RESEARCH ARTICLE

OrCanome: a Comprehensive Resource for Oral Cancer

Deeksha Bhartiya[1], Amit Kumar[1], Harpreet Singh[2], Amitesh Sharma[2], Anita Kaushik[3], Suchitra Kumari[1], Ravi Mehrotra[4]*

Abstract

Oral cancer is one of the most prevalent cancers in India but the underlying mechanisms are minimally unraveled. Cancer research has immensely benefited from genome scale high throughput studies which have contributed to expanding the volume of data. Such datasets also exist for oral cancer genes but there has been no consolidated approach to integrate the data to reveal meaningful biological information. OrCanome is one of the largest and comprehensive, user-friendly databases of oral cancer. It features a compilation of over 900 genes dysregulated in oral cancer and provides detailed annotations of the genes, transcripts and proteins along with additional information encompassing expression, inhibitors, epitopes and pathways. The resource has been envisioned as a one-stop solution for genomic, transcriptomic and proteomic annotation of these genes and the integrated approach will facilitate the identification of potential biomarkers and therapeutic targets.

Keywords: Oral cancer - resource - annotation - genes - biomarkers

Asian Pac J Cancer Prev, **17 (3)**,

Introduction

The incidence of oral and oropharangeal cancers is globally increasing at a very alarming rate. The 2012 report of Globocan shows 1,98,975 cases of incidence of cancer of lip and oral cavity across the world and has a very low survival rate (around 50%) (Ferlay et al., 2015). It is the eleventh most common cancer across the world and is found to be more prevalent in men than in women. Oral cancer, a subset of Head and Neck cancer, is the most prevalent cancers in males in India (Coelho, 2012). A higher prevalence of oral cancer in Indian subcontinent can be associated to the higher exposure of risk factors such as tobacco especially the smokeless form (Mehrotra and Yadav, 2006). Advanced screening and early detection approaches are being followed across the world including India to detect pre-cancerous lesions (Mehrotra and Gupta, 2011; Rajaraman et al., 2015).

Since oral cancer poses a high disease burden globally, studies which unravel the underlying mechanistic insights have started emerging. Over the last few years, studies have been reported which have explored the genic associations to understand the molecular mechanisms. Of late, high-throughput genome-scale studies have also been reported uncovering the mutational landscape of the cancer (India Project Team of the International Cancer Genome, 2013). Apart from this, gene expression analysis involving microarrays have also been reported (Reis et al., 2011; Saeed et al., 2015). Such high throughput studies elevate the volume of the data, thus systematic collection and annotation of the relevant biological information become inevitable.

There have been insufficient efforts in accumulating the data associated with oral cancer. Gadewal and Zingde, 2011 reported the Oral Cancer Gene Database with a total of 374 genes (Gadewal and Zingde, 2011). Mitra et al., 2012 reported Head and Neck and Oral Cancer Database (HNOCDB) with a total of 415 genes (Mitra et al., 2012). These datasets have been compiled on the evidences based on the available literature. Though these databases cover a number of genes involved in oral cancer, a larger proportion of genes studied in oral cancer are completely missing. Moreover, these databases focus mostly on the genomic aspects of these genes and there is no dataset focusing on the various transcriptomic and the proteomic aspects. There is a scarcity of such resources comprising of integrated genomic, transcriptomic and proteomic information of the genes involved in oral cancer.

Targeted gene approaches have picked up a fewer number of genes as compared to the genome scale studies and the available databases have thus missed out a considerable proportion of the genes. With an increase in the abundance of the high-throughput studies, there has been an explosion in datasets (Reuter et al., 2015). Identification of drug targets and chemical inhibitors for these genes becomes highly laborious as these datasets are scattered across different repositories (Thariat et al., 2015). With a growing corpus of genome, proteome and chemical information on oral cancer, there is an urgent need to develop a consolidated database. It has become imperative

*[1]Biomedical Informatics Center, Institute of Cytology and Preventive Oncology, Noida, Uttar Pradesh, [2]Bioinformatics Center, Indian Council of Medical Research, New Delhi, [3]Centre for Bioinformatics, Maharshi Dayanand University, Rohtak, [4]Institute of Cytology and Preventive Oncology, Noida, Uttar Pradesh, India *For correspondence: directoricpo@icmr.org.in, rmehrotra@icmr.org.in*

to utilize an integrated genomic-transcriptomic-proteomic approach to assemble the datasets in a systematic and unified way in order to achieve a better annotation of these genes. A unified repository will enhance the annotation and provide possible biomarkers and therapeutic targets which will aid both the research and clinical community (Nakagawa et al., 2015).

OrCanome has been designed to benefit the researchers as well as clinicians to study the genes involved in oral cancer. At present, it houses 922 genes found to be dysregulated in oral cancer as a start-up resource for biomarker and therapeutic drug target discovery provides their appropriate annotation starting from genome, expression, proteome, pathway, immunological data, active compounds pertinent to oral cancer.

Materials and Methods

Genomic data

Retrieval of genes dysregulated in oral cancer: A systematic search was performed at Gene Expression Omnibus to fetch the expression profiling studies on oral cancer in humans (Barrett et al., 2013). The genes dysregulated in oral cancer were obtained from these studies. Since the gene names were present in different formats in all the studies, Ensembl Gene IDs were fetched for all genes using BioMart from Ensembl Genome Browser (Flicek et al., 2014). The unique gene IDs were used for downstream analysis. Along with these genes, literature mining was performed to include other genes.

Genic information

The genic annotations were fetched from Gencode release 21 (Harrow et al., 2012). This included information regarding the genomic location, strand and gene biotypes as recommended by Gencode.

Transcriptomic Data

Transcripts: The information of the different alternate transcripts of these genes has been obtained from Gencode along with their corresponding genomic location and transcript biotypes.

Expression datasets: The baseline and differential expression of these genes was obtained from Expression Atlas (EMBL-EBI) (Petryszak et al., 2014). Expression atlas provides expression data from functional studies of microarray and RNA sequencing. The baseline expression dataset obtained contains expression levels of each gene in various human tissues from six experiments. The differential expression dataset provides the comparison statistics of the expression levels of the genes in different conditions in various studies.

Pathways and Gene Ontology

The pathways and the gene ontologies corresponding to the genes were fetched using the functional annotation tool of The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Huang da et al., 2009a; Huang da et al., 2009b).

Proteomic Data

Protein Information: The protein IDs for the genes was fetched using Gencode. The corresponding protein description, sequences, PubMed entry were obtained from UniProt release 2015_08 (Consortium, 2015). The information regarding the 3-D structures and targets were extracted from Protein Data Bank (PDB) and ChEMBL (Berman et al., 2002; Bento et al., 2014).

Inhibitors: Chemical compounds showing inhibition against proteins are listed in OrCanome. The inhibitors have been incorporated only if they have been assayed for IC_{50} values for target proteins. The one dimensional structure in the form of SMILES (http://www.daylight.com/smiles/) was generated using Open Babel software and the bioassay data were obtained from the chemical databases such as ChEMBL and BindingDB (Liu et al., 2007; O'Boyle et al., 2011). HTML toolkit from ChemDoodle was integrated for the visualization of chemical compounds (Burger, 2015).

Epitope: One of the most important features of this resource is that we have provided the linear B-cell epitopes predicted using EPMLR web server (Lian et al., 2014). The epitopes have been screened on the basis of their scores and only the ones with values greater than 0.9 have been retained.

Transmembrane Helices: The information about the experimentally validated and predicted transmembrane helices of the proteins was fetched from UniProt release 2015_08. The transmembrane helices were also predicted using TMHMM server 2.0 (Krogh et al., 2001).

Secretory proteins: The information regarding the secretory proteins was obtained from Human Protein Atlas (Berglund et al., 2008; Uhlen et al., 2010; Uhlen et al., 2015).

Literature annotations: A comprehensive literature survey was performed for these genes and proteins and the relevant studies have been provided.

Results

Database architecture

OrCanome is designed to meet the needs of the researchers and clinicians by providing a comprehensive, user-friendly and searchable resource. The database has been developed in MySQL and the interface has been designed using core PhP and HTML. Every gene found to be associated with oral cancer is provided with a separate page with its biologically relevant information on genomic, transcriptomic and proteomic aspects (Figure 1). OrCanome currently houses 922 genes which were found to be associated with oral cancer. These genes fall into different Gencode biotypes such as protein coding genes, antisense, microRNAs, long non-coding RNAs, small nucleolar RNAs and miscellaneous RNAs (misc RNAs).

The biological information at OrCanome is featured as different categories on every gene page. The statistics of the datasets in different categories have been provided in Table 1. The first category includes the basic genomic information which includes the genomic location, strand, Gencode gene ID, and the gene biotype. The second category comprises of the transcript information including the different transcript isoforms along with

their genomic coordinates and biotypes. This is followed by the next two categories of the available datasets of the baseline and differential expression of these genes. The baseline expression data is present for 901 genes while the differential expression is present for 474 genes.

The fifth category includes the details about the gene ontologies and pathways for the corresponding genes and proteins. This includes the three different gene ontologies, the biological process, molecular function and cellular component. The respective InterPro, KEGG pathways and OMIM diseases have also been included (Kanehisa and Goto, 2000; Hamosh et al., 2005; Hunter et al., 2009).

The sixth category comprises of the annotation of the protein products which includes information of 890 proteins along with their name and sequence followed by external links to their corresponding UniProt entries, PDB structures and detailed information from ChEMBL. If the protein is a secretory protein, the information is also included for 183 proteins. External links to the respective databases have also been provided. The seventh category includes the details of the transmembrane helices. This includes the information about the 682 experimentally validated and 672 predicted transmembrane helices in these proteins. The different helices in the protein with their location on the protein sequence have been provided.

The details about the chemical compounds showing inhibitions to the respective proteins have been provided in the eighth category. The respective inhibitor compound ID and compound name have been extracted from ChEMBL. SMILES representing line notation which encodes molecular structures have also been generated and provided. It also includes the respective half maximal inhibitory concentration (IC_{50}) representing the quantitative measure of amount of the inhibitor required to block the protein or function by half. The corresponding literature report has also been provided. 5534 compounds having IC_{50} from 26 targets proteins have been incorporated in the database. The chemical structure of the compounds can also be visualized to identify Hydrogen bond donor/acceptor to aid in drug designing.

The ninth category consists of the predicted linear B-cell epitopes of the corresponding proteins. OrCanome houses epitopes for 207 proteins and provides the rank for each predicted epitope with their score, epitope sequence and the location on the protein. The tenth category includes the literature annotation of the association with oral cancer.

Table 1. Statistics of the Datasets Available at OrCanome

S. No	Dataset	Entries
1	Genes	922
2	Transcript Isoforms	6995
3	Baseline expression data	901
4	Differential expression data	474
5	(a) Gene Ontologies (cellular component)	697
	(b) Gene Ontologies (molecular function)	724
	(c) Gene Ontologies (biological process)	757
6	Pathways	358
7	Proteins	890
	(a) Secretory Proteins	183
	(b) Drug Targets	24
8	Transmembrane Helix	682 (experimental) 682 (predicted)
9	Compounds	5534
10	B-cell epitopes	207

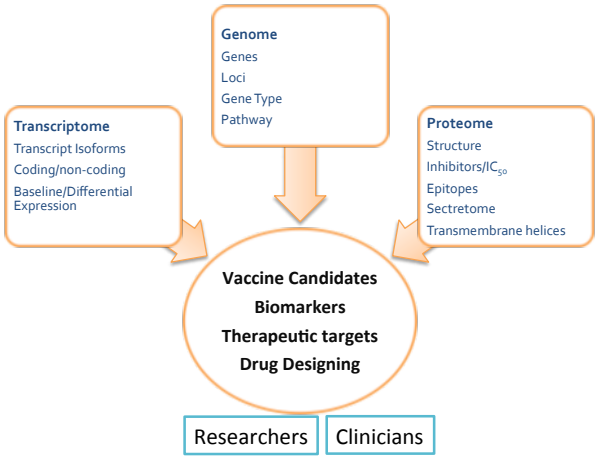

Figure 1. Conceptual framework depicting integrated genomic-transcriptomic-proteomic datasets in OrCanome

Database Access

OrCanome provides user friendly, comprehensive search and browse options. The database is searchable through text/keyword such as gene name, ID, strand, type, transcript ID, inhibitors, epitopes and pathways. The datasets can also be browsed on the basis of the chromosome number, gene ID, gene name and gene type. Advanced search options have also been provided which enable the user to compare different datasets. This facilitates the identification of available inhibitors or epitopes for a gene/protein of interest or will provide expression statistics for the gene. This will aid in identifying potential biomarkers, vaccine candidates or therapeutic targets for oral cancer.

Discussion

The emergence of high throughput genome scale studies has provided us enormous data which needs systematic collection and annotation for identification of biomarkers and therapeutic targets in diseases including cancer. OrCanome aims to serve this purpose by providing genes dysregulated in oral cancer along with their annotations. It is one of the most comprehensive resources for oral cancer genes which provides a one stop solution for genomic, transcriptomic and proteomic information of these genes. OrCanome also provides links to external databases to enhance the interoperability of datasets. The data is organized in ten different categories and the advanced search option enables the user to explore the resource using specific queries.

OrCanome provides the starting point for in depth analysis for the genes associated with oral cancer. Since newer studies related to oral cancer will be emerging and provide better datasets, we aim to include those in our resource on a regular basis. We intend to serve the purposes of both research and clinical communities with

our resource and wish to collaborate with them for data generation and sharing.

Acknowledgements

This work was supported by the project "Second phase of task Force Biomedical Informatics Centers of Indian Council of Medical Research (ICMR)" [grant number BIC/ADHOC/7/2012-13].

References

Barrett T, Wilhite SE, Ledoux P, et al (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res, 41, D991-5.

Bento AP, Gaulton A, Hersey A, et al (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res*, **42**, 1083-90.

Berglund L, Bjorling E, Oksvold P, et al (2008). A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol Cell Proteomics*, **7**, 2019-27.

Berman HM, Battistuz T, Bhat TN, et al (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, **58**, 899-907.

Burger MC (2015). ChemDoodle Web Components: HTML5 toolkit for chemical graphics, interfaces, and informatics. *J Cheminform*, **7**, 35.

Coelho KR (2012). Challenges of the oral cancer burden in India. *J Cancer Epidemiol*, **2012**, 701932.

Ferlay J, Soerjomataram I, Dikshit R, et al (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, **136**, 359-86.

Flicek P, Amode MR, Barrell D, et al (2014). Ensembl 2014. *Nucleic Acids Res*, **42**, 749-55.

Gadewal NS, Zingde SM (2011). Database and interaction network of genes involved in oral cancer: Version II. *Bioinformation*, **6**, 169-70.

Hamosh A, Scott AF, Amberger JS, et al (2005). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, **33**, 514-7.

Harrow J, Frankish A, Gonzalez JM, et al (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, **22**, 1760-74.

Huang da W, Sherman BT, Lempicki RA (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, **37**, 1-13.

Huang da W, Sherman BT, Lempicki RA (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.

Hunter S, Apweiler R, Attwood TK, et al (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res*, **37**, 211-5.

India Project Team of the International Cancer Genome Consortium (2013). Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. *Nat Commun*, **4**, 2873.

Kanehisa M, Goto S (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27-30.

Krogh A, Larsson B, von Heijne G, et al (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-80.

Lian Y, Ge M, Pan XM (2014). EPMLR: sequence-based linear B-cell epitope prediction method using multiple linear regression. *BMC Bioinformatics*, **15**, 414.

Liu T, Lin Y, Wen X, et al (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*, **35**, 198-201.

Mehrotra R, Gupta DK (2011). Exciting new advances in oral cancer diagnosis: avenues to early detection. *Head Neck Oncol*, **3**, 33.

Mehrotra R, Yadav S (2006). Oral squamous cell carcinoma: etiology, pathogenesis and prognostic value of genomic alterations. *Indian J Cancer*, **43**, 60-6.

Mitra S, Das S, Das S, et al (2012). HNOCDB: a comprehensive database of genes and miRNAs relevant to head and neck and oral cancer. *Oral Oncol*, **48**, 117-9.

Nakagawa H, Wardell CP, Furuta M, et al (2015). Cancer whole-genome sequencing: present and future. *Oncogene*.

O'Boyle NM, Banck M, James CA, et al (2011). Open Babel: An open chemical toolbox. *J Cheminform*, **3**, 33.

Petryszak R, Burdett T, Fiorelli B, et al (2014). Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res*, **42**, 926-32.

Rajaraman P, Anderson BO, Basu P, et al (2015). Recommendations for screening and early detection of common cancers in India. *Lancet Oncol*, **16**, 352-61.

Reis PP, Waldron L, Perez-Ordonez B, et al (2011). A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC Cancer*, **11**, 437.

Reuter JA, Spacek DV, Snyder MP (2015). High-throughput sequencing technologies. *Mol Cell*, **58**, 586-97.

Saeed AA, Sims AH, Prime SS, et al (2015). Gene expression profiling reveals biological pathways responsible for phenotypic heterogeneity between UK and Sri Lankan oral squamous cell carcinomas. *Oral Oncol*, **51**, 237-46.

Thariat J, Vignot S, Lapierre A, et al (2015). Integrating genomics in head and neck cancer treatment: promises and pitfalls. *Crit Rev Oncol Hematol*, **95**, 397-406.

Uhlen M, Fagerberg L, Hallstrom BM, et al (2015). Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

Uhlen M, Oksvold P, Fagerberg L, et al (2010). Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*, **28**, 1248-50.

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res*, **43**, 204-12.