Whole-exome sequencing of oral epithelial dysplasia samples reveals an association with new genes

Daniela ADORNO-FARIAS[a]
Jean Nunes dos SANTOS[b]
Wilfredo GONZÁLEZ-ARRIAGADA[c]
Sandra TARQUINIO[d]
Rodrigo Alberto SANTIBÁÑEZ PALOMINOS[e]
Alberto Jesus MARTÍN MARTÍN[f]
Ricardo FERNANDEZ-RAMIRES[g]

[a]Universidad de Chile, School of Dentistry, Oral Medicine and Pathology Department, Santiago, Chile.

[b]Universidade Federal da Bahia – UFBA, School of Dentistry, Laboratory of Oral Surgical Pathology, Salvador, BA, Brazil.

[c]Universidad de los Andes, UANDES, School of Dentistry, Biomedical Research and Innovation Center, Santiago, Chile.

[d]Universidade Federal de Pelotas – UFPel, School of Dentistry, Diagnostic Centre for Oral Diseases, Pelotas, RS, Brazil.

[e]Universidad Mayor – UMAYOR, School of Sciences, Center for Genomics and Bioinformatics, Network Biology Laboratory, Santiago, Chile.

[f]Universidad San Sebastián, School of Engineering, Architecture and Design, Scientific and Technological Center of Excellence Science & Life, Biological Networks Laboratory, Santiago, Chile.

[g]Universidad Mayor – UMAYOR, Center for Precision Oncology, Santiago, Chile.

Declaration of Interests: The authors certify that they have no commercial or associative interest that represents a conflict of interest in connection with the manuscript.

Corresponding Author:
Ricardo Fernandez-Ramires
E-mail: ricardo.ramires@umayor.cl

https://doi.org/10.1590/1807-3107bor-2023.vol37.0016

Submitted: December 7, 2021
Accepted for publication: August 15, 2022
Last revision: September 14, 2022

Abstract: The genetic basis of oral epithelial (OED) is unknown, and there is no reliable method for evaluating the risk of malignant transformation. Somatic mutations are responsible for the transformation of dysplastic mucosa to invasive cancer. In addition, these genomic variations could represent objective markers of the potential for malignant transformation. We performed whole-exome sequencing of 10 OED samples from Brazilian and Chilean patients. Using public genetic repositories, we identified 41 deleterious variants that could produce high-impact changes in the amino acid structures of 38 genes. In addition, the variants were filtered according to normal skin and Native American genome profiles. Finally, 13 genes harboring 15 variants were found to be exclusively related to OED. High-grade epithelial dysplasia samples showed a tendency to accumulate highly deleterious variants. We observed that 62% of 13 OED genes identified in our study were also found in head and neck squamous cell carcinoma. Among the shared genes, eight were not identified in oral squamous cell carcinoma. To our knowledge, we have described for the first time 13 genes that are found in OED in a Latin American population, of which five genes have already been observed in oral squamous cell carcinoma. Through this study, we identified genes that may be related to basal biological functions in OED.

Keywords: Leukoplakia; Whole Exome Sequencing.

Introduction

The transition of normal epithelium to oral epithelial dysplasia (OED) and oral squamous cell carcinoma (OSCC) is the result of the accumulation of genetic and epigenetic alterations.[1] This complex relationship has not yet been clarified at the molecular level, which may explain treatment failures related to these diseases. Molecular stratification is an excellent tool for diagnosing benign and malignant tumors. This characterization uses last-generation technology based on sequencing and identification of typical mutations repeatedly found in the same lesions.[2]

There are many extensive studies on different neoplasms in advanced stages, but very few studies have comprehensively described the genomic changes found in precancerous lesions.[3-5] However, the correct characterization of molecular alterations in potentially malignant oral disorders (PMODs) and the corresponding changes in the microenvironment associated with progression can help contribute to the development of biomarkers for early detection and risk stratification and of preventive interventions to reverse or delay the development of cancer. Considering the complexity and diversity of the changes to be determined, comprehensive methods such as next-generation sequencing (NGS) are necessary.[6]

Whole-genome sequencing of OED lesions was first performed in a study[7] in 2009, where genomic imbalances were demonstrated in the lesions with a high risk of malignant transformation; the study also showed that the genomic profile of low-grade OED lesions that progressed to OSCC more closely resembled that of high-grade OED than that of lesions with the same histopathological diagnosis that did not progress to OSCC.[7] Another study suggested that most genomic alterations that lead to oral cancer occur in prior stages of the condition and are the result of gradual accumulations of random alterations rather than a single event.[8]

In view of the paucity of studies on the application of large-scale sequencing in PMODs and the lack of the use of this technology in the Latin American context, the aim of the present study was to identify genomic alterations in 10 low- and high-grade OED samples obtained from Brazilian and Chilean patients with a clinical diagnosis of leukoplakia using whole-exome sequencing. An understanding of DNA variations in these samples may reveal new genes associated with malignant transformation and new therapeutic targets for OED lesions.

Methodology

Patient samples

The study was approved by the Ethics and Biosafety Committees of the School of Dentistry, University of Chile (Approval nº. 2014/29) and was conducted in full accordance with local ethical guidelines and the principles of the Declaration of Helsinki. The patients were not directly involved in this study. Ten OED samples, six classified as low-grade dysplasia (LGD) and four as high-grade dysplasia (HGD), were selected from the databases of the Pathological Anatomy Service of the University of Chile, Federal University of Pelotas, and Federal University of Bahia. Histopathological diagnosis of OED was confirmed by a specialist using the binary system described by Kujan et al.[9] The selected samples were clinically diagnosed with leukoplakia according to the criteria of Van der Waal.[10] Oral medicine specialists performed the clinical and biopsy assessments.

Genomic DNA extraction

Genomic DNA was extracted from paraffin-embedded tissue samples according to the manufacturer's instructions (Puregene® DNA Purification Tissue Kit; Gentra Systems, Inc., Minneapolis, USA). DNA yield ranged from 0.2 to 2.0 µg. After processing each sample, 20 µL of the solution was obtained, and a 1.0-µL aliquot complemented with 99 µL Milli-Q® water was analyzed using a spectrophotometer (DU-640, Beckman, Palo Alto, USA) to verify the quantity and purity of each DNA sample. The genomic DNA was stored at -80 °C.

Library preparation and sequencing

Whole-exome sequencing of the 10 samples was performed by Macrogen Inc. (Seoul, South Korea). The SureSelect^{XT} Library Prep Kit (Agilent Technologies, Santa Clara, USA) was used for library preparation and exome sequencing of the DNA samples. The sequencing library was prepared by random fragmentation of each DNA sample, followed by 5′ and 3′ adapter ligations. The adapter-ligated fragments were amplified using polymerase chain reaction and purified on a gel. A post-capture classification protocol, SureSelectXT Target Enrichment System for Illumina Version B.2, was used to ensure high efficiency and coverage. The exome libraries were sequenced using 101-bp paired-end reads in a Hiseq-2500 sequencer (Illumina®), with a target sequencing depth of at least 100x. The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) were calculated for 10 samples. The GC content was 48.78%, and Q30 was 95.59%.

Data analysis

Data were analyzed in collaboration with the BioinfoGP group (Spanish National Biotechnology Centre, CNB-CSIC, Madrid, Spain). The GATK workflow was used for variant calling,[11] followed by quantification of the variants detected and comparative statistical analysis by group, based on the presence or absence of each variant, functional annotation, variant filtering, and format of the final data, to generate readable information for the end user.

The quality of the raw sequences was analyzed using the FastQC software.[12] The raw sequences were then aligned with BWA-MEM[6] against the human reference genome (Ensembl release GRCh38.91[13]) using default parameters. MarkDuplicates, BaseRecalibrator, and ApplyBQSR routines from GATK were applied to detect read duplicates and recalibrate alignment qualities. Recalibration was based on the 1000 Genomes Gold Standard provided by GATK.

The HaplotypeCaller and GenotypeGVCF GATK functions were used for SNP/indel calling and genotyping. Annotations for recalibration and variant filtering were also added. Recalibration with the VariantRecalibrator and ApplyVQSR GATK modules was based on HapMap,[14] 1000-genome high-confidence omni SNPs,[15] and dbSNP.[16] Variants from each sample were combined into a single file for comparative analysis. The case-control routine included in SnpSift[17] was used to detect variants with differential occurrence in high- and low-risk samples. P-values were obtained for different genetic models.

Each variant was annotated using the Ensembl Variant Effect Predictor with the option-everything and Condel algorithm plugins. Annotations were obtained from Ensembl, 1000 genomes,[18] Cosmic,[19] ClinVar,[20] ESP,[21] HGMDPUBLI,[22] dbSNP,[16] Gencode,[23] Genebuild,[24] gnomAD,[25] Polyphen,[26] regbuild, SIFT,[27] and Condel.[28]

To determine whether the detected variants were germline variants, the genomic sequences of unrelated subjects were analyzed. The exome sequences of induced pluripotent stem cells from skin biopsies of healthy volunteers (PRJEB11751[29]) were filtered using variant coordinates after transformation to equivalents in the human genome reference GRCh37 (https://genome.ucsc.edu/cgi-bin/hgLiftOver). In contrast, four Native American ancestral individuals (PRJEB24629[30]) were analyzed using GATK, and the obtained variants were contrasted with the coordinates of the variants found in the present study. The pipeline employed MarkDuplicates, HaplotypeCaller, SelectVariants, and VariantFiltration GATK functions, as previously described. Finally, VariantRecalibrator and ApplyVQSR were employed with known variants from dbSNP version 138, 1000 Genomes phase 1, 1000 Genomes OMNI 2.5, HapMap 3.3, Mills gold-standard, and.[31] Axiom Exome Plus. The data were obtained from https://console.cloud.google.com/storage/browser/genomicspublicdata/resources/broad/hg38/v0.

In addition, the effects of the sequence variants were evaluated using the following computer programs: PANTHER, STITCH, and PMut. Specific variant calling of HPV DNA sequences was performed for more than 170 HPV subtypes, including high-risk HPV strains[31]. Data were uploaded to the European Nucleotide Archive (https://www. ebi. ac. uk/ena) under accession number PRJEB42475.

The relationship between the number of variants per sample and per group was analyzed using the chi-square (X^2) and Fisher's exact tests. The correlation between the total number of variants and the degree of dysplasia was determined using Spearman's test. All statistical calculations were performed using GraphPad Prism 6.03 (San Diego, USA), and significance was set at $p < 0.05$.

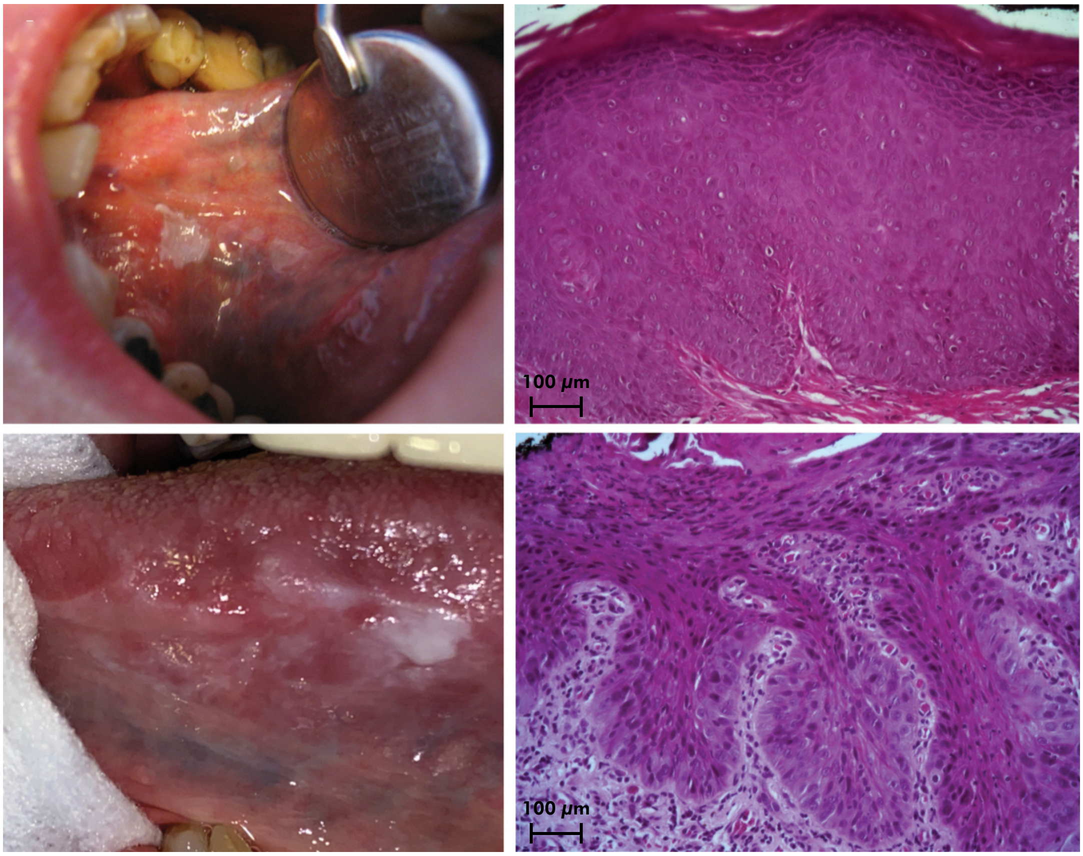
Results

Clinical and histopathological diagnostic data of the samples are presented in Table and Figure 1. Patients had no history of head and neck tumors or genetic diseases. None of the 10 samples was identified by NGS as presenting any of the 170 HPV strains. A total of 3,055,651 variants were identified and analyzed in the 10 OED samples; of these, 90.4% (2,761,210) were single-nucleotide variants (SNVs). Further, 1069 variants showed

Table. Summary of patient characteristics, histopathological diagnosis, and total variants per sample.

ID	Age (years)	Sex	Smoking	Alcohol	Lesion site	Clinic Dx	HistopDx	Country	Biopsy type	Variants	
										SNV	INDEL
										Total	Total
1	56	F	Yes	Yes	Tongue	Homogeneous leukoplakia	LGD	Chile	I	385,557	46,624
2	47	F	No	Yes	Tongue	Verrucous leukoplakia	LGD	Chile	E	324,189	39,567
3	38	F	Yes	Yes	Buccal mucosa	Homogeneous leukoplakia	LGD	Brazil	I	408,071	46,823
4	51	F	No*	No	Palate	Verrucous leukoplakia	LGD	Chile	E	407,367	46,018
5	49	F	Yes	Yes	Tongue	Verrucous leukoplakia	LGD	Chile	E	364,286	39,200
6	52	M	Yes	No	Gingival ridge	Verrucous leukoplakia	LGD	Chile	E	345,893	40,263
7	54	F	Yes	Yes	Gingival ridge	Erythroleuko.	HGD	Chile	I	366,654	41,886
8	69	F	Yes	Yes	Floor of mouth	Homogeneous leukoplakia	HGD	Chile	E	464,411	50,055
9	82	M	No*	No*	Tongue	Erythroleuko.	HGD	Brazil	I	426,368	46,704
10	38	M	Yes	Yes	Buccal mucosa	Homogeneous leukoplakia	HGD	Brazil	I	749,423	72,298

ID: Patient ID/Sample ID; Dx: Diagnosis; Histop.: Histopathological; Erythroleuko., Erythroleukoplakia; LGD: Low Grade Dysplasia; HGD: High Grade Dysplasia; * Quit tobacco/alcohol at least 5 years ago; I, Incisional biopsy; E, Excisional biopsy; SNV, Single Nucleotide Variant; INDEL, insertion and deletion variants.



A Homogeneous leukoplakia (ID 1); B Leukoerythroplakia (ID 7); C Histopathological image LGD (ID 1); D Histopathological image of HGD (ID 9).

Figure 1. Representative clinical and histopathological figures

the highest probability of causing changes with a low, moderate, or high impact on amino acid structures. These variants were responsible for changes in 773 genes, with the impact on amino acid structures being low in 416, moderate in 319, and high in 38.

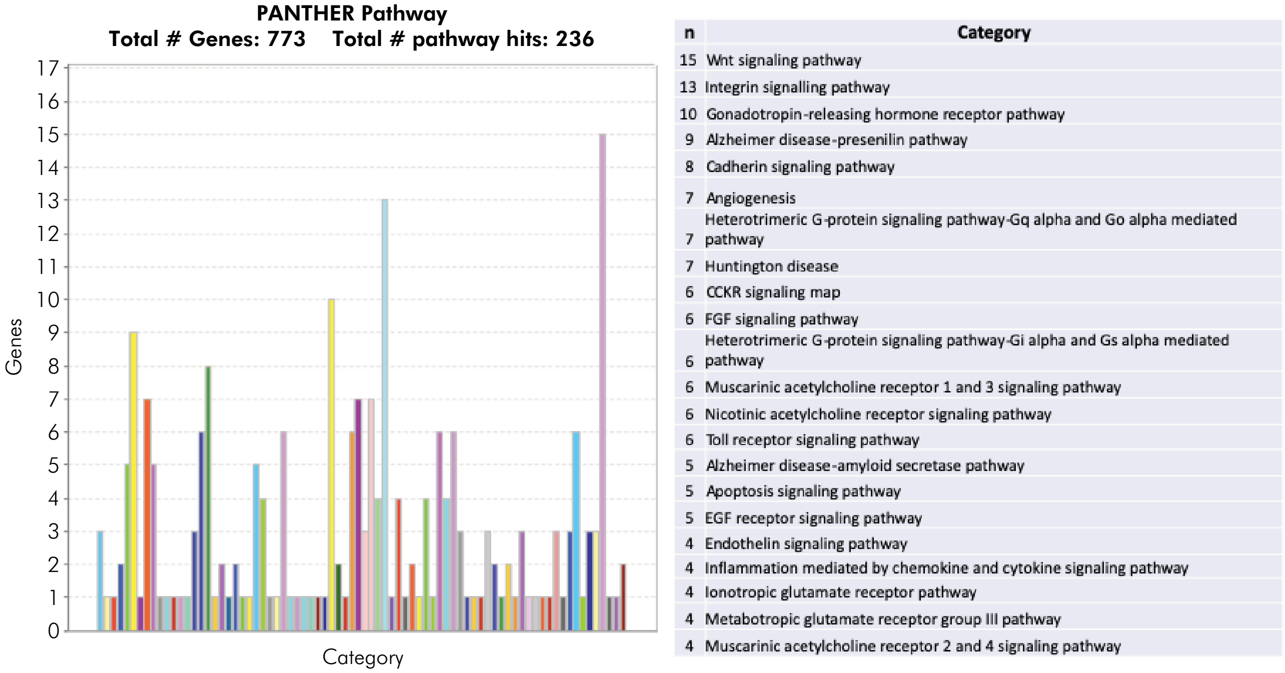
Analysis of all variants found per sample (Table 1) showed that sample 10 from the HGD group exhibited the largest number of SNVs ($p < 0.0029$*) and indels ($p = 0.77$). In contrast, samples 2 and 5 from the LGD group had the smallest number of SNVs ($p > 0.05$) and indels ($p > 0.05$), respectively. The mean total numbers of SNVs (HGD = 50,1714; LGD = 372,560) and indels (HGD = 52,736; LGD = 43,083) were higher in the HGD group than in the LGD group ($p = 0.0003$* and $p = 0.40$, respectively).

Among the 773 altered genes in the OED samples, the molecular functions that clustered with the largest number of genes evaluated were binding and catalytic activities, with 196 (25%) and 171 (22%) genes, respectively. These genes participate in more than 60 signaling pathways; however, the pathways that clustered a larger number of genes were the *Wnt* and integrin signaling pathways, involving 15 and 13 genes, respectively (Figure 2).

Forty-one variants had a high impact on the amino acid structure of 38 genes. Table 2 summarizes all high-impact variants for each sample. The mean number of variants was higher in the HGD group than in the LGD group ($p > 0.05$). In order to exclude other putative germline variants, the normal skin series (HIPSCI) database was accessed[29], remaining 22 genes harboring 24 variants. In addition to the HIPSCI database, the PRJEB24629 series[30] was also analyzed remaining 13 genes harboring 15 variants exclusive to our samples of OED (Table 3).

Of the 15 variants observed in our samples, highlighting six variants that have not been described, 8 (53%) were SNVs and 7 (47%) included a frameshift variant as the calculated functional consequence. Among the 13 genes identified, 23% were detected exclusively in LGD samples, 54%, in HGD samples, and 23%, in both samples.

Finally, considering the mutated genes in head and neck malignant tumors, Table 4 shows the mutated genes shared between OED in the present



n	Category
15	Wnt signaling pathway
13	Integrin signalling pathway
10	Gonadotropin-releasing hormone receptor pathway
9	Alzheimer disease-presenilin pathway
8	Cadherin signaling pathway
7	Angiogenesis
7	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway
7	Huntington disease
6	CCKR signaling map
6	FGF signaling pathway
6	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway
6	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
6	Nicotinic acetylcholine receptor signaling pathway
6	Toll receptor signaling pathway
5	Alzheimer disease-amyloid secretase pathway
5	Apoptosis signaling pathway
5	EGF receptor signaling pathway
4	Endothelin signaling pathway
4	Inflammation mediated by chemokine and cytokine signaling pathway
4	Ionotropic glutamate receptor pathway
4	Metabotropic glutamate receptor group III pathway
4	Muscarinic acetylcholine receptor 2 and 4 signaling pathway

Figure 2. Most representative signaling pathways of the 773 genes associated with variants of low, moderate, and high impact on amino acid structures. Tool PANTHER *Classification System* (http://pantherdb.org) was used.

Table 2. Total variants classified according to the Ensembl variant effect predictor as high impact (n = 41) per sample.

ID Sample	Histopathological Diagnosis	Variant Heterozygous	Variant Homozygous	Total Variant
1	LGD	6	9	15
2	LGD	7	9	16
3	LGD	6	15	21
4	LGD	5	9	14
5	LGD	8	7	15
6	LGD	5	12	17
x̅	LGD	6	10	16
7	HGD	7	11	18
8	HGD	15	6	21
9	HGD	11	17	28
10	HGD	13	13	26
x̅	HGD	11	12	23

study and all types of head and neck cancers, head and neck squamous cell carcinoma (HNSCC), and OSCC.

Discussion

Considering the paucity of studies applying NGS to OED samples in Latin America, the present study contributes to the description of genomic alterations in the exomes of 10 samples from Chilean and Brazilian patients with leukoplakia associated with low- and high-grade OED. Despite the difficulties in obtaining samples with sufficient quality and quantity to perform NGS, the present study included samples from patients who were clinically diagnosed with different types of leukoplakia that were compatible with low- and high-grade OED. The advantage of the present study is the representation of the correlation between genomic data of OED and clinical features.

Most variants identified in the present study were SNVs, in agreement with the literature, which highlights this as the most common alteration in whole-genome or -exome analysis.[32] Although indel variants are less frequent than SNVs, the former are, in general, extremely important in NGS because they are implicated in many constitutional and oncological diseases.[33] SNVs and indel data for each sample were filtered using databases of previously described human genetic variants to remove all known germinal variants. The highest and significant number of average SNVs was observed in the HGD group, in line with the results of studies showing that precancerous lesions are characterized by progressive changes in the DNA sequence, gene expression, and protein structures as well as by microscopic rearrangements.[3,4] In addition, a previous study reported a smaller number of mutations in LGD samples than in HGD and OSCC[8] samples.

Most genes identified in the present study, with a high impact on amino acid structures, are related to metabolic functions such as binding and catalytic activities and participate in the *Wnt* and integrin signaling pathways. Functional dysregulation of the *Wnt* signaling pathway has been shown to promote the development and progression of oral cancer. Therefore, it is an interesting target for treatment strategies for this cancer.[34] Integrins, the main components of cell adhesion, have been implicated in almost all stages of cancer progression, from development of the primary tumor to metastasis.[35]

We identified six new variants, including three SNVs with functional consequences at the splice acceptor and splice donor sites and three deletions that lead to changes in the reading frame. Harboring these variants, the *GAREM1*, *GIPC1*, and *LRRC37A2* genes are associated with mutations described in HNSCC and OSCC.[36]

Table 3. Thirteen genes harboring 15 variants attributed to oral epithelial dysplasia.

Variant	Genes	Position	Type	REF-ALT	HGVSc	HGVSp	ID	S
1	*DRAM2*	Chr1:111,119,614	SNV	C→T	ENST00000477769.1:n.235+1G>A	_	rs644081	7;8
2	*C4orf36*	Chr4:86,892,061	SNV	A→T	ENST00000506308.5:c.-74+2T>A	_	rs10034336	8;10
3	*FAM198B*	Chr4:158,171,600	SNV	C→A	ENST00000296530.12:c.-224-1G>T	_	ND	3
	FAM198B	Chr4:158,171,600	SNV	C→A	ENST00000393807.9:c.-224-1G>T	_	ND	3
	FAM198B	Chr4:158,171,600	SNV	C→A	ENST00000585682.5:c.-224-1G>T	_	ND	3
	FAM198B	Chr4:158,171,600	SNV	C→A	ENST00000592057.1:c.-224-1G>T	_	ND	3
4	*SEPT14P1*	Chr7:63,148,830	DEL.	TG→T	ENST00000458703.1:n.76-2del	_	rs57585159	8;10
5	*CBWD5*	Chr9:65,733,001	DEL.	GTCAA→G	ENST00000377392.9:c.308_311del	ENSP00000366609.6:p.Ile103LysfsTer7	ND	9
	CBWD5	Chr9:65,733,001	DEL.	GTCAA→G	ENST00000377395.8:c.821_824del	ENSP00000366612.4:p.Ile274LysfsTer7	ND	9
	CBWD5	Chr9:65,733,001	DEL.	GTCAA→G	ENST00000382405.7:c.965_968del	ENSP00000371842.3:p.Ile322LysfsTer7	ND	9
	CBWD5	Chr9:65,733,001	DEL.	GTCAA→G	ENST00000429800.6:c.905_908del	ENSP00000405076.2:p.Ile302LysfsTer7	ND	9
	CBWD5	Chr9:65,733,001	DEL.	GTCAA→G	ENST00000430059.6:c.908_911del	ENSP00000397999.2:p.Ile303LysfsTer7	ND	9
	CBWD5	Chr9:65,733,001	DEL.	GTCAA→G	ENST00000489273.1:c.309_312del	ENSP00000417466.1:p.Ile104LysfsTer7	ND	9
6	*CELA1*	Chr12:51,346,623	DEL	CATAA→C	ENST00000293636.1:c.12_15del	ENSP00000293636.1:p.Tyr5AspfsTer14	rs753836828	2;5;6;7;8;9;10
7	*CELA1*	Chr12:51,346,631	INS	A→AAG	ENST00000293636.1:c.7_8insCT	ENSP00000293636.1:p.Val3AlafsTer18	rs370927847	2;5;6;7;8;9;10
8	*CELA1*	Chr12:51,346,632	INS	C→CG	ENST00000293636.1:c.6_7insC	ENSP00000293636.1:p.Val3ArgfsTer22	rs148235680	2;5;6;7;8;9;10
9	*PKD1L3*	Chr16:71,947,511	DEL	CTTTG→C	ENST00000620267.1:c.3695_3698del	ENSP00000480090.1:p.Thr1232ArgfsTer26	ND	7;8;9
10	*LRRC37A2*	Chr17:46,512,902	DEL	TC→T	ENST00000333412.3:c.192del	ENSP00000333071.3:p.Ser65ProfsTer69	ND	10
	LRRC37A2	Chr17:46,512,902	DEL	TC→T	ENST00000576629.5:c.192del	ENSP00000459551.1:p.Ser65ProfsTer69	ND	10
11	*GAREM1*	Chr18:32,136,366	SNV	A→G	ENST00000583696.1:c.*48+2T>C	_	ND	10
12	*PCSK4*	Chr19:1,490,421	SNV	A→C	ENST00000591687.2:n.24+2T>G,	_	rs806528	1;2;3;6
13	*GIPC1*	Chr19:14,483,329	SNV	A→G	ENST00000591245.1:c.-31+2T>C	_	ND	10
14	*RPL13A*	Chr19:49,490,126	SNV	G→T	ENST00000477613.6:n.94-1G>T	_	rs55991145	4;7;9
15	*ZNF83*	Chr19:52,616,816	SNV	C→T	ENST00000595939.5:c.*119-1G>A	_	rs191291023	2

ND: variant not described; REF-AT: reference-change; ID: identification of the variant; Cr.: chromosome; SNV: single nucleotide variant; Del.: deletion; INS.: insertion; HGVSc: name of the coding sequence suggested by the Human Genome Variation Society (HGVS); HGVSp, HGVSc: name of the protein sequence suggested by the Human Genome Variation Society (HGVS); S: samples with the variant. Nomenclatures of the calculated functional consequences of the variants given by the Ensembl program.

Table 4. Similarities of genes with variants identified in OED samples from the present study (n = 15 genes), with the groups of genes mutated in samples of HNC (n = 16807 genes), HNSCC (n = 16099 genes), and OSCC (n = 2656 genes).

Genes with variants in the present study	Mutated genes in HNC*	Mutated genes in HNSCC*	Mutated genes in OSCC*
C4orf36			
CBWD5	X	X	
CELA1			
DRAM2	X	X	X
FAM198B			
GAREM1	X	X	X
GIPC1	X	X	X
LRRC37A2	X	X	X
PCSK4	X	X	
PKD1L3			
RPL13A	X	X	
SEPT14P1			
ZNF83	X	X	X

OED: oral epithelial dysplasia; HNC: head and neck cancer; HNSCC: head and neck squamous cell carcinoma; OSCC: oral squamous cell carcinoma. *Genomic data information obtained through the platform cBioPortal.

Similar to the HGD samples that showed the accumulation of a large number of total variants, the same trend was observed with LGD samples when only high-impact variants were considered; however, this association was not significant. This finding is in agreement with that of a previous study that showed a smaller number of mutations in LGD samples than in HGD samples[8]. Regarding clinical characteristics of patients, it was not possible to establish a relationship with the variants found because the sample size was too small for this type of correlation. The correlation between clinical characteristics and genomic variation has not yet been established.

Regarding the *CELA1* gene, it is important to note that this study detected three variants with a high impact on this gene, which were identified in the same samples with the same type of inheritance. *CELA1*, also known as *ELA1*, encodes elastase-1 and is localized on chromosome 12q13, near the locus for diffuse non-epidermolytic palmoplantar keratoderma. Expression of this gene has been observed in cultured human primary keratinocytes.[37]

It is well known that the distribution of mutations in the genome is not completely random. In the present study, the observation of variants that affected *CELA1* in the same group of samples may be explained by mutation showers, which are not yet fully understood. This phenomenon is characterized by the simultaneous presence of multiple mutations in the same gene or small regions of the chromosomes;[38] these alterations have not yet been explained or associated with cancer; however, an analysis of available mutation catalogs revealed clustered mutagenesis in multiple myeloma and prostate and head and neck tumors.[38]

On comparison of the most severely affected genes identified here with the mutated genes reported by Wood et al.,[8], although they are different variants, a match was found with the mutated *WNK1* gene only in OED samples, with the mutated *MCF2L* gene in OED and OSCC samples, and with the mutated *LAMA5*, *FARP1*, and *SHANK2* genes exclusively in OSCC samples.[8] Observation of this match in only two mutated genes in OED might be explained by the fact that, contrary to the present study, which used clinically and histopathologically representative samples, Wood et al.[8] extracted OED areas from OSCC samples, which may have increased the probability of molecular differences. Although there were fewer mutations in OED samples than in OSCC samples, in that study, most mutations detected in OED samples were also observed in OSCC samples.[8]

In the present study, no normal paired controls were available for Whole-exome sequencing; however, we used bioinformatics methods to remove false-positive variants. To address similarities and differences with normal epithelial tissue, the HIPSCI genome database[29] was analyzed, and the variants were found to be germlines that were not filtered within other public genetic repositories. It is difficult to explain how these variants found in normal skin were not previously found after filtering using tools such as 1000 genomes, cosmic, and dbSNP. O´Huallachain et al.[39] confirmed the presence of a large number of variations in somatic tissues. This can be partly explained by the fact that choosing the relevant tissue for the comparison of genomic profiles might influence the data analysis.

It is also important to understand the history and diversification of human populations at the southern tip of the Americas. The South American population has a unique genetic conformation composed of pre-Columbian and post-colonization genetic signatures. This heterogeneity could play an important role in explaining the number of variants found in this study after variant calling based on international databases, including HIPSCI normal skin genomes.[29] To address this point, we used only the available genome profiles of native Americans representing the pre-Columbian southerners. Interestingly, 38% of the variants not described in either public genetic repositories or the normal skin database were found in the native American genomes, and these could be considered the germline. Of note, these "southern variants" were localized within the same mutated genes referred to in OED previous studies, such as Wood et al.[8] In addition, some genes are considered to harbor a high malignant potential.[8,40]

It is also important to evaluate the roles of these 13 OED genes in malignancy. Similar to The Cancer Genome Atlas (TCGA), which was established for consultations on the genomic diversity of various types of cancer, the Pre-Cancer Genome Atlas project was started in 2016. However, mutated genes for any of the lesions that precede different types of cancers, including OED, are not yet available on these platforms[3,4]. Based on TCGA,[36] 62% of the 13 OED genes identified in our study were also found in HNSCC. Among the shared genes, eight were not identified in OSCC.

It is important to mention that nine of the 11 altered genes identified in more than half of the samples in this study were also mutated in HNSCC and six were mutated in OSCC. In the study of Wood et al.,[8] *SHANK2* and *FARP1* had mutated in OSCC, and *MFC2L*, in OSCC and OED, which were also altered in our samples and in those of other studies on HNSCC[36] and OSCC.[36] Similarities in the genomic profile of OED and cancer have been described for the intestinal, breast, brain, kidney, lung, and skin epithelium, showing that the mutational process can cause clonal evolution from normal to neoplastic cells.[3,4] However, Wood et al.[40] observed subclonal heterogeneity of OSCC in five OED samples and suggested that mutational changes in stages prior to cancer do not predict the onset of invasion.[40]

In 2009, a study demonstrated completely different genomic profiles of OED that progressed to OSCC compared to some other OED that did not progress to cancer despite histological similarities.[7] Despite this observation, 10 years after these discoveries, histopathological diagnosis, including the identification of different stages of OED, continues to be the standard complementary test for outlining the risk of progression and treatment decisions for PMODs. However, this method remains subjective, and diagnostic agreement between pathologists is low. In addition, regardless of their degree, not all OEDs progress to OSCC, and this information cannot be obtained through histopathological analysis. Given this current scenario, our study describes 13 genes harboring 15 variants, providing relevant information for the genomic characterization of OED. Despite the small sample size, the use of a sample comprising a heterogeneous population and an in-depth genomic evaluation method, which currently has an extremely low error rate, allowed us to obtain highly reliable results. However, it is important to complement the main results of prospective multicenter studies using studies with large sample sizes, including validations and healthy controls.

Conclusion

To our knowledge, the present study describes for the first time 13 genes (*DRAM2, C4orf36, FAM198B, SEPT14P1, CBWD5, CELA1, PKD1L3, LRRC37A2, GAREM1, PCSK4, GIPC1, RPL13A, ZNF83)* found in OED samples from Latin American patients that may be related to basal biological functions in OED.

Acknowledgments

All procedures performed in this study were in accordance with the ethical standards of the Scholl Dentistry of the University of Chile (Approval N°. 2014/29) committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. The patients were not directly involved in this study. Funding: FONDECYT nº 11140281 (Ricardo Fernandez-Ramires), Centro Ciencia & Vida, FB210008, Financiamiento Basal para Centros Científicos y Tecnológicos de Excelencia de ANID.

References

1. El-Naggar AK, Chan JK, Grandis JR, Takata T, Slootweg PJ, eds. WHO Classification of head and neck tumours. 4th ed. Lyon: IARC; 2017.
2. Graveland AP, Bremmer JF, Maaker M, Brink A, Cobussen P, Zwart M, et al. Molecular screening of oral precancer [Internet]. Oral Oncol. 2013 Dec;49(12):1129–35. https://doi.org/10.1016/j.oraloncology.2013.09.005
3. Campbell JD, Mazzilli SA, Reid ME, Dhillon SS, Platero S, Beane J, et al. The case for a Pre-Cancer Genome Atlas (PCGA). Cancer Prev Res (Phila). 2016 Feb;9(2):119–24. https://doi.org/10.1158/1940-6207.CAPR-16-0024
4. Spira A, Yurgelun MB, Alexandrov L, Rao A, Bejar R, Polyak K, et al. Precancer atlas to drive precision prevention trials. Cancer Res. 2017 Apr;77(7):1510–41. https://doi.org/10.1158/0008-5472.CAN-16-2346
5. Farah CS, Jessri M, Bennett NC, Dalley AJ, Shearston KD, Fox SA. Exome sequencing of oral leukoplakia and oral squamous cell carcinoma implicates DNA damage repair gene defects in malignant transformation. Oral Oncol. 2019 Sep;96(April):42-50. https://doi.org/10.1016/j.oraloncology.2019.07.005
6. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-prints. 2013.
7. Garnis C, Chari R, Buys TP, Zhang L, Ng RT, Rosin MP, et al. Genomic imbalances in precancerous tissues signal oral cancer risk. Mol Cancer. 2009 Jul;8(1):50. https://doi.org/10.1186/1476-4598-8-50
8. Wood HM, Daly C, Chalkley R, Senguven B, Ross L, Egan P, et al. The genomic road to invasion-examining the similarities and differences in the genomes of associated oral pre-cancer and cancer samples. Genome Med. 2017 Jun;9(1):53. https://doi.org/10.1186/s13073-017-0442-0
9. Kujan O, Oliver RJ, Khattab A, Roberts SA, Thakker N, Sloan P. Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. Oral Oncol. 2006 Nov;42(10):987-93. https://doi.org/10.1016/j.oraloncology.2005.12.014
10. Waal I. Potentially malignant disorders of the oral and oropharyngeal mucosa; terminology, classification and present concepts of management. Oral Oncol. 2009 Apr-May;45(4-5):317-23. https://doi.org/10.1016/j.oraloncology.2008.05.016
11. Van der Auwera G, Carneiro M, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit bestpractices pipeline. Curr Protoc Bioinforma. 2013;43(11.10):1-33.
12. Andrews S. A quality control tool for high throughput sequence data. 2010 [cited 2018 Mar 8]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
13. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018 Jan;46 D1:D754-61. https://doi.org/10.1093/nar/gkx1098
14. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec;426(6968):789-96. https://doi.org/10.1038/nature02168
15. 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. https://doi.org/10.1038/nature11632
16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan;29(1):308-11. https://doi.org/10.1093/nar/29.1.308
17. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. Front Genet. 2012 Mar;3:35. https://doi.org/10.3389/fgene.2012.00035
18. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al.; 1000 Genomes Project Consortium. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct;526(7571):75-81. https://doi.org/10.1038/nature15394
19. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017 Jan;45 D1:D777-83. https://doi.org/10.1093/nar/gkw1121
20. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan;44 D1:D862-8. https://doi.org/10.1093/nar/gkv1222
21. NHLBI Exome Sequencing Project (ESP). Exome Variant Server. 2017 [cited 2018 Mar 3].Available from: http://evs.gs.washington.edu/EVS/
22. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003 Jun;21(6):577-81. https://doi.org/10.1002/humu.10212

23. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep;22(9):1760-74. https://doi.org/10.1101/gr.135350.111
24. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database (Oxford). 2016 Jun;2016:baw093. https://doi.org/10.1093/database/baw093
25. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285-91. https://doi.org/10.1038/nature19057
26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9. https://doi.org/10.1038/nmeth0410-248
27. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001 May;11(5):863-74. https://doi.org/10.1101/gr.176601
28. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011 Apr;88(4):440-9. https://doi.org/10.1016/j.ajhg.2011.03.004
29. Fuente C, Ávila-Arcos MC, Galimany J, Carpenter ML, Homburger JR, Blanco A, et al. Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. Proc Natl Acad Sci USA. 2018 Apr;115(17):E4006-12. https://doi.org/10.1073/pnas.1715688115
30. Chandrani P, Kulkarni V, Iyer P, Upadhyay P, Chaubal R, Das P, et al. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. Br J Cancer. 2015 Jun;112(12):1958-65. https://doi.org/10.1038/bjc.2015.121
31. Shigemizu D, Fujimoto A, Akiyama S, Abe T, Nakano K, Boroevich KA, et al. A practical method to detect SNVs and indels from whole genome and exome sequencing data. Sci Rep. 2013;3(1):2161. https://doi.org/10.1038/srep02161
32. Sehn JK. Insertions and deletions (indels).In: Shashikant K, Pfeifer J, editors. Clinical genomics. Elsevier; 2015. p. 129-150.
33. Shiah SG, Shieh YS, Chang JY. The role of Wnt signaling in squamous cell carcinoma. J Dent Res. 2016 Feb;95(2):129-34. https://doi.org/10.1177/0022034515613507
34. Hamidi H, Ivaska J. Every step of the way: integrins in cancer progression and metastasis. Nat Rev Cancer. 2018 Sep;18(9):533-48. https://doi.org/10.1038/s41568-018-0038-z
35. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015 Jan;517(7536):576-82. https://doi.org/10.1038/nature14129
36. Talas U, Dunlop J, Khalaf S, Leigh IM, Kelsell DP. Human elastase 1: evidence for expression in the skin and the identification of a frequent frameshift polymorphism. J Invest Dermatol. 2000 Jan;114(1):165-70. https://doi.org/10.1046/j.1523-1747.2000.00825.x
37. Chan K, Gordenin DA. Clusters of multiple mutations: incidence and molecular mechanisms. Annu Rev Genet. 2015;49(1):243-67. https://doi.org/10.1146/annurev-genet-112414-054714
38. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci USA. 2012 Oct;109(44):18018-23. https://doi.org/10.1073/pnas.1213736109
39. Wood HM, Conway C, Daly C, Chalkley R, Berri S, Senguven B, et al. The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. J Pathol. 2015 Nov;237(3):296-306. https://doi.org/10.1002/path.4576.