Whole exome sequencing of oral epithelial dysplasias reveals an association with new genes

Daniela Adorno-Fañás  
Universidad de Chile

Sandra Tarquinio  
Universidade Federal de Pelotas

Wilfredo González-Arriagada  
Universidad de Valparaiso

Rodrigo Alberto Santibáñez Palominos  
Universidad Mayor

Alberto Jesus Martín Martín  
Universidad Mayor

Jean Nunes dos Santos  
Universidade Federal da Bahia

Ricardo Fernandez-Ramires  
ricardo.ramires@umayor.cl  
Universidad Mayor

Research Article

Keywords: oral epithelial dysplasia, genomic alterations, whole exome sequencing, next-generation sequencing.

DOI: https://doi.org/10.21203/rs.3.rs-440913/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: The genetic basis of oral epithelial dysplasias is unknown and there is no reliable method of evaluating malignant transformation risk. We understand that somatic mutations are responsible for the transformation of the dysplastic mucosa to cancer. In addition, these genomic variations could represent objective markers to determine potential for malign transformation.

Material and Methods: We performed whole exome sequencing in ten samples of oral epithelial dysplasia of Brazilian and Chilean patients.

Results: Based on public genetic repositories, we identified 41 deleterious variants able to produce high impact changes in amino acid structures of 38 genes. Also, variants were filtered based on normal skin and native American genome profiles. Finally, a total of 13 genes harboring 15 variants were found to be exclusively related to oral epithelial dysplasias. Samples of high-grade epithelial dysplasia showed a tendency to accumulate more deleterious variants. We observed that 62% of 13 OED genes pinpointed in our study were also found in HNSCC. Among the shared genes, eight were not identified in OSCC.

Conclusions: We have described for the first time in a Latin America population 13 genes which are found in oral epithelial dysplasia where 5 genes have already been observed in oral squamous cell carcinoma. This study allowed us to identify genes that may be related to basal biological functions in oral epithelial dysplasias.

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 [1–3]. This complex relationship has not yet been clarified at the molecular level, a fact that possibly explains many failures related to these diseases. An excellent tool for the diagnosis of other benign lesions and malignant tumors is molecular stratification. This characterization uses last-generation technology based on sequencing and the identification of typical mutations, i.e., those repeatedly found in the same lesions 2,3.

In contrast to the extensive research involving different neoplasms in advanced stages, few studies have comprehensively described the genomic changes found in precancerous lesions 4–6. However, the correct characterization of the molecular alterations in potentially malignant oral disorders (PMOD) and the corresponding changes in the microenvironment associated with progression would contribute to the development of biomarkers for early detection and risk stratification, in addition to suggesting preventive interventions to reverse or delay the development of cancer. Considering the complexity and diversity of the changes to be determined, more comprehensive methods such as next-generation sequencing (NGS) are necessary 7–9.

Whole genome sequencing of OED lesions was first performed in 2009. That study demonstrated genomic imbalances in lesions with a higher 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 10. Another study suggested that most genomic alterations that lead to oral cancer occur in prior stages and are the result of gradual random accumulation rather than of a single event 11.

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

Materials And Methods

Patient samples

The study was approved by the Ethics and Biosafety Committees of the School of Dentistry, University of Chile (FONDECYT No. 11140281), and was conducted in full accordance with local ethical guidelines and with the Declaration of Helsinki. The patients weren’t direct involved in this study. Ten samples of OED, 6 classified as low-grade dysplasia (LGD) and 4 as high-grade dysplasia (HGD), were selected from the databases of the Pathological Anatomy Service of University of Chile, Federal University of Pelotas and Federal University of Bahia. The histopathological diagnosis of OED was made by a specialist using the binary system described by Kujan et al 12. The selected samples had a clinical diagnosis of leukoplakia according to the criteria of Van der Waal 13. Oral medicine specialists performed the clinical and biopsy assessments.

Genomic DNA extraction

Genomic DNA was extracted from the paraffin-embedded tissue samples following the instructions of the kit’s manufacturer (Puregene® DNA Purification Tissue Kit, Gentra Systems, Inc., Minneapolis, MN, USA). The DNA yield ranged from 0.2 to 2.0 µg. After processing each sample, 20 µl
showed the highest probability of causing changes with a low, moderate and high impact on amino acid structures. These variants were identified and analyzed in the 10 OED samples; of these, 90.4% (2,761,210) were single nucleotide variants (SNV). There were 1.069 variants that were obtained from Ensembl, 1000 genomes and dbSNP.

Results

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

The quality of the raw sequences was analyzed using the FastQC software. The raw sequences were then aligned with BWA-MEM against the human reference genome (Ensembl release GRCh38.91) using parameters per default. MarkDuplicates, BaseRecalibrator and ApplyBQSR routines from GATK were applied to detect read duplicates and to recalibrate alignment qualities, respectively. 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 of recalibration and variant filtering were added. Recalibration with the GATK routines VariantRecalibrator and ApplyVQSR was based on HapMap, 1000 genomes high confidence omni SNPs and dbSNP. Variants from each sample were then combined into a single file for comparative analysis. The Case Control routine included in SnpSift was used to detect variants with differential occurrence between 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 plugin. Annotations were obtained from Ensembl, 1000 genomes, Cosmic, ClinVar, ESP, HGMDPUBLI, dbSNP, Gencode, Genebuild, gnomAD, Polyphen, regbuild, SIFT, Condel.

To determine if the detected variants are germ-line, genomic sequences of unrelated subjects were analyzed. The exome sequences of induced pluripotent stem cells from skin biopsies of healthy volunteers were filtered employing the 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 were analyzed with the GATK software and the obtained variants were contrasted with the coordinates of the variants found in the present study. The pipeline employed GATK functions MarkDuplicates, HaplotypeCaller, SelectVariants, and VariantFiltration, similarly as described previously. Finally, VariantRecalibrator and ApplyVQSR were employed with known variants from the dbSNP version 138, 1000 Genomes phase 1, 1000 Genomes OMNI 2.5, HapMap 3.3, Mills gold-standard, and Axiom Exome Plus. Data 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 subtypes, including high-risk HPV strains. Data was uploaded at European Nucleotide Archive, https://www.ebi.ac.uk/ena under the accession number PRJEB42475.

The relationship between the number of variants per sample and per group was analyzed using the chi-squared (X²) and Fisher's exact tests. The correlation between the total number of variants and the degree of dysplasia was determined by Spearman's test. All statistical calculations were performed using GraphPad Prism 6.03 (San Diego, CA, USA) and significance was established at p < 0.05.

Results

The clinical and histopathological diagnostic data of the samples are systematized in Table 1 and Fig. 1. 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 (SNV). There were 1,069 variants that showed the highest probability of causing changes with a low, moderate and high impact on amino acid structures. These variants were responsible for changes in 773 genes (Table S1), 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 of the HGD group exhibited the largest number of SNV \( (p < 0.0029^*) \) and indels \( (p = 0.77) \). On the other hand, samples 2 and 5 of the LGD group had the smallest number of SNV \( (p > 0.05) \) and indels \( (p > 0.05) \), respectively. The mean total number of SNV \( (\text{HGD} = 50.1714; \text{LGD} = 372.560) \) and indels \( (\text{HGD} = 52.736; \text{LGD} = 43.083) \) was higher in the HGD group than in the LGD group \( (p = 0.0003^* \text{ and } p = 0.40, \text{ respectively}) \).

Among the 773 altered genes in the OED samples, the molecular functions clustering the largest number of genes evaluated are binding and catalytic activities, with 196 (25%) and 171 (22%) genes, respectively. These genes participate in more than 60 signaling pathways; however, the pathways clustering a larger number of genes are the Wnt and integrin signaling pathways, involving 15 and 14 genes, respectively.

It was estimated that 41 variants had a high impact on the amino acid structure of 38 genes. A list of these variants with detailed information is given in Table S2, highlighting six variants that have not been described. Among the variants with a high impact, 20 (48.9%) were SNV and 19 (46%) include a frameshift variant as the calculated functional consequence. Table 2 summarizes all high-impact variants per sample. The mean number of variants was higher in HGD samples than in the LGD group \( (p > 0.05) \).

Figure S1 shows the genes harboring variants with a high impact on amino acids, chromosome location, samples affected, and genotype. Among all genes identified, 18.4% were detected exclusively in LGD samples, 44.7% only in HGD samples and 36.8% altered genes were identified both samples (Figure S2).

In order to exclude other putative germ-line variants, the normal skin series (HIPSCI) database was accessed \(^{32}\), remaining 22 genes harboring 24 variants (Table S2). In addition to the HIPSCI database, PRJEB24629 series \(^{33}\) was also analyzed remaining 13 genes harboring 15 variants exclusive from our samples of OED (Table 3).

Finally, taking into account described mutated genes in head and neck malignant tumors, Table 4 shows the shared genes identified between the present study with mutated genes in all types of head and neck cancer, head and neck squamous cell carcinoma (HNSSC), and those reported in OSCC.

**Discussion**

Considering the paucity of publications in the literature applying NGS to OED samples, the present study contributes to the description of genomic alterations in the exome of 10 samples from Chilean and Brazilian patients with leukoplaia associated with low- and high-grade OED. In view of the difficulty in obtaining viable tissues and of the small size of PMOD samples for whole genome or exome analysis \(^{10,11,38}\), previous studies employing this technology used OSCC samples and selected OED areas adjacent to the tumor \(^{11,38}\). Since the genomic profile of PMODs is unknown, it is possible that tissues containing OED areas extracted from OSCC samples may exhibit genetic alterations that are not typical of dysplasia lesions. Within this context, despite the difficulties in obtaining samples with sufficient quality and quantity to perform NGS, the present study included samples from patients clinically diagnosed with different types of leukoplaia that were compatible with low- and high-grade OED. The advantage of the present study is the good representativeness in the correlation of the genomic data found in OED with clinical features.

Most variants identified in the present study were SNV, in agreement with the literature which describes this alteration as the most common in whole genome or exome analysis \(^{39}\). Although indel variants were less frequent than SNV, the former are in general extremely important in NGS since they are implicated in many constitutional and oncological diseases \(^{40}\). The SNV and indel data for each sample were filtered using databases of already described human genetic variants in order to remove all known germinal variants. The highest significant average of SNV was observed in the HGD group and this result is supported by studies showing that pre-cancerous lesions are characterized by progressive changes in the DNA sequence, gene expression and protein structures, as well as by microscopic rearrangements \(^{4,5,41,42}\). In addition, a previous studies also reported a smaller number of mutations in LGD sample when compared to HGD and OSCC \(^{11}\).

Most of the genes identified in the present study, with a high impact on the 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. It is therefore an interesting target to elaborate treatment strategies for this cancer \(^{43,44}\). Integrins, the main components of cell adhesion, have been implicated in almost all stages of cancer progression from the development of the primary tumor to metastasis \(^{45}\).

We identified six new variants, including three SNV with functional consequences at splice acceptor and splice donor sites and three deletions that lead to changes in the reading frame. Harboring these variants, the GAREM1, GIPC1 and LRR37A2 genes are associated with mutations described in HNSCC and OSCC \(^{46–49}\).
Like HGD samples that accumulated a larger number of total variants, the same trend was observed when only high-impact variants were considered; however, this association was not significant. This finding agrees with a previous study that reported a smaller number of mutations in LGD samples compared to HGD. Regarding the clinical characteristics of the patients, it was not possible to establish a relationship with the variants found since the sample size was too small for this type of correlation. The correlation between clinical characteristics and genomic variations has not yet been established in the literature.

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. The CELA1 gene, 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 was observed in cultured human primary keratinocytes.

It has long been known that the distribution of mutations in the genome is not completely random. In the present study, the observation of variants that affect CELA1 in the same group of samples may be explained by the phenomenon of mutation showers that is not yet fully understood. This phenomenon is characterized by the simultaneous presence of multiple mutations in the same gene or in small regions of the chromosomes. There are still not many studies that explain or associate these alterations with cancer; however, analysis of available mutation catalogs revealed clustered mutagenesis in multiple myeloma and prostate and head and neck tumors.

Comparing the most severely affected genes identified here with the mutated genes reported in the study of Wood et al., although they are different variants, a match was found with the mutated WNK7 gene only in OED samples, with the mutated MCF2L gene in OED and OSCC samples, and with the mutated LAMAS, FARP1 and SHANK2 genes exclusively in OSCC samples. The observation of this coincidence in only two mutated genes in OED might be explained by the fact that, contrary to the present study that used clinically and histopathologically representative samples, Wood et al. extracted OED areas from OSCC samples, which may increase the probability of molecular differences. In fact, although reporting fewer mutations in OED samples than in OSCC samples, in that study most of the mutations detected in OED were also observed in OSCC.

In the present study, no normal paired controls were available for WES but we used bioinformatics methods to remove false-positive variants. In order to address similarities and differences with normal epithelial tissue, the HIPSCI genome database was analyzed and we could remove variants found to be germ-line that were not filtered within other public genetic repository. It is hard to explain how this variants now found in normal skin were not previously found after filtering using tools such as 1000 genomes, cosmic, dbSNP, etc. O’Huallachain et al. confirmed the presence of a high number of variation in somatic tissue. This can be, in part explained by the theory that choosing the relevant tissue for comparison of genomic profiles might influence data analysis.

It is also important to understand the history and diversification of human populations in the southern tip of the Americas. South American population has an unique genetic conformation composed by pre-Columbian and post colonization. This heterogeneity could play an important role in explaining the number of variants found in this study after performing variant calling based on international databases, including the HIPSCI normal skin genomes. To address this point we have used the only available genome profiles of native Americans representing the pre-Columbian southerners. Interestingly, 38% of variants not described either in public genetic repositories or in the normal skin database were found in the native American genomes and those could be considered, thus, germ-line. It also important to note that these “southern variants” are localized within the same mutated genes referred in previous studies of OED, such as Wood et al. Also, a few genes are considered as harboring a high malignant potential.

It was also important to evaluate the role of this 13 OED genes in malignization. Like The Cancer Genome Atlas (TCGA) that was established for consultations on the genomic diversity of various types of cancer, the Pre-Cancer Genome Atlas (PCGA) project was started in 2016. However, mutated genes for any of the lesions that precede different types of cancer, including OED, are not yet available on platforms. Based on TCGA, we observed that 62% of 13 OED genes pinpointed in our study were also found in HNSCC. Among the shared genes, eight were not identified in OSCC.

It is important to mention that 9 of the 11 altered genes identified in more than half of the samples of this study are also mutated in HNSCC and 6 in OSCC. The study of Wood et al. adds SHANK2 and FARP1, mutated in OSCC, and MCF2L, mutated in OSCC and OED, which were also altered in the present samples and in other studies on HNSCC and OSCC. Similarities in the genomic profile of OED and cancer have been described for intestinal, breast, brain, kidney, lung and skin epithelium, showing that the mutational process can cause the clonal evolution from normal to neoplastic cells. However, Wood et al. observed subclonal heterogeneity with OSCC in five OED samples and suggested that mutational changes in stages prior to cancer do not predict the onset of invasion.

Already in 2009, a study demonstrated completely different genomic profiles of OED that progress to OSCC compared to other OED that do not progress to cancer despite histological similarities. Despite this observation, 10 years after these discoveries, the 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 OED progress to OSCC and this information cannot be obtained by histopathological analysis. Given this current scenario, our study describes 13 genes harbouring 15 variants, providing relevant information on the genomic characterization of OED. Despite the small number of samples, the use of a sample comprising a heterogeneous population and of an in-depth genomic evaluation method, which currently has an extremely low error rate, allowed us to establish highly reliable results. On the other hand, it is important to complement the main data found in prospective multicenter studies using a larger number of samples and including validations and healthy controls.

Declarations

Conflict of interest statement: The authors declare that they have no competing interests.

Ethical approval: All procedures performed in this study were in accordance with the ethical standards of the University of Chile (Approval No. 2014/29) committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The patients weren’t direct involved in this study.

Funding obtained:

Ricardo Fernandez-Ramires: FONDECYT nº 11140281

References

1. El-Naggar, A. K., Chan, J. K. C., Grandis, J. R., Takata, T. & Slootweg, P. J. *WHO Classification of Head and Neck Tumours.* (IARC, 2017).
2. Bradley, G. *et al.* Abnormal DNA content in oral epithelial dysplasia is associated with increased risk of progression to carcinoma. *Br J Cancer* **6**, 822–31 (2010).
3. Graveland, A. P. *et al.* Molecular screening of oral precancer. *Oral Oncol.* **49**, 1129–1135 (2013).
4. Campbell, J. D. *et al.* The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer Prev. Res.* **9**, 119–124 (2016).
5. Spira, A. *et al.* Precancer Atlas to Drive Precision Prevention Trials. *Cancer Res* **77**, 1510–1542 (2017).
6. Farah, C. S. *et al.* Exome sequencing of oral leukoplakia and oral squamous cell carcinoma implicates DNA damage repair gene defects in malignant transformation. *Oral Oncol.* **96**, 42–50 (2019).
7. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Proc. Int. Conf. Intellect. Capital, Knowl. Manag. Organ. Learn.* **20**, 1297–1303 (2010).
8. Andrews, S. A Quality Control tool for High Throughput Sequence Data. *Disponivel em http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2010).
9. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints* (2013).
10. Garnis, C. *et al.* Genomic imbalances in precancerous tissues signal oral cancer risk. *Mol. Cancer* **8**, 1–7 (2009).
11. Wood, H. M. *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.* **9**, 1–14 (2017).
12. Kujan, O. *et al.* Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. *Oral Oncol.* **42**, 987–993 (2006).
13. van der Waal, I. Potentially malignant disorders of the oral and oropharyngeal mucosa; terminology, classification and present concepts of management. *Oral Oncol.* **45**, 317–323 (2009).
14. Van der Auwera, G. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit bestpractices pipeline. *Curr Protoc Bioinforma.* **43**, 1–33 (2013).
15. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
16. Consortium., I. H. The International HapMap Project. *Nature* **426**, 789–96 (2003).
17. 1000 Genomes Project Consortium, Abecasis, G. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
18. Sherry, S. *et al.* dbSNP: the ncbi database of genetic variation. *Nucleic Acids Res.* **29**, 308–11 (2001).
19. Cingolani, P. *et al.* Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Front Genet* **3**, 35 (2012).
20. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol* **17**, 122 (2016).
21. Sudmant, P. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
22. Forbes, S. *et al.* Cosmic: somatic cancer genetics at highresolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
23. Landrum, M. *et al.* Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862-8 (2016).
24. NHLBI Exome Sequencing Project (ESP). Exome Variant Server. *Disponivel em http://evs.gs.washington.edu/EVS/* (2017).
25. Stenson, P. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577–81 (2003).
26. Harrow, J. *et al.* ENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–74 (2012).
27. Aken, B. *et al.* The ensembl gene annotation system. *Database (Oxford)* **2016**, pilt, (2016).
28. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–91 (2016).
29. Adzhubei, I. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 48–9 (2010).
30. Ng, P. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–74 (2001).
31. 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* **88**, 440–9 (2011).
32. PRJEB11751. No Title. omics_ena_project, VI; https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJE (2015).
33. de la Fuente, C. *et al.* Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4006–E4012 (2018).
34. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* **14**, 703–721 (2019).
35. Szkarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380-4 (2016).
36. López-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M. & Gelpí, J. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res.* **45**, W222–W228 (2017).
37. Chandrani, P. *et al.* NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br. J. Cancer* **112**, 1958–1965 (2015).
38. Wood, H. M. *et al.* The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. *J. Pathol.* **237**, 296–306 (2015).
39. Shigemizu, D. *et al.* A practical method to detect SNVs and indels from whole genome and exome sequencing data. *Sci. Rep.* **3**, 1–6 (2013).
40. Sehn, J. *Insertions and Deletions (Indels).* (2015).
41. Hu, T. *et al.* Forward and reverse mutations in stages of cancer development. *Hum. Genomics* **12**, 40 (2018).
42. Srivastava, S., Ghosh, S., Kagan, J. & Mazurchuk, R. The Making of a PreCancer Atlas: Promises, Challenges, and Opportunities. *Trends Cancer* **4**, 523–536 (2018).
43. Shiah, S., Shieh, Y. & Chang, J. The Role of Wnt Signaling in Squamous Cell Carcinoma. *J Dent. Res* **95**, 129–34 (2016).
44. González-Moles, M., Ruiz-Avila, I., Gil-Montoya, J., Plaza-Campillo, J. & Scully, C. Beta-catenin in oral cancer: an update on current knowledge. *Oral Oncol* **50**, 818–824 (2014).
45. Hamidi, H. & Ivaska, J. Every step of the way: integrins in cancer progression and metastasis. *Nat Rev Cancer* **18**, 533–548 (2018).
46. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science (80-. ).* **333**, 1154–7 (2011).
47. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science (80-. ).* **333**, 1157–60 (2011).
48. Pickering, C. *et al.* Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discov* **3**, 770–81 (2013).
49. Cancer Genome Atlas Network, C. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–82 (2015).
50. Talas, U., Dunlop, J., Khalaf, S., Leigh, I. M. & Kelsell, D. P. Human elastase 1: Evidence for expression in the skin and the identification of a frequent frameshift polymorphism. *J. Invest. Dermatol.* **114**, 165–170 (2000).
51. Chan, K. & Gordenin, D. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu Rev Genet* **49**, 243–267 (2015).
52. O’Huallachain, M., Karczewski, K. J., Weissman, S. M., Urban, A. E. & Snyder, M. P. Extensive genetic variation in somatic human tissues. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18018–18023 (2012).

**Tables**

**Table 1.** Summary of patients’ characteristics, histopathological diagnosis and total variants per sample.
| ID | Age (y/o) | Gender | Smoking | Alcohol | Lesion Site | Clinic Dx | HistopDx | Country | Biopsy type | SNV Total | INDEL 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   | 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; y/o., Years old; 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, insertions and deletions variants.

**Table 2** Total variants classified, according to VEP[1], as high impact (n = 41) per sample.

| ID Sample | Histop Dx | Var. Het. | Var. Hom. | Total Var. |
|-----------|-----------|-----------|-----------|------------|
| 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         |
| X         | LGD       | 6         | 10        | 16         |
| 7         | HGD       | 7         | 11        | 18         |
| 8         | HGD       | 15        | 6         | 21         |
| 9         | HGD       | 11        | 17        | 28         |
| 10        | HGD       | 13        | 13        | 26         |

| X         | HGD       | 11        | 12        | 23         |

Histop Dx, Histopathological Diagnosis; Var., Variant; Het., Heterozygous; Hom., Homozygous; LGD, Low Grade Dysplasia; HGD, High Grade Dysplasia.

**Table 3** Thirteen genes harboring 15 variants attributed to oral dysplasia samples
| Var. | Genes   | Position | Type   | REF-ALT | HGVS<sub>c</sub> | HGVS<sub>p</sub> | ID       |
|------|---------|----------|--------|---------|------------------|------------------|----------|
| 1    | DRAM2   | Chr1:111,119,814 | SNV   | C→T | ENST00000477769.1:n.235+1G>A | _ | rs644081 |
| 2    | C4orf36 | Chr4:86,892,061 | SNV   | A→T | ENST00000506308.5:c.-74+2T>A | _ | rs10034336 |
| 3    | FAM198B | Chr4:158,171,600 | SNV   | C→A | ENST00000296530.12:c.-224-1G>T | _ | ND |
| 4    | SEPT14P1 | Chr7:63,148,830 | DEL.  | TG→T | ENST000000458703.1:n.76-2del | _ | rs57581559 |
| 5    | CBWD5   | Chr9:65,733,001 | SNV   | A→T | ENST00000377392.9:c.308+1G>C | _ | ND |
| 6    | CELA1   | Chr12:51,346,623 | SNV   | C→A | ENST00000293636.1:c.-224-1G>T | _ | ND |
| 7    | CELA1   | Chr12:51,346,631 | INS   | A→AG | ENST00000293636.1:c.7+1G>del | _ | ND |
| 8    | CELA1   | Chr12:51,346,632 | INS   | C→G | ENST00000293636.1:c.6+1del | _ | ND |
| 9    | PKD1L3  | Chr16:71,947,511 | DEL.  | CT→T | ENST000000480090.1:n.1232+1T>del | _ | ND |
| 10   | LRRC37A2 | Chr17:46,512,902 | DEL.  | CT→T | ENST00000333342.3:n.192del | _ | ND |
| 11   | GAREM1  | Chr18:32,136,366 | SNV   | A→G | ENST00000583696.1:c.*48+2T>del | _ | ND |
| 12   | PCSK4   | Chr19:1,490,421 | SNV   | A→C | ENST00000293636.1:c.7+1G>del | _ | ND |
| 13   | GIPC1   | Chr19:14,483,329 | SNV   | A→G | ENST00000293636.1:c.12_15del | _ | ND |
| 14   | RPL13A  | Chr19:49,490,126 | SNV   | G→T | ENST00000293636.1:c.12_15del | _ | ND |
| 15   | ZNF83   | Chr19:52,616,816 | SNV   | C→T | ENST00000293636.1:c.12_15del | _ | ND |

Table 4: Similarities of genes with variants identified in OED samples from the present study (n = 17 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 COCE* |
|----------------------------------------|-----------------------|------------------------|------------------------|
| C4orf36                                | X                     | X                      |                        |
| CBWD5                                  | X                     |                        |                        |
| DRAM2                                  | X                     |                        |                        |
| FAM198B                                | X                     |                        | X                      |
| GAREM1                                 | X                     | X                      |                        |
| GIPC1                                  | X                     |                        | X                      |
| LRRC37A2                               | X                     | X                      |                        |
| PCSK4                                  | X                     |                        |                        |
| PKD1L3                                 | X                     |                        |                        |
| RPL13A                                 | X                     |                        |                        |
| SEPT14P1                               | X                     |                        |                        |
| ZNF83                                  | 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 [2]
Figure 1

Representative clinical and histopathological figures. A Homogeneous leukoplakia (ID 1); B Leukoerythroplasia (ID 7); C Histopathological image LGD (ID 1); D Histopathological image of HGD (ID 9).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.jpg
- FigureS2.jpg
- TableS1Jan14.docx
- TableS2Jan14.docx