Comprehensive Somatic Mutation Analysis of Cutaneous Squamous Cell Carcinoma Using Whole-Exome Sequencing

Shumeng Li shumeng.li2@ucdconnect.ie

August 15, 2025

Abstract

Cutaneous squamous cell carcinoma (cSCC) is a common skin cancer, often linked to chronic ultraviolet (UV) exposure [19]. Defects in DNA repair pathways, such as mismatch repair (MMR), may also contribute to its development in a subset of cases [6, 11]. Here, I analysed public whole-exome sequencing data from seven cSCC tumour–normal pairs (P3-P9) to characterise mutation burden, driver genes, and mutational signatures.

I processed raw FASTQ files following GATK4 Best Practices [18], from alignment to variant calling and annotation. The tumour mutational burden (TMB) ranged from 2,752 to 4,729 mutations/Mb, with functional mutations (missense, nonsense, frameshift, splice site) comprising 4–7% of all calls. Cross-referencing with the COSMIC Cancer Gene Census (v102, GRCh38) identified 434 known cancer genes in the cohort, including highly recurrent alterations in *TP53*, *KMT2C*, *ARID1B*, and *NOTCH1*.

Mutational signature analysis identified five de novo signatures. Four matched UV-associated SBS7b (cosine similarity: 0.874–0.926), while one matched SBS15 (0.862), linked to MMR deficiency [1]. Most tumours (5/7) were dominated by SBS7b, while two were dominated by SBS15.

These results confirm that UV-induced DNA damage is the main driver of mutagenesis in cSCC. They also suggest that a small subset of tumours may harbour MMR defects, which could be clinically relevant for MSI testing and treatment planning.

Key words: cutaneous squamous cell carcinoma (cSCC), whole-exome sequencing (WES), somatic variant calling, tumor mutational burden (TMB), mutational signatures, UV-induced DNA damage, mismatch repair deficiency.

1 Introduction

Cutaneous squamous cell carcinoma (cSCC) is the second most common skin cancer worldwide and a major cause of skin cancer–related mortality [19]. While

ultraviolet (UV) radiation is recognised as the primary mutagenic factor, a subset of tumours may also arise from defects in DNA repair pathways [15]. The mutational processes shaping cSCC genomes remain incompletely understood, particularly in relation to distinct biological subtypes.

Whole-exome sequencing (WES) enables comprehensive profiling of somatic alterations, including mutation burden, driver mutations, and mutational signatures [3, 1]. Mutational signatures, which capture the cumulative effects of endogenous and exogenous mutagens, have been widely applied to link tumour genotypes to DNA damage and repair mechanisms [2]. For example, SBS7 subtypes are associated with UV-induced cyclobutane pyrimidine dimer formation, while SBS15 reflects defective DNA mismatch repair (MMR) [1, 8].

Established cancer gene catalogues, such as the COSMIC Cancer Gene Census (CGC), provide curated lists of genes with strong evidence for driver roles across tumour types [17]. Cross-referencing cohort-specific mutation data with these catalogues can help distinguish biologically relevant drivers from large, frequently mutated passenger genes.

In this study, I analysed WES data from seven cSCC tumour–normal pairs. Following the GATK4 Best Practices workflow for variant calling [18], I applied Ensembl VEP for functional annotation [14] and generated Mutation Annotation Format (MAF) files for downstream analysis. I then assessed tumour mutational burden (TMB), identified recurrently mutated cancer genes through COSMIC CGC cross-referencing, and characterised mutational signatures with a focus on UV- and MMR-related processes. The goal is to define the genomic landscape of cSCC and identify potential subgroups driven by distinct mutational mechanisms.

2 Method

2.1 Overall Workflow

The complete analysis workflow is summarised in Figure 1. The workflow consists of four main stages, represented by different colours in the diagram:

- **Pink**: Data acquisition and preprocessing (FASTQ download, QC, alignment, duplicate marking).
- **Green**: Somatic variant calling and annotation (Mutect2, filtering, VEP annotation, MAF conversion).
- **Blue**: Global mutation statistics and visualisation, including tumour mutation burden (TMB), mutation landscape, driver gene identification, and cross-referencing with COSMIC Cancer Gene Census (CGC).
- **Orange**: Mutational signature analysis (96-context catalogue construction, NMF extraction, COSMIC mapping, mechanism grouping).

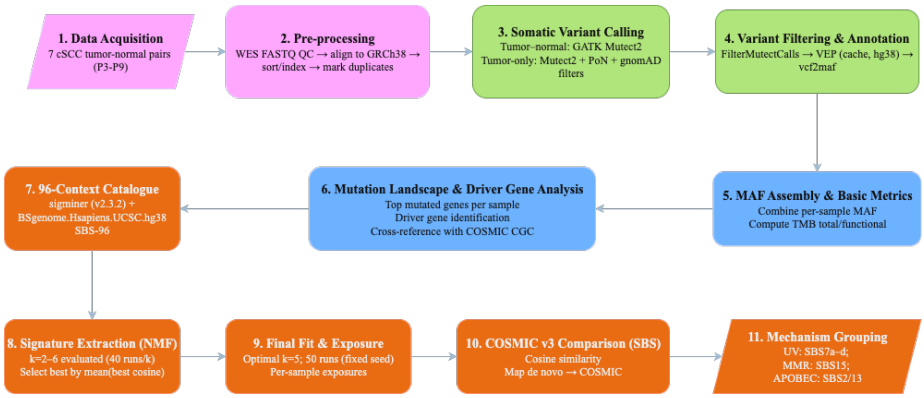


Figure 1: Somatic analysis pipeline from raw WES data to signature-level interpretation. Steps specify sample type, reference genome (hg38), and key tools where applicable.

2.2 Data Acquisition and Pre-processing

I worked with public whole-exome sequencing (WES) data from cutaneous squamous cell carcinoma (cSCC) tumor–normal pairs. The data come from studies by [4] and [7] and are available in the European Nucleotide Archive under the accession PRJNA603106. I downloaded all paired-end FASTQ files and organized them by sample ID.

The goal of this study is to identify and process eight tumor–normal pairs. However, during alignment, two pairs (P2 and P10) failed due to technical issues.

- P2 had mismatched read names, which stopped `bwa mem`.
- P10 had inconsistent SEQ and QUAL lengths, which caused a parse error in `samtools view`. To keep the analysis on track, I excluded these two and continued with the remaining seven pairs (P3–P9).

Instead of using a fully automated workflow such as `nf-core/sarek`, I chose to write my own step-by-step pipeline in shell scripts. This gave me full control over each stage, allowed me to troubleshoot specific issues like the read name mismatches, and made it easier to customize parameters for this dataset. All processing steps followed the GATK4 Best Practices [18], but with flexibility to adapt to the data quality and structure.

All samples were aligned to the GRCh38/hg38 reference genome using `bwa mem` [12]. The SAM files were converted to BAM, sorted, and indexed with `samtools` [13]. I then marked PCR duplicates with GATK `MarkDuplicates`. These steps produced clean, indexed BAM files, ready for somatic variant calling. I kept the same processing settings for all samples to ensure consistency and avoid batch effects.

2.3 Somatic Variant Calling and Annotation

After generating the cleaned BAM files, I performed somatic variant calling using the GATK4 Mutect2 workflow in tumor–normal mode. For each pair, I ran Mutect2 to detect single nucleotide variants (SNVs) and small insertions/deletions (indels). This step produced an unfiltered VCF file containing all candidate somatic variants.

Next, I applied GATK `FilterMutectCalls` to remove likely false positives based on quality metrics and read evidence. The resulting high-confidence variant calls were saved as `.filtered.vcf.gz` files for downstream analysis.

To interpret these variants, I annotated them using the Ensembl Variant Effect Predictor (VEP) in offline mode [14]. The annotation step added information about gene names, predicted functional consequences, and known cancer associations where available. I focused on functional variant classes such as missense, nonsense, frameshift, and splice site mutations, as these are more likely to impact protein function.

Finally, I generated MAF-format tables directly in R. I read each annotated VEP output file (`.vep.tsv`), selected functional variants, mapped the consequence terms to MAF-compatible classifications, and extracted genomic coordinates. I then saved these as MAF files for each sample and also combined them into a cohort-level MAF for downstream visualisation and statistical analysis with `maftools`.

2.4 Downstream Analysis

After generating the combined MAF file, I processed it in R using `maftools`, `dplyr`, and `ggplot2`. I computed tumour mutational burden (TMB) for each sample, reporting both total and functional mutations per megabase. Functional variants included missense, nonsense, frameshift, and splice site changes. I also calculated the percentage of functional mutations per sample.

For visualising the mutation landscape, I generated an oncoplot of the top ten mutated genes and a summary dashboard with `plotmafSummary`. I also produced bar plots of the top mutated genes across all samples. Mutation annotation was further investigated in relation to known cancer driver resources, as described in the next section.

2.5 Cross-referencing with COSMIC Cancer Gene Census

To verify which recurrently mutated genes were established cancer drivers, I cross-referenced my cohort against the COSMIC Cancer Gene Census (CGC; v102, GRCh38), downloaded from `https://cancer.sanger.ac.uk/cosmic/download/cosmic/v102/cancergenecensus`.

First, I summarised the combined MAF at the gene level, recording (i) total mutation count, (ii) number of samples mutated, and (iii) observed variant classes per gene. Gene symbols were normalised to uppercase, and missing or placeholder symbols were removed.

Then I joined the gene summary with the CGC table (`Cosmic_CancerGeneCensus_v102_GRCh38.tsv`) by gene symbol. For each gene, I extracted CGC metadata including Tier, somatic status, and annotated tumour types, and flagged whether the gene appears in CGC. The outputs included full cross-reference tables and visual comparisons of mutation counts between CGC and non-CGC genes.

While these results highlight recurrent alterations in established cancer driver genes, they do not directly reveal the mutational processes shaping these patterns. To address this, I next performed mutational signature analysis.

2.6 Mutational Signature Analysis

I constructed a 96–trinucleotide context SNV catalogue using `sigminer` with the hg38 reference genome. Candidate signature numbers ($k = 2$–6) were evaluated by running non-negative matrix factorisation (NMF) 40 times per k, normalising the extracted de novo signatures, and calculating cosine similarities to COSMIC v3 SBS reference signatures. The optimal k was chosen based on the highest mean of the best cosine similarity across signatures.

Using the selected k, I re-fitted the model with 50 NMF runs (fixed seed) to obtain stable de novo signatures and per-sample exposure profiles. Each signature was mapped to its best COSMIC match, and the dominant signature for each sample was identified. Exposure profiles were further grouped into mechanistic categories (UV: SBS7a–d; APOBEC: SBS2, SBS13; MMR: SBS15) to assess mutational processes. Visualisations included signature profiles, exposure proportions per sample, dominant signature distribution, and mechanism-level contributions.

3 Results

3.1 Tumour Mutation Burden (TMB) Analysis

Tumour mutational burden (TMB) refers to the total number of somatic mutations per megabase of the coding genome and is widely recognised as a genomic biomarker in cancer research and immuno-oncology [3]. Higher TMB levels are often linked to an increased neoantigen load and improved responses to immune checkpoint inhibitors across multiple cancer types [16].

In this section, I calculated the TMB for each sample, reporting both the total and functional mutation counts per megabase. Functional variants were defined as those predicted to alter protein function, including missense, nonsense, frameshift, and splice site changes. I also calculated the percentage of functional mutations relative to the total mutation count for each sample to provide additional context on mutation composition.

Across the seven tumour–normal pairs, the functional TMB ranged from 4.08 to 6.88 mutations/Mb (Table 1).

- Samples P7 and P9 had the highest values, while P6 recorded the lowest.

- The proportion of functional mutations was fairly stable, remaining between 4% and 7% across all samples.

A visual comparison of these values is shown in Figure 2.

Table 1: Tumour mutation burden per sample. Both total and functional TMB are reported, along with the percentage of functional mutations.

Sample	TMB_total	TMB_functional	Functional_TMB_%
P3	4729.553	301.3421	6.37
P4	2752.789	155.5526	5.65
P5	3880.342	221.9474	5.72
P6	3350.763	136.6053	4.08
P7	4014.000	276.0526	6.88
P8	3979.474	233.3421	5.86
P9	4365.737	285.8684	6.55

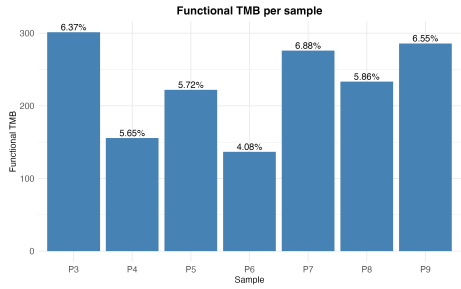


Figure 2: Functional TMB per sample. Bars represent the number of functional mutations per megabase, with percentages shown above each bar.

3.2 Mutation Landscape Summary

The oncoplot of the top ten most frequently mutated genes across the cohort (Figure 3) showed that all seven samples shared the same recurrently altered genes, with each gene mutated in 100% of cases. These included *ARID1B*, *CPEB2*, *IRS2*, *MUC12*, *MUC16*, *MUC19*, *TNRC18*, *TTN*, and *ZFHX3*, along with one unannotated locus. Large genes such as *MUC16* and *TTN* are frequently mutated in WES data due to their genomic size rather than positive selection [10], whereas *ARID1B* and *ZFHX3* have recognised tumour suppressor functions and may contribute to cSCC pathogenesis.

Mutation type distribution in the oncoplot revealed that missense variants were predominant, with fewer nonsense mutations and in-frame insertions/deletions. Several genes, including *ARID1B* and *ZFHX3*, also harboured splice site and multi-hit events, which may indicate biallelic inactivation—a

common mechanism of tumour suppressor gene disruption [9]. The diversity of mutation types suggests the involvement of multiple mutational processes, potentially including ultraviolet (UV) damage and endogenous mechanisms.

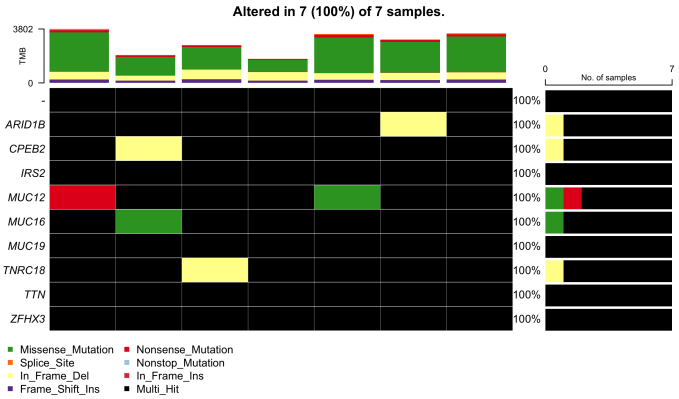


Figure 3: Oncoplot of the top ten mutated genes across all samples. Mutation types are indicated by colour, and alteration frequencies are shown on the right.

The `plotmafSummary` dashboard (Figure 4) confirmed these patterns. Missense mutations were the dominant variant classification, followed by in-frame deletions and frame-shift insertions. SNPs were the primary variant type, with C¿T substitutions as the most frequent SNV class, a signature associated with UV-induced DNA damage in cSCC [2]. Functional mutations accounted for 4–7% of the total mutation load (Table 1), and the variants-per-sample panel indicated a uniformly high mutation count, with a median of 3,068 variants per sample.

To further explore mutation patterns, I plotted the top ten most mutated genes across the entire cohort (Figure 5). This analysis highlighted large structural genes such as *TTN* and *KCNMA1*, which are often highly mutated in WES datasets due to their size rather than positive selection [10]. *CACNA1C*, which ranked second in mutation count, encodes a voltage-gated calcium channel subunit. Although it is not a classical cancer driver, recurrent mutations in this gene have been reported in certain tumour types and may reflect broader genomic instability. Alongside these, several established driver genes, including *TP53* and *ARID1B* [16], also showed high mutation counts. *TP53* plays a central role in DNA damage response and cell cycle regulation, while *ARID1B*, a component of the SWI/SNF chromatin remodelling complex, has been implicated in tumour development across multiple cancer types.

I then focused on cancer-relevant genes by intersecting the mutation set with a curated driver gene list from COSMIC and published studies. The resulting profile (Figure 6) was dominated by *TTN* and *TP53*, consistent with their ranking in the overall top-gene analysis. Other notable drivers included *NOTCH1*, *MUC16*, and *ZFHX3*, all with documented roles in squamous cell

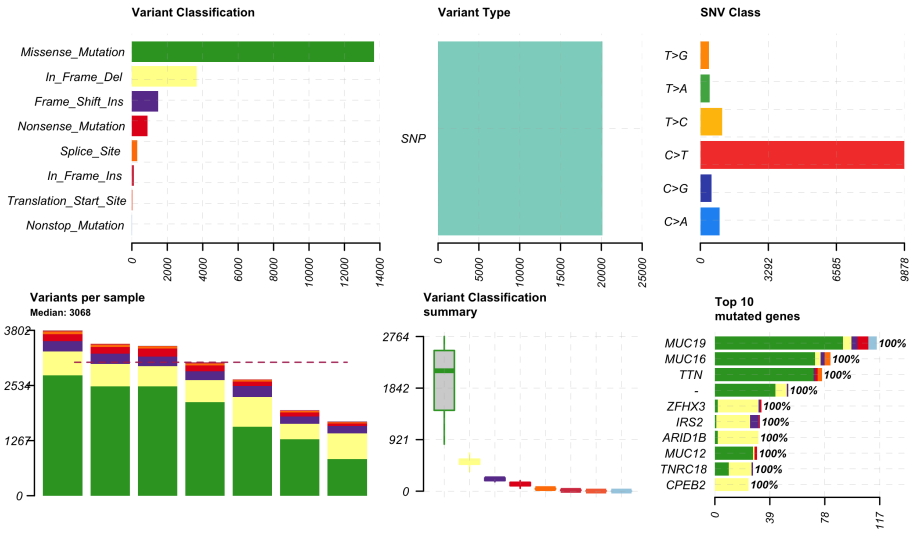


Figure 4: Summary dashboard from `maftools`, showing variant classifications, variant types, SNV classes, variants per sample, and top mutated genes.

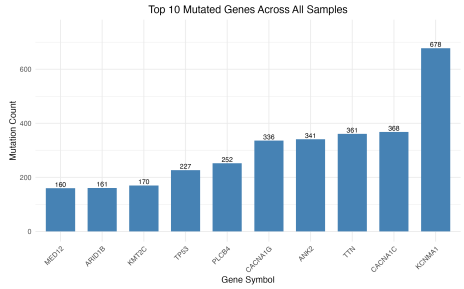


Figure 5: Top 10 most frequently mutated genes across the entire cohort. Bars represent mutation counts. Large structural genes such as *TTN* and *KCNMA1* dominate due to genomic size, while several known drivers, including *TP53* and *ARID1B*, also show high mutation counts.

carcinoma biology. *NOTCH1* mutations are common in keratinocyte-derived tumours and can promote malignant progression, while alterations in *ZFHX3* have been linked to tumour suppressor loss in epithelial cancers. Some high-frequency genes from the overall list, such as *CACNA1C*, did not appear in the curated driver set, highlighting the importance of distinguishing between established drivers and frequently mutated non-driver genes when interpreting mutation landscapes.

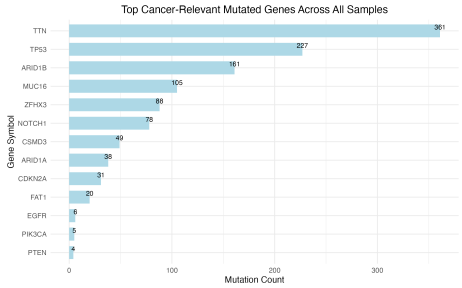


Figure 6: Top mutated cancer-relevant genes identified by intersecting the mutation dataset with a curated driver gene list. *TTN* and *TP53* remain among the most frequently altered, along with other notable drivers such as *MUC16*, *ZFHX3* and *NOTCH1*.

3.3 Cross-referencing with COSMIC Cancer Gene Census

The combined mutation dataset contained 8,828 unique genes, of which 434 were present in the COSMIC Cancer Gene Census (CGC; v102, GRCh38). Among these, several canonical driver genes were highly recurrent (Table 2). *TP53* was the most frequently mutated CGC gene (227 mutations across all 7 samples), followed by *KMT2C* (170 mutations), *ARID1B* (161 mutations), and *MED12* (160 mutations). Other notable CGC genes included *PLEC*, *KMT2D*, and *MUC16*, all mutated in at least six samples.

Table 2: Top five most frequently mutated COSMIC Cancer Gene Census (CGC) genes in the cohort.

Gene	**Mutations**	**Samples_mutated**	**Tier**	**CGC_TumourType**
TP53	227	7	1	Breast, colorectal, lung, sarcoma, prostate, adrenocortical, glioma, multiple other tumours
KMT2C	170	7	1	Medulloblastoma
ARID1B	161	7	1	Breast, hepatocellular carcinoma, clear cell ovarian carcinoma
MED12	160	7	1	Uterine leiomyoma, fibroadenoma, phyllodes tumour
PLEC	141	6	2	Breast carcinoma

The top 20 mutated genes in the cohort comprised both CGC and non-CGC members (Figure 7). Large structural genes such as *KCNMA1* and *TTN*, al-

though not in CGC, were among the most mutated, likely reflecting mutation accumulation due to gene size rather than selective pressure. In contrast, established drivers such as *TP53*, *KMT2C*, and *ARID1B* represent biologically relevant alterations with potential roles in squamous cell carcinoma biology.

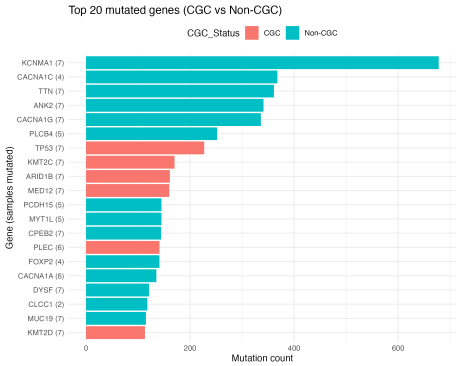


Figure 7: Top 20 most frequently mutated genes in the cohort, classified by COSMIC Cancer Gene Census (CGC) status. Numbers in brackets indicate the number of samples in which each gene was mutated. Red bars: CGC genes; blue bars: non-CGC genes.

3.4 Mutational Signature Analysis

To determine the optimal number of de novo signatures, I evaluated k values from 2 to 6 using non-negative matrix factorisation (NMF). For each k, 40 independent runs were performed, and the resulting signatures were normalised and compared to COSMIC v3 SBS reference signatures using cosine similarity. The mean of the highest similarity for each de novo signature was used as the evaluation metric (Figure 8). The highest mean(best cosine) score was achieved at $k = 5$, which was selected for downstream analysis.

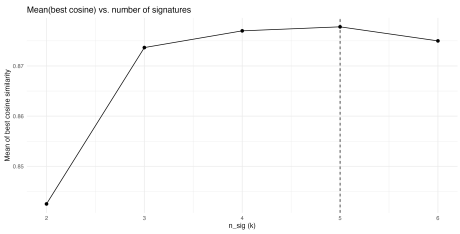


Figure 8: Evaluation of k (number of signatures) using mean(best cosine similarity) to COSMIC v3 SBS reference signatures. The optimal $k = 5$ (dashed line) achieved the highest average similarity.

Mutational signature analysis revealed four UV-associated signatures (Sig1–Sig4) closely matching COSMIC SBS7b (cosine similarity: 0.874–0.926), char-

acterized by predominant C>T mutations at dipyrimidine contexts, particularly T[C>T]T and C[C>T]T (Figure 9). This pattern aligns with UV-induced cyclobutane pyrimidine dimer (CPD) formation in cutaneous squamous cell carcinoma (cSCC) [1]. The fifth signature (Sig5, similarity: 0.862) matched SBS15, a hallmark of defective DNA mismatch repair (MMR), showing relatively balanced proportions of C>T and C>A mutations across diverse trinucleotide contexts [8]. While SBS7b's prevalence supports chronic UV exposure as the primary mutagenic source, the co-occurrence of SBS15 suggests possible MMR deficiency in a subset of clones, warranting further investigation of microsatellite instability (MSI) status.

Table 3: Best COSMIC match for each de novo signature, including cosine similarity scores.

DeNovo	COSMIC	Similarity
Sig1	SBS7b	0.926
Sig2	SBS7b	0.917
Sig3	SBS7b	0.874
Sig4	SBS7b	0.903
Sig5	SBS15	0.862

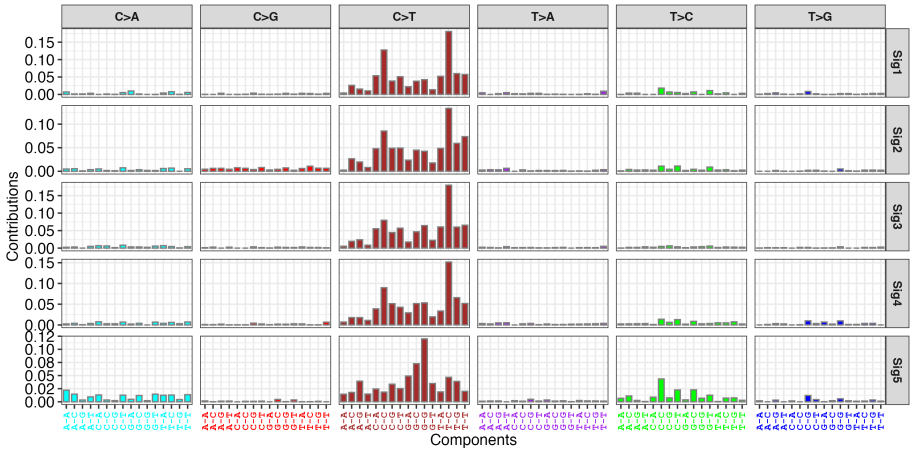


Figure 9: 96–trinucleotide context profiles of the five extracted de novo signatures, matched to COSMIC v3 SBS reference signatures.

Per-sample signature exposures (Figure 10) revealed two distinct groups of mutational processes.

- Five tumours (P3, P5, P7-P9) were dominated by UV-associated SBS7b (around 71–100% contribution), with P9 exhibiting complete reliance on this signature (Table 4).

- The remaining two (P4, P6) showed predominant SBS15 activity (54.4% and 83.4%, respectively), suggesting divergent mutagenic mechanisms.

SBS7b-high tumours had uniformly high mutation counts (median: 3,068 variants), consistent with their cutaneous origin and chronic UV exposure [15]. In contrast, SBS15-dominant cases displayed moderate mutation loads, possibly indicating MMR deficiency without hypermutation, potentially due to limited time for mutation accumulation or lower tumour proliferation rates [5]. The observed mutual exclusivity of dominant signatures implies distinct pathogenic pathways, where SBS15 may mark a rare cSCC subset with defective DNA repair.

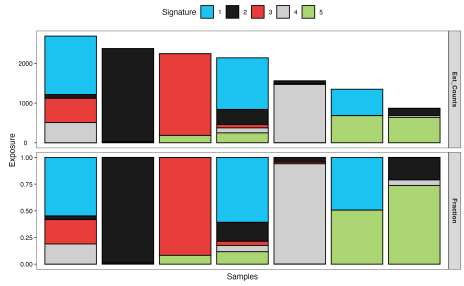
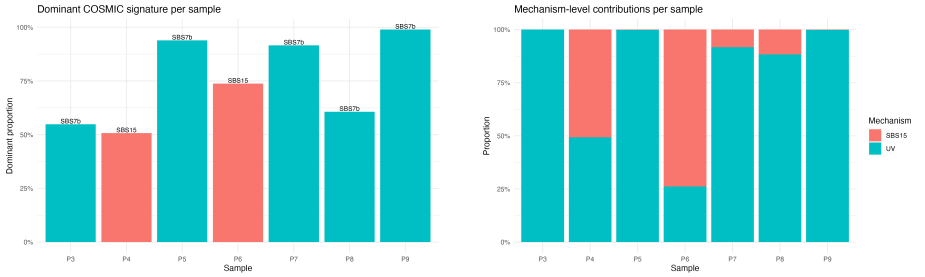


Figure 10: Per-sample exposures of the five de novo signatures. Top: estimated mutation counts. Bottom: proportion of total mutations.

Table 4: Dominant signature per sample, with both de novo and COSMIC annotation, and their proportional contributions.

Sample	Dominant_DeNovo	Dominant_COSMIC	Proportion
P3	Sig1	SBS7b	0.928
P4	Sig3	SBS15	0.544
P5	Sig1	SBS7b	0.738
P6	Sig3	SBS15	0.834
P7	Sig1	SBS7b	0.817
P8	Sig1	SBS7b	0.711
P9	Sig2	SBS7b	1.000

Grouping the exposures into mechanistic categories (UV: SBS7a–d; MMR: SBS15) revealed that UV-induced damage accounted for 70-95% of mutations in UV-dominant samples, whereas MMR deficiency contributed 75-85% in P4 and P6. No APOBEC-associated signatures (SBS2, SBS13) were detected in this cohort (Figure 11).



(a) Dominant COSMIC signature per sample. Each bar represents the dominant mutational signature (SBS7b or SBS15) and its proportion in the sample.

(b) Mechanism-level contributions per sample, grouping signatures into UV-related (SBS7a–d) and MMR deficiency (SBS15).

Figure 11: (A) Distribution of dominant COSMIC mutational signatures across samples and (B) corresponding contributions when signatures are grouped into mechanistic categories. The left panel highlights whether each sample is dominated by UV-induced (SBS7b) or MMR-deficiency (SBS15) processes, while the right panel shows the relative contribution of each mechanism within samples.

4 Conclusion

My analysis of 7 cSCC tumour–normal pairs integrated mutation burden assessment, driver gene analysis, and mutational signature profiling to dissect the molecular landscape of the disease.

The combined dataset contained 8,828 unique genes, 434 of which were listed in the COSMIC Cancer Gene Census (CGC). Several canonical drivers, including *TP53*, *KMT2C*, and *ARID1B*, were highly recurrent across samples, whereas large structural genes such as *KCNMA1* and *TTN* were frequently mutated but not recognised as CGC drivers. Cross-referencing with the CGC enabled the prioritisation of biologically relevant alterations, providing a framework for interpreting downstream mutational processes.

Mutational signature analysis revealed two dominant processes. Most tumours showed high contributions from the UV-associated SBS7b signature, consistent with chronic sun exposure as the primary driver of cSCC [15]. In contrast, two tumours were dominated by SBS15, a hallmark of mismatch repair deficiency [8], suggesting a rare MMR-deficient subtype.

These findings have both biological and clinical relevance. The strong enrichment of UV-associated mutations reinforces public health measures for UV protection and early detection in high-risk populations. The identification of MMR-deficient cases supports further molecular testing, including MSI analysis, as these tumours may be candidates for immune checkpoint inhibitor therapy [11].

Overall, this study demonstrates the value of integrating driver mutation

cross-referencing with COSMIC CGC and mutational signature profiling to refine our understanding of cSCC aetiology and identify subgroups with potential therapeutic implications.

Data and Code Availability

All analysis scripts, processed data, and figure outputs for this study are available at: `https://github.com/Shumeng-Li/Research-Rroject-2-DNAseq`

References

[1] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Meier Huang, Alvin W Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

[2] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Borresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500:415–421, 2013.

[3] Zachary R. Chalmers, Christopher F. Connelly, David Fabrizio, Lisa Gay, Sandeep M. Ali, Robert Ennis, Alexa Schrock, Brittany Campbell, Adam Shlien, Juliann Chmielecki, Feng Huang, Yuxuan He, James Sun, Samuel Jones, Victor E. Velculescu, Garrett M. Frampton, Philip J. Stephens, and David Lipson. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, 9(1):34, 2017.

[4] Limin Chen, Darwin Chang, Bishal Tandukar, Delahny Deivendran, Joanna Pozniak, Noel Cruz-Pacheco, Raymond J. Cho, et al. Stmut: A framework for visualizing somatic alterations in spatial transcriptomics data of cancer. *Genome Biology*, 24(1):273, 2023.

[5] Isidro Cortes-Ciriano, James J K Lee, Rui Xi, Dilip Jain, Young Seok Jung, Lu Yang, Dmitry Gordenin, Leszek J Klimczak, Cheng Zhang, David S Pellman, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Science*, 356(6335):eaam9078, 2017.

[6] Yevgeny Ionov, Marco A Peinado, Suren Malkhosyan, Darryl Shibata, and Manuel Perucho. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, 363(6429):558–561, 1993.

[7] Andrew L. Ji, Adam J. Rubin, Kim Thrane, Sizun Jiang, David L. Reynolds, Robin M. Meyers, Margaret G. Guo, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514.e22, 2020.

[8] Josef Jiricny. The multifaceted mismatch-repair system. *Nature Reviews Genetics*, 14(7):406–420, 2013.

[9] Alfred G. Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

[10] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, and et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499:214–218, 2013.

[11] Dung T Le, Jennifer N Durham, Knut N Smith, Hui-Yi Wang, Bryan R Bartlett, Lenny K Aulakh, Shuai Lu, Heather Kemberling, Carol Wilt, Brian S Luber, et al. Pd-1 blockade in tumors with mismatch-repair deficiency. *Science*, 357(6349):409–413, 2017.

[12] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[13] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[14] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17:122, 2016.

[15] Divya Perera, Rebecca C Poulos, Arman Shah, David Beck, John E Pimanda, and Jason W H Wong. Uv-induced dna damage and mutagenesis in skin cancer. *Journal of Clinical and Translational Research*, 5(2):1–14, 2020.

[16] Robert M. Samstein, Chloe-Han Lee, Alexander N. Shoushtari, Matthew D. Hellmann, Ronglai Shen, Yelena Y. Janjigian, David A. Barron, Ahmet Zehir, Eric J. Jordan, Antonio Omuro, Thomas Kaley, Sarah M. Kendall, Robert J. Motzer, A. Ari Hakimi, Matthew H. Voss, Paul Russo, Jamie E. Chaft, Charles M. Rudin, Gregory J. Riely, Nadeem Riaz, David M. Hyman, Barry S. Taylor, Michael F. Berger, Luc G.T. Morris, David B. Solit, and Timothy A. Chan. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics*, 51:202–206, 2019.

[17] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sam A. Ward, Ian Dunham, and Simon A. Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

[18] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tim Jordan, Khalid Shakir, David Roazen, Jeremiah Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43(1):11.10.1–11.10.33, 2013.

[19] Zena C Venables, Philippe Autier, Tamar Nijsten, Christopher S M Wong, Sinéad M Langan, Bernard Rous, Keiran S Thomas, Poornima Moayyedi, Julia Newton-Bishop, Joanne R Chalmers, et al. Epidemiology of cutaneous squamous cell carcinoma of the skin: a systematic review and meta-analysis. *British Journal of Dermatology*, 177(2):373–381, 2017.