Integrative analysis of genomic and transcriptomic data of normal, tumour and co-occurring leukoplakia tissue triads drawn from patients with gingivobuccal oral cancer identifies signatures of tumour initiation and progression

A Ghosh *et al. J Pathol* DOI: 10.1002/path.5900

Supplementary materials and methods

Supplementary Figures S1–S8

Supplementary materials and methods

Reference numbers refer to the main text list

DNA and RNA extraction, library preparation and sequencing

RNA and/or DNA were isolated from the collected tissues using AllPrep DNA/RNA Mini Kit (QIAGEN, Hilden, Germany). DNA concentration was measured using NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA). The $OD_{260/280}$ of each DNA sample exceeded 1.8. About 45 Mb of the protein coding region of the human genome was captured using Nextera Exome Enrichment Kit (Illumina, San Diego, CA, USA). Whole exome sequencing was performed on all cDNA library pools using Illumina HiSeq-2500 at $\geq$100x depth of coverage (supplementary material, Table S2).

Assessment of quality and concentration of extracted RNA samples (leukoplakia, tumour and adjacent normal) were performed using Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clare, CA, USA) and NanoDrop 2000 (Thermo Fisher Scientific). All isolated RNA samples were of $OD_{260/280} \geq 2$ and RIN (RNA Integrity Number) ≥ 7. Removal of rRNA (ribosomal RNA) was performed using Ribo-Zero Magnetic Kit (Epicentre, Illumina). Post QC, sequencing libraries were prepared using TruSeq RNA Sample Preparation Kit (Illumina). RNA of acceptable quality could be isolated from tissues of 22 (of 28) patient. Each cDNA library pool was sequenced on HiSeq-2500 for a minimum of 100 million reads (supplementary material, Table S3).

DNA sequence data alignment and post-alignment processing

From whole exome sequencing data, high quality reads (90% bases of each read with quality value > 20 and each read with < 5% N bases), determined using a FASTQ filtering software developed by us, were retained. Adapter contaminations were removed using trimmomatic-0.27 [90]. Filtered adapter-free FASTQ files were evaluated by FastQC-0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for standard quality checks. BWA-MEM [22] (version 0.7.17) was used with default parameters (additionally, -M, -T 1 was used) to align the quality filtered FASTQ files against human genome reference sequence (hg19 with decoy sequence, Source:

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz). PCR duplicates were removed using PICARD (https://github.com/broadinstitute/picard) (version 2.17.11). Local indel realignment and base-quality score recalibration were performed using GATK [24] (version 3.8) with known 1000G, dbSNP, hapmap, Mills and 1000G gold standard indel sites (included with GATK-bundle; Source: ftp://ftp.broadinstitute.org/bundle/hg19/) as per the best practices protocol to generate the BAM files. The processed BAM files were filtered further for unmapped, multi-mapped, and low quality reads (quality < 40) using samtools [91] (version 1.8). Coverage and insert size were estimated and standard QC of BAMs was performed using Qualimap [92] (version 2.2.1) tool. Finally analysis ready BAMs were prepared after checking for any cross sample contamination by using GATK tools [24] (version 4) GetPileupSummaries and CalculateContamination and gnomAD (https://gnomad.broadinstitute.org/) population VCF (af-only-gnomad.raw.sites.b37.vcf.gz) (supplementary material, Table S18).

Somatic variant calling and variant filtration

Somatic variants were detected for patients (exome sequence data; n=28) using 4 variant callers : GATK-4 Mutect2 [24], MuSE [25], strelka-2.8.4 [26] and BaseByBase-variant-caller [2]. Variants that are flagged as “PASS” by any one of those callers and detected by at least two of the callers, were accepted for further analysis. Oxo-G [93] associated variants were removed using pair-orientation-bias information generated by GATK-4 Pileup [24]. Variants with strand bias were removed from the variant callset. In order to minimise alignment artifacts, variants that reside within low complexity regions (reference sequence spanning 50 bp upstream and downstream of the variant, mapped to multiple locations with $\geq$90% sequence similarity computed with ncbi-blast-2.7.1 [94]) of genome were removed. Variants within a homopolymer

region (variant base homopolymer ≥5) were also removed. Only high confidence somatic variants with depth of coverage ≥10 in tumour/leukoplakia and variant allele count ≥3 and frequency ≥ 0.1 were retained (for somatic mutations that were present in both tumour and leukoplakia tissue of the same individual, were recovered if variant allele frequency in at least leukoplakia lesion or tumour is ≥0.05).

Detailed functional annotations were generated for the set of variants that were retained, using Oncotator-1.9.9.0 [35] (version- 2018Apr16) and ANNOVAR [34] with the latest oncotator database available as of March, 2020.

Germline SNPs were removed from the list of all variants, to retain only somatic variants. We removed a variant if its allele frequency was ≥ 0.01 in any of the 1000G populations [27] or GA100K [28] populations. Any germline variant jointly called from the data on blood samples using GATK-4 HaplotypeCaller joint genotyping method [24] was also removed (panel-of-normal filter). IGV [29] verification for all detected mutations in TCGA-pan-cancer [95] genes was performed by at least two data analysts. Mutations in known TCGA-pan-cancer genes were also detected and verified from mRNA data that were available for 22 patients. Some variants (n=25) that failed to meet the above-mentioned criteria (depth ≥10, variant containing reads ≥ 3, variant allele frequency ≥ 0.1, strand bias) but were located within a known TCGA pan-cancer driver gene were manually recovered if supportive evidence were available from RNAseq data. Somatic mutation rate per Mb of genome was obtained for tumour and leukoplakia tissue of each patient using MutSig2CV algorithm [96]. Further all exonic non-silent mutations in known OSCC-GB driver genes [2] [*TP53*, *CASP8*, *FAT1*, *PIK3CA*, *NOTCH1*, *KMT2B*, *HRAS*, *ARID2*, *CDKN2A*, *USP9X*, *EPHA2*, *HLA-B*] and two TCGA-HNSC [48] driver genes [*FBXW7*, *TGFBR2*] present in these samples were validated from RNAseq alignment data that were available for 79% of patients (22 of 28 patients).

Detection of mutation signatures

Mutation signatures were estimated from data on somatic single nucleotide variants of tumour and leukoplakia lesions, by Signal Analyse (version 2) [32] package (https://signal.mutationalsignatures.com/analyse2) with 100 bootstraps and *p*-value cut-off of 0.05 for signature fitting. The reference mutational signature specific to head and neck cancer was used as described in recent publication by Degasperi *et al.* [32] to estimate the contributions of each detected mutational signatures in our samples. The contributions of detected signatures in

each tissue and the cosine similarities of the observed and matched model were calculated using algorithms encoded in Signal Analyse. We have validated relevant results with an independent algorithm encoded in SignatureAnalyzer (v.1.1) tool (https://software.broadinstitute.org/cancer/cga/msp).

Estimation of progression time of an oral leukoplakia to a malignant tumour

To estimate the time for an oral potentially malignant lesion to progress to a malignant tumour, we used a method described by Noe *et al.* [33]. The number of additional non-synonymous mutations acquired (m_i) in the tumour of the i^{th} patient is modelled as a Poisson distribution with mean m_i. The average number of additional mutations gained m_i will be equal to $\boldsymbol{\lambda}T_i$, if the mutation rate is $\boldsymbol{\lambda}$ and time to acquire m_i mutations is T_i. The prior for T_i were assumed to be a Gamma distribution [33]. The mutation rate ($\boldsymbol{\lambda}$) was taken to be equal for all patients and constant throughout the time of progression. Since $\boldsymbol{\lambda}$ is unknown, we used a realistic range of values (1 to 10 mutations per year) for $\boldsymbol{\lambda}$ to estimate time. The model was implemented in JAGS (version 4.3.0) [https://mcmc-jags.sourceforge.io/].

Germline variant calling and detection of rare germline variants in DNA repair genes

Germline variants for each patient were detected using GATK-4 HaplotypeCaller joint calling algorithm [24]. To identify only rare variants, all variants having population allele frequency $\geq$ 0.001 in any of the 1000G populations [27] or GA100K [28] were removed. We also removed polymorphic variants with allele frequency (a) $\geq$ 0.01 in ExAC [97] or, (b) $\geq$ 0.001 in 50 normal Indian individuals (in-house data). Rare exonic non-silent variants in genes belonging to all known DNA repair pathways [KEGG [98]] were extracted for further analysis (annotated using ANNOVAR). The exonic non-silent rare germline variants in DNA repair genes (https://www.genome.jp/kegg/) detected from blood samples were further validated from data on paired tumour and leukoplakia tissues. IGV was manually screened for further filtration of variants.

Detection of somatic copy number alterations

Somatic copy number alterations were profiled in sets of blood, leukoplakia and tumour samples of the 28 OSCC-GB patients with concomitant presence of oral leukoplakia using whole exome

sequencing (WES) data. WES data were segmented using the GATK-4 [24] and significantly altered somatic genomic regions were identified using GISTIC-2.0 [30] algorithm. Broad arm-level deletion estimates were obtained from the GISTIC [30] analysis.

To validate the findings from exome-CNV analysis, for each of a random subset of 15 patients blood, leukoplakia lesion and tumour samples were genotyped by Illumina genome-wide genotyping array technology (2.5 Million). Illumina GenomeStudio was used to infer raw signal intensities from the image files. Genome-wide signal data were segmented using the ASCAT-2.5 [31] R package. Significantly altered somatic genomic regions were identified using the GISTIC [30] algorithm. Broad arm level deletion estimates were obtained from the GISTIC [30] analysis.

RNA read mapping and differential gene expression analysis

Briefly, a suite of packages was used for QC and read-mapping that included FastQC-0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), STAR-2.6.0c [36] and HTSeq-0.11.0 [37] as per GDC RNAseq-analysis recommendation [https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/].
RNAseq reads were mapped using STAR [36] aligner against Homo sapiens GRCh37 DNA primary assembly (Source: ftp://ftp.ensembl.org/pub/release-79/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.dna.primary_assembly.fa.gz). 100 million mRNA reads were obtained from 76% samples with at least 50 million reads for every sample. HTSeq [37] was used for generation of counts using Homo_sapiens.GRCh37.87.gtf (Source: ftp://ftp.ensembl.org/pub/grch37/current/gtf/homo_sapiens/Homo_sapiens.GRCh37.87.gtf.gz).
Protein coding genes that expressed at least one transcript in over 90% of patients in any set of tissues (normal/leukoplakia/oral tumour) were considered for differential gene expression analysis. DESeq2 [38] was used for identification of differentially expressed genes in leukoplakia and in tumour separately using adjacent normal as the control set in both cases. A gene was considered as differentially expressed if its |log2 fold change| exceeded 1 and the FDR-corrected p-value was <0.1.

Cellular contamination is not the cause of observed sharing of somatic mutations between leukoplakia and tumour

It may be argued that somatic mutations are shared between the leukoplakia lesion and tumour tissue of a patient because the tissue surgically resected from the leukoplakia lesion also contained some tumour tissue and is therefore "contaminated." We have investigated this possibility by considering frequencies of shared somatic mutants with those that are specific to the leukoplakia or the tumour. If the sharing of somatic mutations is because of contamination of leukoplakia and tumour tissues, then it is expected that (a) all somatic mutations of tumour and leukoplakia, especially for those with high cellular fraction (i.e., high VAF) will be shared between the two tissues, and (b) VAFs of shared somatic mutations will be lower than mutations that are unique to either leukoplakia or tumour. We note that a similar approach was adopted in a recent study on malignant progression in neoplastic pancreatic cysts [33]. Detailed results of our investigation are presented in supplementary material, Figure S7 and S8. To summarise, we found that most somatic mutations (median 56.3%) and chromosomal arm-level alterations (median 80%) were unique to tumours and not shared even with leukoplakia lesions located in proximity. Similarly, about a quarter of somatic mutations (median 26.7) found in leukoplakia lesions were absent in adjacent tumours. We observed that 61.45% (median value) of somatic variants with high allele frequencies (i.e., belonging to the upper quartile of overall allele-frequency distribution in tumours) occur exclusively in the tumour. Similarly, 34.31% of high-frequency somatic variants of leukoplakia were exclusive to leukoplakia. We validated these observations from RNA sequence reads. We also found that VAFs of shared somatic mutations were not significantly lower ($p \geq 0.05$, Kolmogorov–Smirnov test performed for each patient separately) than unique mutations in the (a) leukoplakia tissues of 77% patients, and (b) tumours of 91% of patients for each of whom the tumour were located in the proximity of leukoplakia. Thus, our observed results cannot be explained by invoking contamination between surgically resected tissues of a patient.

The evidence that most *CASP8* mutations are confined to the protease domain in OSCC-GB patients is not a chance finding

In our OSCC-GB patient cohort with concomitant presence of leukoplakia lesions, we found most tumour *CASP8* somatic mutations (81.25%) in its protease domain (supplementary material, Figure S4). This is in contrast to the observations from the larger OSCC-GB ICGC-India patient cohort [2] (n=177) where oral cancer patients both with and without leukoplakia lesions were recruited. Therefore, we sought to test whether the observation of all mutations in CASP8 being confined to the protease domain can be simply due to chance. For this, we randomly sampled 28 oral cancer patients (to match the sample size of this study) from the ICGC OSCC-GB patient

cohort and noted whether all (only if at least ten mutations appeared in the sample) *CASP8* mutations from the sampled patients are in the protease domain. We repeated this experiment 1000 times and calculated the proportion of times when 80% or more mutations in CASP8 were observed in the protease domain; the observed proportion was 0.067. Thus, the clustering of mutations in the protease domain of *CASP8* cannot be attributed to chance.

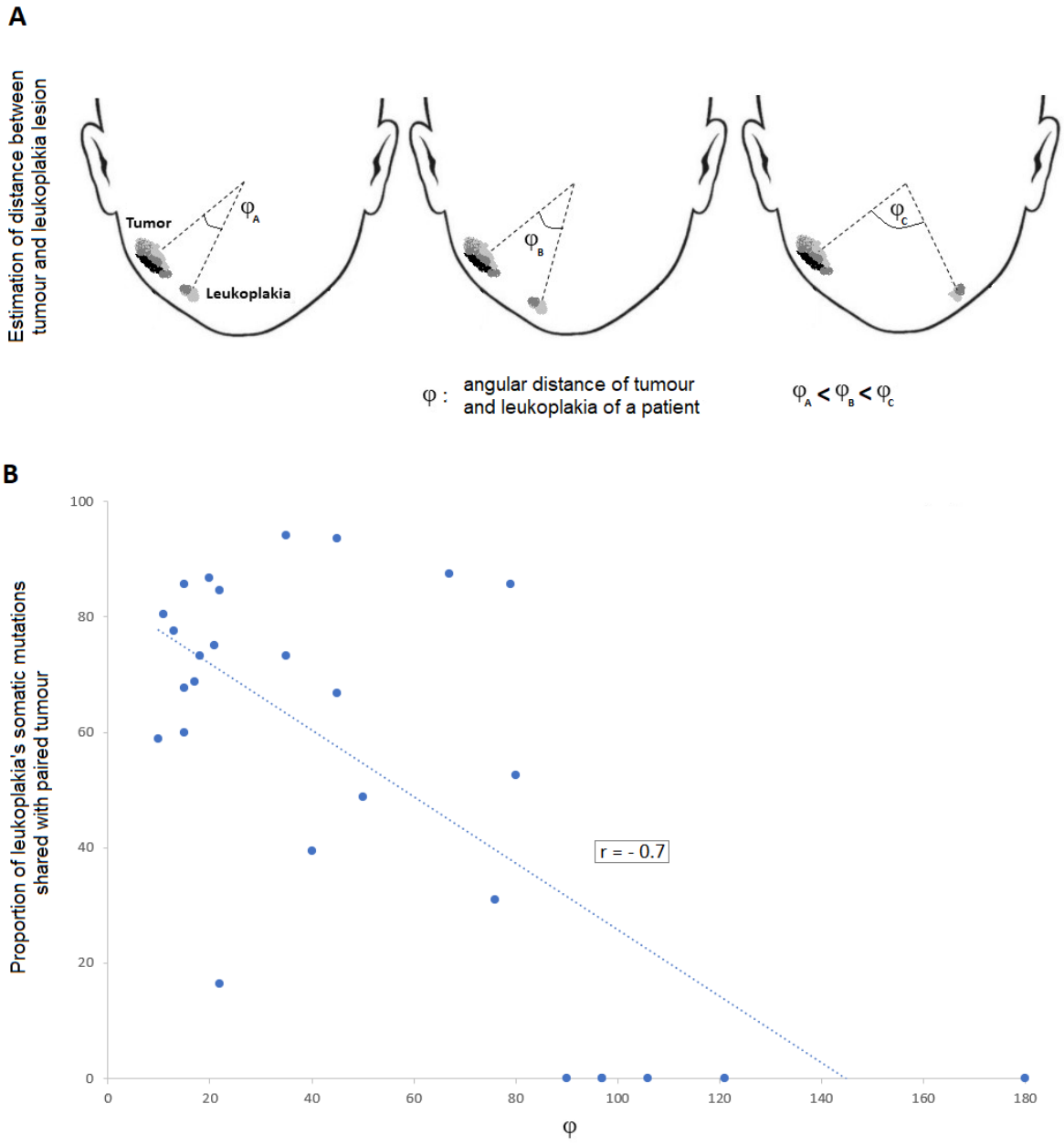

Figure S1. Relationship between the proximity of lesions and the sharing of somatic mutations.

(A) Estimation of distance between oral tumour and leukoplakia lesion present within same patient by measuring angular distance between the two tissue locations. (B) Proportion of somatic mutations in leukoplakia shared with primary tumour negatively correlates with the angular distance between primary tumour and leukoplakia lesion (φ).

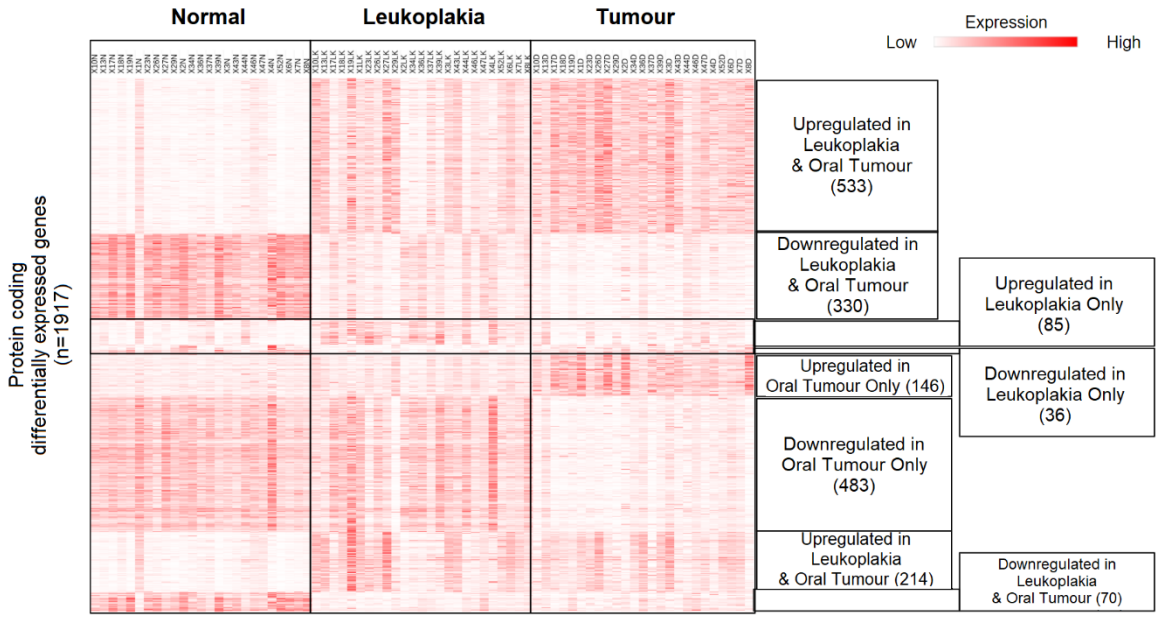

Figure S2. Differential gene expression from mRNA sequencing data showed distinct patterns in oral tumour and leukoplakia lesions.

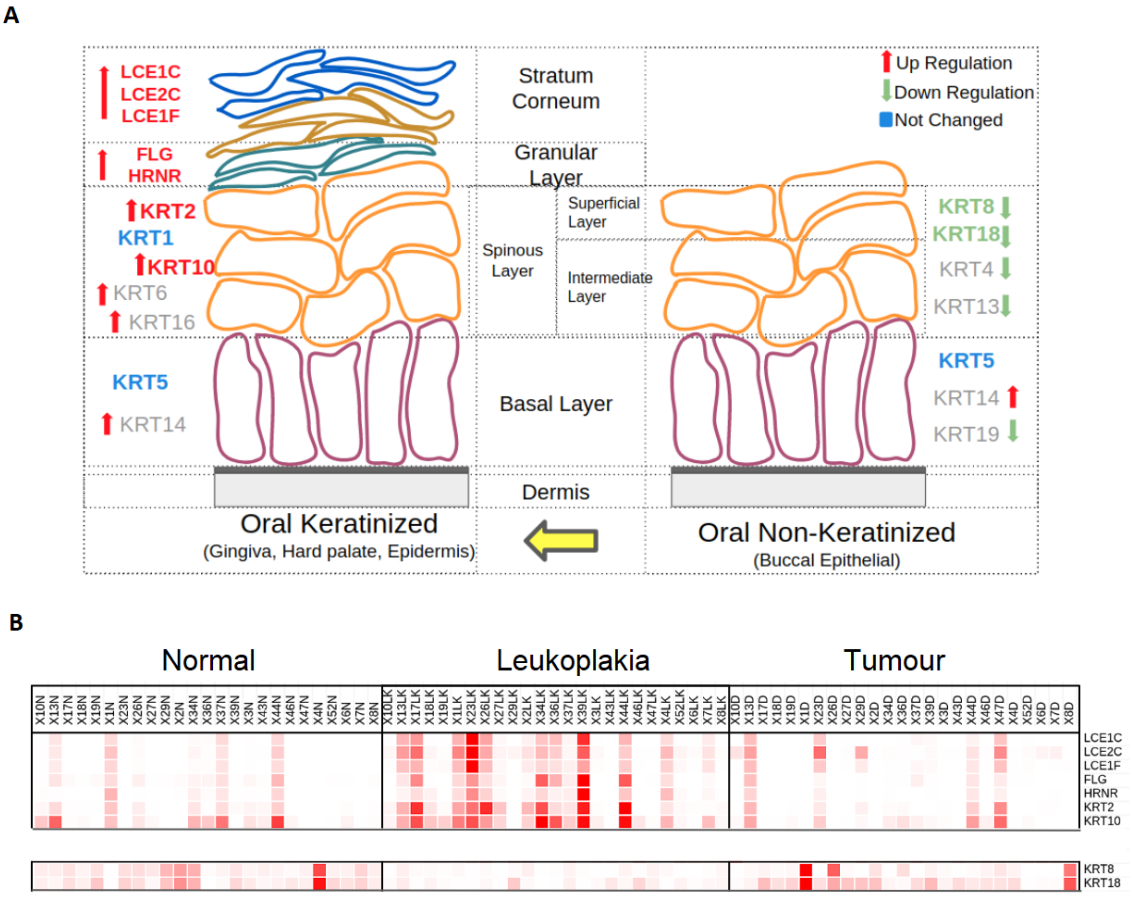

Figure S3. The predominantly-overexpressed genes of oral hairy leukoplasias and oral tumours.

(A) Analysis of mRNA expression data (25 patients) revealed overexpression of genes responsible for epithelial hyperkeratinization in oral leukoplakia lesion. Late cornified envelope family of genes (*LCE1C*, *LCE2C*, *LCE1F*), filaggrin (*FLG*), hornerin (*HRNR*), *CASP14* and keratins that are responsible for hyperkeratosis (*KRT10*, *KRT2*) were significantly upregulated in oral leukoplakia lesion as compared to adjacent normal as well as primary tumours. On the other hand, the non-keratinised epithelium determining keratins (*KRT8*, *KRT18*) which generally expresses in buccal mucosa, were down regulated in leukoplakia lesions. Collectively, these results indicated the hyperkeratinised nature of oral leukoplakia lesions which is also supported by clinical observations. (B) In contrast to leukoplakia, expression of matrix metallopeptidases (*MMP1*, *MMP9*, *MMP10*, *MMP13* etc), *SPINK6*, *IL24*, *CCNA*, *IDO1*, *VEGFC*, *FN1* etc. were significantly higher in tumours which were also observed to be upregulated in several cancer types.

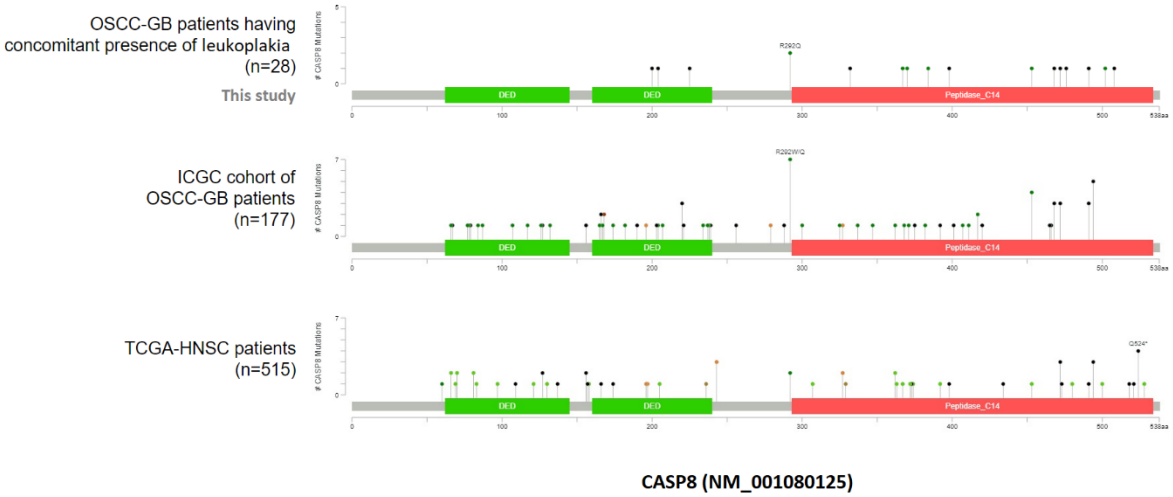

Figure S4. CASP8 mutations in this study, the ICGC-India study, and the TCGA-HNSC collection.

A key initiator protein of the apoptosis pathway encoded by the *CASP8* gene (Pro-Caspase 8) consists of 2 major domains: 1) the death effector domain (DED) and 2) the protease domain. Among the OSCC-GB patients, most CASP8 alterations (81.25%) were present in the protease domain. To exclude the possibility of this observation being solely due to chance, we carried out a resampling study using a larger (n=177 OSCC-GB patients) dataset from our ICGC-India study. We estimated that the chance of observing 80% or more mutations in the protease domain of CASP8 is only ~7%. The protease domain is responsible for activating downstream Caspases to trigger apoptosis. It is evolutionary much more conserved than the DED domain of CASP8. Therefore, alterations in the protease domain are more pathogenic. This is in contrast to the findings from larger patient cohorts, such as, ICGC-India-OSCC-GB (N=177) and TCGA-HNSC (N=515), in which patients were recruited without attention to the existence of leukoplakia lesions in the oral cavity. Among these patients, mutations in tumours occurred in both DED and protease domains of CASP8.

Figure S5. Relative proportions of anti-tumour (CD8+ cytotoxic T cells, dendritic cells), immune suppressive (CD4+ memory activated T cells) and inflammatory (activated mast cells and M0 macrophages) cell types in normal tissue, leukoplakia and tumour from the same patients.

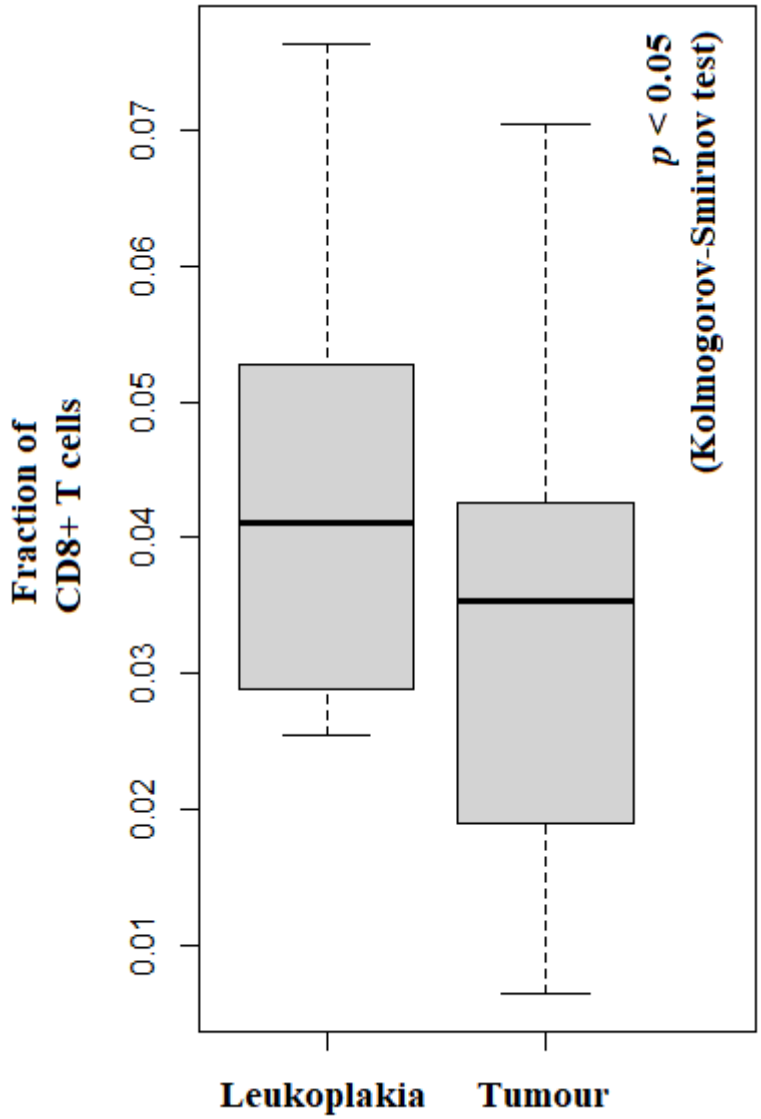

Figure S6. The abundance of CD8+ T cells estimated in each tissue sample from tumour and leukoplakia using an orthogonal algorithm encoded in EPIC to validate the findings from CIBERSORT results.

The estimation from EPIC also showed the infiltration of CD8+ T cells was significantly ($p < 0.05$, Kolmogorov-Smirnov test) lower in tumours as compared to leukoplakia; among 68% patients CD8+ T cell abundance was lower in tumour than co-existing leukoplakia tissue.

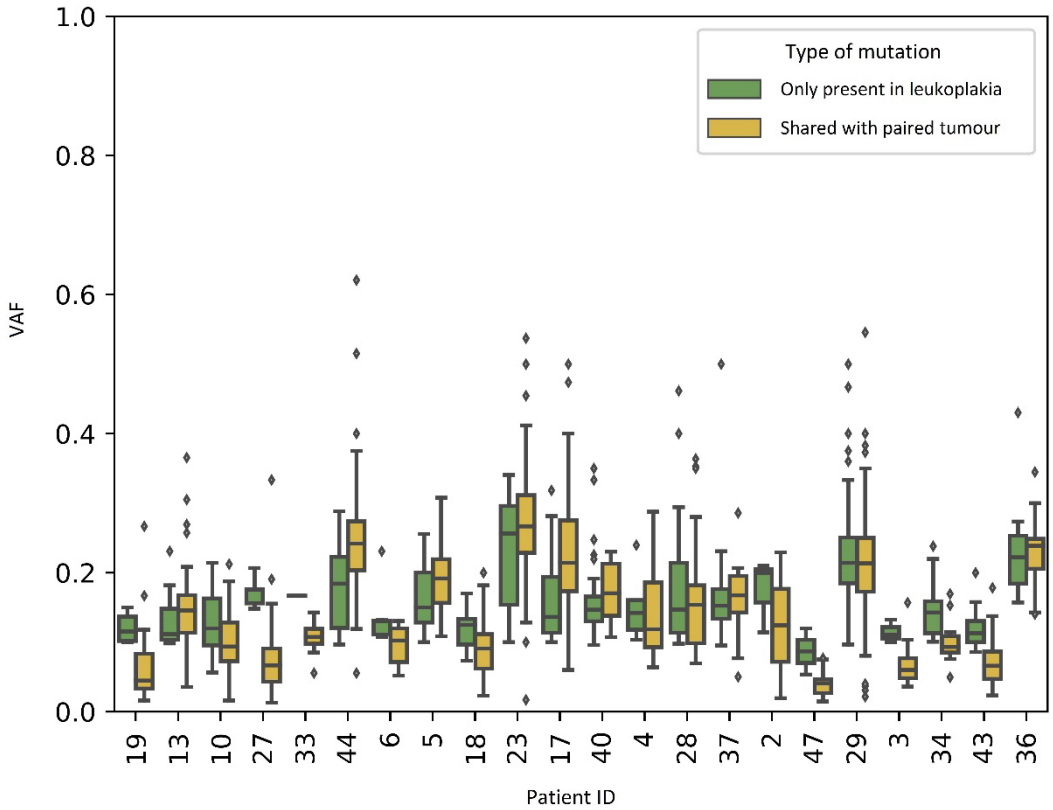

Figure S7. Distributions of variant allele frequencies (VAF) of somatic mutations in leukoplakia that are unique to leukoplakia (green) and shared with nearby tumours (yellow).

A median of 26.7 % somatic SNVs found in leukoplakia lesions that are unshared with tumour. All high frequency mutations (having higher VAF) in leukoplakia lesions were not found in nearby tumours.

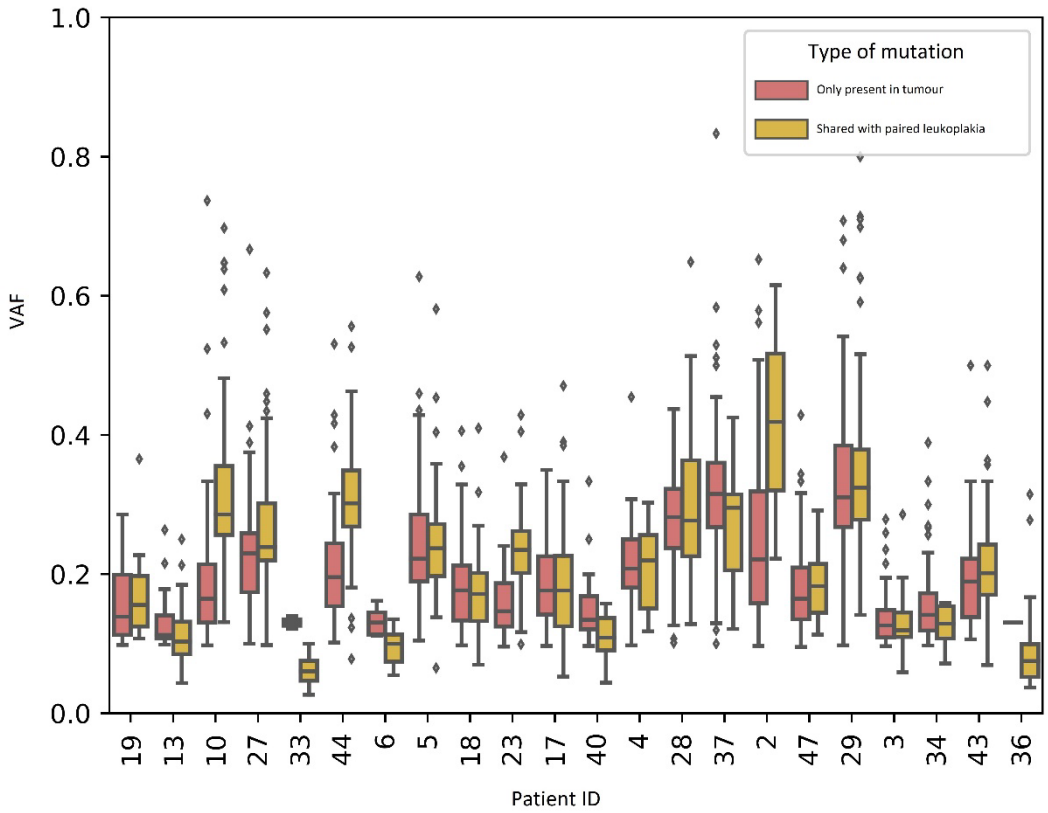

Figure S8. Patient-wise variant allele frequencies (VAF) of somatic mutations in tumours that are unique to tumours (red) and shared with adjacent leukoplakia lesions (yellow).

Overall, a median fraction of 56.3% somatic SNVs found in tumour were unshared with the leukoplakia lesion in the same patient. All high frequency mutations (having higher VAF) in tumour were absent in nearby leukoplakia tissues.