3

METHODOLOGY

Cervical cancer progresses slowly, initially as a precancerous condition known as dysplasia or intraepithelial neoplasia, which may take years to progress to cancer. Early detection at this stage makes the disease highly curable. Advances in medical technology, such as improved screening methods, have significantly enhanced detection rates (Sebutsoe *et al*., 2024). Still, there is no standard comprehensive cancer registry in Tamil Nadu, so that data from neighboring states or regional registries is often used. Less awareness of the signs, associated risks, and prevention strategies of cervical cancer contributes to the high death rates in India, particularly in Tamil Nadu. Healthcare access and other aspects, including sociodemographic, reproductive, and behavioural factors, critically affect prevention and control. Research such as Sunilkumar *et al*. (2024), Yadav *et al*. (2023) and Kadian *et al*. (2021) have pointed out the fact that the examination of sociodemographic and reproductive profiles is important to explain the determinants of cervical cancer in India. Advancements in technologies including NGS, including WES, have been able to describe the mutational profiles, driver genes, and the nature of genetic variants in cancer pathways (Horak *et al*., 2016).

These insights enable targeted therapies and enhance the understanding of cervical cancer mechanisms across ethnic groups. This study addresses the lack of specific epidemiological data in Tamil Nadu by examining the sociodemographic, reproductive, and clinical profiles of advanced cervical cancer patients, underlining the importance of public awareness and molecular profiling in combating this disease.

3.1. Layout of this study

The methodology of this study was designed in a structured sequence of four phases, each contributing to a comprehensive understanding of cervical cancer awareness and treatment outcomes among women in Tamil Nadu, particularly in both rural and urban areas.

In **Phase I**, a survey was conducted among women to assess their knowledge of HPV infection, vaccination, risk factors, symptoms, and screening techniques. This phase aimed to capture the baseline awareness of cervical cancer-related issues among women. The survey identified gaps in knowledge, particularly among women with limited understanding of the disease, which served as a foundation for further research.

Phase II focused on gaining a better understanding of the clinical aspects and treatment outcomes associated with cervical cancer patients. This phase also involved a detailed assessment of the clinical profile, treatment regimens, and sociodemographic characteristics of patients diagnosed with cervical cancer at Sri Ramakrishna Hospital in Coimbatore, Tamil Nadu. This provided a real-world view of the patient population and their treatment journey, offering crucial insights into the disease's impact on the community.

Phase III delved into the genetic reinforcements of cervical cancer, where whole-exome sequencing was performed on cervical cancer patients to identify potential novel gene mutations. This phase aimed to explore the disease's molecular landscape, offering the potential to discover new genetic markers that could aid in better diagnosis, prognosis, and treatment strategies.

To validate the findings from Phase III, **Phase IV** utilized Sanger sequencing to confirm the novel gene mutations identified through whole-exome sequencing. This approach ensured the accuracy and reliability of the genetic data. Together, these four phases constituted a cohesive methodology designed to enhance the understanding of cervical cancer from both clinical and molecular perspectives while addressing gaps in public knowledge and treatment efficacy.

Phase 1

3.2. Study design and study population

An exploratory, questionnaire-based survey was conducted in adherence to the ethical guidelines of Sri Ramakrishna Hospital, Coimbatore, between May and November 2020. The study included female college students aged 17 to 54 years who had no cervical cancer. The method of recruitment was via the internet and social media networks, including emails, and WhatsApp. From the 2,215 persons who completed the screening for eligibility, 70 were turned away, while 2,145 gave their consent to be part of the study. Upon removing duplicates, 2,100 responses were obtained. The participants responded to the survey through Google forms. For clarification, the participants were given an English version alongside a Tamil version of the questionnaire for easy reading.

3.2.1. Data collection

The survey was conducted by distributing questionnaires (Annexure I) through a Google Form, allowing for efficient data collection and management. Participants were provided with a link to the form, which they could access via their smartphones, computers, or other internet-enabled devices, making the process both accessible and convenient. The questionnaire was designed to cover key areas of interest, such as knowledge of HPV infection, vaccination, risk components, warning signs, and screening techniques for cervical cancer.

To ensure the quality and completeness of the data, we implemented measures to guarantee that all registered participants completed the questionnaire in full. Reminders were sent out to those who had started but not yet finished the survey, and follow-ups were conducted to maximize participation. This systematic approach ensured that the dataset was comprehensive, reliable, and representative of the target population, forming a solid foundation for subsequent phases of the study.

3.2.2. Ethical approval

The ethical approval for this study was acquired from the Avinashilingam Institute for Home Science and Higher Education for Women, located in Coimbatore, Tamil Nadu, India (Annexure II). All participants gave their consent before the data was gathered. Additionally, participants were encouraged to extend invitations to other women in their families to participate. It was underlined that all information gathered would be utilised exclusively for the study in order to preserve confidentiality.

3.2.3. Statistical analysis

The responses collected from the survey were thoroughly documented and analysed to extract meaningful insights. Once the data was compiled, it was analysed using Microsoft Excel and GraphPad Prism 8. Microsoft Excel was used for initial data organization, where the responses were categorized and sorted based on key variables. Basic calculations like percentages and averages were derived in Excel to provide an overview of the findings.

Fisher's exact test was employed for the statistical comparison between rural and urban populations. By applying Fisher's exact test, the study aimed to determine whether there were any significant differences between the knowledge and awareness levels of women from rural and urban areas regarding cervical cancer and its prevention. A P-value of less than 0.05 was set as the threshold for statistical significance between the groups.

Through this rigorous analysis, we were able to identify key disparities about cervical cancer awareness among rural and urban populations, which could inform targeted educational interventions and healthcare strategies moving forward.

Phase II

3.3. Study design, setting, and case selection

This hospital-based study presents a descriptive, cross-sectional investigation focusing on cervical cancer patients diagnosed with the disease and seeking treatment at Sri Ramakrishna Hospital in Coimbatore, Tamil Nadu, India. The study was conducted for three years, from January 2020 to December 2022, with the objective of collecting comprehensive data on the experiences, challenges, and treatment outcomes of cervical cancer patients cross different age groups.

In this study, the participants were specifically chosen to reflect a wide age range, and in particular, to include more of the younger patients. People under 30 years of age were included with a minimum age of 29 to understand the implications of cervical cancer in this relatively under represented group. It also enrolled cervical cancer survivors aged 50-69 years, as well as those older than 80 years, with the maximum age limited to 85 years. Age inclusion at a broad spectrum in this case ensured that different phases of life showed variations in how they respond to treatment and their health-related quality of life.

This study focuses on a comprehensive understanding of cervical cancer patients' experiences to identify gaps in knowledge, barriers to effective treatment, and factors influencing health outcomes. The findings are expected to contribute to the existing body of literature on cervical cancer in Tamil Nadu, India, with the ultimate goal of improving prevention, early detection, and treatment strategies for patients across the age spectrum.

3.3.1. Sample size and sampling method

The sample size for this research study was meticulously designed to ensure statistical reliability and validity. The sample size calculation was based on an anticipated response ratio of 50%, a confidence level of 95%, and a margin of error of 5%. Using these parameters, the required sample size for the study was determined to be 381 individuals. This number was calculated to provide sufficient

power to detect meaningful differences and trends within the population under study while maintaining a balance between feasibility and accuracy.

There was no problem of surplus responses as it ensured that the findings were representative of the larger population.

Random sampling was used as the main tool in selecting participants. This method ensures that every member of the larger population has a fair chance of being included in the study. This minimizes selection bias and increases the sample's representativeness. The inclusion and exclusion parameters were applied to ensure that the final sample was relevant and in line with the objectives of the research.

Inclusion criteria

1. The study population included 391 females diagnosed with cervical cancer.
2. Patients from the Tamil Nadu population.

Exclusion criteria

1. Those who declined to participate in the survey.
2. Patients who were diagnosed other than cervical cancer, including cancer of the uterus, endometrial cancer, or breast cancer.
3. Patients not belonging to Tamil Nadu.

3.3.2. Preparation of questionnaire

The questionnaire developed for this study was designed to systematically gather essential information regarding cervical cancer patients' demographic, clinical, and treatment-related characteristics. To make the questionnaire relevant and effective, it was carefully drafted. This process enriched the content validity of the questionnaire, making it comprehensive enough to cover all key areas of interest.

The questionnaire (Annexure III) was divided into four sections which focused on specific aspects of the backgrounds and health status of participants:

1. Demographic Questionnaire: This section of the questionnaire was used to gather fundamental demographic information from the respondents. Some of

the questions in this section included age, occupation, diet, family history of cervical cancer, and menstrual history.

2. Habits, Behavior, and Case History: This section asked the participants questions related to habits and behaviors that might influence their health regarding cervical cancer. Significant questions in this part included the loss of weight or appetite, exposure to radiation earlier, history of surgery, and whether they received traditional or native treatment, among other questions regarding smoking.

3.Clinical Profile: The third section of the questionnaire was regarding the clinical profiles of cervical cancer patients. It contained questions on comorbidities, symptoms presented by patients, and histological features of the cervical cancer diagnosed.

4. Treatment Modalities: The last section of the questionnaire addressed the various treatment modalities that the cervical cancer patients underwent. It documented responses regarding the types of treatment provided, which may include surgery, radiation, chemotherapy, or palliative care options.

5. Survival analysis: Survival analysis was performed to analyze the survival differences between the three treatment modalities in cervical cancer patients. The Kaplan-Meier method was used to estimate survival probabilities over time for three groups of treatment: combination therapy, chemotherapy alone, and radiation alone. Survival time was defined as the time period from the initiation of treatment to either death, which is the event, or the last follow-up, which is censored. Data were analyzed using R software to give graphical representations and accurate estimations of survival patterns. Some of the potential confounding variables include age, comorbidities, and stage of the disease were considered but not fully adjusted due to data limitations.

These findings would go on to advance knowledge into the patient population and would elucidate specific improvements to be sought after in delivering health care as a result and help eventually design interventions leading to improved results for cervical cancer patients.

3.3.3. Patient's response

The study meticulously documented responses related to sociodemographic characteristics and clinical history from a cohort of 391 cervical cancer patients. Further, the study obtained treatment history for a total of 229 cervical cancer patients. It was able to capture information on the types of therapies the patients received. Of the 229 patients, the survival history of 138 patients was adequately documented. These patients were followed for a given period to determine their outcomes. Survival data were analyzed by Kaplan-Meier technique, which estimates survival probabilities at various points in time. By focusing on this subset, the study could evaluate the comparative efficacy of treatment options such as combination therapy, chemotherapy alone, and radiation therapy alone.

3.3.4. Ethical approval

The study received ethical approval from the Ethics Committee of Sri Ramakrishna Hospital, Coimbatore, Tamil Nadu, India (Annexure IV). Before participating in the survey, the participants were briefed about the objectives and purpose of the study. Informed consent was obtained from all participants prior to their participation. Consent was given voluntarily, with participants agreeing to answer the survey questions based on a clear understanding of the research aims. The collected data was securely stored, and access was restricted to maintain confidentiality. By upholding these ethical standards, the study ensured the integrity of its findings while respecting the rights and privacy of the participants.

3.3.5. Statistical analysis

The data collected were analysed using SPSS Version 24.0 (SPSS Inc., Chicago, IL, USA). Descriptive statistics, including percentages, were computed. Descriptive analysis was conducted, with a p-value of <0.05 considered statistically significant. The Chi-square test was used to examine associations between relevant variables.

After gaining insights into the levels of awareness, associated risk factors, and preferred treatment options for cervical cancer, we proceeded to collect tumor samples from patients visiting the hospital. However, in the case of cervical cancer,

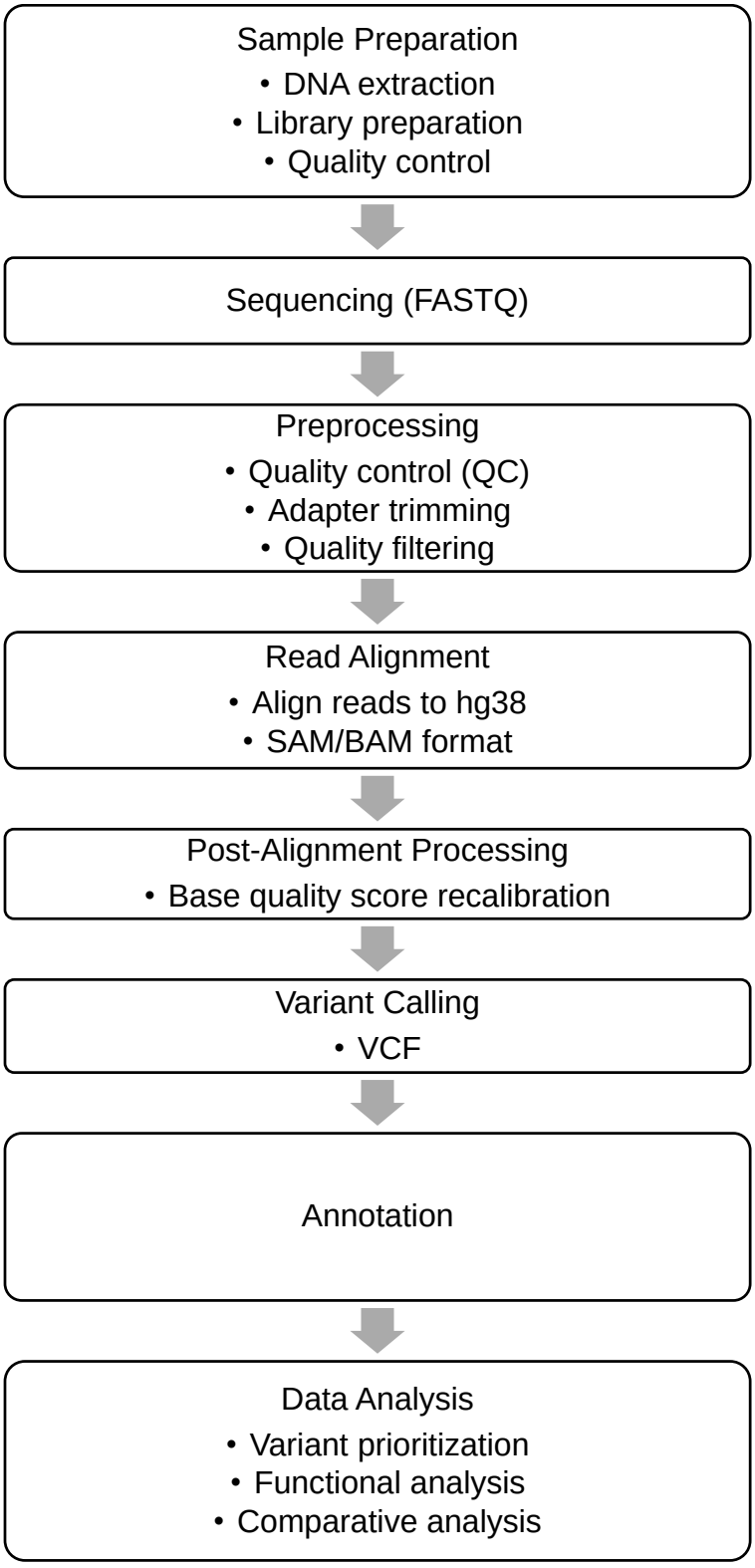
patient awareness is generally low, and many patients seek medical attention only when the disease has reached an advanced stage. As a result, we were able to obtain tumor samples from only five patients for our whole exome sequencing (WES) studies. By focusing on this group, we aim to investigate the genetic mutations and variations associated with cervical cancer, potentially shedding light on the underlying mechanisms of disease progression and identifying any novel biomarkers that could guide future therapeutic interventions.

Phase III

3.4 Patient selection and sample collection

This study focused on five patients, designated as CC1 to CC5, diagnosed with squamous cell cervical cancer at the Oncology division of Sri Ramakrishna Hospital, Coimbatore. The inclusion criteria required histologically confirmed cases of squamous cell cervical cancer, availability of biopsy samples, and informed consent from the patients. Comprehensive patient profiles were documented, encompassing essential information such as age, histology type, nutritional status, anemia and jaundice status, blood type, presenting symptoms, medical history, dietary habits, and comorbidities. This thorough data collection aimed to enhance understanding of cervical cancer's clinical characteristics and factors. Figure 1 provides a detailed illustration of the workflow implemented in phase III of the study. This phase involves a series of systematic steps designed to analyse and interpret the clinical data collected from the five patients diagnosed with squamous cell cervical cancer. The workflow encompasses key processes such as sample collection, data preparation, and analysis methodologies.

Figure 1. Whole exome sequencing (WES) pipeline: from sample preparation to data analysis



3.4.1. DNA isolation and quality assessment

DNA extraction from squamous cell cervical cancer biopsy samples was performed using the QIAamp DNA Mini Kit (Qiagen, CA, USA), which is widely recognized for its efficiency in purifying high-quality DNA from various types of biological samples. The procedure was carried out strictly in accordance with the manufacturer's protocol (as detailed in Appendix I), ensuring consistency and reliability in the DNA isolation process.

After isolation, the DNA was eluted in a volume of 50-100 µl of elution buffer, a buffer specifically designed to maintain the stability of the extracted DNA. The eluted DNA samples were then stored at a temperature of -20°C to prevent degradation, ensuring that the DNA remained intact for further analysis.

To assess the amount and concentration of DNA isolated, two major methods were used. First, a Nanodrop spectroscopy measure was carried out to determine DNA quantity and purity (Appendix II). The quality of the DNA was further checked through agarose gel electrophoresis as shown in Appendix III.

It also allowed for an overall assessment of the quantity and quality of extracted DNA through combining Nanodrop spectroscopy and agarose gel electrophoresis, such that the sample was suitable for downstream applications including whole exome sequencing.

3.4.2. Library preparation

The preparation of DNA libraries for next-generation sequencing was done using Twist Library Preparation EF Kit, a highly efficient system from Twist Bioscience, CA, USA. The library preparation steps were done according to the provided manufacturer protocol for detailed preparation steps (Appendix IV), thus ensuring high-quality and reproducible libraries for the sequencing runs.

During library preparation, DNA fragments were produced, and the generated libraries were carefully sized to ensure that they fall within the 300-500 base pair range. This size range is ideal for most sequencing platforms because it strikes a balance between efficient sequencing and accurate read alignment during

downstream bioinformatic analysis. It is important to maintain specific size range to obtain high-quality sequencing data and to maximize the coverage of targeted regions during sequencing.

After the preparation of the libraries, the quality of the DNA libraries was evaluated based on two critical parameters: library concentration and insert size.

1. Library Concentration (ng/µl): The concentration of the DNA libraries was measured to ensure that sufficient DNA was available for sequencing.
2. Insert Size in base pairs-This is defined as the DNA fragments length, and it usually includes the bases between the sequences used for preparation of the adapter during library preparation. A significant deviation in insert size would impact the quality of sequencing, alignment accuracy, and overall integrity of the data.

These validation steps were instrumental in ascertaining that the DNA libraries were good quality and qualified for sequencing at appropriate concentrations with fragment sizes suitable for reliable and reproducible sequencing data. All these stringent quality control measures by the research group ensured that NGS results were robust, accurate, and indicative of the true genetic landscape of the studied samples.

3.4.3. Next-Generation Sequencing (NGS) - Whole Exome Sequencing (WES)

WES was performed on the Illumina NovaSeq 6000, which is one of the latest and most efficient sequencing technologies that are high-throughput and highly accurate. The use of this platform ensures full and high-quality coverage of the exome to ensure detection of common and rare genetic variants that could be linked to cervical cancer in the cohort under study. The sequencing was performed in Ahmedabad, India, strictly adhering to the standard WES protocol, which is outlined in Appendix V. This ensures consistency and reproducibility in the sequencing process.

3.4.4. Base calling and quality control

The base calling and quality control (QC) processes in the Whole Exome Sequencing (WES) workflow are critical for ensuring high-quality and reliable sequencing data (Appendix VI). For this study, these processes included several steps aimed at removing artifacts, improving the accuracy of sequence reads, and preparing the data for downstream analysis. The following tools and methods were employed for base calling, quality filtering, and alignment:

1. Adapter Removal: During sequencing, adapters are ligated to DNA fragments to facilitate their recognition by the sequencing machine. After sequencing, these adapter sequences can still be present in the raw data and need to be removed before further analysis.
2. Trimming of Low-Quality Bases: Low-quality bases—typically found at the ends of sequencing reads—can introduce errors into the data if left unfiltered. This step minimizes errors in variant calling and increases the reliability of the sequencing results.
3. Elimination of Redundant Reads and Contaminants: Reads can be duplicated in the course of sequencing due to PCR amplification. The duplicates do not add any more information and will tend to cause overcalling variants, so duplicates were identified and removed.
4. Alignment of Sequences Using Burrows-Wheeler Aligner (BWA): After the data were cleaned and trimmed, the high-quality sequence reads were aligned to the human reference genome using the Burrows-Wheeler Aligner (BWA), which is one of the most popular alignment tools, known for its efficiency and accuracy in handling large datasets.
5. Generation of Output Using SAM Tools: The reads were stored in Sequence Alignment Map (SAM) format after alignment, which is a standard file format for storing large-scale sequence alignment data. These files were then processed and converted using the SAM tools suite into Binary Alignment Map (BAM) files, being a compressed version of the SAM files, thereby easier to

store and manipulate (Appendix VII). SAM tools were also used to sort, index, and filter the aligned reads further improving the quality of the dataset.

The final BAM files contained the final cleaned and aligned reads, which were ready for downstream analyses, including variant calling and mutation identification.

3.4.5. NGS data analysis

NGS data quality control parameters were assessed, including total reads, reads passed filters (%), and data reduction during filtering (Mb). Bioinformatics tools were employed for somatic alterations, variant calling and annotation, and gene expression analysis.

3.4.6. Somatic data processing and variant filtering in squamous cell cervical cancer analysis

We analysed somatic single-nucleotide variations (SNVs) in five squamous cell cervical cancer samples, calculating mutation frequencies and identifying a total of variants. An allele frequency filter (<0.05) was applied to focus on high-confidence mutations, followed by categorization into exonic, intronic, untranslated region (UTR), and splice site variants. A read depth filter (>20 reads) was used for improved data reliability, and variants were cross-referenced with the dbSNP database.

3.4.7. Transition and transversion analysis in cervical cancer mutations

To analyse nonsynonymous single nucleotide polymorphisms (nsSNPs) in squamous cell cervical cancer, a comprehensive methodological approach was employed to investigate their potential role in disease pathogenesis. Nonsynonymous mutations with high confidence were filtered from the set, requiring a higher read depth to increase the confidence in the variant's detection. All mutations were classified into the following types of nucleotide substitution: transitions, such as purine-to-purine or pyrimidine-to-pyrimidine substitutions, and transversions, including purine-to-pyrimidine or vice versa substitutions, for the analysis of patterns of nucleotide changes.

In order to understand the mechanisms of mutational processes, the different types of nucleotide substitutions were analyzed and their frequencies calculated.

The data after processing were depicted graphically in mutation spectrum plots to illustrate the distributions and patterns of these substitutions.

3.4.8. Nonsynonymous mutation profiling in squamous cell cervical cancer: a comparative analysis with COSMIC data

In this study, we focused only on nonsynonymous mutations that change the amino acid sequences of proteins and are likely to have an impact on their functions. We excluded synonymous mutations and intronic variants in order to carry out a more accurate analysis of mutations with functional implications. The identified nonsynonymous mutations from our dataset were then compared to squamous cell cervical cancer data from the COSMIC (Catalogue of Somatic Mutations in Cancer) database (Tate *et al*., 2019). This comparison aimed to validate and contextualize the findings from our study by contrasting the mutation rates and counts in our cohort with those available in the COSMIC database.

The study involved analyzing nonsynonymous mutations in several key genes. For each gene, the mutation rate and total number of mutations identified in our cohort were compared with the corresponding data in the COSMIC database. This comparison allowed for the identification of consistent mutation patterns across both datasets and provided insights into the genetic landscape of squamous cell cervical cancer. By comparing our findings with a larger, well-established database, we aimed to enhance the reliability and relevance of our results, thereby contributing to a better understanding of the genetic alterations driving this cancer type.

3.4.9. Mutation spectrum analysis

We adopted a multifaceted approach utilizing various predictive algorithms to assess the impact of nonsynonymous single nucleotide polymorphisms on protein structure and function. Nonsynonymous SNPs, which result in amino acid changes in proteins, can significantly affect protein function, stability, and interactions. Therefore, understanding their potential pathogenicity is critical for unravelling their role in diseases, particularly cancer.

The algorithms we employed, SIFT, PolyPhen-2, LRT, MutationTaster, FATHMM, RadialSVM, and Logistic Regression (LR), are widely used for predicting the functional impact of genetic variants, particularly missense mutations (Appendix XIII-XIV). Each tool utilizes distinct methodologies to assess the likelihood that a given variant affects protein function or contributes to disease (Dong *et al*., 2015).

SIFT (Sorting Intolerant From Tolerant) predicts variant effects based on evolutionary conservation of amino acid residues and their chemical properties, classifying substitutions as deleterious or tolerated depending on their impact on protein function. Similarly, PolyPhen-2 evaluates amino acid substitutions by combining sequence conservation, structural analysis, and known functional sites, providing scores that range from benign to probably damaging. LRT (Likelihood Ratio Test) focuses on evolutionary constraints by analysing sequence conservation at variant sites, categorizing mutations as deleterious, neutral, or unknown. MutationTaster incorporates nucleotide conservation, splice site predictions, and functional annotations to classify variants as either disease-causing or benign polymorphisms. FATHMM (Functional Analysis Through Hidden Markov Models) predicts if a mutation is neutral or harmful by combining hidden Markov models with evolutionary conservation. Radial Support Vector Machines (SVMs) uses a machine learning technique to create probabilistic predictions about the impact of variants by combining factors such as protein structure, biochemical properties, and sequence conservation. Logistic Regression (LR) employs statistical modelling to predict variant effects by combining conservation scores, structural information, and functional annotations. In summary, these tools offer complementary insights into the functional consequences of genetic variants and aid in the identification of potentially pathogenic mutations, thus furthering our understanding of their role in squamous cell cervical cancer.

In addition to the above-mentioned tools, several other scores were used to evaluate the functional relevance of genetic variants in squamous cell cervical cancer. The VEST3 score assesses the potential pathogenicity of a variant by analyzing its evolutionary conservation pattern. CADD phred combines multiple

annotations to assign a score, indicating how likely a variant is to affect gene functions. GERP++_RS measures sequence conservation to identify evolutionary constraints, with higher scores indicating regions more likely to be functionally important.

The PhyloP46way placental and PhyloP100way vertebrate scores assess the evolutionary conservation of sequences across different species, providing insights into the variant's importance in placental mammals and vertebrates. Lastly, SiPhy 29way logOdds scores variants based on their conservation across 29 species, offering a probability that a variant is deleterious. These tools together provided a comprehensive analysis of variants, helping to identify those that could contribute to the development of squamous cell cervical cancer.

3.4.10. Functional enrichment analysis

To understand the biological importance of the mutations found in our study, we used the DAVID bioinformatics tool (Sherman *et al*., 2022) to perform Gene Ontology (GO) and KEGG pathway enrichment analyses. These analyses identify the functions and pathways affected by genetic changes, which gives insight into their role in cellular and molecular processes.

Gene Ontology is a standardized vocabulary that describes gene functions across different species. Using GO enrichment analysis, we attempted to uncover the significant biological processes and cellular components associated with the genes harboring the identified mutations.

1. Biological Processes: Enriched biological processes are recognized to provide an insight into the cellular functions affected by mutations that could correlate with disease mechanisms or tumorigenesis.
2. Cellular Components: GO describes the location within the cell where gene products are active, such as organelles, complexes, and cellular structures. The identification of which cellular components are affected can shed light on how mutations interfere with normal cellular functions, potentially contributing to cancer progression.

3. Molecular Functions: Analyzing enriched molecular functions helps to identify the specific biochemical roles of the mutated genes and how these changes may influence cellular signaling and metabolism.

3.4.11. Protein-Protein Interaction network analysis

To further deepen the biological understanding of the results obtained from the mutations identified in our study, we created an extensive PPI network based on the instructions from Zhou et al. (2019) using the Network Analyst tool. This analysis was crucial for understanding how the mutated genes interact with one another and their roles within broader biological contexts.

PPI networks are graphical representations of protein interactions within a biological system. Each node in the network represents proteins, and their interactions are indicated by the edges connecting them. The analysis of the PPI network will provide insight into the functional relationships among the proteins encoded by the mutated genes and identify key players within these networks.

1. Data Source and Methodology: The first step towards building the PPI network involved gathering interaction data from reliable databases such as STRING, BioGRID, and IntAct.
2. Visualisation: The results of the interaction data were input into Network Analyst to visualize and analyze the network. The PPI network was visualised to give a more intuitive understanding of the interaction between the mutated proteins.

3.4.12. Hub gene verification using the UALCAN database

Construct a PPI network to focus on genes that are significantly interacted and might interfere with biological processes. These hub genes were determined based on the significance of their interaction within the PPI network. Statistical significance of gene expression was determined by applying appropriate tests for differences between conditions, with thresholds for significance used to select genes for validation.

The methodology for this study involved analysing gene expression and identifying potential biomarkers for squamous cell cervical carcinoma in patients aged 50–55 years. Five cervical cancer patients in the study were confirmed to have squamous-type cervical cancer within this specific age group. Due to the unavailability of normal sample sequencing data within the study's collection, gene expression data from the UALCAN database (Chandrashekar*et al*., 2022), which integrates TCGA and GTEx datasets, was utilized. This was the reason for choosing this criteria from the UALCAN database, allowing for comparison across three conditions: Normal vs. Primary Tumor, Squamous Cell vs. Normal, and Normal vs. Age (41–60 years).

The UALCAN platform was employed to validate the expression levels of hub genes across the study conditions. This platform enabled comparisons with external datasets, ensuring a robust evaluation of gene expression and addressing the limitations of the study's dataset. The workflow incorporated data filtering to focus on meaningful expression patterns and excluded genes with non-significant findings. This methodological framework facilitated the comprehensive analysis of gene alterations, with a focus on identifying biomarkers relevant to the progression and characterization of squamous cell cervical carcinoma.

Phase IV

3.5. Mutational profiling and sanger sequencing confirmation for novel variants

3.5.1. Identification of novel variants

The study aimed to identify and validate novel nonsynonymous variants associated with squamous cell cervical cancer using whole-exome sequencing. The sequencing data were initially processed to filter out common variants, focusing on identifying potential novel genetic alterations. A total of 2389 variants were identified in the exome data, with 182 nonsynonymous variants found to be unreported in the dbSNP database. These unreported variants could represent novel mutations associated with SCC. To prioritize variants for further analysis, the study matched the identified variants with the ClinVar database (Landrum *et al*.,

2014). The ClinVar database contains previously reported variants with clinical significance. From the 182 identified unreported variants, those previously classified in ClinVar as benign, of uncertain significance, or pathogenic were noted, and the rest were selected for further examination. Due to sample size limitations, the study did not validate all the identified variants but selected a subset for validation.

These eight nonsynonymous mutations in seven genes were chosen for validation based on their potential relevance to cancer biology. Genomic locations of each mutation, such as the chromosome and position, along with nucleotide changes and exon positions including resultant amino acid substitutions, were taken into consideration. The number of sequencing reads for each mutation was further considered to ensure enough coverage for variant calling. Higher read depths for mutations indicated greater validity to support validation. These variants were considered for their implications in cancer-related biological processes and the potential to be explored in future functional studies.

3.5.2. Mutation spectrum analysis

To confirm the effects of novel nonsynonymous SNPs on protein structure and function, we applied a set of prediction algorithms: SIFT, PolyPhen-2, LRT, MutationTaster, FATHMM, RadialSVM, and Logistic Regression (LR) (Appendix XIII-XIV). These are programs that evaluate the pathogenicity of genetic variants by estimating their potential influence on protein function.

By using these algorithms, we would be able to identify and prioritize potentially damaging mutations among the novel SNPs identified in our study, thus contributing to understanding their role in disease pathology. This multifaceted approach enables a comprehensive evaluation of variant pathogenicity, facilitating further investigation into their clinical significance in squamous cell cervical cancer.

3.5.3. Primer design

To validate the identified variants, specific primers designed for targeted polymerase chain reaction (PCR) amplification. This was done using Primer3Plus,

a widely used software tool that facilitates primer design based on user-specified parameters (Untergasser *et al*., 2007).

The initial phase of primer design requires identifying the target gene and the specific variant region within that gene. It is essential to focus on the exact nucleotide sequence where the mutation occurs to ensure that the primers will amplify the desired segment of DNA. This targeted approach is critical for achieving accurate validation of the variants identified during the sequencing process.

Setting parameters for primer design

Once the target region was defined, several key parameters must be specified to optimize primer design:

- **Amplicon size**: The desired size of the PCR product, typically ranging from 100 to 500 base pairs. A smaller amplicon size is often preferable for high specificity and efficient amplification, especially when dealing with low-template DNA samples.
- **Length**: Primers should generally be between 18-24 base pairs long. This length strikes a balance between specificity and binding stability, allowing the primers to anneal effectively to the target sequence without forming nonspecific products.
- **GC content**: The guanine-cytosine (GC) content of the primers should ideally be between 40-60%. GC-rich primers tend to have higher melting temperatures (Tm), which is beneficial for specificity during the annealing phase of PCR. Maintaining an appropriate GC content helps ensure the primers effectively bind to the target sequence without forming secondary structures.
- **Melting temperature (Tm)**: The melting temperature of the primers should be between 55-65°C. Tm is crucial for determining the annealing conditions of the PCR. The Tm of both forward and reverse primers should be similar to ensure they bind to the target DNA at the same temperature during the PCR cycling process.

Optimizing primer performance

After designing the primers, it is important to conduct further evaluations to confirm their efficacy. This involved:

- ***In silico* analysis**: Using software tools to assess primer specificity against the reference genome, checking for potential off-target bindings that could lead to nonspecific amplification (Appendix XV).
- **Experimental testing**: Performing PCR with the designed primers using genomic DNA from relevant samples to evaluate amplification efficiency, specificity, and yield.

Primers can be designed and optimized, and this guarantees reliable amplification of the regions of interest harbouring the found variants. Subsequent validation may involve sequencing PCR products or carrying out further studies for confirmation of mutation presence and impact.

3.5.4. Touch down PCR amplification

Touchdown PCR is a variation of the regular PCR, which enhances specificity. Denaturation of DNA takes place at 94°C; it then undergoes a touchdown cycle of 10-15 cycles with stepwise lowering of the annealing temperature, such as from 65°C to 55°C, then continues with 20-25 cycles at a specific temperature. Each iteration contains the following steps: denaturation at 94°C for 30 seconds, annealing for 30 seconds, and extension at 72°C for 1-2 minutes. This protocol minimizes nonspecific binding and maximizes specificity, making it ideal for further analysis. The PCR products were followed by gel electrophoresis to verify the amplicon's presence and size. This confirms that the PCR reaction has amplified the target regions as Appendix XVI explains.

3.5.5. Sanger sequencing

The PCR products were then subjected to sanger sequencing to confirm the presence of the genetic variants from each sample and ensure its accuracy. The PCR products were purified and prepared for sanger sequencing using standard protocols to obtain optimal results.

After completing the Sanger sequencing procedure (Appendix XVII), the produced chromatograms from this step were closely read to confirm the variant calls that were done at the above stages of genetic analysis. A chromatogram for every peak visualizes the nucleotide sequence; peaks in the chromosome are DNA bases. The quality of these peaks were assessed in terms of characteristics such as peak height, sharpness, and overlapping of signals, which may suggest errors in sequencing. Comparing observed variants with reference sequences would enable the researchers to know the accuracy of the data acquired through sequencing.

This validation step is important because it confirms the mutations identified are real and not artifacts of the sequencing procedure. Results from this analysis will make significant contributions to understand the genetic landscape of squamous cell cervical cancer, providing a basis for possible future applications in diagnostics and targeted therapies.

The following chapter gives a thorough presentation of the observations and results obtained by applying the research methodology described herein.