scientific reports

Check for updates

OPEN

Feasibility of long-read sequencing to identify molecular alterations in an Indonesian cohort of locally advanced to advanced nasopharyngeal cancer

Handoko[1,2]✉, Marlinda Adham[3], Lisnawati Rachmadi[4], Demak Lumban Tobing[5], Asmarinah[6], Fadilah[7,8], Wei Dai[9,10], Anne Wing Mui Lee[10] & Soehartati A Gondhowiardjo[2]✉

Nasopharyngeal carcinoma (NPC) is prevalent in Southeast Asia, particularly in Indonesia. Despite advances in treatment, patients with advanced NPC face poor outcomes. Examining the NPC mutational landscape is crucial for understanding its biology and enable potential new therapeutic strategy. To characterize the landscape of single nucleotide variants (SNVs), structural variants (SVs), copy number variations (CNVs), and short tandem repeats (STRs) in locally advanced to advanced NPC within an Indonesian cohort using long-read sequencing. Six fresh-frozen nasopharyngeal biopsy samples were collected from the NPC biobank. DNA was extracted and sequenced using Oxford Nanopore's Promethion 2 Solo long-read sequencer. Structural and small variants were identified and annotated. The SNVs, SVs, and CNVs were categorized based on predicted effects, and key findings were validated using external RNA-seq data. Copy number loss genes were checked against the Tumour Suppressor Gene database (TSGene v2.0). Genetic findings were correlated with patient clinical histories. Approximately 4.4 to 5.1 million SNVs were identified per sample, with 0.023% categorized as high consequence. Notable tumour suppressor genes, such as LIMD1 and *CNDP2*, were frequently mutated. Around 30,000 to 41,599 SVs were detected per sample. High-consequence tumour suppressor gene SVs were identified in *EPHA3*, *CASP8*, *DMBT1*, *ZFHX3*, and *IRF5* gene. Common copy number tumour suppressor gene loss observed in *RNH1*, *H19*, *CDKN1C*, and others, suggesting their role in NPC carcinogenesis. Copy number gains were found in potential oncogenes such as *Y RNA*, *LTO1*, and *FADD*. Pathogenic short tandem repeats (STRs) in *PABPN1* and *RFC1* were identified in three samples, presenting a novel association with NPC. NPC sample which exhibited significant genomic instability had the shortest survival, potentially linked to multiple defective DNA repair genes. This study utilized long-read sequencing to identify a complex spectrum of genetic alterations, including numerous SVs and potentially pathogenic STRs, in Indonesian NPC. Extensive DNA repair gene defects, primarily complex SVs detectable by long reads, were observed and highly possibly associated with poor survival. These findings underscore the potential of long-read sequencing for uncovering clinically relevant mutations in NPC.

Keywords Nasopharyngeal carcinoma, Mutational landscape, SNVs, SVs, CNVs, Indonesia, Tumour suppressor genes, DNA repair, Copy number variation

[1]Doctoral Program in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia. [2]Department of Radiation Oncology, Faculty of Medicine, Universitas Indonesia, Cipto Mangunkusumo National General Hospital, 10430Jl. Salemba Raya No.4, DKI, Jakarta, Indonesia. [3]Department of Otorhinolaryngology - Head and Neck Surgery, Faculty of Medicine, Universitas Indonesia, Cipto Mangunkusumo National General Hospital, Jakarta, Indonesia. [4]Department of Anatomical Pathology, Faculty of Medicine, Universitas Indonesia, Cipto Mangunkusumo National General Hospital, Jakarta, Indonesia. [5]Department of Clinical Pathology, Dharmais Cancer Hospital, Jakarta, Indonesia. [6]Medical Biology Department, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia. [7]Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia. [8]Bioinformatics Core Facilities, Faculty of Medicine, Indonesian Medical Education and Research Institute (IMERI), Universitas Indonesia, Jakarta, Indonesia. [9]Department of Clinical Oncology, Centre of Cancer Medicine,

School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong (SAR), PR China. [10]Clinical Oncology Center, University of Hong Kong – Shenzhen Hospital, Shenzhen, People's Republic of China. ✉email: handoko12@ui.ac.id; gondhow@gmail.com

Nasopharyngeal carcinoma (NPC) is a malignancy with distinct geographic distribution, having high prevalence in Southeast Asia, particularly in Indonesia[1]. Despite advances in therapeutic strategies, patients with locally advanced to advanced NPC often face poor outcomes, underscoring the need for a deeper understanding of the disease's molecular landscape[2]. The mutational profile of NPC, including single nucleotide variations (SNVs), structural variations (SVs), and copy number variations (CNVs), has been increasingly recognized as crucial in influencing tumour behaviour, response to treatment, and overall survival (OS).

The underlying basic of cancer is the alteration in the genomic sequence, leading to various downstream changes in protein level. Genetic sequence in certain important regions such as the regions that code for DNA repair is of upmost importance to be preserved. Alteration in that important regions very often results in catastrophic condition, including the development of cancer. The alteration can take in various forms from small nucleotide changes until large segment of nucleotide alteration. The alteration can be a simple small nucleotide variant (SNV), involving only 1 base change until a complex structural variation (SV) consisting of deletion, insertion, duplication, translocation, and the mix of all the possibilities.

The SNVs are readily detectable by any next generation sequencing (NGS) technology. However, SVs are more difficult to detect because it spans a larger segment of nucleotides[3]. Furthermore, SVs can also be very complex for example in translocation with triplication[4]. This complicates the process of detecting SVs. While not as extensively studied as SNVs, SVs are now increasingly recognized for their significant role in cancer development and progression. There's a growing body of research and publications on SVs. Primarily due to the availability long read sequencers, for example NGS from Oxford Nanopore or PacBio, which can better capture major SVs[5].

Nasopharyngeal cancer (NPC) is a type of solid cancer with relatively lower somatic mutation frequency compared to other solid cancers[6]. The somatic mutation referred here is generally SNVs. Compared to SNVs, SVs are generally less frequent in NPC, but their impact can be more profound due to their larger size and potential to disrupt entire genes or regulatory regions. NPC is also unique from other solid cancer due to its association with Epstein Barr Virus (EBV), especially those NPC in endemic region[1,7–9].

Recent studies have highlighted the heterogeneity of NPC at the molecular level, revealing a complex interplay of genetic alterations that drive oncogenesis and impact clinical outcomes[10]. However, much of the existing research has focused on populations outside of Indonesia, leaving a gap in our understanding of the specific mutational patterns prevalent in Indonesian patients. Given the unique genetic and environmental factors influencing NPC in this region, it is imperative to explore the mutational landscape of Indonesian NPC cases to uncover potential biomarkers that could guide personalized treatment approaches[11].

In this study, we present a comprehensive analysis of the mutational landscape of locally advanced to advanced NPC in an Indonesian cohort. By examining SNVs, SVs, CNVs, STR and correlating these findings with patient clinical histories, including overall survival and response to treatment, we aim to identify genetic alterations that may serve as prognostic indicators or potential therapeutic targets. This research not only contributes to the global understanding of NPC but also addresses the pressing need for region-specific data that could enhance clinical management and improve patient outcomes in Indonesia.

Method

Samples were clinical nasopharyngeal biopsy which have been confirmed histo-pathologically to be nasopharyngeal cancer. Samples were fresh frozen biopsy tissue from Department of Radiation Oncology Cipto Mangunkusumo General National Hospital – Faculty of Medicine, Universitas Indonesia NPC biobank. The storage of fresh tissue samples was in -80 °C liquid nitrogen tank. Ethical approval was obtained prior to this study from Faculty of Medicine, Universitas Indonesia ethical committee in accordance with the Declaration of Helsinki. All patients contributed their biospecimen had been informed and had prior consent obtained.

Six fresh frozen samples were retrieved. The samples minimal weight was 20 mg, thawed in room temperature. Then tissue homogenization was done mechanically using scalpel with 1–2 drops of phosphate-buffered saline (PBS). If the tissue has not disintegrated well, sonicator was used to better disintegrate the tissue. The DNA was then extracted with Qiagen DNA Mini Kit following the manufacturer instructions. The extracted DNA was assessed for quantity and purity using Qubit fluorometer and Nanodrop spectrophotometers, respectively.

Following DNA extraction, whole genome DNA sequencing was carried out using long read sequencer Promethion 2 Solo from Oxford Nanopore. The library preparation was carried out with Ligation sequencing DNA V14 (SQK-LSK114) protocol. The sequencing was done with Promethion flow cell version R10.4.1 and using kit 14 chemistry. One Promethion flow cell was used one for each sample. The sequencing ran until the flow cells pores were all used up completely (around 72–120 h). Reloading of sample DNA library to the flow cell was done 2–3 times, whenever the pores occupancy has gone below around 20%.

The raw data obtained was in the format of POD5. The raw data underwent basecalling with super accuracy mode dna_r10.4.1_e8.2_400bps_sup@v4.2.0 model using Dorado basecaller. Alignment of tumour DNA was done with human GRCh 38 reference genomes using Minimap2[12,13]. Alignment to Epstein Barr Virus (EBV) DNA / human herpes virus type 4 reference genome (NC_007605.1) was also done to confirm the presence EBV DNA within the NPC samples[7,8,14]. Reads with quality score below Q10 were excluded from alignment process.

The aligned tumour DNA sequence in BAM format underwent further structural variant calling using Sniffles2[13,15], small variant calling using Clair3 with coverage / read depth of 10 and above as the filter threshold[16]. Variant annotation of SV and SNV was performed using SnpEff[17]. Predicted functional consequences (high, moderate, low, modifier) were assigned based on Ensembl definitions. 'High consequence' variants (e.g.,

frameshift, stop gained, splice acceptor/donor) are predicted to cause loss-of-function. Variants were cross-referenced with the TSGene v2.0 database from University of Texas[18], ClinVar, and the OncoVar database (compiling known driver mutations from The Cancer Genome Atlas / TCGA and International Cancer Genome Consortium / ICGC). Copy number variants (CNV) calling was also done with QDNASeq with bin size of 15kb[19]. Short tandem repeat (STR) genotyping was done with Straglr[20]. It is important to note that formal ACMG classification guidelines were not applied for de novo pathogenicity assessment; interpretations rely on predicted functional impact and overlap with known cancer variants/genes in public databases.

The SNV transcript variant type was categorized based on its predicted consequences, either high, moderate, low, or modifier impact. The high impact was very likely to have the gene function altered. All high impact variants genes were filtered for tumour suppressor gene and validated toward an independent transcriptome data to determine whether it was downregulated or not. The transcriptome data used was from nasopharyngeal cancer transcriptome profiling RNA Seq (data obtained from European Bioinformatics Institute, available from https://www.ebi.ac.uk/gxa/experiments-content/E-GEOD-12452/resources/DifferentialSecondaryDataFiles.Microarray/normalized-expressions).[21,22]. The SNVs result was also assessed for possibility of oncogenic variants. We matched our SNV data toward OncoVar database which contained known oncogene from pan cancer TCGA and ICGC database[21]. We matched the exact gene and exact transcript variation position.

The copy number loss genes were further checked whether they are known tumour suppressor genes or not toward TSGene version 2.0. Furthermore, those copy number loss tumour suppressor gene were validated using similar independent nasopharyngeal cancer transcriptome profiling RNA Seq from European Bioinformatics Institute[22,23]. The copy number loss tumour suppressor genes with significantly lower expression in NPC tissue relative to normal nasopharyngeal tissue (p value < 0.05 and FDR value 0.05) were reported and they were likely to the driver of carcinogenesis in NPC. The copy number gain was also checked against the independent transcriptome profiling data to externally validate the possible upregulation. Summary of the bioinformatics is shown in Fig. 1.

Further analysis of SVs and SNVs VCF data were done specifically on targeted genes. We assess key genes commonly involved in DNA repair process include *ATM*, *BRCA1*, *BRCA2*, *CDK12*, *CHEK2*, *PALB2*, *POLE*, *RAD51B*, *TP53*, group of *FANC* genes *(FANCA*, *FANCC*, *FANCD2*, *FANCE*, *FANCF*, *FANCG*, *FANCI*, *FANCL* and *FANCM*), and group of MSH genes (*MSH2*, *MSH3*, *MSH4* and *MSH6*) as described in the study of SV in

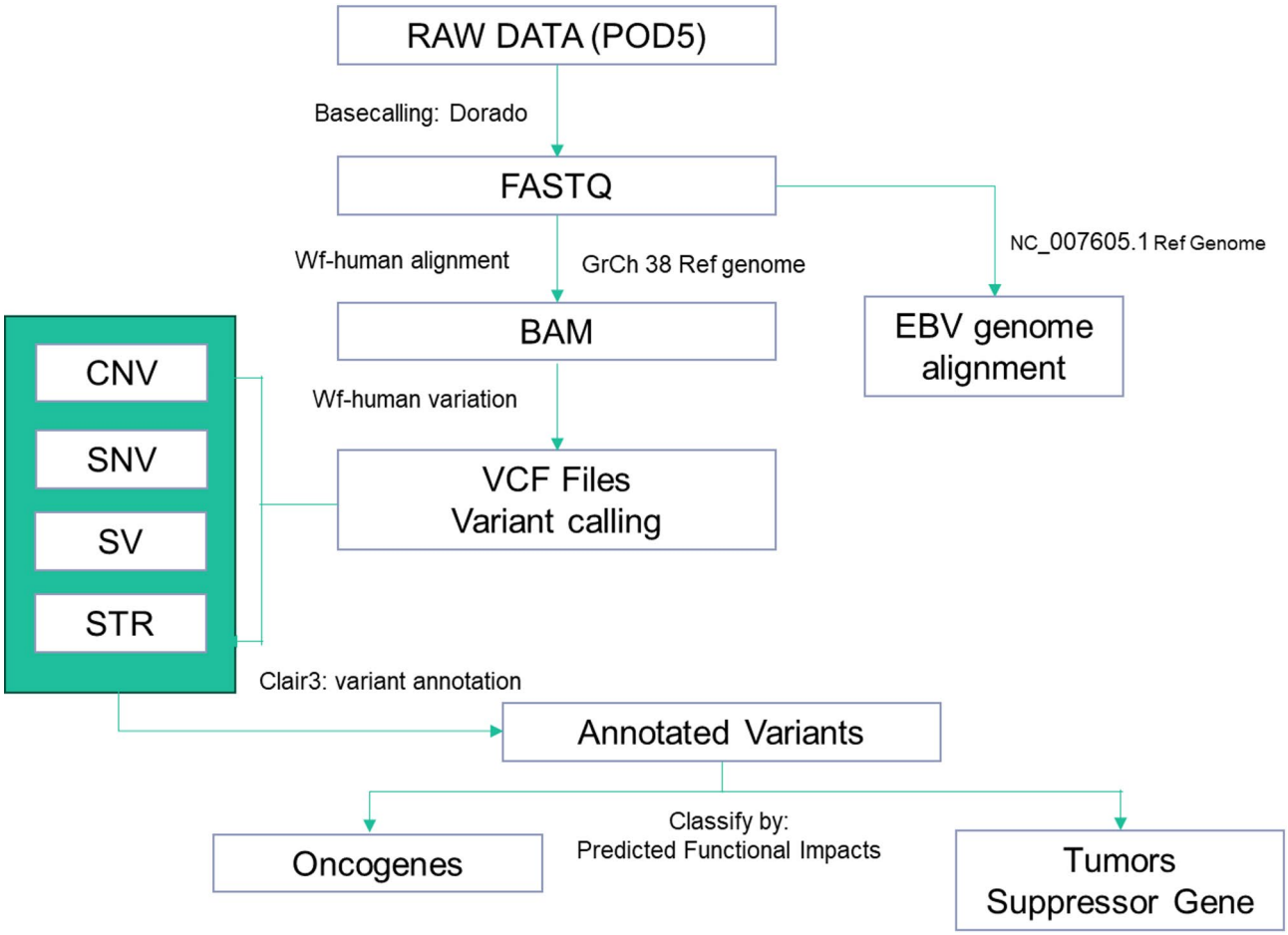

Fig. 1. Summary of bioinformatics workflow.

human cancer genomes[24]. Assessment of the variants was done first for all variant types, then filtered to exclude less clinically relevant variants that has low, moderate and uncertain consequences (modifier).

The patients' clinical history was retrieved from hospital electronic medical record. Information collected including initial patients' demographic, staging based on the American Joint Committee on Cancer (AJCC) version 8, treatment record, histopathology report, all imaging reports, clinical follow ups and overall survival. All patient's clinical information was anonymised. All patients did not have prior radiation or chemotherapy. The clinically relevant variants were tabulated and a correlation with patients' clinical history were done.

Result

NPC small variants (SNVs) mutational landscape

There were around 4.4 to 5.1 million SNVs across whole NPC genome from all samples. Around 2.5–2.8% of those SNVs were occurring in non-intron and non-intergenic region. Based on the predicted consequences, on average across all samples, there were 0.023%, 0.265%, 0.346%, and 99.367% of predicted high, moderate, low and modifier consequences, respectively. The distribution of SNVs across all samples by predicted consequences were summarized in Fig. 2. There were few SNVs categorized as others. The others category was SNVs that fit into multiple SNV types, the detail was available as Supplementary Data 1.

The SNVs predicted to have high consequences were very likely to have lost its function. They were all checked for tumour suppressor genes against TSGene version 2.0 database. The *LIMD1* and *CNDP2* were two tumour suppressor genes detected to harbour frameshift variant or splice acceptor variant that were categorized as high consequences and found in all our NPC samples. When all those detected high consequences tumour suppressor gene were checked against independent transcriptome data, it was found that *LIMD1* and *CLDN23* gene were downregulated. Those genes were potentially altered gene that may drive NPC progression and warrant further experimental validation. Other relatively frequent tumour suppressor gene variants that were detected in at least half of our NPC samples and categorized as high consequences were *PKL5* and *TMPRSS11A* gene. The complete list of SNV variants with high consequences and annotated for tumour suppressor gene was available in Supplementary Data 2.

The SNVs from all samples were also annotated toward ClinVar database. We detected *PERM1* (NM_001291366.2:c.2330T > C) and *APOA2* (NM_001643.2:c.-323 C > T) mutation across all our NPC samples. Those mutation was considered pathogenic by ClinVar. Furthermore, the *PERM1* c.2330T > C (p.Val777Ala) missense mutation was also commonly reported in upper aerodigestive tissue, head and neck squamous cell carcinoma in COSMIC database. The SNV data was also checked toward published public known oncogene database based on TCGA and ICGC pan cancer driver mutation data (OncoVar). We detected the oncogene that have an exact match toward the gene and mutation position.

The detected oncogene were *PDE5A* [NM_001083.4:c.1465G > A(p.Glu489Lys)], *ABCB7* [NM_004299.6:c.2059G > A(p.Glu687Lys)], and *HERC2* [NM_004667.6:c.14305 C > T(p.Pro4769Ser)] for sample S1, S5, and S6 respectively that matches TCGA oncogene database. We did not detect any matched oncogene in sample S2, S3, and S4 that match TCGA oncogene database. Toward ICGC oncogene

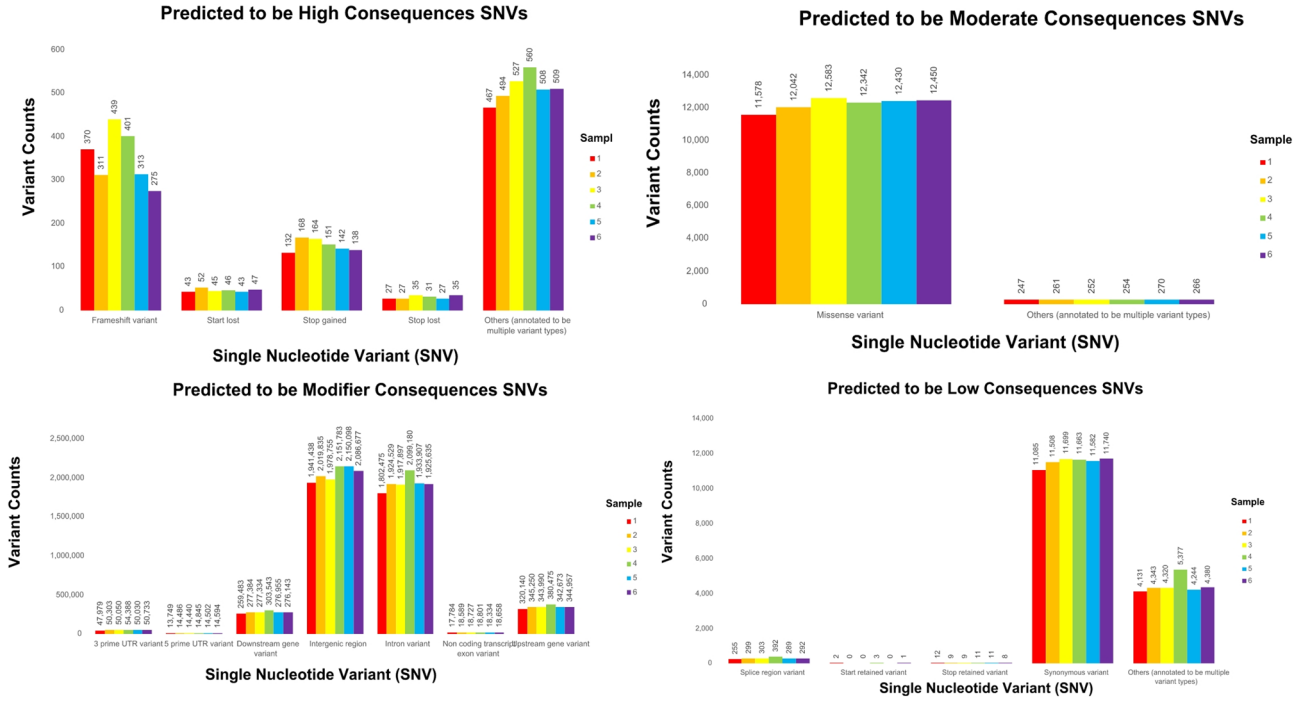

Fig. 2. Distribution of the SNVs across all samples by predicted consequences.

database, we detected oncogene *ESX1* [NM_153448.4:c.1015G>C(p.Val339Leu)] in sample S1 and S3. We also detected additional oncogene *SLC37A1* [NM_018964.4:c.1209 C>T(p.Leu403Leu)] and *DCAF12L2* [NM_001013628.3:c.1001 C>T(p.Pro334Leu)] in sample S3. In sample S6, we detected *RBMXL3* [NM_001145346.2:c.3145 A>G(p.Arg1049Gly)]. No oncogene matches ICGC oncogene database for sample S2, S4, and S5.

NPC structural variants (SVs) mutational landscape

There were around total of 30,000 SVs detected from each NPC sample, except for sample S1, it has total 41,599 SVs. This disproportionately higher total number of SV in S1 compared to other samples was due to very high number of breakend translocation. The other types of SVs were relatively similar across all samples. The breakdown of the number of SVs by type from all samples is shown in Table 1.

From the SV variant type, among those variants that were predicted to have high consequences, we detected the EPHA3 gene transcript ablation variant as the most common tumour suppressor gene variant, detected in 4 out of 6 NPC samples. Other tumour suppressor gene SV variant with high consequences detected were variants in *CASP8*, *DMBT1*, *ZFHX3*, and *IRF5* gene. Those genes were predicted to have loss its function. For sample S1, there were lots of tumour suppressor gene with high consequences variant type. Together with very high number of translocations in S1, this indicated a very unstable genome in S1, but not observed in other samples. The detail of SVs detected categorised as high consequences across all samples is available in Supplementary Data 3.

NPC copy number variations (CNVs) mutational landscape

The CNV calling was done using QDNASeq with 15 kb bin size to get the optimal gene level loss or gain with acceptable noise. All the copy number gene loss was checked for tumour suppressor gene against TSGene version 2.0. We found copy number loss consistently in RNH1, H19, CDKN1C, PHLDA2, ZBTB4, MAP4K1, SIRT2, LRIG1, TRIM31, IER3, HBP1, EPHB6, ZYX, EPHA1, DEFB1, CLDN23, MTSS1, and ZNF185 genes from all our NPC samples. Other common tumour suppressor genes found to have copy number loss were SRPX, BTK, FHL1, and RPL10 found in sample S1, S2, S3, and S5. The sample S6 has additional CDKN2A and CDKN2B copy number gene loss. Those listed tumour suppressor copy number loss gene have been further independently validated toward published NPC RNA seq data and confirmed to be commonly known to be downregulated.

All the copy number gain genes were also validated toward an independent NPC RNA Seq data to confirm they are upregulated. Sample S3 and S6 have lots of validated genes with copy number gains. Sample S4 has no single validated gene with copy number gain. The Y RNA gene was the most common gene with copy number gain presence in all samples except sample S4. Other common copy number gain genes found include *LTO1*, *FADD*, and *FRG1CP* gene. The overall copy number plot across all sample is shown in Fig. 3. The detail gene level copy number loss and gain across all samples is available as Supplementary Data 4.

NPC short tandem repeat (STR) mutation

Among 6 NPC samples, we detected pathogenic STR only in sample S4, S5 and S6. We found abnormal STR in *PABPN1* and *RFC1* gene. Sample S4 and S6 had *PABPN1* gene STR, while sample S5 and S6 had *RFC1* gene STR. The *PABPN1* gene GCG trinucleotide was repeated up to 16 times, while normal repeat was supposed to be around 10 times or less. The *RFC1* gene AAAAG nucleotide was repeated up to 600 times, while the normal repeat was supposed to be around 10 times. Those genes STR has been known to be associated with cerebellar ataxia, neuropathy, vestibular areflexia syndrome for *RFC1* STR and oculopharyngeal muscular dystrophy for *PABPN1* STR, but they have not been reported in NPC yet. Our finding here reported a possible association between STR and NPC that worth further exploration. The report of detected STR is available as Supplementary Data 5.

Assessment of mutation in DNA repair genes

Since we found genomic instability in sample S1, we then did further in-depth analysis focusing on DNA repair genes. When the DNA repair genes are mutated or loss, then this finding support the result of genomic instability. We found sample S1 to had various defective DNA repair genes which was annotated to have high impact from the SV variant calling. No other NPC samples with defective high impact DNA repair genes defect. The sample S1 had predicted defective *ATM*, *BRCA1*, *CDK12*, *PALB2*, *POLE*, *RAD51B*, *FANCA*, *FANCC*, *FANCD2*, *FANCE*, *FANCM*, and *MSH2* DNA repair genes from the SV calling process. All of them are breakend translocation with variant type ranging from transcript ablation, gene fusion with frameshift variant until bidirectional gene fusion. The distribution of all DNA repair genes variants was shown in Fig. 4. The Fig. 4 was showing all type of

	Translocation	Deletion	Duplication	Insertion	Inversion
S1	15,647	12,264	17	13,642	29
S2	57	13,279	7	16,677	17
S3	95	13,681	14	17,172	35
S4	67	13,465	12	16,338	27
S5	61	12,906	11	16,017	23
S6	123	13,538	22	16,961	33

Table 1. Summary of SVs count by SV types across all samples.

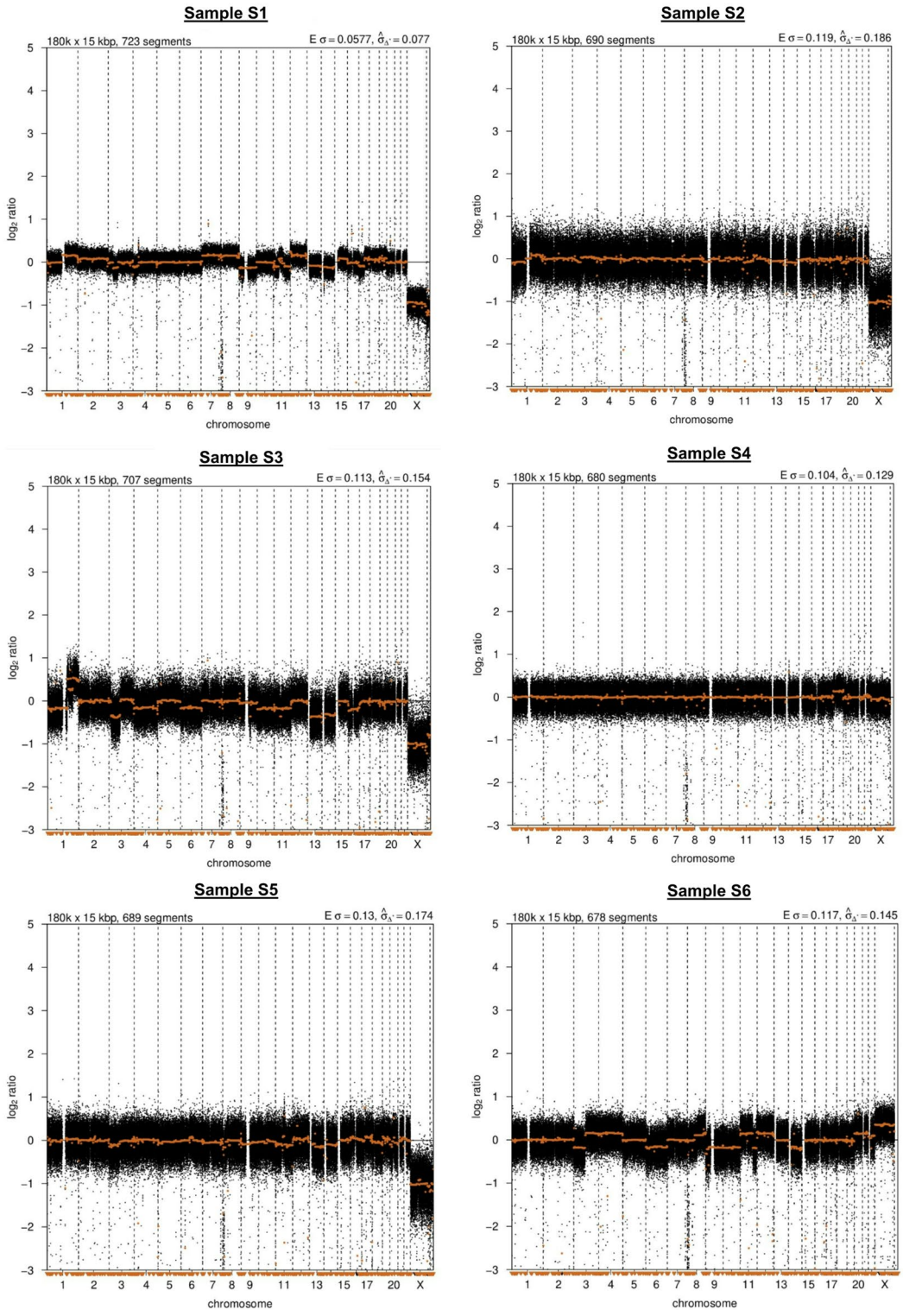

Fig. 3. Overall CNV plot across all NPC samples (sample 4 and 6 having 2 pair of X chromosomes were female, while others were male).

variants from all key DNA repair genes. The sample S4 had breakend translocation, but they were all located at non-coding region, thus not considered to have high impact. Then from SNV result, we also assessed the DNA repair genes variants. We only found high impact DNA repair genes variants in MSH3 gene for sample S2, S4 and S6. The variant type in those MSH3 gene was disruptive inframe deletion. There were no other DNA repair genes high impact variants detected in other samples.

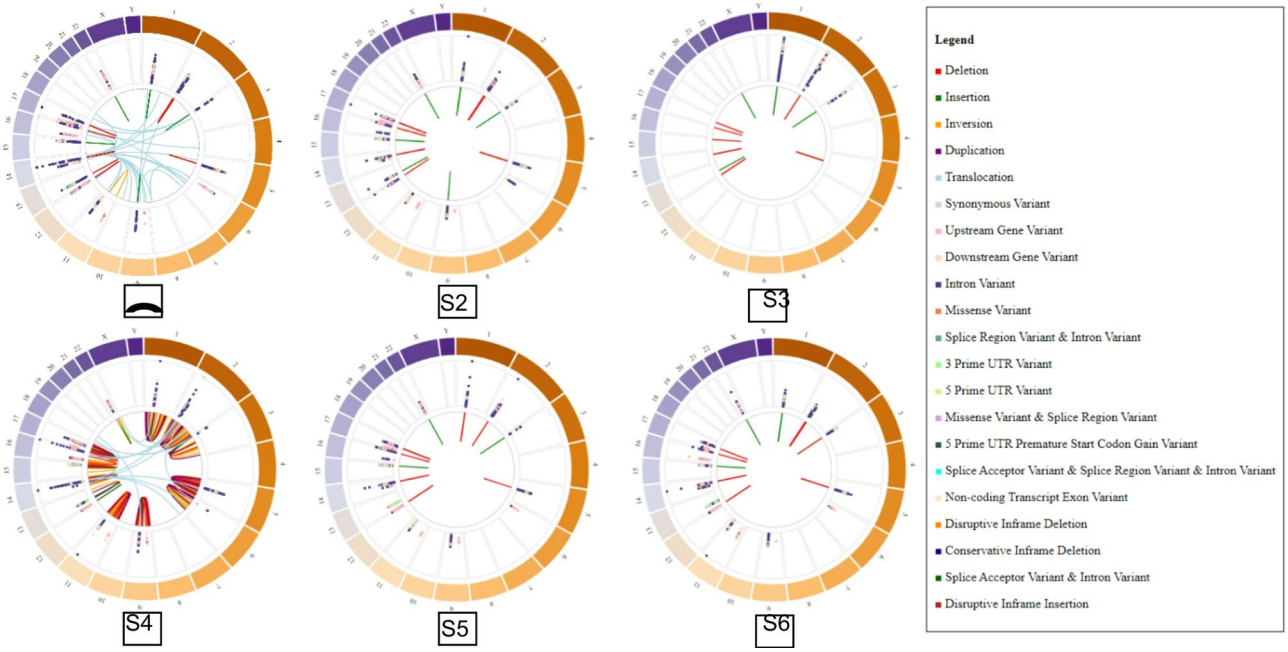

Fig. 4. Circle plot showing all SVs and SNVs in DNA repair genes.

Sample	Gender	Age (years)	T	*N*	M	Stage	Treatment	Response	Overall Survival
S1	Male	58	4	3	0	IVA	IC + CRT	Partial locoregional response + Progressive lung metastasis	18 months
S2	Male	43	2	3	0	IVA	IC + CRT	Complete locoregional response, bone metastasis 8 months after CRT	26 months (alive)
S3	Male	63	3	1	0	III	CRT	Complete response	32 months (alive)
S4	Female	69	2	2	0	IVA	CRT	Near Complete Response, then local progress in 6 months	23 months (censored: lost contact)
S5	Male	49	4	3	1	IVB	Chemo 2 course + RT	Partial Response	30 months
S6	Female	56	2	3	1	IVB	Chemo 2 course + CRT	Partial Response	43 months

Table 2. Summary of samples' clinical history. IC = induction chemotherapy. CRT = concurrent chemoradiotherapy. RT = radiotherapy. Chemo 2 course = 2 course of chemotherapy.

Our comprehensive variant calling also detected moderate, low and modifier variants. We detected 4 out of 6 samples to harbour *TP53* gene missense variants (moderate functional impact) in NM_000546.6:c.215 C > G(p.Pro72Arg) genomic location. This *TP53* missense variant has also been reported in COSMIC database (COSMIC gene mutation ID: COSV52708746) to be presence in various cancers in alignment with our findings. Other common DNA repair genes variants detected are *BRCA1* missense variant in various genomic location such as NM_007300.4:c.4900 A > G(p.Ser1634Gly); NM_007300.4:c.3548 A > G(p.Lys1183Arg); NM_007300.4:c.3113 A > G(p.Glu1038Gly), *BRCA2* missense variant in various genomic location such as NM_000059.4:c.7397T > C(p.Val2466Ala); NM_000059.4:c.1114 A > C(p.Asn372His), etc. We also detected variants in various non-coding regions such as intron variants, upstream or downstream gene variants which theoretically has less direct potential to cause functional consequences, but increasing evidence suggesting some specific non-coding area might have a role in driving cancer and they are currently under reported[25]. The complete list of all the DNA repair genes predicted functional impacts across all samples are available in **Supplementary Data 6**.

Clinical history

We sequenced 6 NPC samples, with 4 samples in locally advanced stage (AJCC 8th edition stage III to IVA) and 2 samples in advanced stage (AJCC 8th edition stage IVB). The summary of the samples' characteristics, diagnosis stage, treatment administered until follow up of treatment response was summarized in Table 2. The treatment administered to all samples had been reviewed independently by the peer oncologists and deemed standardised as per NPC national guideline. All sequencing samples were successfully mapped against EBV reference genome, suggesting the presence of NPC in all our samples.

Sample S1 had the shortest survival among all other samples, despite not being in metastatic disease at diagnosis. No other significant co-morbidities were presence in all samples. When correlating the mutational profile of all samples, it is likely that there is a strong association between sample S1 worst survival and detected multiple defective DNA repair genes. Sample S2, S4, and S6 only have 1 predicted defective DNA repair genes which is MSH3 gene. However, there was no defective MSH6 gene, and it is known that MSH6 gene can cooperatively handle loss of MSH3 gene. This prevents any DNA mismatch repair from occurring.

Discussion

Cancer exhibits remarkable variation in its genetic makeup across different populations. Even the same cancer type can present with strikingly different mutational landscapes when compared between geographical regions. This phenomenon is influenced by a complex interplay of environmental factors as well as genetic predispositions specific to different ethnicities. As a result, targeted therapies designed for a specific population might not be equally effective in another, highlighting the importance of understanding these regional disparities in cancer genomics. Our study described an in-depth analysis of the mutational landscape of locally advanced to advanced NPC in an Indonesian cohort, shedding light on significant genetic alterations and their potential clinical implications. The findings from this research contribute to the broader understanding of NPC's molecular underpinnings and highlight several key areas of interest.

This study utilized long-read sequencing to perform an in-depth analysis of the mutational landscape in six Indonesian patients with locally advanced to advanced NPC. Compared to previous NPC sequencing studies, often utilizing short-read technology, our use of long reads facilitated the detection of numerous structural variants and potentially pathogenic STR expansions, which may be underestimated by other methods[6,26]. A primary limitation is the small sample size, which precludes definitive conclusions about the prevalence of these mutations in the Indonesian population or statistically robust comparisons with other cohorts. However, the data provides a valuable snapshot of the types of genomic alterations, including complex structural variants and short tandem repeats, that occur in Indonesian NPC and demonstrates the utility of long-read sequencing for their detection.

The frameshift and slice acceptor variant detected in *LIMD1* and *CNDP2* tumour suppressor gene suggests possible disruption of their function. The frameshift variant shifts the reading frame of DNA sequence leading to completely changed amino acid sequence. The splice acceptor variant affects the sequence at the end of an intron, where splicing occurs, leading to removal of entire exon or retainment of intron, which likely also disrupt the gene function. The *CNDP2* is a known tumour suppressor gene reported in various cancer including gastric cancer and urothelial cancer[27–29]. Our finding of disrupted *CNDP2* gene in all our NPC samples suggest that this gene may also be responsible for NPC carcinogenesis.

The *LIMD1* downregulation is observed in many other solid cancers including lung cancer, non-Hodgkin lymphomas, gastric colorectal carcinomas, and breast cancer[30,31]. Higher expression in *LIMD1* suggests better overall survival in those multiple cancers[30]. However, EBV *LMP1* oncoprotein is known to regulate *LIMD1* through *IRF4* and NFκB signalling in EBV-transformed cells[30–32]. *LIMD1* deletion is found to impairs *LMP1* signalling and functions, therefore in EBV associated malignancy including NPC, it is likely that *LIMD1* oppositely acts as a tumour promoter[31]. The frequent presence of *PKL5* and *TMPRSS11A* variants in NPC also indicates potential targets for future therapeutic strategies and require further exploration.

Annotation against the ClinVar database revealed variants in PERM1 (NM_001291366.2:c.2330T > C) and APOA2 (NM_001643.2:c.-323 C > T) across all six samples. While ClinVar lists assertions for these variants, the evidence for pathogenicity, particularly for PERM1, is limited on current entries. The PERM1 missense variant (p.Val777Ala) has been reported in head and neck squamous cell carcinoma in the COSMIC database[33]. However, given their presence in all samples in our small cohort and the limited pathogenicity evidence for PERM1, it is possible these represent common germline variants in this population, potentially benign or of uncertain significance. The *APOA2* mutation was not reported yet in COSMIC database[34,35]. This was likely due to *APOA2* mutation occurring in upstream non-coding region. The COSMIC database contained numerous cancer sequences, but majority were not whole genome sequencing, therefore non-coding region might be underrepresented.

PERM1 (NM_001291366.2:c.2330T > C) mutation has been known to cause renal tubular epithelial cell apoptosis / neutrophil inclusion bodies, while *APOA2* (NM_001643.2:c.-323 C > T) mutation has been known to cause familial hypercholesterolemia. Both *PERM1* and *APOA2* gene function are related to lipid metabolism[36,37]. Based on those gene function, it was unlikely that mutation on those genes is the initial event in NPC carcinogenesis. If this is a somatic mutation, then it is more plausible that those mutations are likely conferring a significant growth or survival advantage to cells in the early stages of NPC carcinogenesis. This advantage could lead to the preferential expansion of those mutant cells, making them more likely to form detectable tumours. Similar to the case of low-grade glioma, where *IDH1*/2 mutation (involve in Krebs cycle metabolism) are also very common and turned out to be clinically prognostic[38]. Their functional impact and relevance to NPC, particularly concerning lipid metabolism pathways, require further investigation in larger cohorts with germline analysis and functional validation.

In this study, we examined somatic single nucleotide variations (SNVs) in all our NPC samples by comparing detected mutations with established oncogene databases, including TCGA and ICGC. The results highlighted the presence of known oncogenes in a subset of samples, supporting the potential role of these genes in cancer progression. Notably, we identified *PDE5A*, *ABCB7*, and *HERC2* as matched oncogenes in samples S1, S5, and S6, respectively, in alignment with the TCGA database. These findings suggest that *PDE5A* mutations may contribute to tumorigenesis through mechanisms involving the cGMP signalling pathway[39,40], while *ABCB7* mutations could disrupt iron homeostasis, a critical factor in various cancers[41,42]. Similarly, the HERC2 mutation aligns with its known function in DNA repair, which may drive oncogenic processes[43,44].

Further analysis using the ICGC database revealed additional oncogenes, such as *ESX1*, detected in samples S1 and S3. *ESX1* has been implicated in cancer through its role in transcriptional regulation, and mutations like Val339Leu may alter its activity[45]. In addition, we observed *SLC37A1* and *DCAF12L2* mutations in sample S3. *SLC37A1*, involved in phosphate transport[46], and *DCAF12L2*, a component of the ubiquitin ligase complex[47], may play novel roles in oncogenesis, although their exact implications in cancer remain to be further elucidated. Furthermore, the identification of *RBMXL3* in sample S6 suggests potential disruptions in RNA splicing, a pathway frequently dysregulated in cancer[48,49].

Some samples did not show any detectable oncogenes matching the TCGA / ICGC pan-cancer mutation database. The absence of detectable mutations does not necessarily indicate the lack of oncogenic potential in these samples. Several factors could explain this outcome. First, the mutation profile of these samples may involve genes or mutations not yet well-characterized in current databases, particularly for rare cancers or subtypes that have not been extensively studied. Alternatively, non-SNV alterations, such as copy number variations (CNVs), epigenetic changes, or mutations in regulatory regions outside the coding sequence, may drive oncogenesis in these samples.

Overall, these oncogenic findings underscore the utility of integrating comprehensive oncogene databases, such as TCGA and ICGC, to better understand the mutational landscape in cancer. Our results point to several novel oncogenes that may contribute to cancer development, providing potential targets for future therapeutic interventions. However, further functional studies are required to confirm the oncogenic potential of these mutations and to explore their specific roles in different cancer types.

The potential role of structural variants (SVs) in NPC development has started to be reported since a decade ago. The detection of SV used to be a very complex process. Valouev et al. did an extensive study to detect SV in just 2 NPC samples with a very resource intensive screening and validation[50]. They were only able to detect a handful of breakpoints that indicate gene fusions SV[50]. Those detected gene fusions were later validated in larger NPC cohorts (192 other NPC samples) with fluorescent in situ hybridization (FISH) but the frequency of positivity was very low, just 3 detected gene fusion out of 192 screened samples for *MAML2* gene fusion[50]. Unlike SNV such as missense mutation which can be consistently detected in various samples, SV in the form of gene fusion can be very complex. A gene fusion can be formed with slightly different genes, causing the exact similar gene fusion to be rarely presence. However, we are aware that functional consequences of SVs are much greater than SNV and now they are becoming much easier and accurately to be detected with the long read sequencer[51], as the one we used in our study.

This study's findings on structural variants (SVs) provide crucial insights into the genomic instability characteristic of NPC. Sample S1, which exhibited the shortest survival, displayed significant genomic instability, characterized by a high number of translocations and predicted high-consequence SVs affecting multiple DNA repair genes spanning pathways like homologous recombination (BRCA1, PALB2, RAD51B) and DNA damage response (ATM, CDK12). This instability may contribute to the aggressive nature of the disease observed in this patient, as reflected in the poor overall survival. The frequency of overall DNA repair defects genes in NPC has been reported to be around 28%.52 However, extensive genome-wide SV was not so common, but it was shown to be associated with a deleterious outcome, consistent with our finding[52,53]. While a direct causal link between these specific alterations and the poor clinical outcome cannot be established from a single case, the observation suggests that complex, deleterious DNA repair defects, identifiable through long-read sequencing, may correlate with aggressive disease phenotypes in NPC. Further studies are needed to confirm this association.

DNA repair genes play a crucial role in maintaining the integrity of the human genome. They are responsible for identifying and correcting damage to the DNA molecules that encode the genome. In human cells, both normal metabolic activities and environmental factors such as radiation can cause DNA damage. Many of these lesions cause structural damage to the DNA molecule and can alter or eliminate the cell's ability to transcribe the gene that the affected DNA encodes[54]. Other lesions induce potentially harmful mutations in the cell's genome, which affect the survival of its daughter cells after it undergoes mitosis. The DNA repair process is constantly active as it responds to damage in the DNA structure. When normal repair processes fail, and when cellular apoptosis does not occur, irreparable DNA damage may occur, including double strand breaks and DNA cross linkages. This can eventually lead to malignant tumours, or cancer as per the two-hit hypothesis[55].

All key DNA repair genes in our study have been shown to be linked to various cancers[24,56]. It has also been linked with cancer aggressiveness and inform poor survival[57]. Even detection of those defective genes such as *BRCA2*, *ATM*, and *CDK12* in circulation were prognostic in a circulating tumour DNA prostate cancer study[57]. *FANC* genes are also important genes required for DNA repair and stabilization of replication fork, with as many as 65% of cancer in NIH GDC portal harbour *FANC* genes mutation[58,59]. The loss of *FANC* genes have been confirmed in pre-clinical study to foster expansion of oncogenic cells[60]. Genes involved in homologous recombination double strand break DNA repair process such as *BRCA2*, *RAD51B*, *FANC* gene groups are very critically important because if defective, it means cells loss the ability for error free DNA repair and genomic instability occurs which later trigger various cancers[59–62].

Copy number variation analysis further validated the presence of gene losses in known tumour suppressors like *RNH1*, *H19*, and *CDKN1C*, across all samples. Other integrative analysis of NPC genome also showed enriched copy gene loss in region coding for the tumour suppressor[52,63]. The consistent loss of these genes supports their potential role as drivers of carcinogenesis in NPC. The validation of these findings against independent transcriptomic data enhances the robustness of this study and suggests these genes could serve as biomarkers for prognosis or targets for novel therapies.

On the other hand, the identification of copy number gains in certain genes, especially in *Y RNA*, *LTO1*, and *FADD*, highlights potential oncogenes that may contribute to tumour progression in NPC. The *Y RNAs* are essential for the initiation of DNA replication. They interact with chromatin and other proteins to facilitate the replication process[64]. The *Y RNA* has been found to be significantly overexpressed up to 13 fold in many

cancers[64,65]. These findings warrant further exploration to understand their functional roles and implications for targeted therapy.

The detection of pathogenic STRs in *PABPN1* and *RFC1* genes, particularly in samples S4, S5, and S6, presents a novel association between STRs and NPC. The repeated GCG trinucleotide in *PABPN1* and the significantly amplified AAAAG nucleotide in RFC1 have known associations with neurological syndromes but have not previously been linked to NPC[66,67]. Dysregulated *PABPN1* has been shown to promote bladder cancer aggressiveness[68]. While reduced expression of *RFC1* was shown to be correlated with lower overall survival in colorectal cancer patients[69]. This discovery opens new avenues for research into the role of STRs in NPC development and progression.

While the initial mutagenic events of NPC may be random, the high frequency of those pathogenic mutations suggests a strong selective pressure for this specific alteration in the context of NPC development. It's not that those mutations are the only path to NPC formation, but rather that they represent a particularly effective path in the specific cellular and molecular context of NPC carcinogenesis. This phenomenon highlights the complex interplay between random mutational events and the selective pressures of the cellular environment in cancer evolution. Our findings, albeit limited in sample sizes, but comprehensive in analysis may pave a clearer path for ongoing further research, and important in shedding light on our understanding on NPC carcinogenesis processes. There is a need for such comprehensive mutational analysis on a larger scale to validate the observed potential association with the clinical history found here.

The interpretation of the variants must be approached cautiously due to several study limitations. Firstly, we lacked matched germline controls, which limits our ability to distinguish somatic from germline variants. Secondly, no quantitative tumour purity assessment was performed, which may affect the reliability of variant allele frequency interpretation. Furthermore, while we utilized public oncogene and tumor suppressor databases, many of the variants identified, particularly in STRs and non-coding regions, remain of uncertain clinical significance. This reflects broader challenges in variant annotation for underrepresented populations, such as Indonesians, in global databases.

Despite these constraints, our study demonstrates that long-read sequencing offers a viable solution for comprehensive mutational screening in NPC. With a single flow cell per sample, the sequencing process was completed within 7 days, and the total reagent and flow cell cost per sample was estimated at approximately USD $1,200, lower than many clinical WGS workflows. These attributes make the platform particularly attractive in LMIC settings. Moreover, this study underscores the importance of establishing local genomic reference databases and increasing integration of multi-omics approaches in future research.

In conclusion, this study provides valuable insights into the genetic landscape of locally advanced to advanced NPC in an Indonesian population. The identification of high-consequence genetic alterations, the implications of SVs and CNVs, and the novel association of STRs with NPC contribute to the growing understanding of this disease. These findings underscore the importance of region-specific research in NPC and highlight potential biomarkers and therapeutic targets that could improve patient outcomes in Indonesia and beyond.

Data availability

The data has been submitted to the SRA database with BioProject SRA number PRJNA1060446.The reviewer link to view the data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE272926The reviewer token to view the data: enubueiyprsbjgt.

Received: 30 September 2024; Accepted: 5 June 2025
Published online: 01 July 2025

References

1. Handoko, Louisa, M., Permata, M. & Gondhowiardjo, S. Deciphering driver of nasopharyngeal Cancer development. *Oncol Rev* **16**, 10654. https://doi.org/10.3389/or.2022.10654 (2022).
2. Gondhowiardjo, S. A. et al. *Tackling Resist. Cancer Immunotherapy: What Do We Know? Molecules* **25**(18), 4096. https://doi.org/10.3390/molecules25184096 (2020).
3. Balachandran, P. & Beck, C. R. Structural variant identification and characterization. *Chromosom Res. Int. J. Mol. Supramol Evol. Asp Chromosom Biol.* **28**, 31–47 (2020).
4. Hamdan, A. & Ewing, A. Unravelling the tumour genome: the evolutionary and clinical impacts of structural variants in tumourigenesis. *J. Pathol.* **257**, 479–493 (2022).
5. Yang, L. A practical guide for structural variation detection in the human genome. *Curr. Protoc. Hum. Genet.* **107**, e103 (2020).
6. Dai, W., Zheng, H., Cheung, A. K. L. & Lung, M. L. Genetic and epigenetic landscape of nasopharyngeal carcinoma. *Chin. Clin. Oncol.* **5**, 16 (2016).
7. Gondhowiardjo, S. A. et al. Epstein–Barr virus (EBV) viral load in tumor cells did not predict tumor extensiveness in nasopharyngeal Cancer. *Microbiol. Res. (Pavia)*. **12**, 150–156 (2021).
8. Gondhowiardjo, S. A. et al. Tumor microenvironment predicts local tumor extensiveness in PD-L1 positive nasopharyngeal cancer. *PLoS One*. **15**, e0230449 (2020).
9. Adham, M. et al. Epstein-Barr virus DNA load in nasopharyngeal brushings and whole blood in nasopharyngeal carcinoma patients before and after treatment. *Clin. Cancer Res.* 2175–2187. https://doi.org/10.1158/1078-0432.CCR-12-2897 (2013).
10. Handoko et al. First Indonesian nasopharyngeal Cancer whole epigenome sequencing identify tumour suppressor CpG methylation. *Biologics* **19**, 1–13 (2025).
11. Adham, M. et al. Nasopharyngeal carcinoma in indonesia: epidemiology, incidence, signs, and symptoms at presentation. *Chin. J. Cancer*. **31**, 185–196 (2012).
12. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
13. Zhou, A., Lin, T. & Xing, J. Evaluating nanopore sequencing data processing pipelines for structural variation identification. *Genome Biol.* **20**, 237 (2019).
14. de Jesus, O. et al. Updated Epstein-Barr virus (EBV) DNA sequence and analysis of a promoter for the BART (CST, BARF0) RNAs of EBV. *J. Gen. Virol.* **84**, 1443–1450 (2003).

15. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02024-y (2024).
16. Su, J., Zheng, Z., Ahmed, S. S., Lam, T. W. & Luo, R. Clair3-trio: high-performance nanopore long-read variant calling in family trios with trio-to-trio deep neural networks. *Brief Bioinform* **23**, bbac301. https://doi.org/10.1093/bib/bbac301 (2022).
17. Cingolani, P. et al. 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. (Austin).* **6**, 80–92 (2012).
18. Zhao, M., Sun, J. & Zhao, Z. TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.* **41**, D970–D976 (2013).
19. Smolander, J. et al. Evaluation of tools for identifying large copy number variations from ultra-low-coverage whole-genome sequencing data. *BMC Genom.* **22**, 357 (2021).
20. Chiu, R., Rajan-Babu, I. S., Friedman, J. M. & Birol, I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.* **22**, 224 (2021).
21. Wang, T. et al. OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Res.* **49**, D1289–D1301 (2021).
22. Dodd, L. E. et al. Genes involved in DNA repair and Nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. *Cancer Epidemiol. Biomarkers Prev. Publ Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **15**, 2216–2225 (2006).
23. Sengupta, S. et al. Genome-wide expression profiling reveals EBV-associated Inhibition of MHC class I expression in nasopharyngeal carcinoma. *Cancer Res.* **66**, 7999–8006 (2006).
24. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
25. Boscolo-Rizzo, P. et al. TERT promoter mutations in head and neck squamous cell carcinoma: A systematic review and meta-analysis on prevalence and prognostic significance. *Oral Oncol.* **140**, 106398 (2023).
26. Chow, K. Y. Epigenomic landscape study reveals molecular subtypes and EBV-associated regulatory epigenome reprogramming in nasopharyngeal carcinoma. *EBioMedicine* **86**, 104357 (2022).
27. Zhang, Z. et al. Underexpressed CNDP2 participates in gastric cancer growth Inhibition through activating the MAPK signaling pathway. *Mol. Med.* **20**, 17–28 (2014).
28. Vosoughi, A. et al. Common germline-somatic variant interactions in advanced urothelial cancer. *Nat. Commun.* **11**, 6195 (2020).
29. Singh, S. et al. Proteomic signatures of diffuse and intestinal subtypes of gastric Cancer. *Cancers (Basel)* **13**(23), 5930. https://doi.org/10.3390/cancers13235930(2021).
30. Wang, L., Sparks-Wallace, A., Casteel, J. L., Howell, M. E. A. & Ning, S. Algorithm-Based Meta-Analysis reveals the mechanistic interaction of the tumor suppressor LIMD1 with Non-Small-Cell lung carcinoma. *Front. Oncol.* **11**, 632638 (2021).
31. Wang, L. & Ning, S. New look of EBV LMP1 signaling landscape. *Cancers (Basel)* **13**(21), 5451. https://doi.org/10.3390/cancers13215451 (2021).
32. Wang, L. et al. LIMD1 is induced by and required for LMP1 signaling, and protects EBV-transformed cells from DNA damage-induced cell death. *Oncotarget* **9**, 6282–6297 (2018).
33. Martin, D. et al. The head and neck cancer cell oncogenome: a platform for the development of precision molecular therapies. *Oncotarget* **5**, 8906–8923 (2014).
34. Sondka, Z. et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
35. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
36. Huang, C. Y. et al. PERM1 regulates genes involved in fatty acid metabolism in the heart by interacting with PPARα and PGC-1α. *Sci. Rep.* **12**, 14576 (2022).
37. Jafari Azad, B., Yaseri, M., Daneshzad, E. & Koohdani, F. Interaction between Apo A-II -265T > C polymorphism and dietary total antioxidant capacity on some oxidative stress and inflammatory markers in patients with type 2 diabetes mellitus. *Br. J. Nutr.* **128**, 13–29 (2022).
38. Han, S. et al. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. *Br. J. Cancer*. **122**, 1580–1589 (2020).
39. Ahmed, W. S., Geethakumari, A. M. & Biswas, K. H. Phosphodiesterase 5 (PDE5): Structure-function regulation and therapeutic applications of inhibitors. *Biomed. Pharmacother*. **134**, 111128 (2021).
40. Ravi, S. et al. An update to hallmarks of Cancer. *Cureus* **14**, e24803 (2022).
41. Kim, J. Y., Kim, J. K. & Kim, H. ABCB7 simultaneously regulates apoptotic and non-apoptotic cell death by modulating mitochondrial ROS and HIF1α-driven NFκB signaling. *Oncogene* **39**, 1969–1982 (2020).
42. Pearson, S. A. & Cowan, J. A. Evolution of the human mitochondrial ABCB7 [2Fe-2S](GS)(4) cluster exporter and the molecular mechanism of an E433K disease-causing mutation. *Arch. Biochem. Biophys.* **697**, 108661 (2021).
43. Liu, Y. et al. HERC2 promotes inflammation-driven cancer stemness and immune evasion in hepatocellular carcinoma by activating STAT3 pathway. *J. Exp. Clin. Cancer Res.* **42**, 38 (2023).
44. Sala-Gaston, J. et al. HERC ubiquitin ligases in Cancer. *Cancers (Basel)* **12**6, 1653 (2020).
45. Tan, J. et al. Identification and Analysis of Three Hub Prognostic Genes Related to Osteosarcoma Metastasis. *J. Oncol.* 6646459 (2021).
46. Meng, Z. et al. A prospective diagnostic and prognostic biomarker for hepatocellular carcinoma that functions in glucose metabolism regulation: solute carrier family 37 member 3. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1869**, 166661 (2023).
47. Lidak, T. et al. CRL4-DCAF12 ubiquitin ligase controls MOV10 RNA helicase during spermatogenesis and T cell activation. *Int. J. Mol. Sci.* **22**(10), 5394 (2021).
48. Prieto, C. et al. Transcriptional control of CBX5 by the RNA-binding proteins RBMX and RBMXL1 maintains chromatin state in myeloid leukemia. *Nat. Cancer*. **2**, 741–757 (2021).
49. Liu, L., Liu, X., Bi, W. & Alcorn, J. L. A primate-specific RNA-binding protein (RBMXL3) is involved in glucocorticoid regulation of human pulmonary surfactant protein B (SP-B) mRNA stability. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **320**, L942–L957 (2021).
50. Valouev, A. et al. Discovery of recurrent structural variants in nasopharyngeal carcinoma. *Genome Res.* **24**, 300–309 (2014).
51. Tang, H. et al. Fusion genes in cancers: biogenesis, functions, and therapeutic implications. *Genes Dis.* **101536** https://doi.org/10.1016/j.gendis.2025.101536 (2025).
52. Bruce, J. P. et al. Whole-genome profiling of nasopharyngeal carcinoma reveals viral-host co-operation in inflammatory NF-κB activation and immune escape. *Nat. Commun.* **12**, 4193 (2021).
53. Handoko, Adham, M., Rachmadi, L., Wibowo, H. & Gondhowiardjo, S. A. Cold tumour phenotype explained through whole genome sequencing in clinical nasopharyngeal cancer: A preliminary study. *ImmunoTargets Ther.* **13**, 173–182 (2024).
54. Eaton, J. A. & Zidovska, A. Structural and dynamical signatures of local DNA damage in live cells. *Biophys. J.* **118**, 2168–2180 (2020).
55. Lipsick, J. A. History of Cancer research: tumor suppressor genes. *Cold Spring Harb Perspect. Biol* **12**(2), a035907 (2020).
56. Magraner-Pardo, L., Laskowski, R. A., Pons, T. & Thornton, J. M. A computational and structural analysis of germline and somatic variants affecting the DDR mechanism, and their impact on human diseases. *Sci. Rep.* **11**, 14268 (2021).
57. Warner, E. et al. BRCA2, ATM, and CDK12 defects differentially shape prostate tumor driver genomics and clinical aggression. *Clin. cancer Res. Off J. Am. Assoc. Cancer Res.* **27**, 1650–1662 (2021).
58. Liu, W. et al. Fanconi anemia pathway as a prospective target for cancer intervention. *Cell. Biosci.* **10**, 39 (2020).

59. Parsa, F. G. et al. Fanconi Anemia pathway in colorectal cancer: A novel opportunity for diagnosis, prognosis and therapy. *J Pers. Med* **12**(3), 396 (2022).
60. Pawlikowska, P. et al. FANCA deficiency promotes leukaemic progression by allowing the emergence of cells carrying oncogenic driver mutations. *Oncogene* **42**, 2764–2775 (2023).
61. Bonilla, B., Hengel, S. R., Grundy, M. K. & Bernstein, K. A. RAD51 gene family structure and function. *Annu. Rev. Genet.* **54**, 25–46 (2020).
62. Yoshimura, A., Imoto, I. & Iwata, H. Functions of breast Cancer predisposition genes: implications for clinical management. *Int J. Mol. Sci* **23**(13), 7481 (2022).
63. Hu, C. et al. A global view of the oncogenic landscape in nasopharyngeal carcinoma: an integrated analysis at the genetic and expression levels. *PLoS One*. **7**, e41055 (2012).
64. Gulia, C. et al. Y RNA: an overview of their role as potential biomarkers and molecular targets in human cancers. *Cancers (Basel)* **12**(5), 1238 (2020).
65. Christov, C. P., Trivier, E. & Krude, T. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br. J. Cancer*. **98**, 981–988 (2008).
66. Dominik, N. et al. Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. *Brain* **146**, 5060–5069 (2023).
67. Boivin, M., Charlet-Berguerand, N. & Trinucleotide, C. G. G. Repeat diseases: an expanding field of polyglycine proteins?? *Front. Genet.* **13**, 843014 (2022).
68. Chen, L. et al. PABPN1 regulates mRNA alternative polyadenylation to inhibit bladder cancer progression. *Cell. Biosci.* **13**, 45 (2023).
69. Misbah, M., Kumar, M., Najmi, A. K. & Akhtar, M. Identification of expression profiles and prognostic value of RFCs in colorectal cancer. *Sci. Rep.* **14**, 6607 (2024).

Author contributions
All authors contributed equally.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-06096-5.

Correspondence and requests for materials should be addressed to H. or S.A.G.