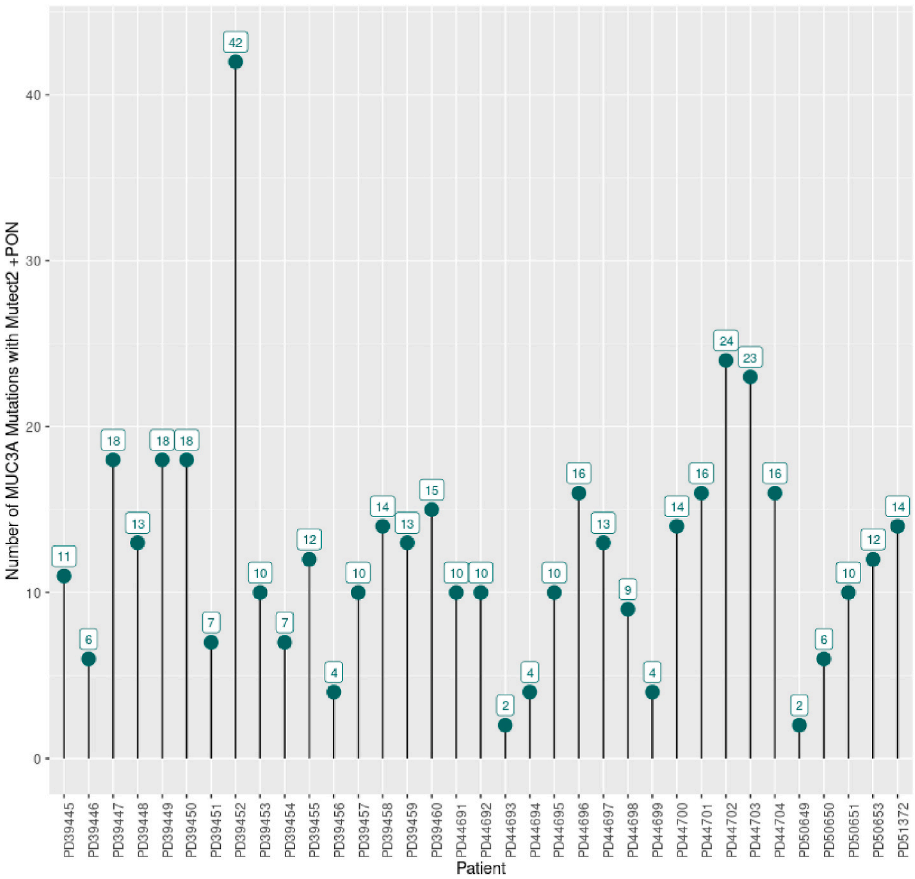

Fig. 2. Lollipop plot of the number of *MUC3A* mutations per patient, detected using Mutect2 with bcbio-nextgen pipeline.

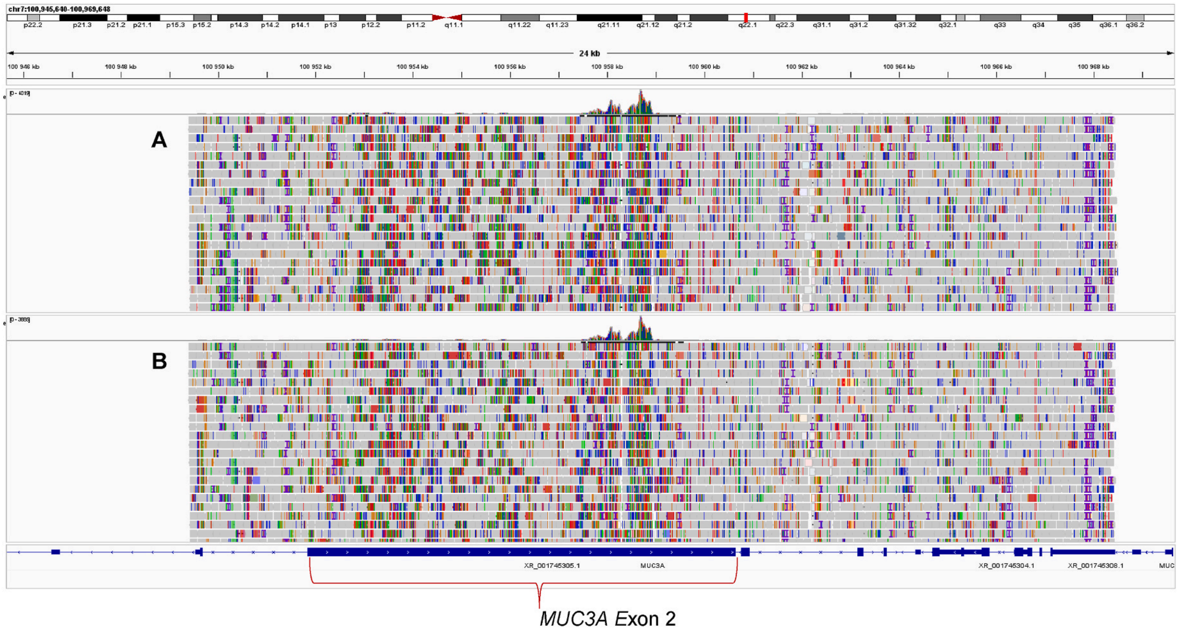

Fig. 3. A representative IGV survey of the *MUC3A* gene on chromosome 7, showing excessive noise in both the A) tumour and B) matched normal sample, concentrated over exon 2. Samples were aligned against the GRCh38 human reference genome.

and troubleshooting of the software packages before the installation and setup ran without errors. When performing an extensive search of the database and VCF output files to identify all genes through the cohort where multiple instances of HIGH impact mutations were observed, the results provided an interesting finding where *MUC3A*, a gene not previously described as associated with ESCC, appeared to be the most highly mutated gene (with HIGH impact mutations) with 258 detected mutations across the cohort, 96 % of which were frameshift variants.

Table 3
Medium (MED) and High impact mutations identified in *MUC3A* using Vardict and Mutect2 variant callers, with and without the Pass filter and PON, respectively.

Patient Number	Vardict (MED/HIGH)		Mutect2 (MED/HIGH)	
	No Pass Filter	PASS Filter	No PASS Filter	PON PASS Filter
PD39445	233/5	2/0	72/11	0/0
PD39446	248/0	1/0	19/6	0/0
PD39447	241/5	3/0	70/18	0/0
PD39448	248/18	6/0	58/13	0/0
PD39449	239/0	1/0	111/18	0/0
PD39450	239/0	2/0	89/18	0/0
PD39451	239/5	3/0	72/7	0/0
PD39452	251/10	3/0	182/42	0/0
PD39453	228/12	4/0	37/10	0/0
PD39454	265/6	10/0	88/7	0/0
PD39455	266/16	7/0	47/12	0/0
PD39456	264/17	4/0	69/4	0/0
PD39457	260/13	4/1	46/10	0/0
PD39458	256/7	8/0	61/14	0/0
PD39459	241/2	8/0	87/13	0/0
PD39460	258/7	4/0	90/15	0/0
PD44691	259/15	8/0	62/11	0/0
PD44692	253/1	4/1	54/10	0/0
PD44693	243/9	6/0	49/2	0/0
PD44694	268/0	3/0	41/4	0/0
PD44695	248/2	3/2	53/10	0/0
PD44696	262/11	5/0	61/16	0/0
PD44697	248/15	6/0	53/13	0/0
PD44698	254/7	4/0	48/9	0/0
PD44699	264/7	3/0	34/4	1/0
PD44700	254/6	4/1	70/14	0/0
PD44701	240/2	4/0	38/16	0/0
PD44702	243/6	6/0	139/24	0/0
PD44703	249/7	1/1	73/23	0/0
PD44704	257/8	7/0	74/16	0/0
PD50649	NA	NA	32/2	0/0
PD50650	261/11	4/3	52/6	0/0
PD50651	NA	NA	48/10	0/0
PD50653	NA	NA	74/12	0/0
PD51372	NA	NA	50/13	0/0

*NA = Data Not Available.

Mutations in the *MUC3A* gene far exceeded the numbers detected for other more commonly mutated genes. This transmembrane mucin gene has been found to be highly expressed in various epithelial cells of the intestines (Kitamoto et al., 2010; Wang et al., 2020), whose protein product is reported to be involved in cellular protection through barrier function as well as in intracellular signal transduction pathways for regulation of inflammation, cell adhesion, cellular differentiation, and apoptosis (van et al., 2017). When investigating these findings in the laboratory using PCR, none of the mutations could be confirmed. This can be due to the nature of the *MUC3A* gene, which is notoriously difficult to PCR amplify due to the very large number of tandem repeats in the genomic sequence or due to false positives in the bioinformatic data.

However, an IGV investigation of this gene and its genomic sequences confirmed a high degree of complexity of the gene at an individual read level, (all samples showed a high level of background noise and technical artefacts in both tumour and normal samples), casting doubt on the validity of the *MUC3A* mutations. The likelihood of false positives became apparent, and reanalysis using the PON approach and Mutect2 variant caller was implemented. All mutations previously identified using Vardict variant caller were subsequently filtered out, and we were able to conclude that they had indeed been false positives. However, a new set of *MUC3A* mutations was identified and further questions and speculations around their validity arose given the improbability of these mutations occurring in all patients of the cohorts, as well as the fact that no previous studies have reported this gene as playing any role in ESCC.

A study conducted by Bian et al. (2018) (Bian et al., 2018) sought to compare the performance of a number of different variant callers, including Vardict and Mutect2 with bcbio-nextgen software. They questioned whether different callers might perform differently on different parts of the genome and whether GC content might affect analysis results. Their investigations showed that differences in true positives between callers were small, but the number of false positives varied greatly. Furthermore, callers experienced diminishing accuracy when exposed to increasing levels of data complexity and that sequencing properties such as read depth, read quality, strand bias, and varying allele frequencies can challenge a given caller's ability to accurately detect mutations. Their results showed that Vardict produced the highest number of true positives, but in a trade-off, also produced a high number of false positives, while Mutect2 was among the best-performing tools for detecting true positives and controlling for false positives (Bian et al., 2018). When investigating our own results, we found that depending on the impact of mutations, Vardict was found to have fewer false positives for HIGH impact mutations, but much

Table 4
Medium (MED) and High impact mutations identified in *TP53* using Vardict and Mutect2 variant callers, with and without the Pass filter and PON, respectively.

Patient Number	Vardict (MED/HIGH)		Mutect2 (MED/HIGH)	
	No Pass Filter	PASS Filter	No PASS Filter	PON PASS Filter
PD39445	0/1	0/1	0/1	0/1
PD39446	0/1	0/0	0/1	0/1
PD39447	0/0	0/0	1/0	1/0
PD39448	1/0	0/0	0/0	0/0
PD39449	1/2	0/2	0/2	0/2
PD39450	1/1	0/1	0/1	0/1
PD39451	2/0	1/0	1/1	1/0
PD39452	2/0	1/0	1/0	1/0
PD39453	0/1	0/1	0/1	0/1
PD39454	0/1	0/1	0/1	0/1
PD39455	0/1	0/1	0/1	0/1
PD39456	0/1	0/1	0/1	0/1
PD39457	1/1	0/1	0/1	0/1
PD39458	2/0	1/0	1/0	1/0
PD39459	1/1	0/1	0/1	0/1
PD39460	0/0	0/0	0/1	0/0
PD44691	2/1	1/1	1/0	1/0
PD44692	0/1	0/1	0/1	0/1
PD44693	1/1	0/1	0/1	0/1
PD44694	1/1	0/1	0/1	0/0
PD44695	0/0	1/0	0/0	0/0
PD44696	1/1	0/1	0/1	0/1
PD44697	1/0	1/0	1/0	1/0
PD44698	2/0	1/0	1/0	1/0
PD44699	0/1	0/1	0/1	0/1
PD44700	3/0	2/0	2/0	2/0
PD44701	1/0	0/0	0/0	0/0
PD44702	0/1	0/1	0/1	0/1
PD44703	2/1	1/1	1/1	1/1
PD44704	0/0	0/0	0/0	0/0
PD50649	NA	NA	1/0	1/0
PD50650	1/0	0/0	0/0	0/0
PD50651	NA	NA	1/0	1/0
PD50653	NA	NA	1/0	1/0
PD51372	NA	NA	0/1	0/1

*NA = Data Not Available.

Table 5
Optimised conditions for the first cluster of *MUC3A* mutations in exon 2.

Primer Pair	PCR Conditions		Product Size
Cluster 1 Forward:	95 °C: 4 min		889bp
5′- TAA GTA CAC TCA GCA CTC CTA -3′	95 °C: 45 s 60 °C: 45 s	40 cycles	
Cluster 1 Reverse:	72 °C: 1 min		
5′- GAG ATC ATG GAT GTA GAA GTT ACC -3′	72 °C: 7 min 4 °C: 7 min		

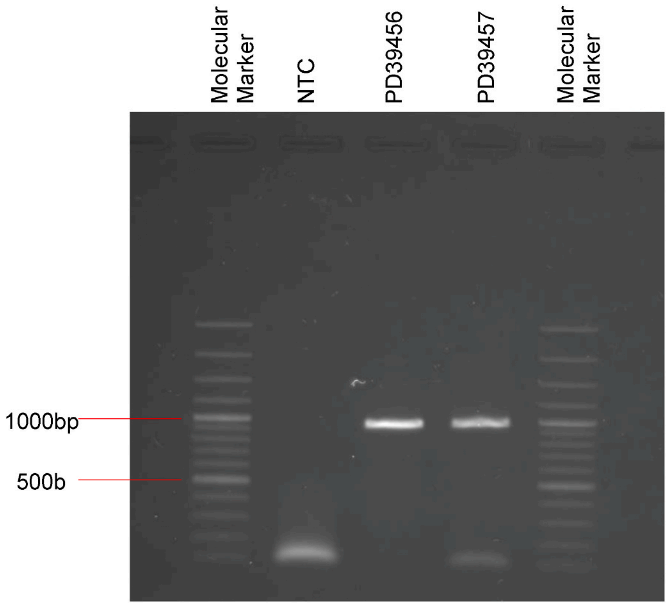

Fig. 4. Visualization of PCR products using primers for cluster 1 *MUC3A* mutations in patients PD39456 and PD39457. Post-PCR, products were electrophoresed through a 1 % agarose gel for 35 min at 100V and visualised under UV light for 20 s using Novel juice (Bio-Helix, LD001-1000, Taipei, Taiwan). Expected amplification size are bands shown in the region of 889bp. NTC indicates no-template control.

higher false positives for MED impact mutations when compared to Mutect2. This interesting finding suggests that when researchers are looking for HIGH impact mutations without using the PON approach, then the Vardict variant caller would be preferable.

MUC3A is a gene with a high degree of genomic complexity, especially within the second exon, which has subsequently led to difficulties in variant calling of true positive variants, and is greatly affected by the variant callers used. The value of NGS data is wholly dependent on valid methods of interpretation and the accurate analysis and identification of mutations. Eliminating the presence of false positives is imperative and, in this instance, extensive further investigations would be needed to conclude whether the new putative *MUC3A* mutations identified in the reanalysis with the PON approach were indeed true positives. Biologically, it is improbable that all 35 patients in the cohort would have mutations in this gene. Furthermore, the high volume of mutations detected in *MUC3A* raises the question of why no previous ESCC studies have identified and reported mutations in this gene. When we incorporated the PASS filter using both Vardict and Mutect2+PON approaches, all *MUC3A* mutations detected were false positives.

When attempting to validate the WGS *MUC3A* mutations using the Mutect2 variant caller, all initial *MUC3A* mutations were filtered out, and an entirely new set of mutations was identified. This indicates that the initial mutations were in fact false positives. It is also highly likely that most, if not all the new *MUC3A* mutations identified are also spurious. In contrast to *MUC3A*, the operation of calling variants by both Vardict and Mutect2 tools appears to be less complex and simpler for the *TP53* gene. In a substantial number of cases, both tools identified an equivalent number of mutations, exhibiting only minor discrepancies. For instance, patient PD39446 was reported as having a high-impact mutation by both callers without filtering. However, following filtering with PASS filter, Vardict discarded the mutation, while Mutect2 + PON retained it. Given the high mutation prevalence of *TP53* in esophageal cancer patients, which can reach up to 80 %, and the fact that the patients in our study are all at an advanced stage, it is possible that most of them have mutated *TP53*. Consequently, the mutation that was retained may also be a true positive, although this requires further laboratory validation to ascertain its validity. PD44691 has an opposite

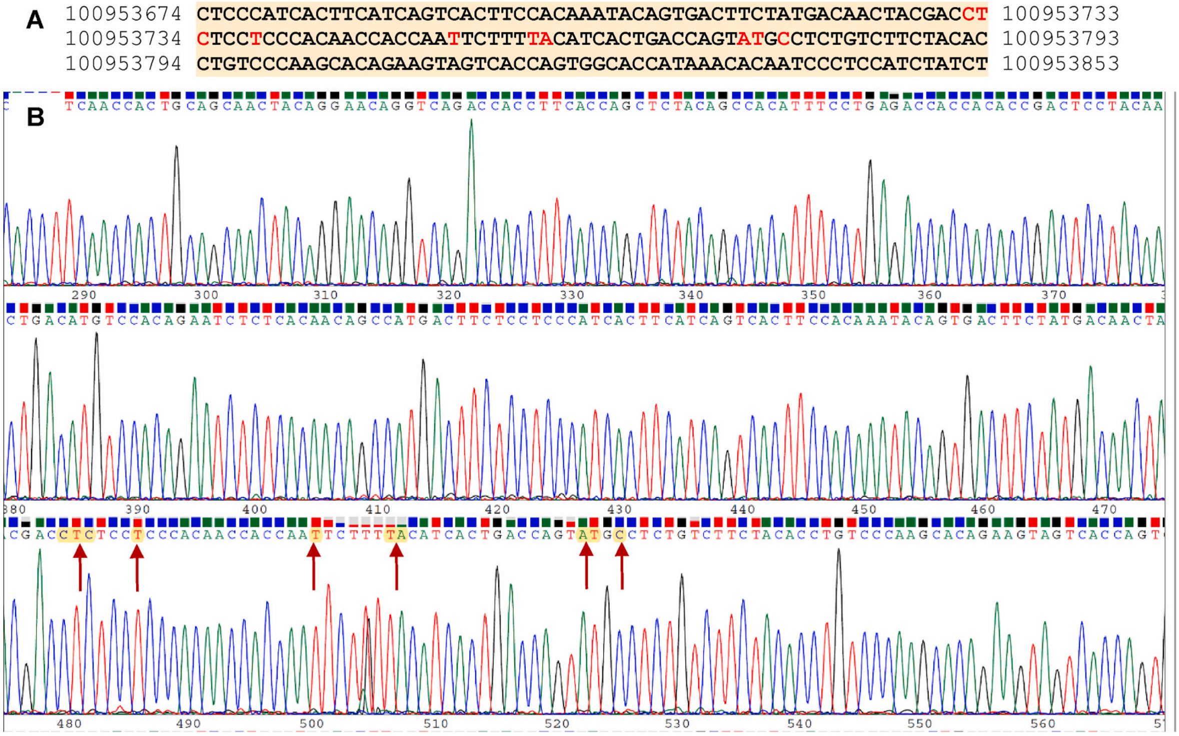

Fig. 5. **A**) *MUC3A* reference sequence showing the locations of the cluster 1 mutations determined from bioinformatic data for patient PD39456 in red. **B**) Chromatogram of the PCR product sequence for patient PD39456 using the primer set for cluster 1 mutations. The annotated sequence is shown above each line of corresponding peaks and the nucleotides indicated by arrows represent the positions where the identified variants should be located, as identified in Table 6. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 6
Cluster 1 variants identified in the *MUC3A* gene for patient PD39456 through bioinformatic analysis. Positions indicated correspond to reference genome GRCh38, and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Chromosome	Position	Gene	Ref	Alt	Impact
PD39456	chr7	100953731	*MUC3A*	CTC	T	Frameshift
	chr7	100953737	*MUC3A*	T	TGG	Frameshift
	chr7	100953752	*MUC3A*	T	TA	Frameshift
	chr7	100953758	*MUC3A*	TA	T	Frameshift
	chr7	100953774	*MUC3A*	AT	A	Frameshift
	chr7	100953777	*MUC3A*	C	CA	Frameshift

scenario than the previous patient (PD39446), where a high-impact mutation was retained by Vardict PASS filter, but Mutect2+PON discarded it. This mutation may also be a true positive, and as such, it should not be disregarded. As demonstrated by the two aforementioned scenarios, it is imperative to utilize multiple variant callers in order to achieve more accurate results and potentially identify rare variants that may be discarded by one variant caller but retained by another or other callers. It is, therefore, recommended that these scenarios be tested in a laboratory setting.

Improper alignment to the reference genome can also significantly influence the discovery of false positives, although this discordance has been improved over recent years. However, SNVs (in our case, frameshift mutations) still remain a challenge with NGS data, and various tools clearly display divergent outcomes (Kumaran et al., 2019). In recent years, tools specific for SNV detection have been developed yet their underlying algorithms of error models and assumptions for identifying mutations frequently result in diverse variant calling across tools (Xu et al., 2014). Short-read alignment tools are also commonly challenged by encountering reads that map to multiple locations in the reference genome (Treangen and Salzberg, 2011). Thus, the tools and methods one selects for analysis and variant calling on NGS data are critically important, given their heavy influence on mutational calling when the aim is for high sensitivity and specificity.

It is interesting to note that the detection of variants via bioinformatic pipelines can only prioritize novel findings of mutations and genes for functional testing. In this way, they can only identify mutations as drivers of tumorigenesis, which, it is advised, need laboratory confirmation (Gonzalez-Perez et al., 2013). Thus, researchers should perceive bioinformatic tools as predictors rather than validation, and laboratory confirmation should always be performed. Field (2022) reported that an increasingly popular approach among bioinformatic researchers is to run multiple calling tools and apply a consensus approach to minimize potential biases within single tools. High-quality variant data for specificity or sensitivity sets can be obtained using this approach (Field, 2022).

For future investigations, it would be pertinent to investigate different software packages and make use of different variant caller tools in combination as the different structural nuances associated with each of the four groups of genomic alterations (SNV, indels, CNV, SV) exclude the possibility of using one single versatile bioinformatic tool for identifying all variants within all four groups (Jennings et al., 2017).

5. Conclusion

NGS technologies have revolutionized genetic studies, diagnostics, and treatment strategies. However, the exponential increase in the volume of data generated necessitates robust and reliable analytical pipelines, particularly as reports in literature suggest variant calling tools may vary in consistency.

Initial analysis of our dataset identified 258 somatic mutations in *MUC3A* in 86 % of the sample cohort using Vardict variant caller, 98 % of which were frameshift mutations. However, no experimental validation could confirm these mutations, and upon reanalysis incorporating the PON-based Mutect2 approach, more than 400 alternative *MUC3A* variants were identified with the original Vardict mutations filtered out completely. Given the known false-positive rate for variant calling in repetitive genomic regions together with *MUC3A*'s complex structure, we conclude these likely represent technical artefacts or false positives rather than genuine somatic events.

Our findings demonstrate that the use of a single variant caller approach can yield numerous false positives in structurally complex genes. We recommend that future investigations should implement multi-caller consensus approaches (minimum of 3 variant callers with ≥2 caller concordance required), incorporate matched normal samples with PON filtering, and include quantitative laboratory validation to confirm computational findings. Without such rigorous pipeline set-up and validation, somatic mutations cells in genomically complex regions such as *MUC3A* should be interpreted with extreme caution.

CRediT authorship contribution statement

Victoria A. Patten: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Hocine Bendou:** Writing – review & editing, Supervision, Software, Investigation, Data curation. **Denver T. Hendricks:** Writing – review & editing, Supervision, Investigation, Data curation, Conceptualization. **Christopher G. Mathew:** Writing – review & editing, Resources, Funding acquisition. **Wenlong Carl Chen:** Writing – review & editing, Resources. **Joanna C. Fowler:** Writing – review & editing, Software, Formal analysis, Data curation. **Roshan K. Sood:** Writing – review & editing, Data curation. **Philip H. Jones:** Writing – review & editing, Software, Data curation. **M. Iqbal Parker:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

AI declaration

During the preparation of this work the author used Claude Next Generation AI Assistant (https://claude.ai) in order to streamline some paragraphs for better reading flow and sentence structure. After using this tool/service, the author critically reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Professor M. I. Parker reports financial support was provided by South African Medical Research Council. Professor M.I. Parker reports financial support was provided by Newton Fund. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by grant #046 (M.I.P., C.G.M) from the South African Medical Research Council (SAMRC) received from the South African National Department of Health and the Medical Research

Council of the United Kingdom (UKMRC) with funds from the United Kingdom Government's Newton Fund (MRCRFA-SHIP 01–2015).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amolm.2025.100097.

References

GATK 2023. https://gatk.broadinstitute.org/hc/en-us. (Accessed 13 March 2023).

Anglesio, M.S., Papadopoulos, N., Ayhan, A., et al., 2017. Cancer-associated mutations in endometriosis without cancer. N. Engl. J. Med. 376, 1835–1848. https://doi.org/10.1056/NEJMoa1614814.

bcbio-nextgen 1.2.9 documentation n.d. https://bcbio-nextgen.readthedocs.io/en/latest/index.html (accessed June 10, 2022).

Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., et al., 2019. Applications and analysis of targeted genomic sequencing in cancer studies. Comput. Struct. Biotechnol. J. 17, 1348–1359. https://doi.org/10.1016/J.CSBJ.2019.10.004.

Bian, X., Zhu, B., Wang, M., et al., 2018. Comparing the performance of selected variant callers using synthetic data and genome segmentation. BMC Bioinf. 19, 1–11. https://doi.org/10.1186/s12859-018-2440-7.

Boutayeb, A., 2006. The double burden of communicable and non-communicable diseases in developing countries. Trans. R. Soc. Trop. Med. Hyg. 100, 191–199. https://doi.org/10.1016/J.TRSTMH.2005.07.021.

Chapman B, Kirchner R, Pantano L, et al. bcbio/bcbio-nextgen: v1.2.5 2021. https://doi.org/10.5281/ZENODO.4429770.

Cingolani, P., Platts, A., Wang, L.L., et al., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80–92. https://doi.org/10.4161/fly.19695.

CreateSomaticPanelOfNormals (BETA) – GATK n.d. https://gatk.broadinstitute.org/hc/en-us/articles/9570531313563-CreateSomaticPanelOfNormals-BETA (accessed March 13, 2023).

de-Graft Aikins, A., Unwin, N., Agyemang, C., et al., 2010. Tackling Africa's chronic disease burden: from the local to the global. Glob. Health 6. https://doi.org/10.1186/1744-8603-6-5.

Dietlein, F., Weghorn, D., Taylor-Weiner, A., et al., 2020. Identification of cancer driver genes based on nucleotide context. Nat. Genet. 52 (2), 208–218. https://doi.org/10.1038/s41588-019-0572-y, 2020;52.

Ferndale, L., Moodley, M., Chen, W.C., et al.. Processing and analysis of tissue samples from esophageal cancer patients in an African setting. Https://HomeLiebertpubCom/Bio 2022. https://doi.org/10.1089/BIO.2021.0030, 20, 185-194.

Field, M.A., 2022. Bioinformatic challenges detecting genetic variation in precision medicine programs. Front. Med. 9, 806696. https://doi.org/10.3389/fmed.2022.806696.

Futreal, P.A., Coin, L., Marshall, M., et al., 2004. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183. https://doi.org/10.1038/NRC1299.

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. https://doi.org/10.48550/arxiv.1207.3907.

GATK. Mutect2 – GATK 2022. https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2. (Accessed 13 March 2023).

GEMINI. GEMINI: gemini 0.20.1 documentation 2022. https://gemini.readthedocs.io/en/latest/index.html. (Accessed 13 June 2022).

Gonzalez-Perez, A., Mustonen, V., Reva, B., et al., 2013. Computational approaches to identify functional genetic variants in cancer genomes. Nat. Methods 10 (8), 723–729. https://doi.org/10.1038/nmeth.2562, 2013;10.

Goswami, N., 2024. A dual burden dilemma: navigating the global impact of communicable and non-communicable diseases and the way forward. International Journal of Medical Research 12, 65–77. https://doi.org/10.55489/IJMR.123202412.

Griffith, M., Miller, C.A., Griffith, O.L., et al., 2015. Optimizing cancer genome sequencing and analysis. Cell Syst. 1, 210–223. https://doi.org/10.1016/j.cels.2015.08.015.

Guimera, R.V., 2011. bcbio-nextgen: automated, distributed next-gen sequencing pipeline. EMBnet J 17, 30.

Gum, J.R., Ho, J.J.L., Pratt, W.S., et al., 1997. MUC3 human intestinal mucin: analysis of gene structure, the carboxyl terminus, and a novel upstream repetitive region. J. Biol. Chem. 272, 26678–26686. https://doi.org/10.1074/JBC.272.42.26678.

Jaiswal, S., Natarajan, P., Silver, A.J., et al., 2017. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. N. Engl. J. Med. 377, 111–121. https://doi.org/10.1056/NEJMoa1701719.

Jennings, L.J., Arcila, M.E., Corless, C., et al., 2017. Guidelines for validation of next-generation sequencing–based oncology panels: a joint consensus recommendation of the association for molecular pathology and college of American pathologists. J. Mol. Diagn. 19, 341–365. https://doi.org/10.1016/J.JMOLDX.2017.01.011.

Kitamoto, S., Yamada, N., Yokoyama, S., et al., 2010. Promoter hypomethylation contributes to the expression of MUC3A in cancer cells. Biochem. Biophys. Res. Commun. 397, 333–339. https://doi.org/10.1016/J.BBRC.2010.05.124.

Kumaran, M., Subramanian, U., Devarajan, B., 2019. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinf. 20, 1–11. https://doi.org/10.1186/s12859-019-2928-9.

Lai, Z., Markovets, A., Ahdesmaki, M., et al., 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 44. https://doi.org/10.1093/NAR/GKW227 e108–e108.

Layer, R.M., Chiang, C., Quinlan, A.R., et al., 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, 1–19. https://doi.org/10.1186/gb-2014-15-6-r84.

Lee, P.Y., Costumbrado, J., Hsu, C.Y., et al., 2012. Agarose gel electrophoresis for the separation of DNA fragments. JoVE J., e3923 https://doi.org/10.3791/3923.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589. https://doi.org/10.1093/BIOINFORMATICS/BTP698.

Liu, X., Han, S., Wang, Z., et al., 2013. Variant callers for next-generation sequencing data: a comparison study. PLoS One 8, e75619. https://doi.org/10.1371/JOURNAL.PONE.0075619.

Martincorena, I., Raine, K.M., Gerstung, M., et al., 2017. Universal patterns of selection in cancer and somatic tissues. Cell 171, 1029–1041.e21. https://doi.org/10.1016/J.CELL.2017.09.042.

Matejcic, M., Mathew, C.G., Iqbal Parker, M., 2019. The relationship between environmental exposure and genetic architecture of the 2q33 locus with esophageal cancer in South Africa. Front. Genet. 10, 447060. https://doi.org/10.3389/FGENE.2019.00406/BIBTEX.

Namini, M., Bhagya, G., Sharma, M., 2025. Personalized approaches to lung cancer treatment: a review of targeted therapies, pharmacogenomics, and combination strategies. Aspects of Molecular Medicine 5, 100073. https://doi.org/10.1016/J.AMOLM.2025.100073.

Nguyen, H.D., Vu, G.H., Kim, W.K., 2024. Identification of biomarkers and molecular mechanisms implicated in genetic variations underlying Alzheimer's disease pathogenesis. Aspects of Molecular Medicine 3, 100045. https://doi.org/10.1016/J.AMOLM.2024.100045.

Ofori, E.G., Kyei, F., Tagoe, E.A., et al., 2025. Mutational analysis of antibiotic resistance genes in Helicobacter pylori from Ghanaian dyspepsia patients: implications for treatment strategies. Aspects of Molecular Medicine 5, 100078. https://doi.org/10.1016/J.AMOLM.2025.100078.

Opoku-Agyeman, P., Ameyaw, P., Bruku, S., et al., 2025. Prevalence of pathogenic genetic variants associated with familial hypercholesterolemia in Ghanaian children. Aspects of Molecular Medicine 5, 100067. https://doi.org/10.1016/J.AMOLM.2025.100067.

Paila, U., Chapman, B.A., Kirchner, R., et al., 2013. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput. Biol. 9, e1003153. https://doi.org/10.1371/JOURNAL.PCBI.1003153.

Patten, V., Bendou, H., Hendricks, D., et al., 2023. The analysis of genetic aberrations in South African oesophageal squamous cell carcinoma patients. http://hdl.handle.net/11427/38531.

Pereira, R., Oliveira, J., Sousa, M., 2020. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J. Clin. Med. 9, 132. https://doi.org/10.3390/JCM9010132, 2020;9:132.

Pertea, M., Shumate, A., Pertea, G., et al., 2018. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 19, 1–14. https://doi.org/10.1186/s13059-018-1590-2.

Pratt, W.S., Crawley, S., Hicks, J., et al., 2000. Multiple transcripts of MUC3: evidence for two genes, MUC3A and MUC3B. Biochem. Biophys. Res. Commun. 275, 916–923. https://doi.org/10.1006/bbrc.2000.3406.

Stratton, M.R., Campbell, P.J., Futreal, P.A., 2009. The cancer genome. Nature 458, 719–724. https://doi.org/10.1038/nature07943, 7239 2009;458.

Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13 (1), 36–46. https://doi.org/10.1038/nrg3117, 2011;13.

Umobong, E., Ugwu, I., Gbaa, Z., et al., 2025. An overview of the current state of cancer diagnosis and treatment in Sub-Saharan Africa. Journal of American Medical Science and Research 4, 18–24. https://doi.org/10.51470/AMSR.2025.04.02.18.

van Putten, J.P.M., Strijbis, K., 2017. Transmembrane mucins: signaling receptors at the intersection of inflammation and cancer. J. Innate Immun. 9, 281–299. https://doi.org/10.1159/000453594.

Vestergaard, L.K., Oliveira, D.N.P., Høgdall, C.K., et al., 2021. Next generation sequencing technology in the clinic and its challenges. Cancers 13, 1751. https://doi.org/10.3390/CANCERS13081751, 2021;13:1751.

Wang, J., Zhou, H., Wang, Y., et al., 2020. Serum mucin 3A as a potential biomarker for extrahepatic cholangiocarcinoma. Saudi J. Gastroenterol. 26, 129. https://doi.org/10.4103/sjg.SJG_447_19.

Xu, H., DiCarlo, J., Satya, R.V., et al., 2014. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genom. 15, 1–10. https://doi.org/10.1186/1471-2164-15-244.