CancerSmell:Molecular Basis of Chemosensation in Cancer Cells

By-
Siddhant Kalra
MT18241

Under the supervision of
Dr. Gaurav Ahuja

Indraprastha Institute of Information Technology Delhi
May 2020

CancerSmell:Molecular Basis of Chemosensation in Cancer Cells

By-
Siddhant Kalra
MT18241

Submitted
in partial fulfillment of the requirements for the degree of
Master of Technology

To
Indraprastha Institute of Information Technology Delhi
May 2020

Certificate

This is to certify that the thesis titled **"CancerSmell:Molecular Basis of Chemosensation in Cancer Cells"** being submitted by Siddhant Kalra to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. Gaurav Ahuja   May 2020

Department of Compuational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi 110 020

Acknowledgment

List of Abbreviations

Abbreviation	Description
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
ARMS	Alveolar rhabdomyosarcoma
Astr	Astrocytoma
BAC	Bronchoalveolar carcinoma
BRCA	Breast cancer
CESC	Cervix cancer
CML	Chronic myelogenous leukemia
CRC	Colorectal cancer
ESCC	Esophageal squamous cell carcinoma
GBM	Glioblastoma
Glioma	Glioma
GSVA	Gene set variation analysis
HNSC	Head and neck cancer
LUAD	Lung adenocarcinoma
MRCLS	Myxoid liposarcoma
NSCLC	Non-small cell lung cancer
OC	Ovarian carcinoma
OD	Oligodendroglioma
ORs	Olfactory receptors
PDAC	Pancreatic ductal adenocarcinoma
PORs	Pseudo olfactory receptors
PRAD	Prostate cancer
RCC	Renal cell carcinoma
SKCM	Melanoma
T2R	Taste receptor type 2
TAARs	Trace amine-associated receptors
TCGA	The Cancer Genome Atlas
V1R	Vomeronasal receptors type 1
V2R	Vomeronasal receptors type 2

List of Figures

	subtypes of breast carcinoma. C) Whisker plot representing the median expression of the breast carcinoma-associated ORs in the indicated conditions	
14	Correlation plot depicting the expression similarities between breast carcinoma-associated ORs.	31
15	UMAPs represent the cellular expression of ORs in breast carcinoma along with the density histograms depicting the normalized expression of the indicated ORs in single-cell breast carcinoma dataset.	32
16	Graphical illustration depicting the number of cells and reliably detected ORs in healthy and malignant breast epithelial cells.	33
17	Percentage bar graph depicting the relative proportion of detected ORs in the indicated healthy and malignant epithelial cells. The Venn diagram on the right depicts the number of overlapping ORs in the indicated conditions.	35
18	Bar graph depicting the correlation between the indicated ORs expression and GSVA scores.	36
19	Schematic representation depicting the strategy employed for differential gene expression analysis, in which the malignant cells were segregated based on median OR expression.	37
20	Heatmap depicting cluster wise enrichment of the prominent biological functions	38
21	(Above) Representative gene ontologies depicting the functional importance of BRCA associated ORs in key biological/molecular processes. (Below) Metascape plot depicting the cluster information of the single-cell breast carcinoma dataset.	39
22	A) Monocle generated pseudo temporal trajectory of malignant breast epithelial cells. Arrowheads indicating the direction of the cellular differentiation. B) UMAP plot representing the number of active OR genes per malignant cell. C) UMAP plot depicting the relative cellular stemness across all malignant breast carcinoma cells. D) UMAP plot depicting the tumor stages of the cells across the pseudo temporal trajectory.	41
23	A) Scatter plot depicting the correlation between cellular stemness and cell differentiation time points (pseudotime). B) Scatter plot depicting the correlation between mean expression of breast carcinoma-associated ORs and cell differentiation time points (pseudotime). C) Scatter plot depicting the correlation between cellular frequencies of expressed ORs in breast carcinoma and cancer cell differentiation time points (pseudo time).	42
24	(Top) Monocle generated pseudo temporal trajectory of malignant breast epithelial cells of luminal A molecular subtype, depicting the decrease in cellular stemness during the cellular differentiation time course. Arrowheads indicating the direction of cellular differentiation. (Bottom) UMAP represents the decrease in the number of active OR genes per malignant cell (luminal A subtype) during the pseudo temporal trajectory.	43

36	Heatmap depicting the cosine similarity between the scRNA sequencing datasets derived OR centric signatures and stratified TCGA patients cohorts.	59
37	(Top) Kaplan-Meier plot depicting patient's survival in the indicated group segregated based on cosine similarity towards OR enrichment signatures, inferred from tissue-derived scRNA sequencing profiling of malignant cells. (Bottom) Comparison of the tumor stages between two prominent clusters representing TCGA patients cohorts with distinct survival.	60
38	(Top) Kaplan-Meier plot depicting patient's survival in the indicated group segregated based on cosine similarity towards OR enrichment signatures, inferred from single-cell transcriptional profiling of circulating tumor cells. (Bottom) Comparison of the tumor stages between two prominent clusters representing TCGA patients cohorts with distinct survival.	61
39	(Top) Kaplan-Meier plot depicting patient's survival in the indicated group segregated based on cosine similarity towards OR enrichment signatures, inferred from transcriptomic profiling of breast carcinoma patient-derived xenograft dataset. (Bottom) Comparison of the tumor stages between two prominent clusters representing TCGA patients cohorts with distinct survival.	62

List of Tables (Main)

Table No.	Legend	Page No
1	The number of Functional and Pseudo Olfactory receptors across different species.	2
2	The table contains information about the datasets used in this study. It also contains information about the total number of receptors expressed in each dataset.	24
3	The table contains information about the different ORs expressed in different types of samples.	34
4	The table contains information about the pseudo time, OR frequency, OR expression, Stemness, and Differentiation for Luminal A population of breast carcinoma dataset.	46

List of Tables (Supplementary)

Table No.	Legend	Access URL
1	The table contains information about the datasets used in this study. It also contains the information about the number and identity of the expressed chemoreceptors (functional ORs, Pseudo ORs, V1Rs, T1Rs, T2Rs and TAARs) identified in these datasets.	URL
2	The table contains information about the functional relevance of the cancer associated ORs. It contains the correlation values for the OR expression and the Gene set variation analysis (GSVA) scores of the indicated tumor related signatures.	URL
3	The table contains information about the upregulated and downregulated genes in different clusters of BRCA.	URL
4	The table contains the information about the parameters used for pseudo temporal analysis by Monocle on breast carcinoma and healthy breast epithelial datasets.	URL

URL=**https://drive.google.com/drive/u/0/folders/1azQ5tq2gQOdzVQhgUyaP2vgpOElyhHEO**

Table of Contents

INTRODUCTION

Olfactory receptors

Olfactory receptors (ORs) are also known as smell receptors, are the seven-transmembrane proteins that constitute the largest gene family in the mammalian genome and are responsible for perceiving the odor molecules[1]. Because of their role in activating heterotrimeric G-proteins, they belong to the family of G-protein coupled receptors (GPCR). Their basic structure of any typical olfactory receptor consists of seven hydrophobic transmembrane domains, a disulfide bond between the conserved cysteines in the extracellular loops, a conserved glycosylation site in the N-terminal region, and several amino acid sequence patterns that are conserved in the OR family[2]. The sequence of amino acids forming these proteins is of crucial importance because it leads to proper folding of ORs in the plasma membrane so that the ORs can act in binding odorants and in coupling with the necessary G-proteins. Olfactory receptor proteins are positioned inside the cell membrane in such a way that one end projects outside the cell and the other end projects within the cell[3,4]. This allows a chemical outside the cell, such as an odorant molecule, to interact with and bring about changes in the cellular machinery without entering the cell.

The olfactory cascade takes place when the molecule of a particular shape fits into a corresponding "pocket" in the receptor molecule. A change in a single amino acid can change the pocket shape and thus alter the chemicals that fit in the pocket[5]. There are a large number of different odor receptors, with as many as 1,000 in the mammalian genome which represents approximately 3% of the genes in the genome[6,7]. However, not all of these potential odor receptor genes are expressed and functional[8]. According to an analysis of data derived from the Human Genome Project, humans have approximately 400 functional genes coding for olfactory receptors, and the remaining 600 candidates are pseudogenes[9]. The number of receptors across different species can be summarized in **Table 1**.

Table 1: The number of Functional and Pseudo Olfactory receptors across different species [9].

Species	Number of intact ORs	Number of pseudogenes
Human	400	600
Chimpanzee	397	693
Dog	859	315
Opossum	613	898
Platypus	350	404

A large number of different olfactory receptors provide a system for discriminating between as many different odors as possible. The single olfactory receptor can be activated by different numbers of similar odorant structures. Furthermore, most odors activate more than one type of odor receptor[10]. Since the number of combinations and permutations of olfactory receptors is very large, the olfactory receptor system is able to recognize and differentiate between a very large number of odorant molecules. Such diversity of OR expression maximizes the capacity of olfaction. Both monoallelic OR expression in one neuron and maximum OR expression variability in the neuron population is important to olfactory sensing precision and sensitivity. Thus, olfactory receptor activation is a dual-objective design problem[11,12].

The mammalian olfactory receptor system consists of hundreds of different types of highly diverse receptors which are expressed in distinct subcompartments of the nose[13]. Based on their structure and topographic distributions, these receptors can be classified into several distinct receptor families such as odorant receptor (ORs), vomeronasal receptors (V1Rs and V2Rs), trace amine-associated receptors (TAARs), formyl peptide receptors (FPRs), and the membrane guanylyl cyclase (GC-D)[14,15].

Odorant receptors can be categorized into three classes: A, B, and C. Their average length is about 320 ± 25 amino acid residues. The difference in their length arises due to variation in N and C terminal residues. ORs are identified based on their conserved motifs such as LHTPMY in the first intracellular loop, MAYDRYVAIC motif present at the end of transmembrane domain 3 (TM3), SY motif present at the end of TM5, FSTCSSH motif present at the beginning of TM6 and PMLNPF which is present in the TM7[16]. The ORs are further classified into Class I and Class II receptors. The Class I receptors are those which are found in aquatic animals like fishes whereas Class II receptors belong to the non-fish species such as mammals, birds, and amphibians[17].

Vomeronasal receptors (VRs) are classified into two major groups: V1Rs and V2Rs. V1Rs belong to class A of GPCRs and have highly diverse sequences. Only TM3 is the conserved domain in them. V1Rs are expressed in the apical region of the Gαi2-positive layer of the VNO[18]. V2Rs belong to class C of GPCRs and have a large N-terminal extracellular domain which is joined to the transmembrane part of the receptor protein via cysteine-rich linker region[9].

Trace amine-associated receptors (TAARs) were the accidentally identified group of GPCRs that have a distinct sequence motif. They are activated by trace amines such as β-phenylethylamine, p-tyramine, tryptamine, and octopamine. Formyl peptide receptors (FPRs) are present on a single chromosome which is the 19th chromosome in humans and 17th chromosome in the mouse[9]. Their coding sequence has a length of about 350 amino acids and is intronless. These receptors are activated by formylated peptides. Membrane Guanylyl cyclase GC-D receptors are another category of receptors that are activated by gas and other ligands like uroguanylin and guanylin[19].

Based on the classification, the olfactory receptors follow the official Human Genome Project (HUGO) for their nomenclature. The nomenclature follows the **OR_nX_m** format where: OR represents the root name, n represents the family, single letter code

denoting the subfamily is placed in the place of X and m is an integer representing an individual family member. For example, OR5A1 is an OR gene of family 5, subfamily A and is the first isoform of this family[9,13].

Olfactory receptors outside non-olfactory tissue

Olfactory receptors are the principal chemosensory systems in many organisms and detect a diverse range of environmental cues important for survival including food, toxins, predators, and prey. Recent literature shows that the ORs express ectopically in non-olfactory tissues. These extra-nasally expressed ORs have different effects on cell biological functions owing to their versatility in activating different molecular and cellular signaling mechanisms[20,21].

<u>Expression in Tissues and Signalling Pathways</u>

The olfactory receptor OLFR544 is reported to express in Mouse White and brown adipose tissue cells where it causes induction of lipolysis[22]. Other reports claim that OR51E1 is expressed in Human cardiomyocytes and is responsible for the reduction of contraction force of the explanted heart. Similarly, OR51E1 and OR11H7 are reported to Increase Ca2+ transients and cAMP levels when expressed in Kidney tissue cells and proximal tubule HK-2 cells of Humans[6]. OR2F1, OR2W1, decreases the serotonin secretion once expressed in pulmonary neuroendocrine cells and human tracheobronchial epithelial cells[23]. The olfactory receptors are known to be expressed in various tissues such as lungs, prostate, skin, adipose, liver, gut, salivary glands[24,25]. They have also been reported to have a role in embryogenesis[20].

<u>Sperm Chemotaxis</u>

Transcriptome analysis has shown that testis is the second most OR-transcript rich tissue apart from the nose[25]. Mouse olfactory receptor OLFR16, when activated by its ligand Lyral, helps in sperm chemotaxis[23,24,26,21]. In humans when olfactory receptor OR1D2 which is expressed in spermatozoa when activated by its ligand Bourgeonal

directs sperm chemotaxis[27,28]. Apart from them two olfactory receptors OR7A5 and OR4D1 control the swimming speed of the sperm[20,29].

Kidney Function and blood pressure regulation

Mouse olfactory receptor OLFR78 which is expressed in renal juxtaglomerular apparatus cells, when activated by its ligand SCFAs is responsible for renin secretion and thus regulates the control of blood pressure[30,31]. SCFAs when absorbed in the blood, play an important role in different physiological processes which include inflammation and metabolism[32].

Hypoxia sensing

OLFR78, is a mice receptor and is expressed in oxygen sensitive areas such as heart and lung[33,34], and OR51E2, present in humans when activated by SCFAs reduces the airway obstruction. Similarly the receptors OR1D2 induces cell contraction and the release of IL-8 while OR2AG1 inhibits histamine induced contractions in lungs[35–37].

Muscle regeneration

OLFR16 when activated by Lyral, causes an increase in the intracellular cAMP levels and promotes myocyte migration, branching and myogenesis, thus playing a role in muscle regeneration[38,39]. OR2H2 when expressed in human skeletal myoblasts when activated by aldehyde 13-13 results in the reduction of myoblast fusion[40].

Obesity and fat reduction

ORs expression is also linked to obesity for eg OR7D4 have been reported with the eating behaviours and reduction in levels of adiposity[36]. The ORs such as OR4P4, OR4S2 and OR4C6 have also been reported to have a role in obesity[41,42].

Wound healing, hair growth and skin hypopigmentation

Receptor OR2AT4 when activated by Sandalore increases keratinocyte proliferation, migration, regeneration thus playing a role in wound healing[43,44]. This receptor also promotes the growth of hair follicles thus reducing hair loss. OR2A4 promotes cytokinesis, cell proliferation, migration and regeneration of primary melanocytes when activated by its ligand cyclohexyl salicylate[45,46]. OR51B5 when activated by isononyl alcohol shows properties similar to OR2A4[21].

Function as therapeutic potential

The activation of the ORs outside their native environment has been proved to have a physiological and pathophysiological role[20]. The olfactory receptor OR10H1 is a potential biomarker for bladder cancer. When activated by sandranol, it helps to control the tumour cell proliferation[47]. The receptor OR51B6 is a promising therapeutic target for colon cancer. Once this receptor is activated by troenan it decreases cell proliferation and migration[48–50]. Similarly OR2AT4 is a target for wound healing and OLFR544 has a role in fat reduction[51].

Olfactory receptors in Cancer

The ectopic expression of ORs have been earlier reported to have a clinical relevance[52]. Their expression in non-olfactory tissues plays an important role in physio-molecular processes such as sperm chemotaxis[29], wound healing[20,39], muscle regeneration[39] and even in cancer. According to the previous reports the ORs have high expression in cancerous tissue as compared to their normal state[53]. For example OR51E2 is upregulated in the case of prostate cancer. Similarly breast cancer shows the upregulation of OR2B6, OR2C3 and OR10H1 are upregulated in melanoma and urinary cells respectively.

OR1A2 when activated by citronellal has been reported to play a role in cell proliferation in hepatocellular carcinoma. Similarly in colorectal cancer, when the receptor OR51B4 is activated by Troenan, the cancer cell undergoes apoptosis[54]. Moreover in prostate cancer, activation of OR51E2 by endogenous agonist, 19-hydroxyandrostenedione resulted in neuroendocrine trans differentiation revealing the functional implication of ORs in diverse physio-molecular processes[54,55]. When the same receptor is activated by β-ionone in melanoma cells, it leads to the activation of anti-proliferative and pro-apoptotic pathways[56].

The activation of OR2J3 by helional induces apoptosis in non small cell lung cancer. Notably, due to their expression specificity in the tumor-state, a handful of ORs have been identified as potential biomarkers e.g. OR51E1 in small intestine neuroendocrine carcinomas, OR7C1 in colorectal cancer, PSGR in prostate cancer, and OR2B6 in breast carcinoma[20,21,57]. All these collectively reinforce the idea of the potential therapeutic and diagnostic function of these ectopic olfactory receptors[54,55].

Materials and Methods

Cancer Smell: A Computational workflow for the detection of chemosensory receptor activation status from scRNA-sequencing datasets

Cancer Smell computational workflow was used to identify the potential chemosensory receptor gene activation status in cancer datasets. It contains inbuilt information about human-specific chemosensory receptor repertoire, which was manually curated from public databases i.e. The Human Olfactory Data Explorer, Horde database (https://genome.weizmann.ac.il/horde/) and (https://www.uniprot.org). Cancer Smell takes a raw TPM matrix as input in which the genes and cell Id information are arranged in rows and columns, respectively. To ensure that the input data matrix represents single cells, Cancer Smell utilizes "scrublet" (0.2.1), on python (3.7.0). Notably, the cells which qualify the singlet criteria were used in the subsequent downstream analysis. Cancer Smell utilizes zFPKM (1.8.0), a Bioconductor package for the estimation of the gene activation status. It uses a recommended cutoff i.e. zFPKM value >-3 for the active genes. For the downstream steps, Cancer smell utilizes Seurat (v 3.1.1) for data scaling, normalization, dimension reduction, and clustering. A deeper insight of each step is mentioned below **(Figure 1)**.

Data Collection

The PanTumor single-cell RNAseq (non-normalized) TPM matrix was obtained from CancerSea (http://biocc.hrbmu.edu.cn/CancerSEA/home.jsp). A total of 12839 cells across various cell lines and 29690 cells across multiple tumors were used for the analysis. The Cancer Smell pipeline was implemented only on those datasets in which the minimum number of malignant cells was at least 60. In addition to the malignant cells, the healthy breast epithelial cells data was obtained from this given **link**. By using Human Olfactory Data Explorer (horde), (https://genome.weizmann.ac.il/horde/) the list of ORs (391 Functional ORs + 466 Pseudo Ors) and their corresponding fasta sequences were taken. Through a literature survey, the list of V1R, T1R, T2R, TAAR was also obtained and used in the analysis.
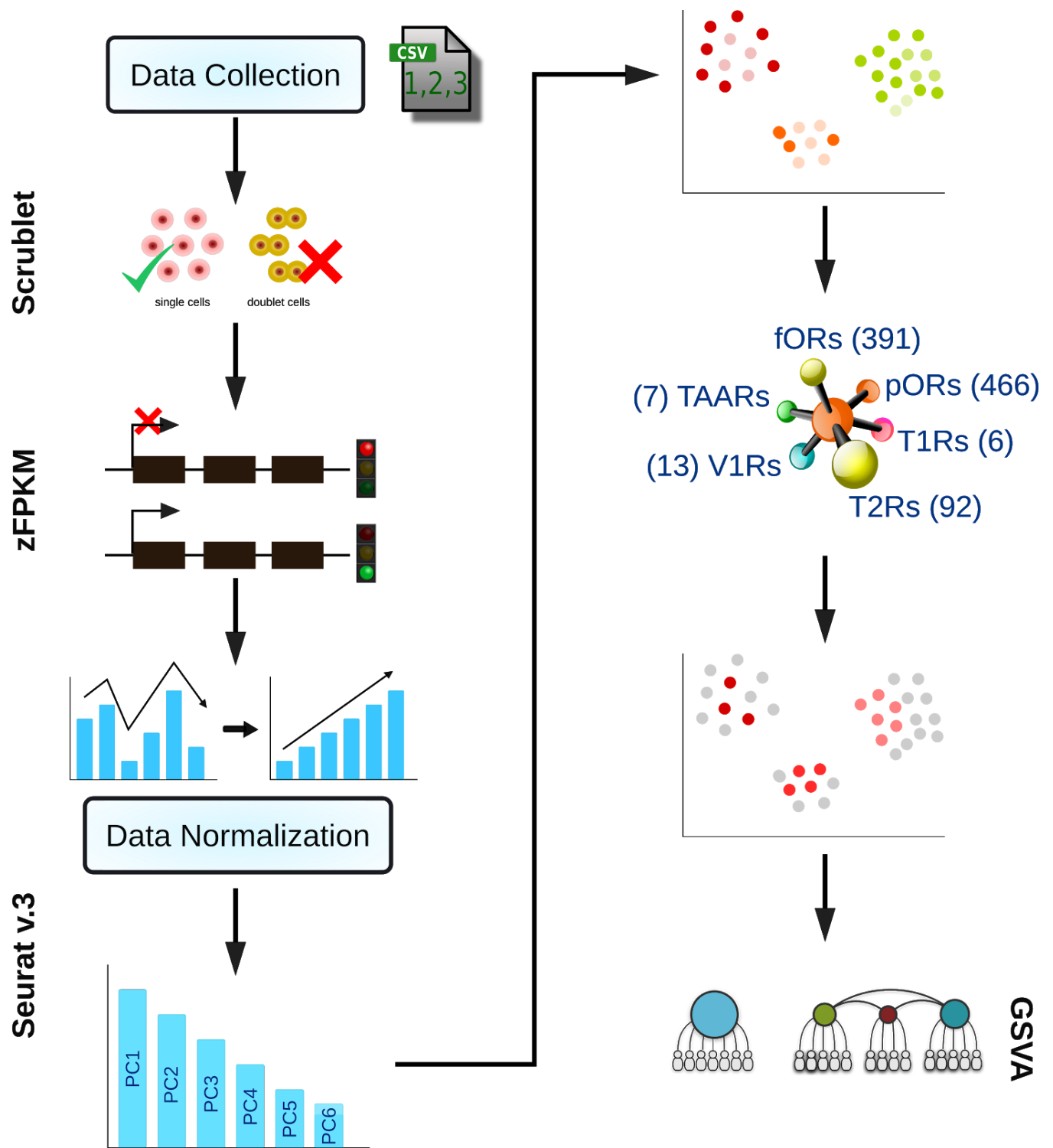
Figure 1: Schematic representation of the computational workflow (Cancer Smell) to systematically estimate the activation status of the chemosensory receptors at the single cell resolution.

Data Processing

The analysis pipeline inputs a raw TPM matrix that contains genes in rows and cell Ids in columns. The Ensembl ids were converted to gene names in R using the "org.Hs.eg.db" library.

Doublet Analysis

To ensure that the single cell is actually single, doublet analysis was performed using the python (3.7.0) package "scrublet" (0.2.1) (https://github.com/allonkleinlab/scrublet). This package inputs a raw TPM or FPKM matrix which has cell ids in rows and genes in columns and returns the output in the form of a CSV file that contains Cell Ids, Doublet Score and Doublet State (True or False). It generates the simulated samples based on the data points and projects them on bimodal distribution. The cells having doublet state equals to TRUE were removed and the remaining cells were further used for analysis.

zFPKM cutoff

To identify which genes are active and which are inactive, zFPKM package (1.8.0) was installed through Bioconductor in R version 3.6. The genes that have the median zFPKM value >-3 were considered as active genes and rest were removed. Reference recommends using zFPKM > -3 to select expressed genes. This has been validated with encoding open/closed promoter chromatin structure epigenetic data on six of the ENCODE cell lines. This package works well for gene-level data using FPKM or TPM values.

Seurat Implementation

The Seurat version 3.1.1 was used for scaling, normalization, dimension reduction, and clustering.

Object Creation

The input file (TPM matrix) which contains active genes in rows and non-doublet cells in columns was converted into a Seurat object by using the inbuilt function CreateSeuratObject.

Normalization and variable features

Data was log normalized and then the FindVariableFeatures function was used to identify the variable features in the data set.

Dimension Reduction

Before reducing the dimensions, the data were linearly transformed using the ScaleData function. After this transformation RunPCA function was used to perform Principal component analysis.

Clustering

KNN graph-based approach using the euclidean distance was used in the inbuilt function FindCluster of Seurat. The clusters were further optimized by using the Louvain algorithm.

Biomarker Identification

The function FindAllMarkers identifies all the possible biomarkers for each cluster.

UMAPs

The detected functional ORs, pseudo-OR's, V1R's, T1R's, T2R's, TAARS were highlighted in the UMAPs.

<u>Cell Classification</u>

Based on the zFPKM values, the cells were classified as Positive or Negative based on the expression of ORs. If the count of zFPKM value >-3 for a particular cell is greater than 0 then that cell was classified as Positive.

Gene Set Variation Analysis

It is a non-parametric unsupervised method used for the functional and pathway enrichment of the chemosensory receptors. The GSVA package version 1.34.0 was installed through Bioconductor. The analysis was performed using fourteen gene signatures namely Angiogenesis, Apoptosis, Cell Cycle, Differentiation, DNA damage, DNA repair, EMT, Hypoxia, Inflammation, Invasion, Metastasis, Proliferation, Quiescence, and Stemness.

Phylogeny tree construction

The phylogeny tree was constructed using SeaView software (version 1:4.6.4-1). The fasta sequences of only functional ORs were loaded and were used to construct the phylogenetic tree. The sequences were aligned using Muscle and PhyML algorithm was used to construct this phylogenetic tree. Further manipulations in the tree were done using an interactive tree of life (iTOL) (https://itol.embl.de/).

Ideogram

The ORs identified across pan tumors and cancer cell lines were further analysed based on their chromosomal distributions. Through WashU Epigenome Browser (version 46.2). The Hg38 genome of humans was selected as a reference. Two different datasets were uploaded, the first one entire functional horde ORs (391) and the other was the identified ORs across various cell lines and tumors (68). The two-color scheme was used to discriminate between cancer-associated (red) and non-cancer associated ORs.

Metascape

Differential gene expression analysis using a custom-build script, utilizing the Wilcox-test for the estimation of statistical significance, if any. Genes showing significant up-or-down regulation (fold change cutoff +-4; p-value < 0.05) were used as an input for the Metascape (http://metascape.org/), along with the cluster information.

Pseudo Temporal Ordering

Cell fate decisions and differentiation trajectories were reconstructed with the Monocle 3 package. This computational workflow utilizes reverse graph embedding based on a user-defined gene list to generate a pseudo time plot that can account for both branched and linear differentiation processes. In order to understand the changes in the OR expression or their cellular frequencies during breast cancer cell differentiation, for which the degree of stemness in each individual malignant cell for the determination of the starting point for the trajectory formation was computed. Ordering of the cells in the 2D space was achieved by using an inbuilt function of Monocle 3. The cell that has the highest value of stemness was set as the starting point of the trajectory. Estimation of the cellular OR frequency and their mean expression pattern were plotted along the pseudo time.

Classification of bulk tumor profiles based on single-cell derived OR Signature

Generation of a fine-grained gene expression signature that harbors information about both the OR expression and their relative cellular frequencies from single-cell sequencing datasets was performed as follows. The Expression data matrix (TCGA and single-cell breast cancer datasets) were log2 transformed (log2(TPM+1)) and the estimation of the cellular OR frequencies was performed based on zFPKM cutoff (>-3). This criterion for the estimation of active/non-active ORs in each malignant cell was applied which resulted in the binary matrix for downstream clustering (if zFPKM =< -3 is 0 and otherwise 1). Clustering on the resulting OR binary matrix was performed using cutting on the dendrogram. In most cases, > 2 clusters (groups of malignant cells)

segregated based on cellular OR frequencies were obtained.. A single cell-based OR signature was generated by averaging all cells within a cluster. The projection of the obtained signature on the TCGA bulk tumor profile was determined using pairwise cosine similarity, which was calculated by the given formula.

$$cosine\ similarity(A, B) = \frac{A.B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A and B are the two independent vectors.

Lastly, to segregate the bulk patient samples based on single-cell inferred OR signature, the clustering on the cosine similarity matrix was performed by using cutting on the dendrogram. Survival estimation of the patients in each resulting group was performed using the Kaplan-Meier estimator.

Statistical analysis

Graphical illustrations are plotted with R Stats packages. The P-value cut-off used in this study is 0.05. *, **, *** and **** in the figures refer to P-values ≤ 0.05, ≤ 0.01, ≤ 0.001 and ≤ 0.0001, respectively.

Code Availability

The entire computational pipeline with complete explanation and working code for every analysis performed has been provided on the Github repository under the name Cancer Smell. (https://github.com/siddhant18241/Cancer-smell).

An R package enabling OR-based stratification of the patient's cohort using a single-cell expression-based gene signature (ORsurv: https://github.com/krishan57gupta/ORsurv).

Results

Widespread activation of the Olfactory receptor gene repertoire in malignant cells.

In Cancer cells, it is a well known established fact that every layer whether it is the genome, metabolome or proteome becomes highly dysfunctional and starts expressing some transcripts that were not supposed to be expressed there, one such class is the olfactory receptors. Their activation in various cancer types has previously been reported in the literature, in most cases, the assessment of the OR transcripts is achieved by profiling the bulk samples, which has its own drawbacks like it is unable to reveal the important information related to cell type and also the expression of the rare transcript is lost. To address and overcome this problem, the CancerSmell workflow that has been developed takes into consideration the activation status of the chemosensory olfactory and taste receptors across pan tumors at the single-cell resolution **(Figure 2)**.

Overall this study evaluated the expression of 42529 malignant cells in 49 scRNASeq datasets, across 22 different tumor types. Once the datasets were obtained, the primary task was to actually make sure that only the non-doublet and active genes are present there. To ensure this scrublet and zFPKM score was used respectively. With the help of this workflow, we have developed the largest Olfactory receptor gene repertoire at a single-cell resolution **(Figure 3)**.
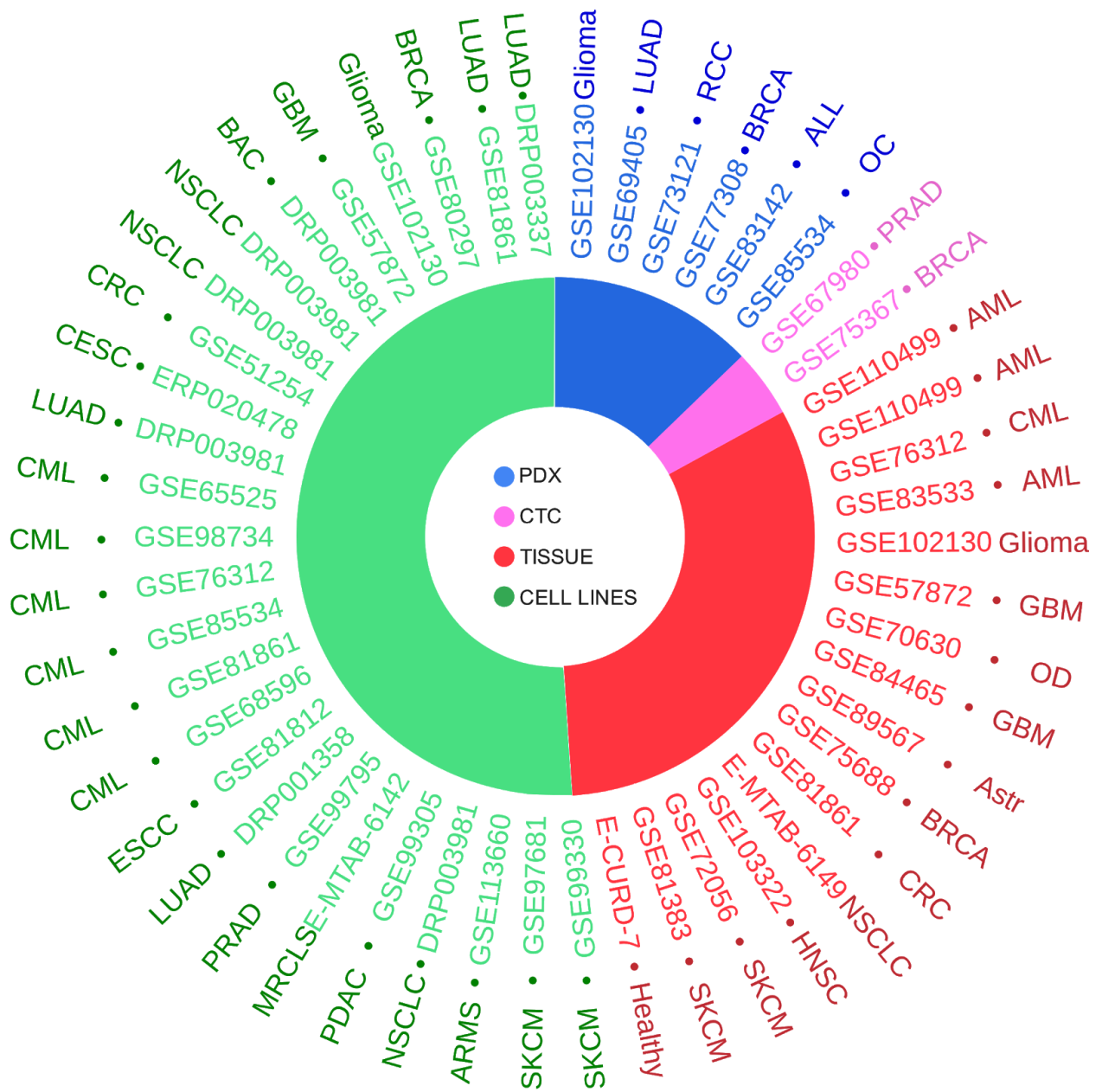
Figure 2: Graphical illustration of the scRNA-seq datasets used in this analysis
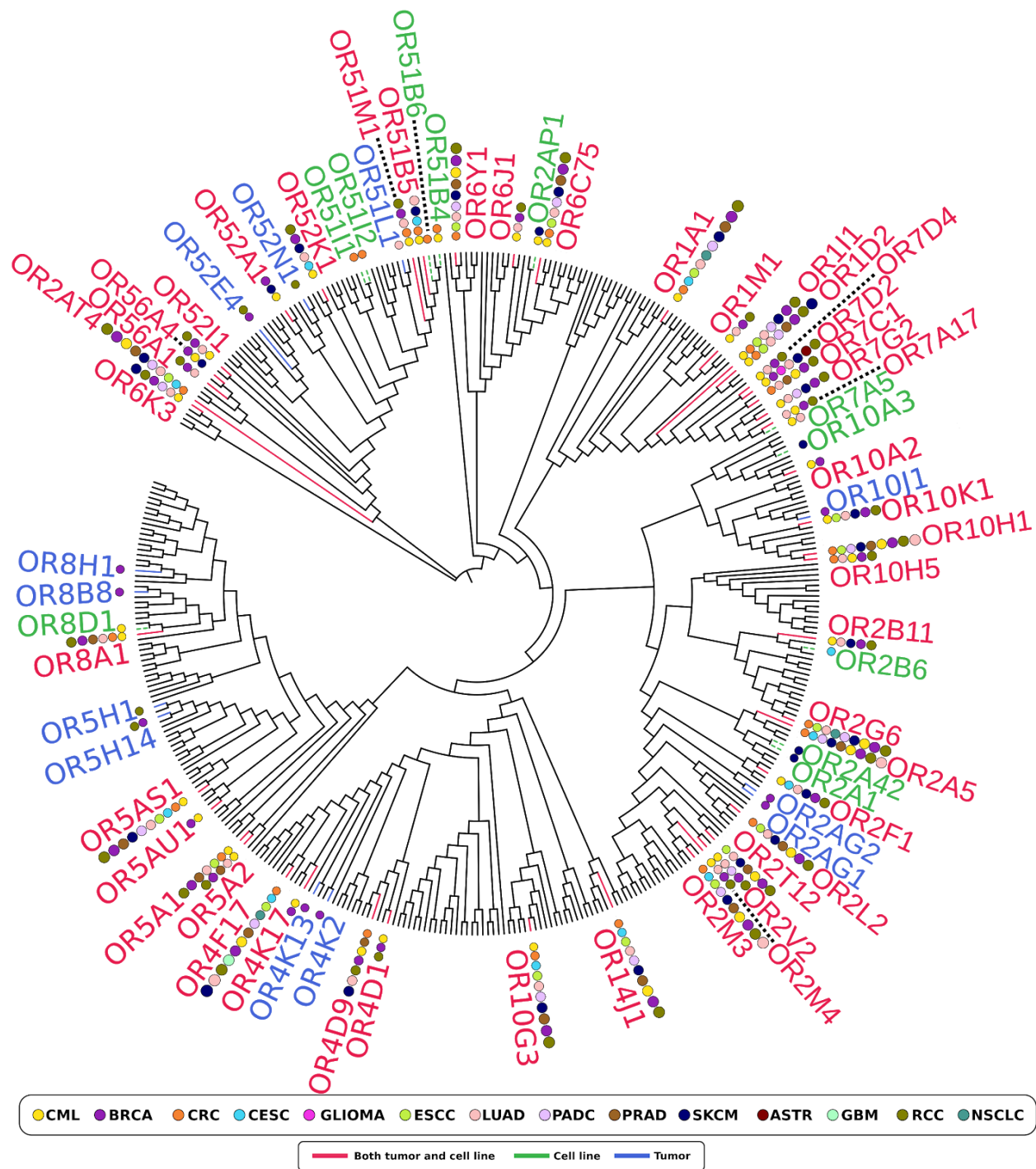
Figure 3: Phylogenetic tree representing the ORs sequence relationships and their detection status in 14 distinct cancer types. Colored circles represent the tumor type and branch color represents the source information.

In order to determine the specificity and exclusivity of the identified ORs, we looked at their distribution across various tumor types and found out that certain tumors are expressing only one type of receptor while others are expressing more than one for eg few ORs like OR8B8 and OR8H1 were only expressed in malignant breast cells while others like OR1A1 was found to be expressed in eleven different tumor types **(Figure 4, 5)**. Their results are summarised in the form of chord plots.
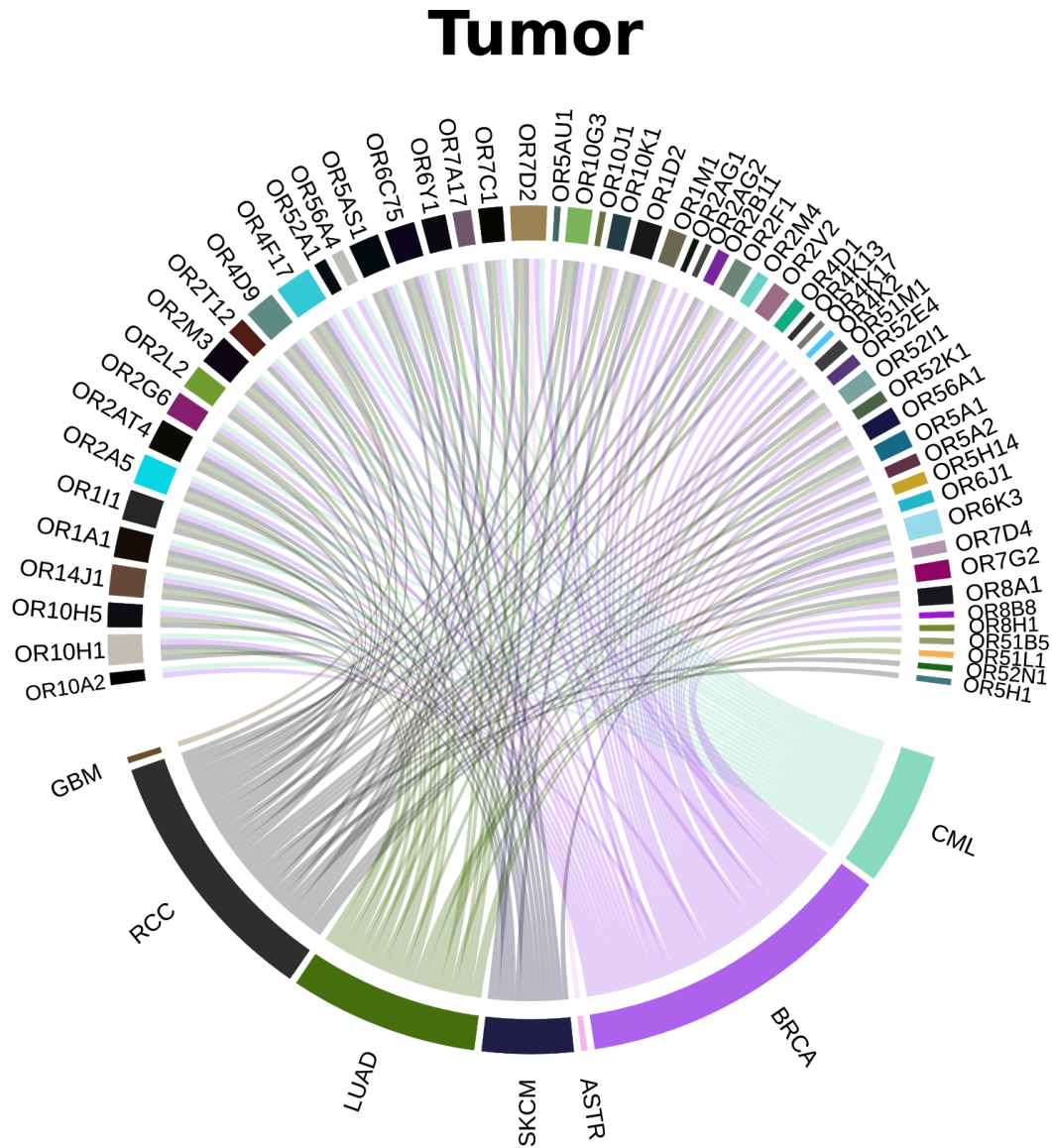
Figure 4: Chord diagram representing different types of tumors expressing different ORs at a single-cell resolution.

Cell Lines
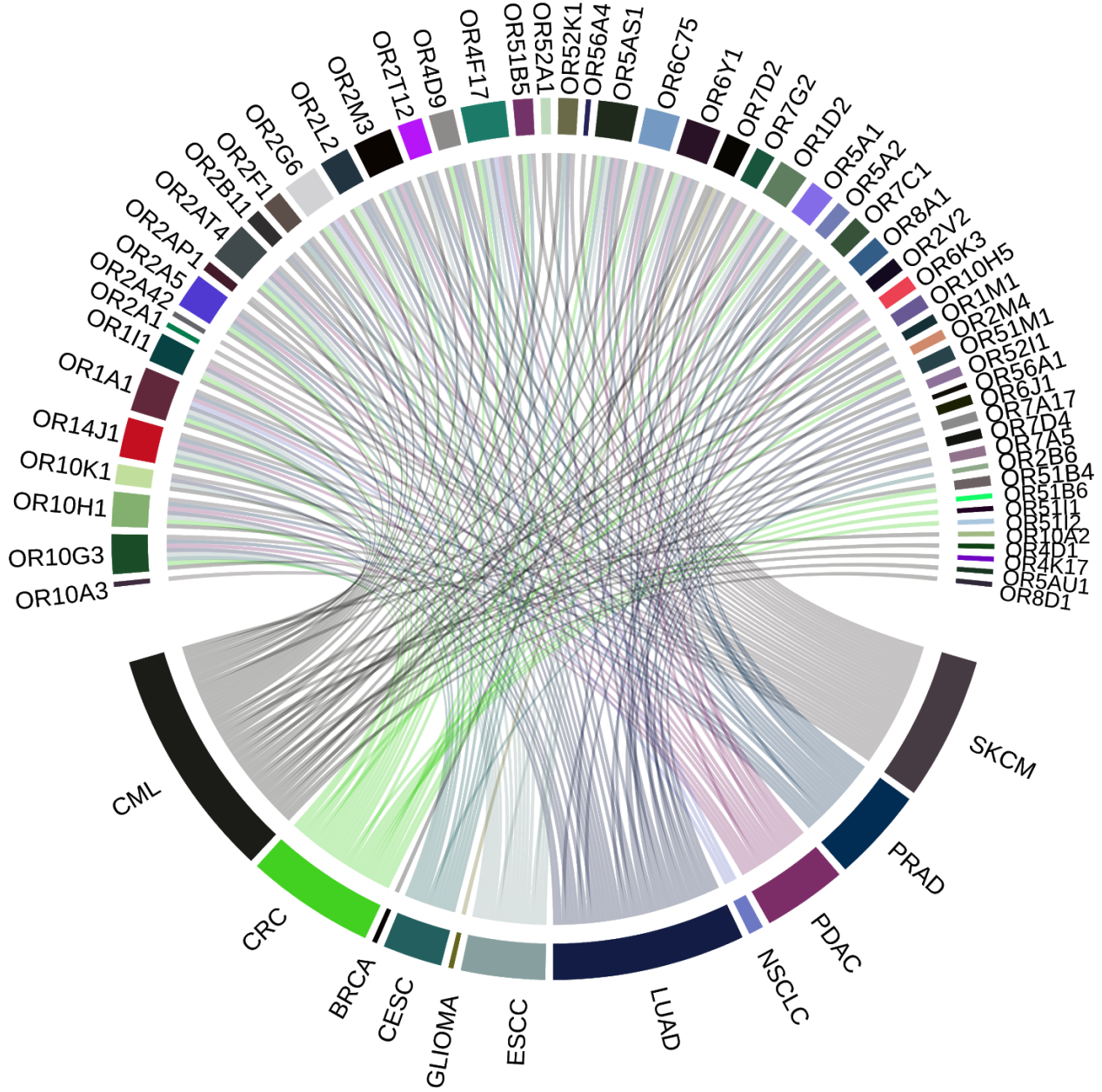
Figure 5: Chord diagram representing different types of cell lines expressing different ORs at a single-cell resolution.

In order to understand if ORs exhibit a certain bias towards particular genomic loci, we mapped these receptors with the Hg38 human reference genome in the WashU Epigenome Browser. The ideogram depiction **(Figure 6)** is one the best way to understand the chromosomal distribution of these receptors. The results clearly indicate that chromosome 11 was highly enriched with them also the distribution of these receptors can be seen only on 10 out of 21 chromosomes.
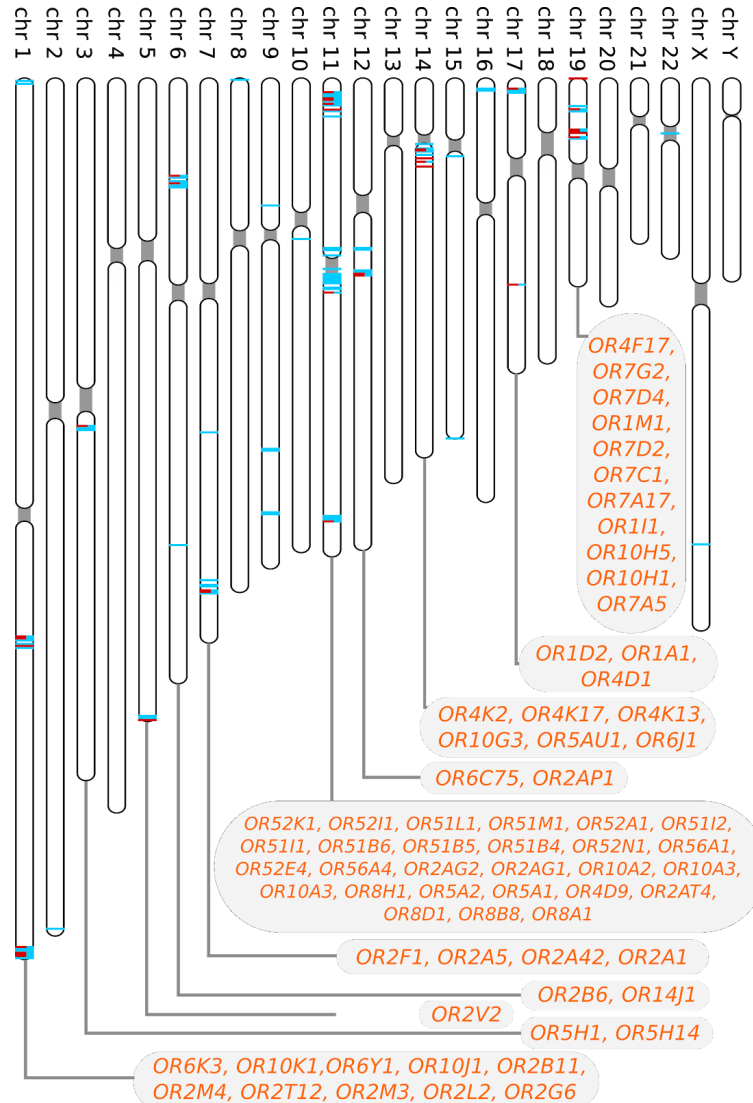
Figure 6: Ideogram illustrating the chromosomal location of all the functional ORs in humans. Red lines represent functional ORs reliably detected in malignant cells (tumor or cell lines) whereas cyan lines represent Ors that are not detected in any tumor-types or cell lines under investigation.

Apart from functional olfactory receptors various other receptors family-like pseudo olfactory receptors, Trace Amine associated receptors (TAARs), Vomeronasal type 2 receptors (V2Rs) and taste receptors were also evaluated for their expression. The results show that only the ORs expression was significantly rich in contrast to others **(Figure 7)**. The ORs also show a high degree of variability in their expression in different tumors. Some of the examples are shown here. Taste receptor, TAS1R3 was found to be expressed only in the Hela cell line of Cervix cancer. Vomeronasal receptors, VN1R1 and VN1R2 were found to be expressed in the WM983B cell line of Melanoma **(Figure 8)**. None of the dataset showed the expression of pseudo olfactory receptors.
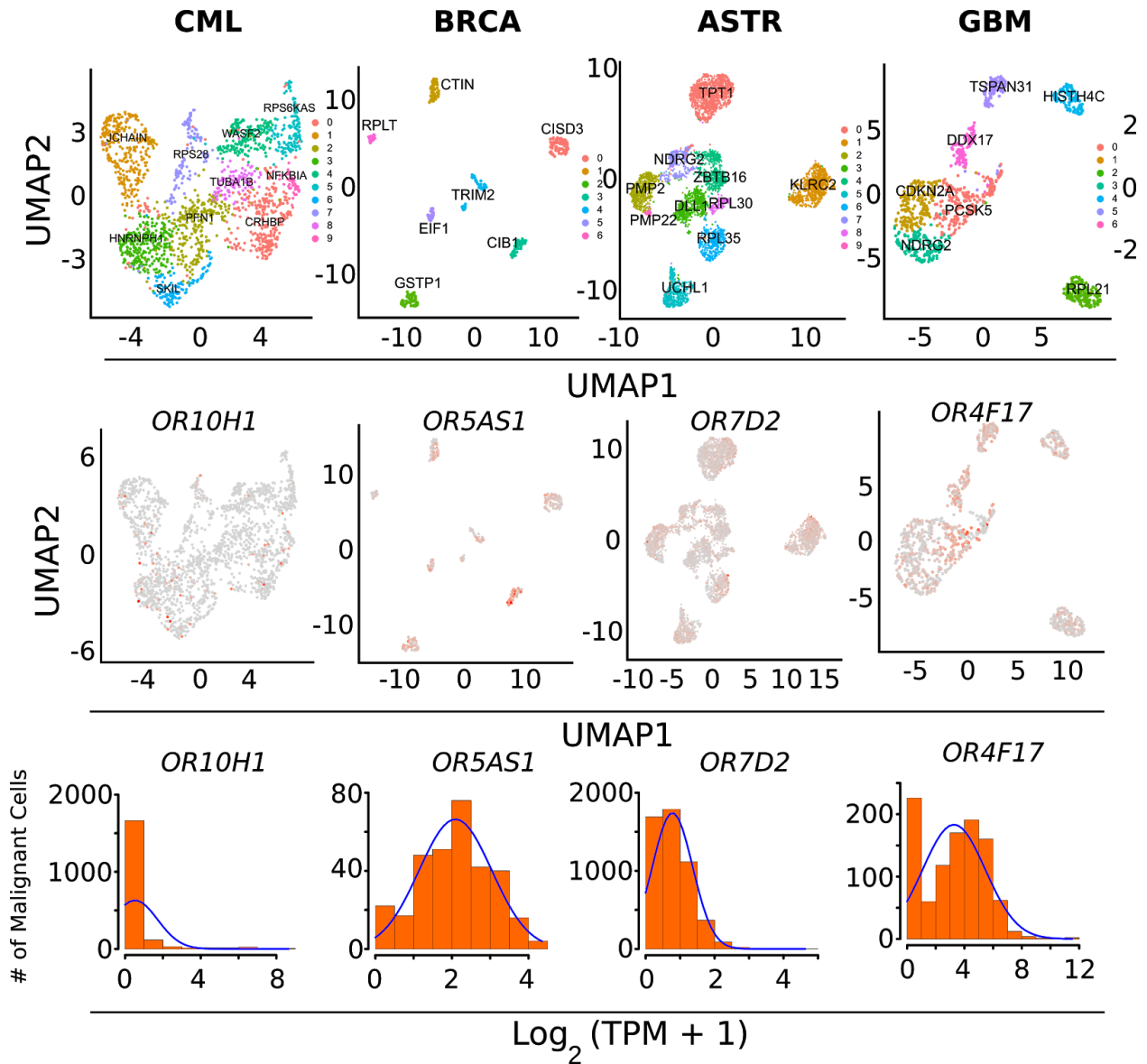
Figure 7: (Top) UMAP illustrates distinct cellular clusters in the indicated single-cell tumor datasets. Markers for each cluster are indicated in the text. (Middle) UMAPs depicting the relative expression of the representative ORs in the indicated single-cell tumor dataset. (Bottom) Density histograms depicting the normalized expression of the indicated ORs.
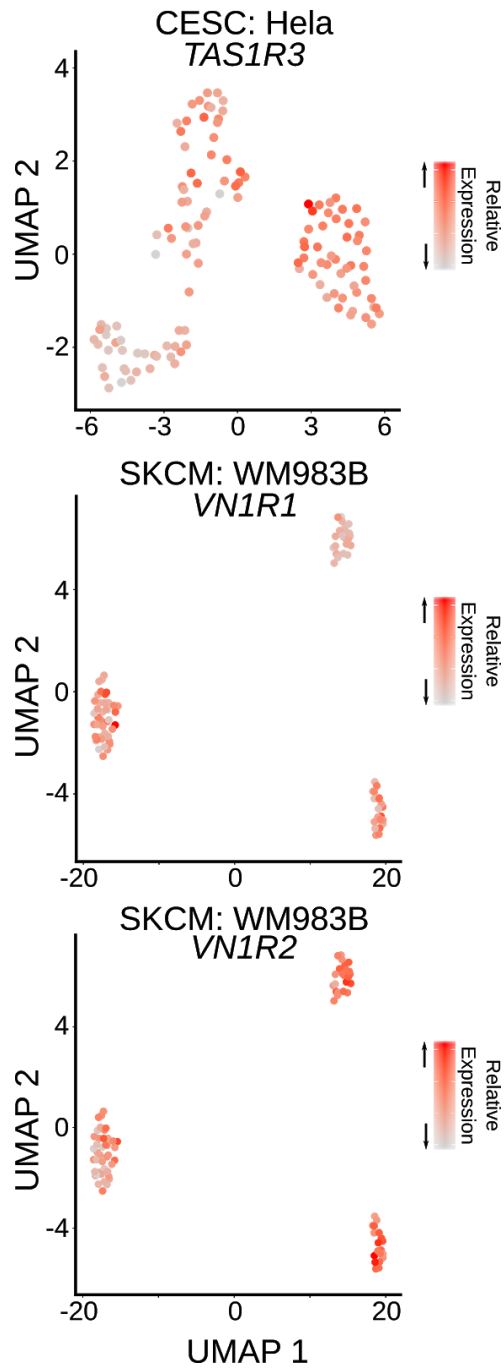
Figure 8: UMAP illustrates distinct cellular clusters and relative expression of the representative ORs in the indicated single-cell tumor dataset.

The total count of which type of receptor was found to be expressed in which type of tumor is mentioned in the **Table 2**. The detailed table (consisting the names of the receptor expressed) for the same, can be found in the supplementary Table 1 **(SI_Table1)** by accessing through this URL

Table 2: The table contains information about the datasets used in this study. It also contains information about the total number of receptors expressed in each dataset.

Accession	Type	ORs	pORs	T1Rs	T2Rs	TAAR	V1Rs
GSE83142	Acute lymphoblastic leukemia	0	0	0	0	0	0
GSE110499	Acute myeloid leukemia	0	0	0	0	0	0
GSE110499	Acute myeloid leukemia	0	0	0	0	0	0
GSE83533	Acute myeloid leukemia	0	0	0	0	0	0
GSE76312	Chronic myelogenous leukemia	22	0	0	0	0	0
GSE81861	Colorectal cancer	0	0	0	0	0	0
GSE77308	Breast cancer	16	0	0	0	0	0
GSE75688	Breast cancer	53	0	0	0	0	0
GSE75367	Breast cancer	19	0	0	0	0	0
GSE89567	Astrocytoma	1	0	0	0	0	0
GSE84465	Glioblastoma	1	0	0	0	0	0
GSE57872	Glioblastoma	0	0	0	0	0	0
GSE102130	Glioma	0	0	0	0	0	0
GSE102130	Glioma	0	0	0	0	0	0
GSE70630	Oligodendroglioma	0	0	0	0	0	0
GSE103322	Head and neck cancer	0	0	0	0	0	0
GSE73121	Renal cell carcinoma	43	0	0	0	0	0
GSE69405	Lung adenocarcinoma	31	0	0	0	0	0
E-MTAB-6149	Non-small cell lung cancer	0	0	0	0	0	0
GSE85534	Ovarian carcinoma	0	0	0	0	0	0
GSE67980	Prostate cancer	0	0	0	0	0	0
GSE72056	Melanoma	1	0	0	0	0	0
GSE81383	Melanoma	14	0	0	0	0	0

GSE113660	Alveolar rhabdomyosarcoma	0	0	0	0	0	0
DRP003981	Bronchoalveolar carcinoma	0	0	0	0	0	0
GSE65525	Chronic myelogenous leukemia	0	0	0	0	0	0
GSE98734	Chronic myelogenous leukemia	46	0	0	0	0	0
GSE76312	Chronic myelogenous leukemia	17	0	0	0	0	0
GSE68596	Chronic myelogenous leukemia	0	0	0	0	0	0
GSE85534	Chronic myelogenous leukemia	1	0	0	0	0	0
GSE81861	Chronic myelogenous leukemia	2	0	0	0	0	0
GSE51254	Colorectal cancer	26	0	0	0	0	0
GSE80297	Breast cancer	1	0	0	0	0	0
ERP020478	Cervix cancer	13	0	1	0	0	0
GSE57872	Glioblastoma	0	0	0	0	0	0
GSE102130	Glioma	1	0	0	0	0	0
GSE81812	Esophageal squamous cell carcinoma	17	0	0	0	0	0
DRP001358	Lung adenocarcinoma	38	0	0	0	0	0
DRP003337	Lung adenocarcinoma	0	0	0	0	0	0
DRP003981	Lung adenocarcinoma	3	0	0	0	0	0
GSE81861	Lung adenocarcinoma	1	0	0	0	0	0
DRP003981	Non-small cell lung cancer	3	0	0	0	0	0
DRP003981	Non-small cell lung cancer	2	0	0	0	0	0
DRP003981	Non-small cell lung cancer	1	0	0	0	0	0
GSE99305	Pancreatic ductal adenocarcinoma	17	0	0	0	0	0
GSE99795	Prostate cancer	19	0	0	0	0	0
GSE99330	Melanoma	0	0	0	0	0	0
GSE97681	Melanoma	31	0	0	0	0	2
E-MTAB-6142	Myxoid liposarcoma	0	0	0	0	0	0

Once we obtained the expression values of receptors under study across multiple tumors, we computed the activation and deactivation status of the cell using zFPKM scores **(Figure 9)**. The cells had a median zFPKM score >-3 we considered to be active and the rest were inactive. This cutoff of -3 is only for humans and has been reported in the literature after validation on ENCODE data.
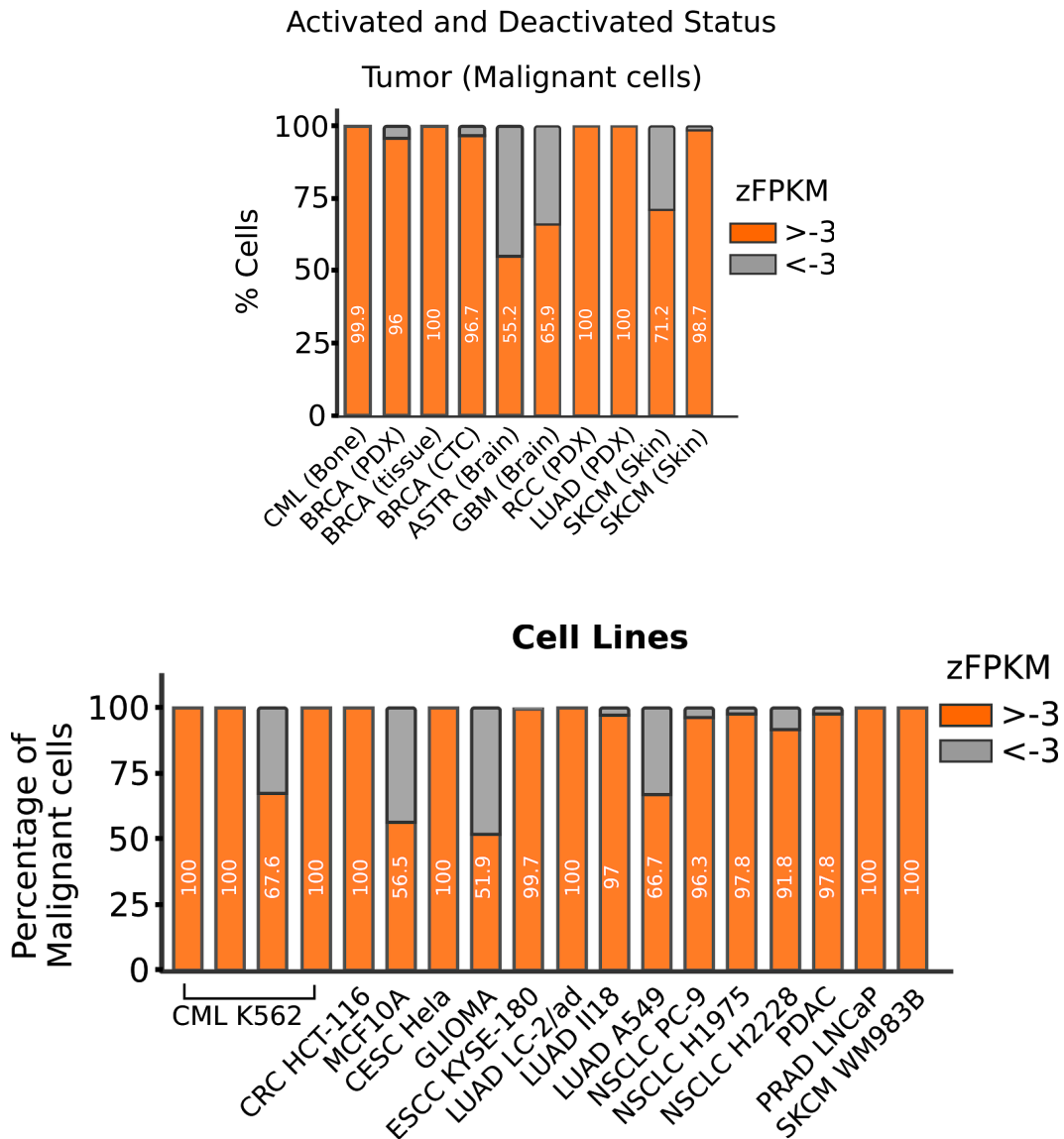
Figure 9: Bar graph representing the percentage of OR-positive malignant cells in the indicated tumor single-cell datasets. zFPKM algorithm was used for the determination of OR activation status.

Overall with this workflow, we have identified 68 OR-Tumor pairs. Among them, only 9 were reported for their expression in the previous cancer literature. Thus we can say that Cancer Smell workflow has identified 59 novel OR-Tumor pairs **(Figure 10)**.

Status in Cancer
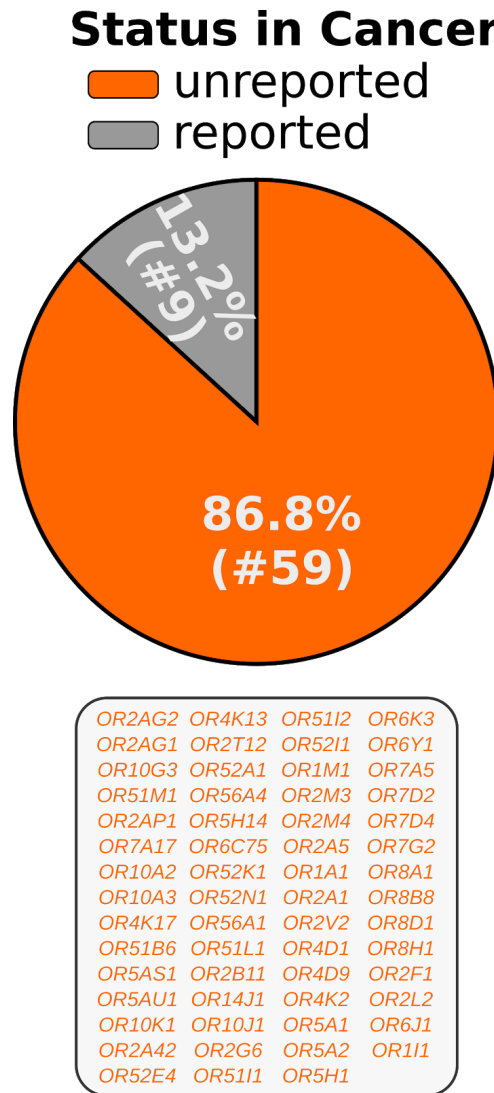
Figure 10: Estimation of the Olfactory repertoire at the single-cell resolution revealed previous unreported cancer-associated olfactory receptors, represented here as Pie chart.

Now in order to identify what role these receptors are playing there, Gene Set Variation Analysis (GSVA) was performed against fourteen cancer signatures. The results indicate that a significant proportion of the detected ORs are implicated in the process such as stemness, metastasis, invasion, and differentiation **(Figure 11)**. Few ORs, when expressed, were involved in the pathways that were causing tumor progression while some were expressed in the pathways responsible for the reduction of tumor progression. The GSVA scores for receptors along with their signatures is given in Supplementary Table 2 **(SI_Table2)** which can be accessed through this URL
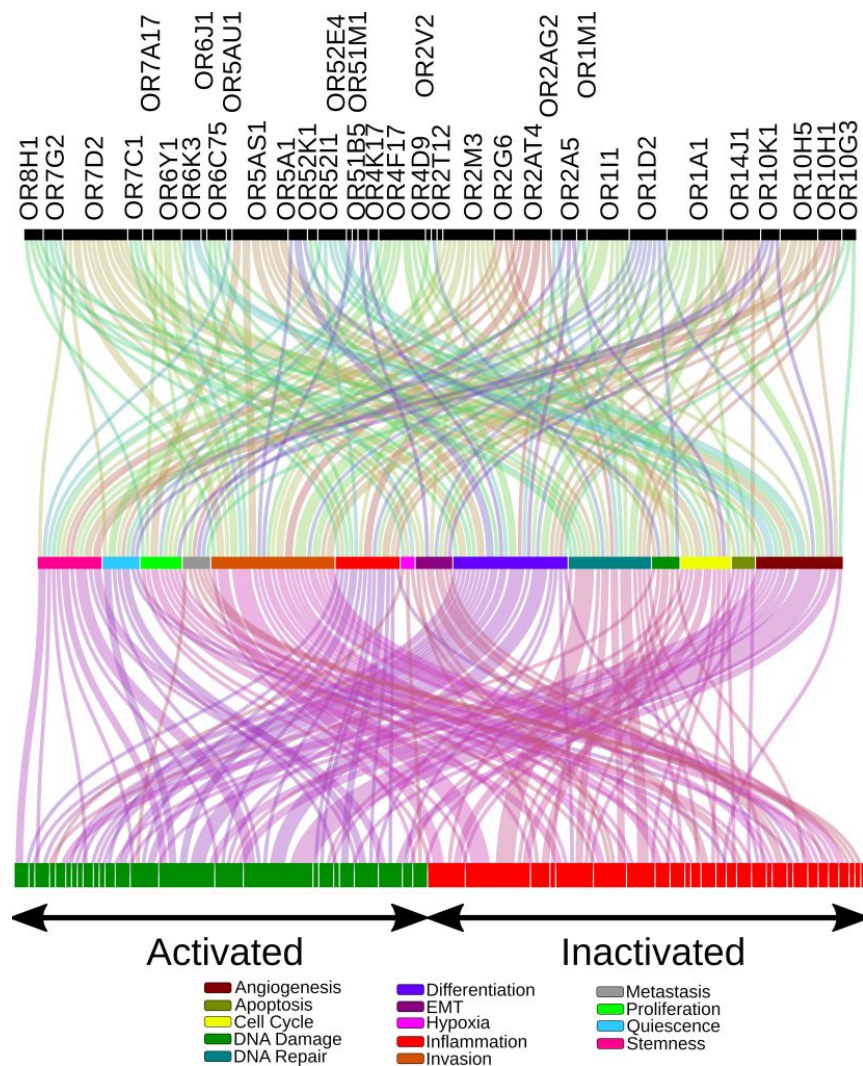
Figure 11: Gene Set Variation Analysis (GSVA) revealed the functional relevance of the indicated olfactory receptors in tumor-related processes, indicated here as the Alluvial plot.

Aberration of the "one-receptor one cell rule" in cancer

In a normal condition, the OR gene expression is tightly regulated and thus in its native environment, OR expression is restricted to "one-receptor one-cell rule". In order to test if this is applicable in the malignant cells, we estimated the cellular frequency of co-expressed ORs. For a particular cell if the zFPKM score of the OR was greater than -3 that suggests that the receptor is active in that particular cell. Our results reveal that a cancerous cell does not obey "one-receptor one-cell rule" as it expresses multiple receptors at once, except in the case of tumors related to the nervous system **(Figure 12)**.
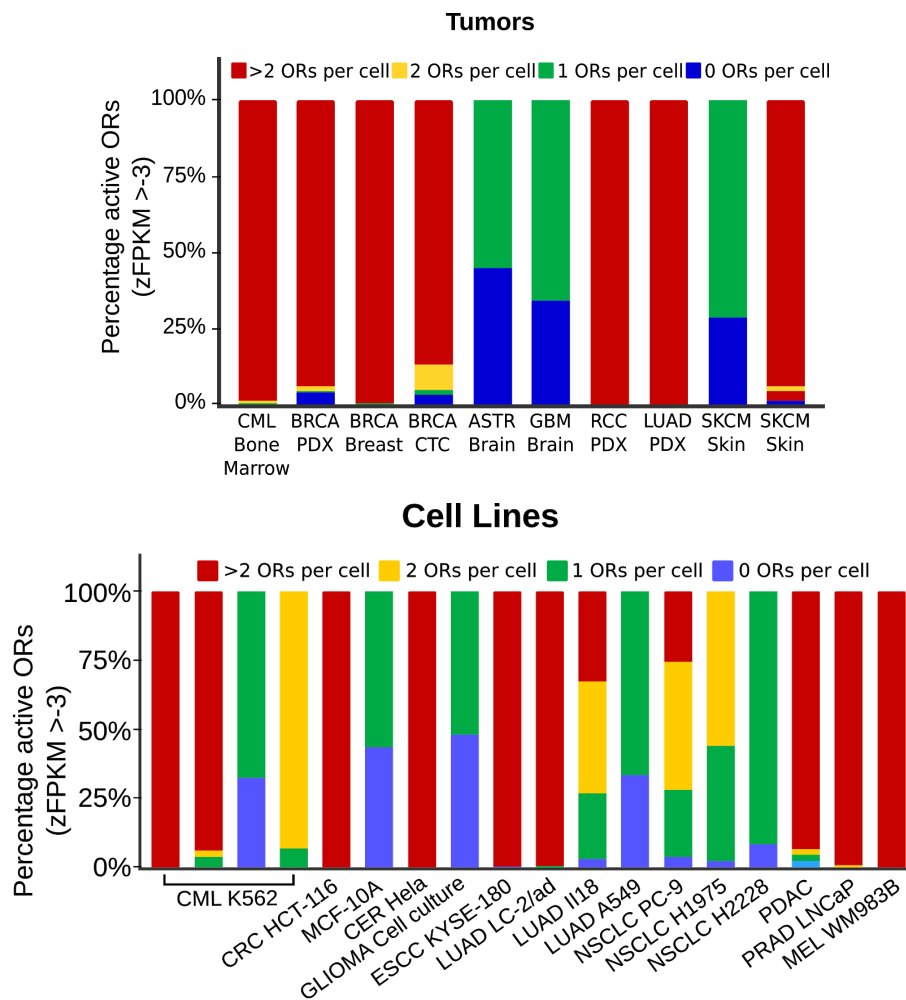
Figure 12: Cellular frequency of expressed ORs largely varies across multiple tumor types, depicted here as a percentage bar graph.

The Breast cancer dataset was further deeply analyzed as in our results, it showed the maximum relative abundance of the OR transcripts. We overlay the different clusters formed with their molecular subtypes to identify whether the certain molecular type is more responsible for expressing high amounts of ORs. After applying the Mann-Whitney pairwise test with Bonferroni corrected p-values, no significant result was obtained between the OR expression and the molecular subtype **(Figure 13)**.
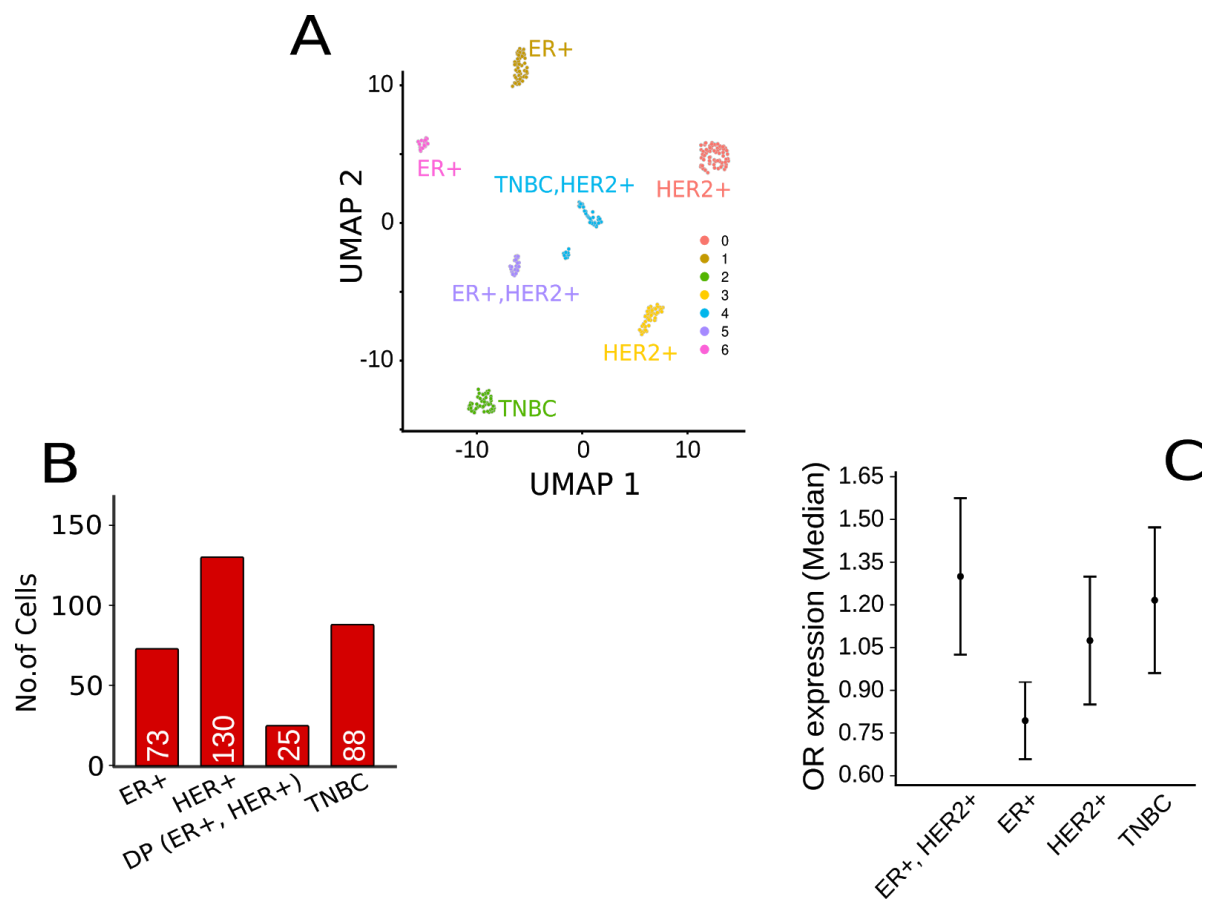
Figure 13: A) UMAP plot depicting the distinct cellular clusters segregated based on their transcriptomes overlayed with the molecular identity of different clusters. B) Bar graph representing the number of cells in different molecular subtypes of breast carcinoma. C) Whisker plot representing the median expression of the breast carcinoma-associated ORs in the indicated conditions

Further, a correlation analysis was performed to check which ORs have a tendency to express together **(Figure 14)**.
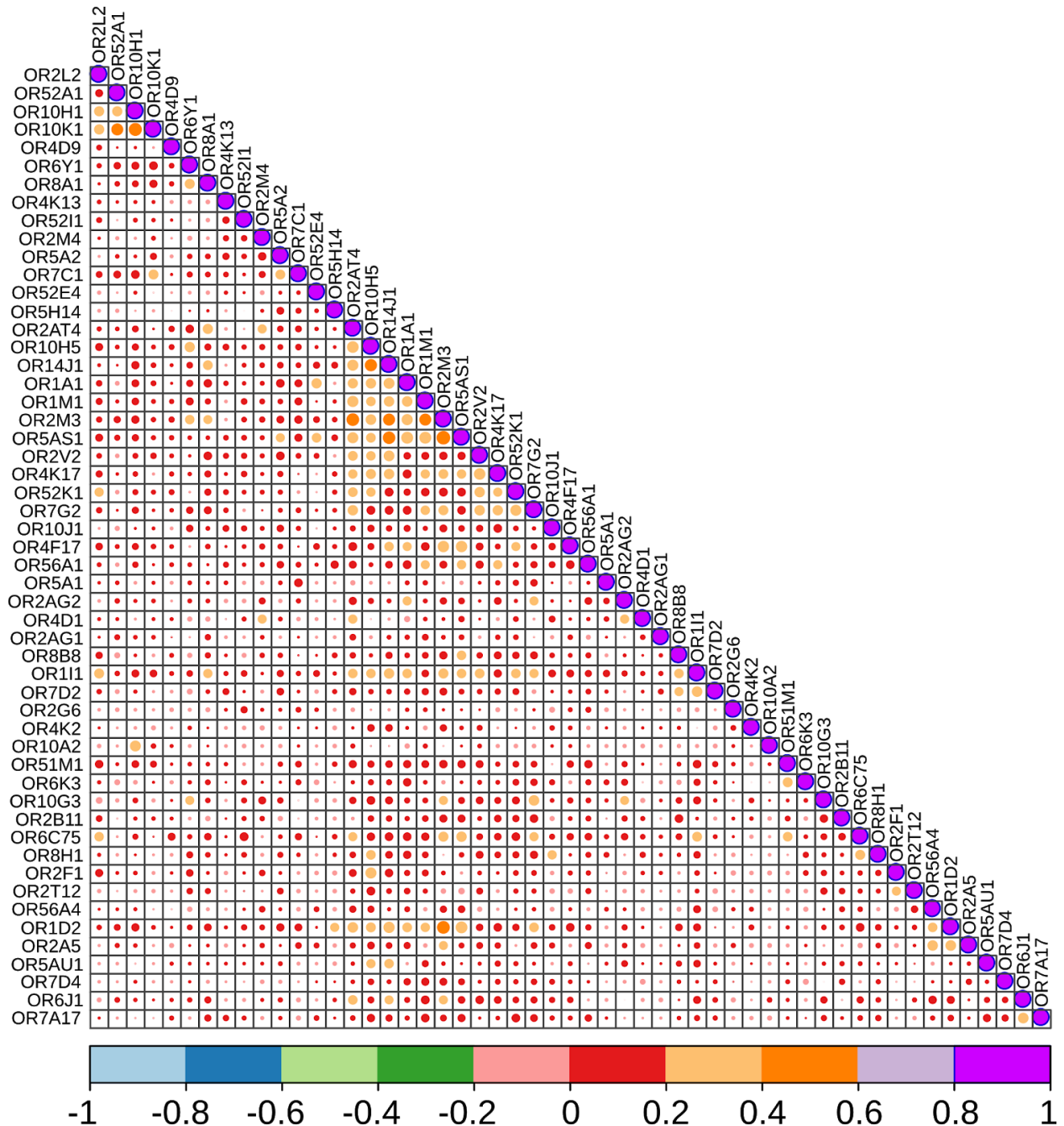
Figure 14: Correlation plot depicting the expression similarities between breast carcinoma-associated ORs.

One example of the co-expressed ORs is shown below. These results further clearly indicate that the tumor cell does not follow the one-receptor one cell rule. The Red arrow points towards the cell expressing OR2M3 while the green arrow points towards the cell expressing OR1A1. Few cells indicated with green arrows are the ones which are expressing both of these receptors **(Figure 15)**.
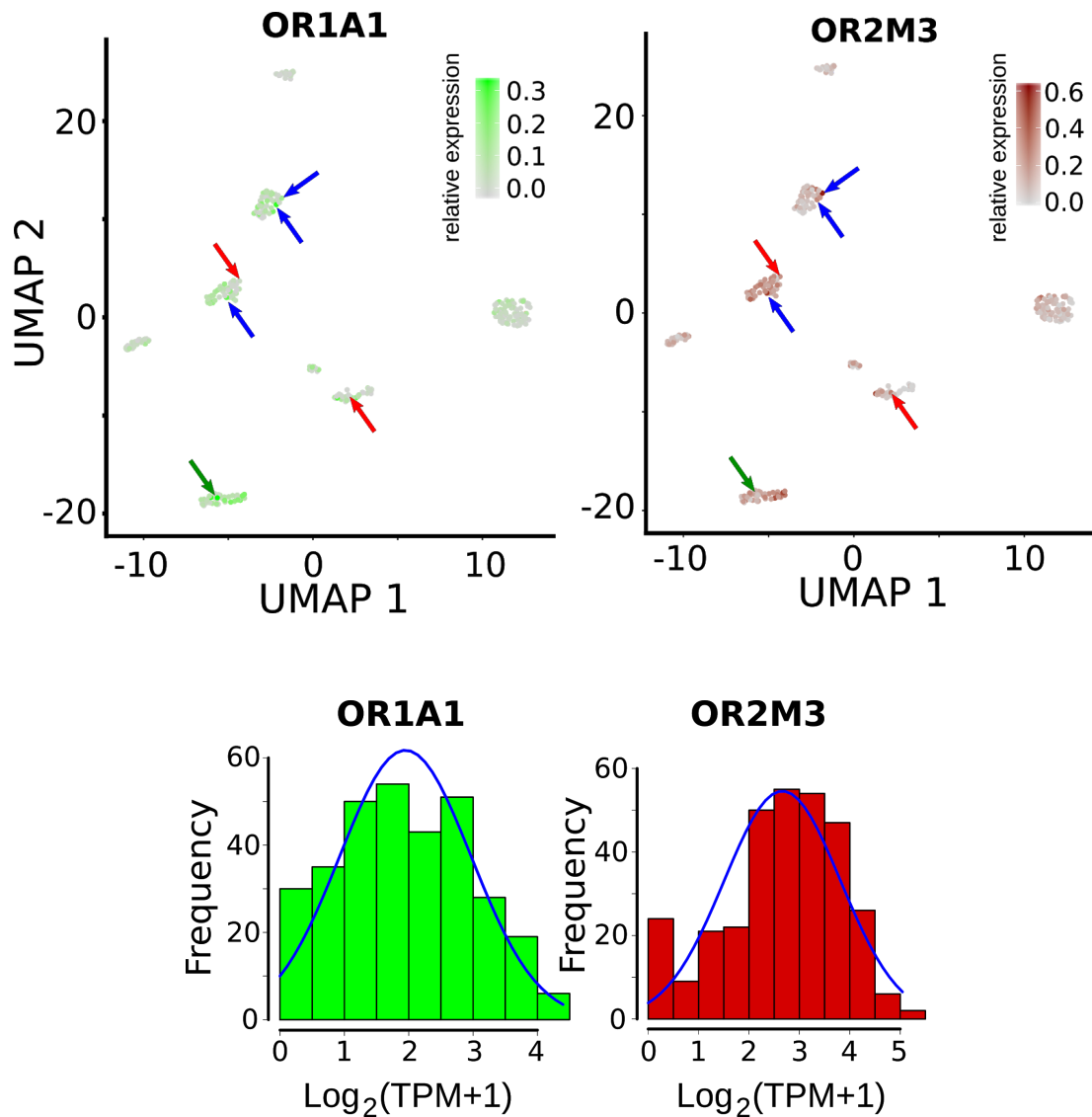
Figure 15: UMAPs representing the cellular expression of two representatives ORs in breast carcinoma along with the density histograms depicting the normalized expression of the indicated ORs in single-cell breast carcinoma dataset.

Past reports evaluating the activation status of chemosensory receptors in healthy tissues revealed the activation status of chemosensory receptors in healthy tissues revealed the activation of multiple ORs. Therefore we examined the exclusivity of the ORs, expressed in tumors. Comparative analysis revealed that the number of activated OR genes are systematically higher in the malignant state. Only a sub-fraction of these (14 out of 53) were detected in healthy cells **(Figure 16)**.

Breast Epithelial Cells

Healthy   Cancerous

(E-CURD-7)

Cells - 799   # Cells - 767

ORs: 14   # ORs:
                                    PDX   - 16
                                    Tissue - 53
                                    CTC   - 19

Figure 16: Graphical illustration depicting the number of cells and reliably detected ORs in healthy and malignant breast epithelial cells.

These results indicate that the expression of the ORs is more related to the tumor cells type. The different types of ORs that are expressed in different samples like PDX, Tissue, CTC, and Normal are summarized in **Table 3**. Also, the Venn diagram representation portrays the common number of ORs in these different conditions **(Figure 17)**.

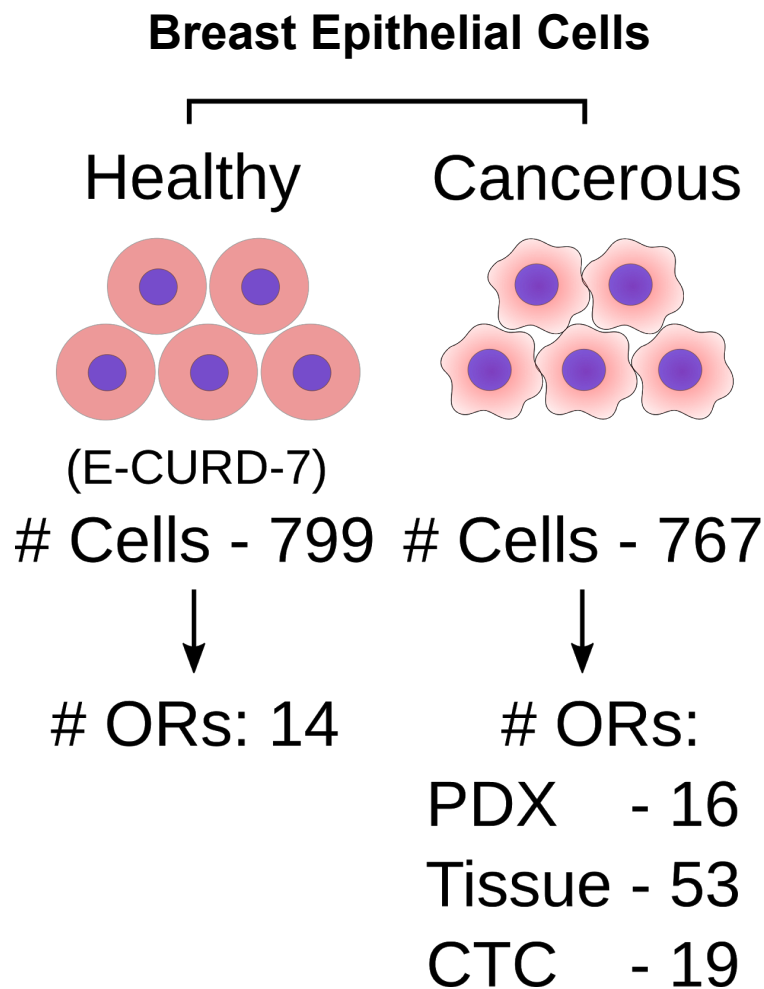Table 3: The table contains information about the different ORs expressed in different types of samples.

Sample	ORs expressed
Normal	OR10H1, OR10H5, OR14J1, OR1A1, OR1I1, OR2AT4, OR2M3, OR4D9, OR4F17, OR5A1, OR5AS1, OR6C75, OR7C1, OR7D4
PDX	OR10H1, OR10H5, OR14J1, OR1A1, OR1I1, OR2A5, OR2G6, OR2L2, OR2M3, OR2T12, OR4D9, OR4F17, OR5AS1, OR5AU1, OR6C75, OR7C1
Tissue	OR10A2, OR10G3, OR10H1, OR10H5, OR10J1, OR10K1, OR14J1, OR1A1, OR1D2, OR1I1, OR1M1, OR2A5, OR2AG1, OR2AG2, OR2AT4, OR2B11, OR2F1, OR2G6, OR2L2, OR2M3, OR2M4, OR2T12, OR2V2, OR4D1, OR4D9, OR4F17, OR4K13, OR4K17, OR4K2, OR51M1, OR52A1, OR52E4, OR52I1, OR52K1, OR56A1, OR56A4, OR5A1, OR5A2, OR5AS1, OR5AU1, OR5H14, OR6C75, OR6J1, OR6K3, OR6Y1, OR7A17, OR7C1, OR7D2, OR7D4, OR7G2, OR8A1, OR8B8, OR8H1
CTC	OR10G3, OR10H1, OR10H5, OR14J1, OR1A1, OR1I1, OR2A5, OR2AT4, OR2G6, OR2L2, OR2M3, OR2T12, OR4F17, OR52I1, OR56A4, OR5AS1, OR6C75, OR6Y1, OR7C1
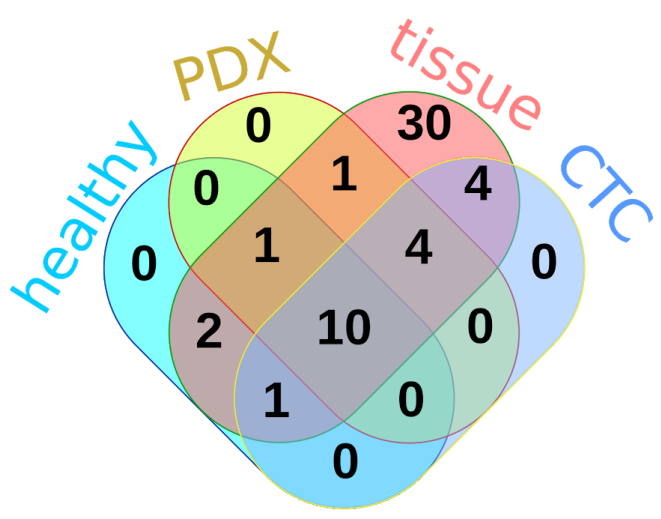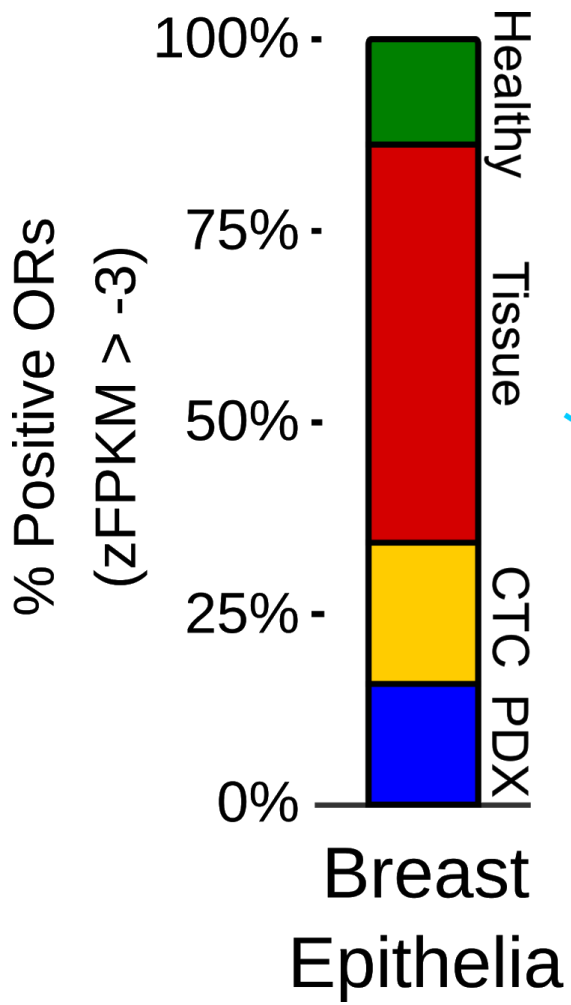
Figure 17: Percentage bar graph depicting the relative proportion of detected ORs in the indicated healthy and malignant epithelial cells. The Venn diagram on the right depicts the number of overlapping ORs in the indicated conditions.

The fourteen receptors which were common in these conditions were further inspected for the coherence of OR activation with well-known cancer related functional states. The pro-tumor signatures such as invasion, metastasis, proliferation and DNA damage show an inverse relation with the OR expression **(Figure 18)**. Some ORs are positively correlated with stemness, differentiation, and angiogenesis, implying their contrasting implications in the activation of cancer-related biological pathways.
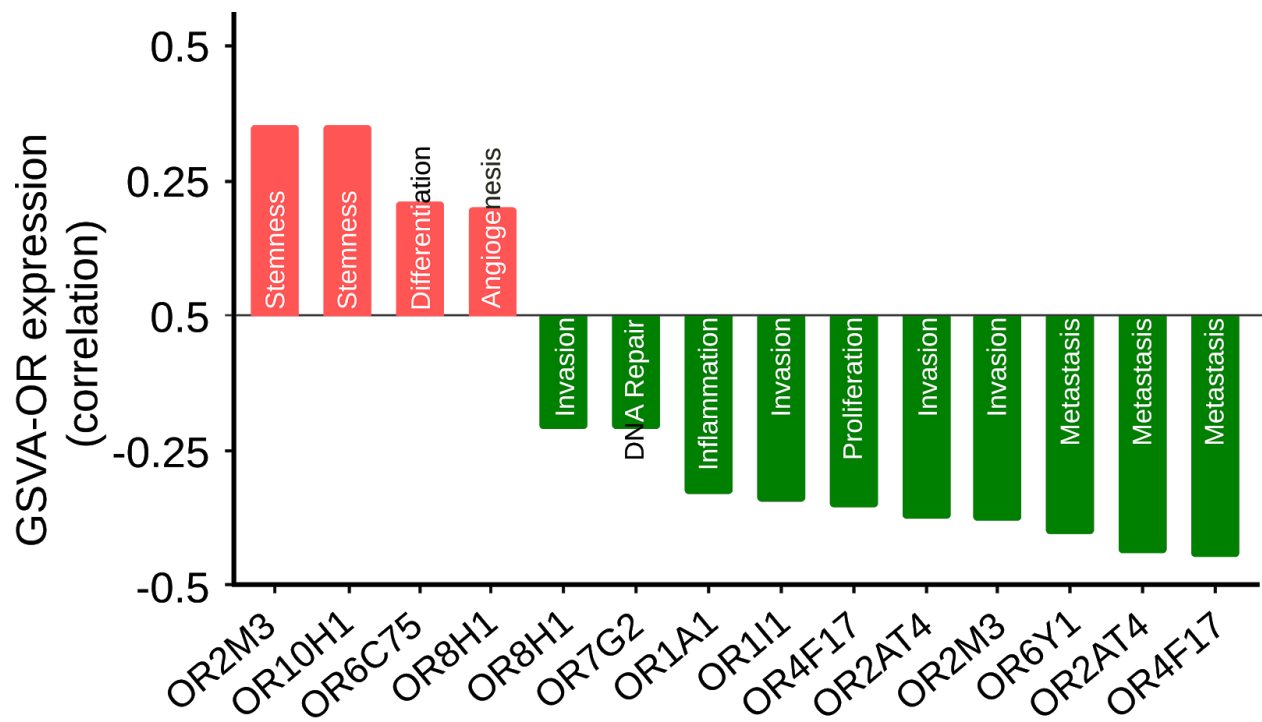
Figure 18: Bar graph depicting the correlation between the indicated ORs expression and GSVA scores.

Some previous reports have linked the OR activation with multiple non-canonical molecular processes. To check these pathways the breast cancer dataset was initially split cluster wise resulting in seven different sub-datasets. Cell wise median expression value was computed within each cluster and based on a cutoff on that median, cells were labeled as "high" and "low" (median >=10, "High" else "low") **(Figure 19)**. After that Wilcox test was performed for each cluster between the high and low category cells. P-value cutoff (< 0.05) was applied and the resulting genes that have fold change >=4 were considered as upregulated while the genes having fold change <=-4 were considered as downregulated. The entire list of upregulated and downregulated genes within each cluster is mentioned in Supplementary Table 3 **(SI_Table3)**. To access it, click on this URL
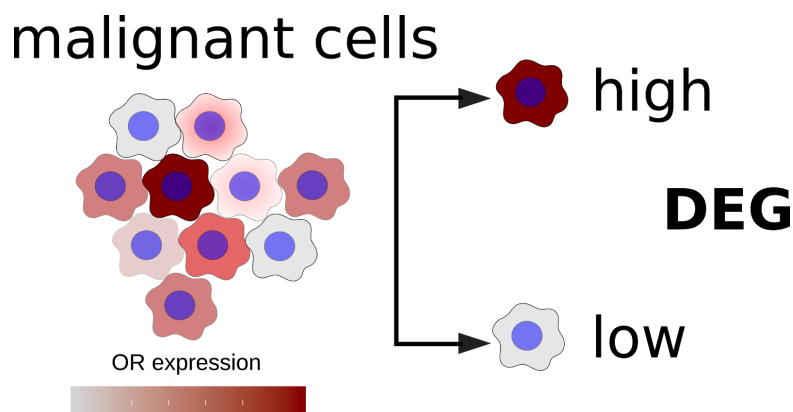
Figure 19: Schematic representation depicting the strategy employed for differential gene expression analysis, in which the malignant cells were segregated based on median OR expression.

In order to identify the genes within each cluster belongs to which pathway or protein network, metascape analysis was performed. Metascape is a free gene annotation and analysis resource which helps to identify enriched biological ontologies and pathways. Key molecular processes thus retrieved included regulation of cell cycle, transcriptional or translational regulation, autophagy etc **(Figure 20, 21)**.
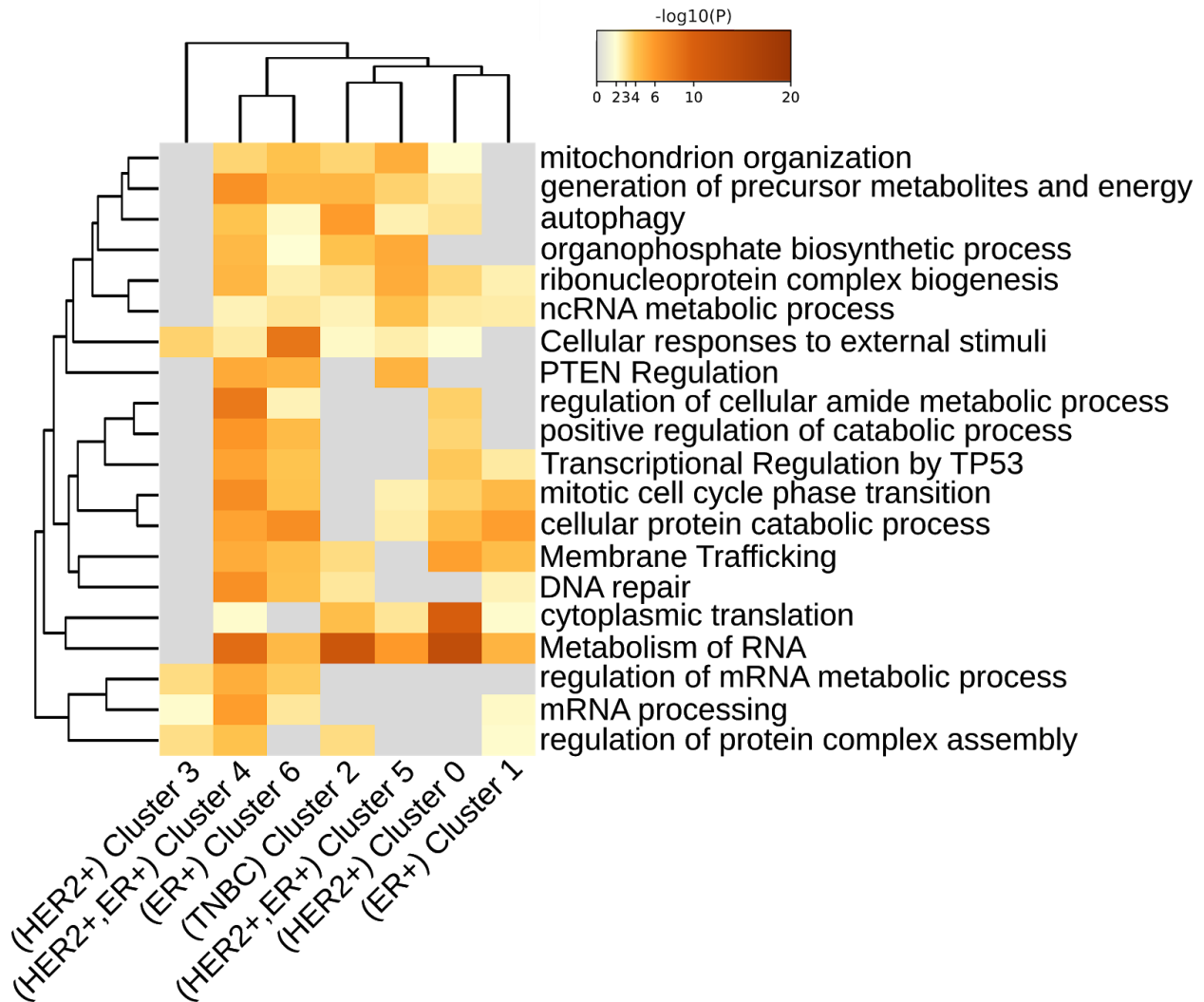
Figure 20: Heatmap depicting cluster wise enrichment of the prominent biological functions.

p-value < 0.05

Cluster 0 (HER2+)   Cluster 4 (HER2+,ER+)
Cluster 1 (ER+)   Cluster 5 (HER2+,ER+)
Cluster 2 (TNBC)   Cluster 6 (ER+)
Cluster 3 (HER2+)
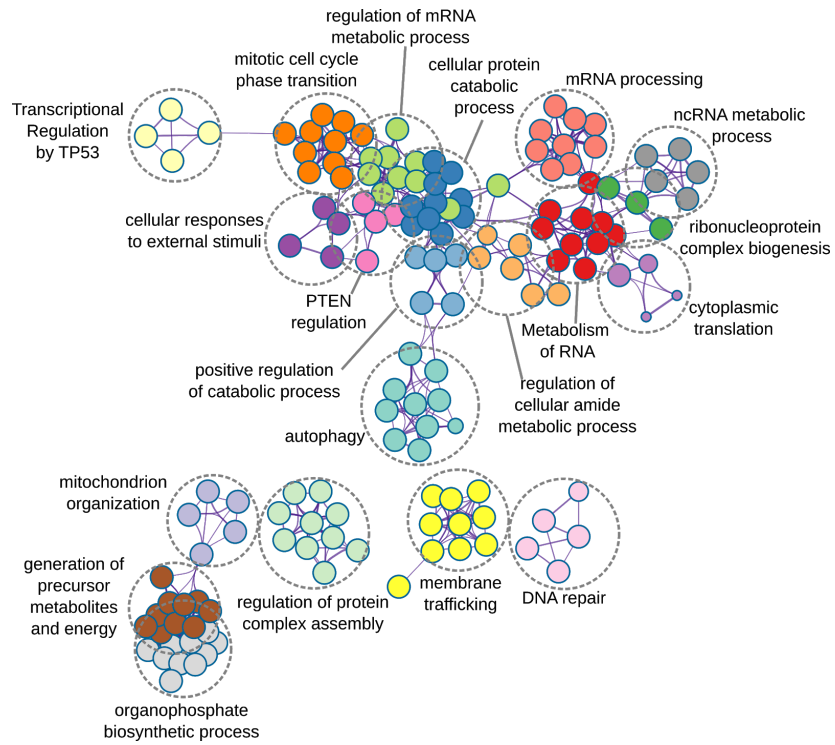
Figure 21: (Above) Representative gene ontologies depicting the functional importance of BRCA associated ORs in key biological/molecular processes. (Below) Metascape plot depicting the cluster information of the single cell breast carcinoma dataset.

Decreased cellular OR frequency and expression during cancer cell differentiation.

Cancer cells undergo extensive epigenetic reprogramming during intratumor cellular (de)differentiation. We wanted to understand the change in the cellular OR activation frequency in the context of cancer-cell differentiation. For which pseudo temporal analysis was performed on a Breast cancer dataset using Monocle. The dataset contains several patients and different molecular types. The patient labelled with BC05 was removed from the analysis since he underwent the treatment. For the pseudo temporal analysis one needs to specify the starting point to the algorithm. The starting point in this case was determined by the GSVA scores of the stemness and differentiation. The difference of stemness and differentiation for each cell was computed and those cells having the value greater than the median value of this difference were assigned to have a stemness property while others were assigned the differentiation property as it is a well known fact that stemness and differentiation are inversely proportional to each oth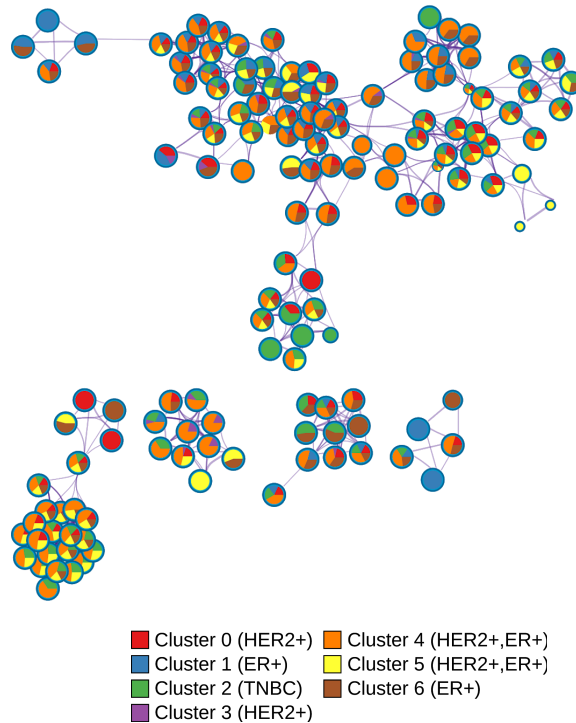er. The group of cells that have been assigned for their stemness property were further divided into three categories (low, moderate and high stemness) based on their expression. The cell that was ultimately assigned with "high stemness" was considered as the starting point for the pseudo temporal analysis. Monocle yielded three main branches elucidating the emergence of differential pathological stages entailing molecular subtypes **(Figure 22)**.
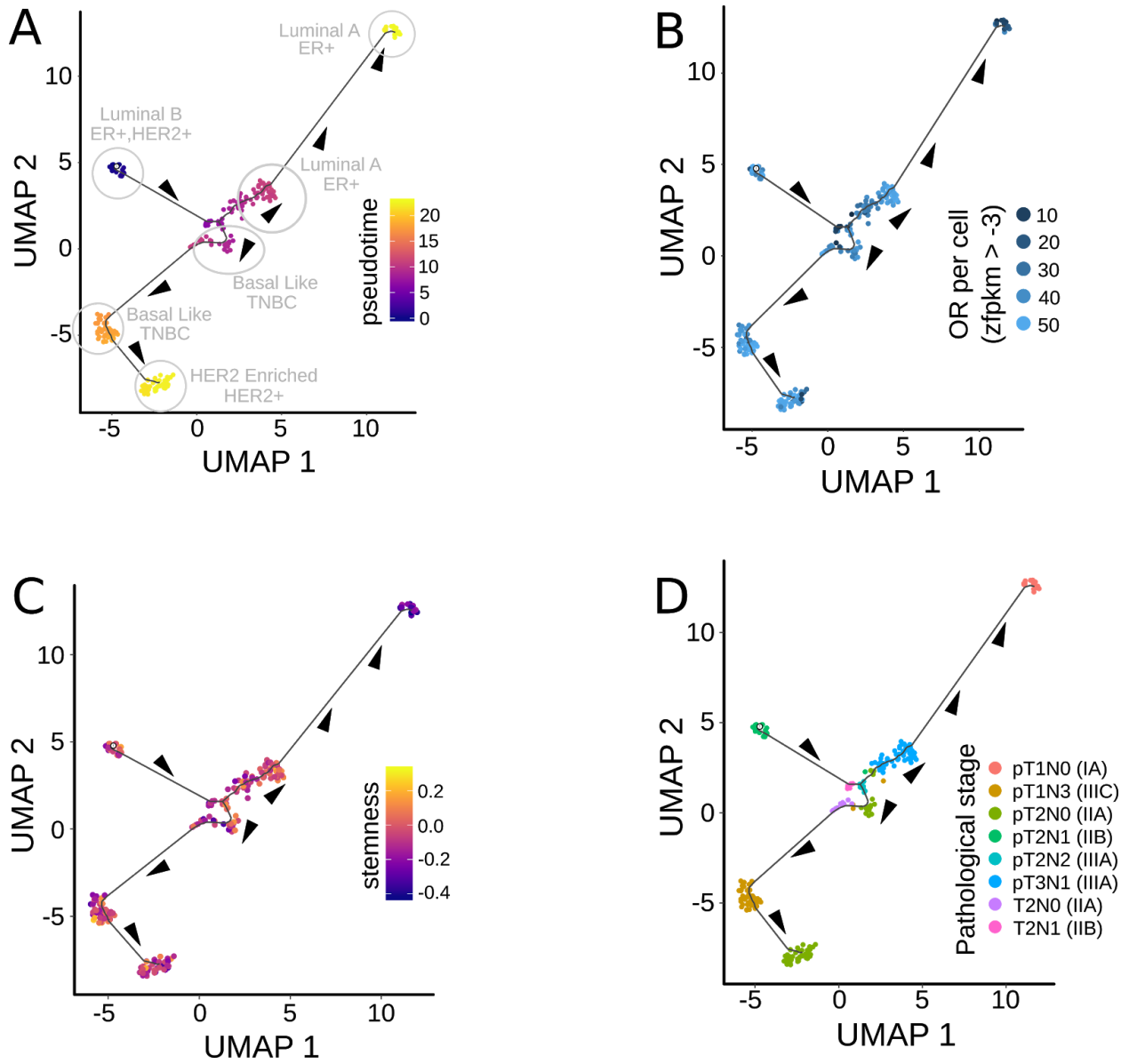
Figure 22: A) Monocle generated pseudo temporal trajectory of malignant breast epithelial cells. Arrowheads indicating the direction of the cellular differentiation. B) UMAP plot representing the number of active OR genes per malignant cell. C) UMAP plot depicting the relative cellular stemness across all malignant breast carcinoma cells. D) UMAP plot depicting the tumor stages of the cells across the pseudo temporal trajectory.

A negative correlation (R=-0.27) was observed with a high significant p value (p value<= 0.0001) between stemness and pseudotime, indicating that the cell is losing its stemness property over time. However the OR expression increases with a minor correlation (R=0.2) with a significant p value along with pseudo time. On the contrary, no significant correlation (R=-0.021, p value =0.75) was observed between the OR frequency per cell along with the pseudo time **(Figure 23)**. The exact details of the values is mentioned in Supplementary Table 4 **(SI_Table4)**, which can be accessed through the given URL.

Figure 23: A) Scatter plot depicting the correlation between cellular stemness and cell differentiation time points (pseudotime). B) Scatter plot depicting the correlation between mean expression of breast carcinoma associated ORs and cell differentiation time points (pseudotime). C) Scatter plot depicting the correlation between cellular frequencies of expressed ORs in breast carcinoma and cancer cell differentiation time points (pseudo time).

Upon a closer inspection of individual subpopulation of Luminal A cells, we observed that the number of ORs per cell was declining as the cell progressed through time. Again the cells which have high stemness value within the Luminal A population were used as a starting point for pseudo temporal analysis **(Figure 24)**. Moreover it was also observed that the OR expression decreases along with the pseudotime **(Figure 25)**.
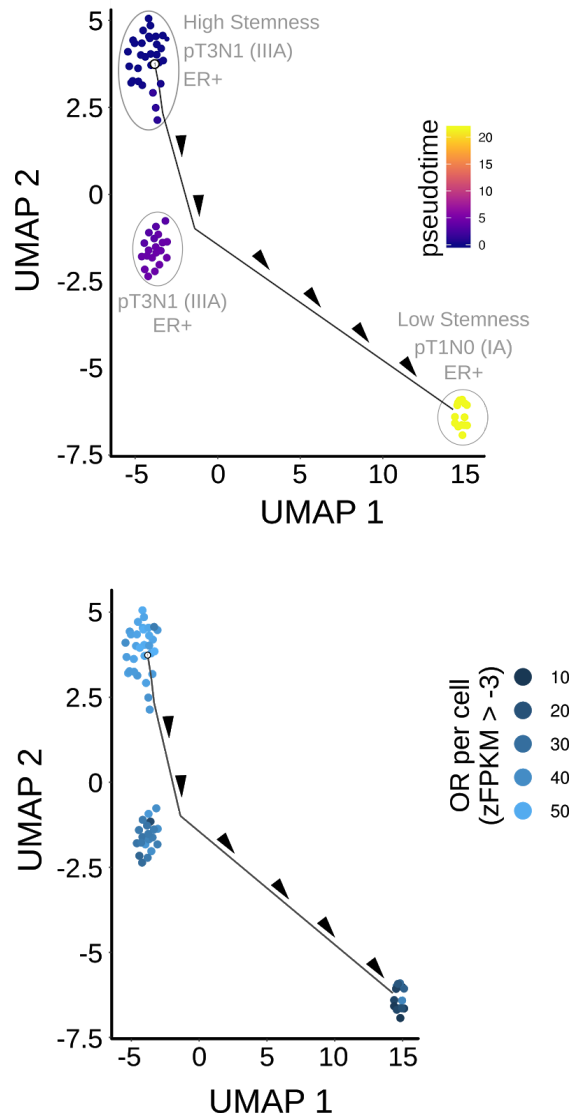
Figure 24: (Top) Monocle generated pseudo temporal trajectory of malignant breast epithelial cells of luminal A molecular subtype, depicting the decrease in cellular stemness during cellular differentiation time course. Arrowheads indicating the direction of cellular differentiation. (Bottom) UMAP represents the decrease in the number of active OR genes per malignant cell (luminal A subtype) during pseudo temporal trajectory.
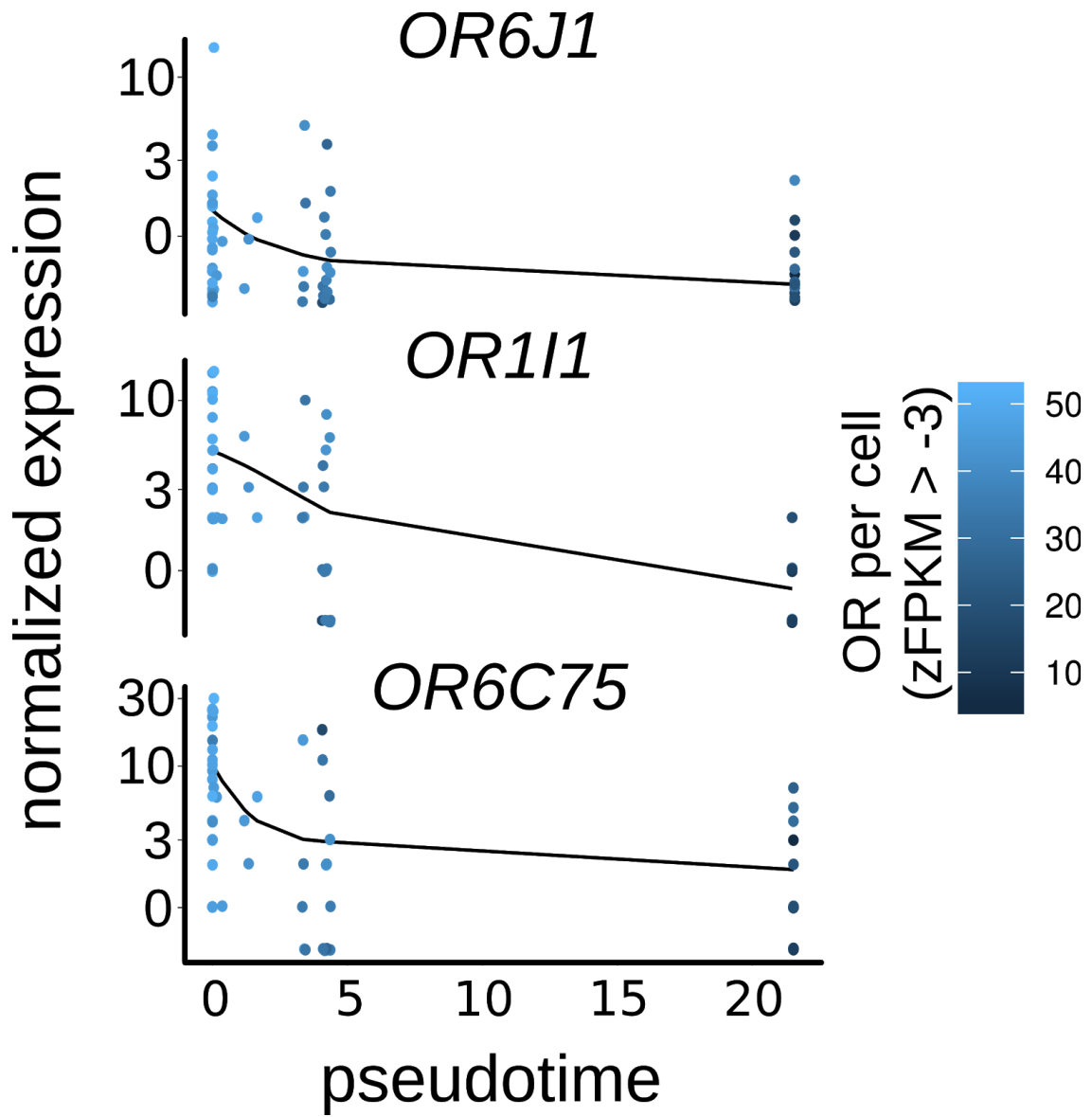
Figure 25: Scatter plots depicting an overall decrease in the expression of representative ORs along the pseudo temporal trajectory in the indicated conditions.

Similar results like that of combined molecular subtypes was observed in Luminal A subpopulation too. Cellular stemness has a negative correlation with the pseudotime (R=-0.77, p value <= 0.0001). The active OR count per cell also decreases with time (R=-0.85, p value <=0.0001). A strong positive correlation (R=0.55, p value <= 0.001) was observed between active OR count per cell and cellular stemness **(Figure 26)**.
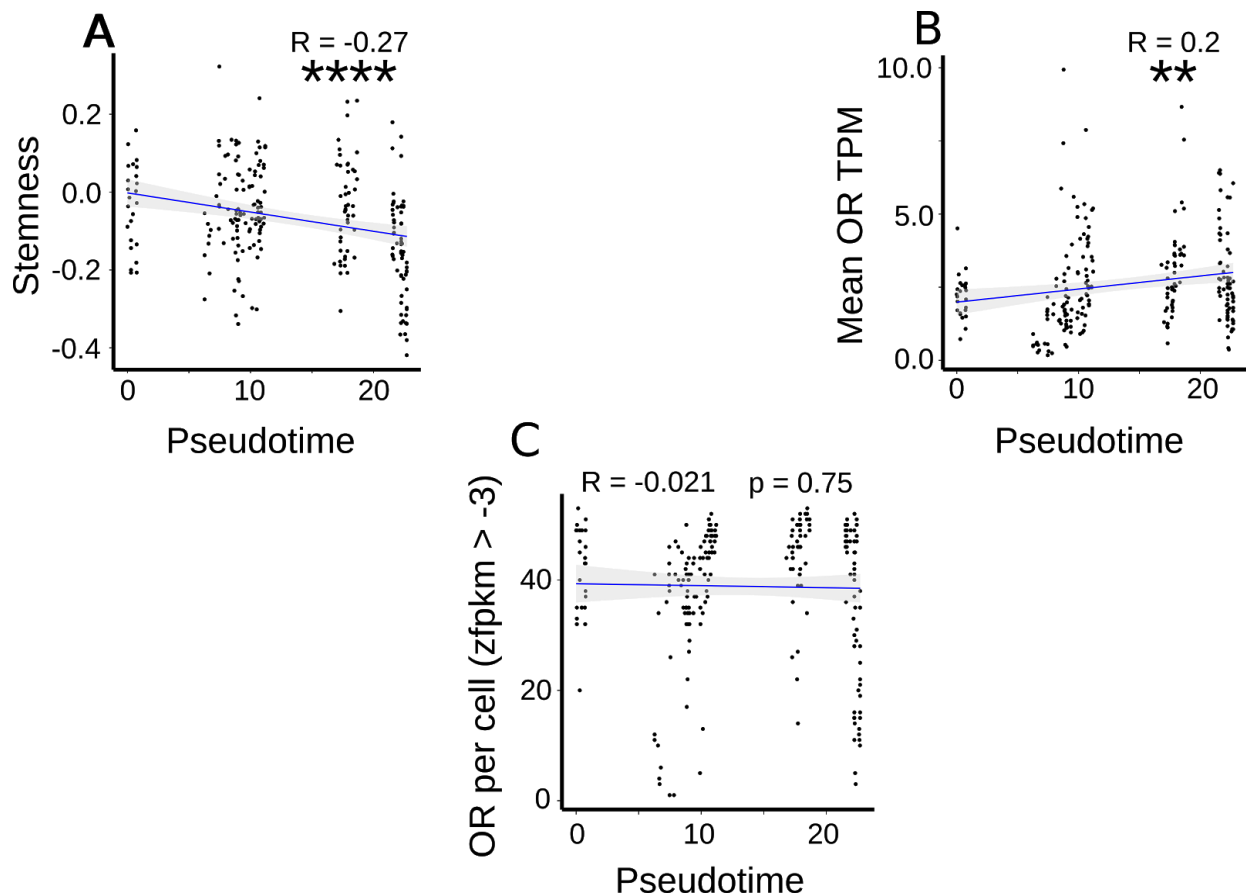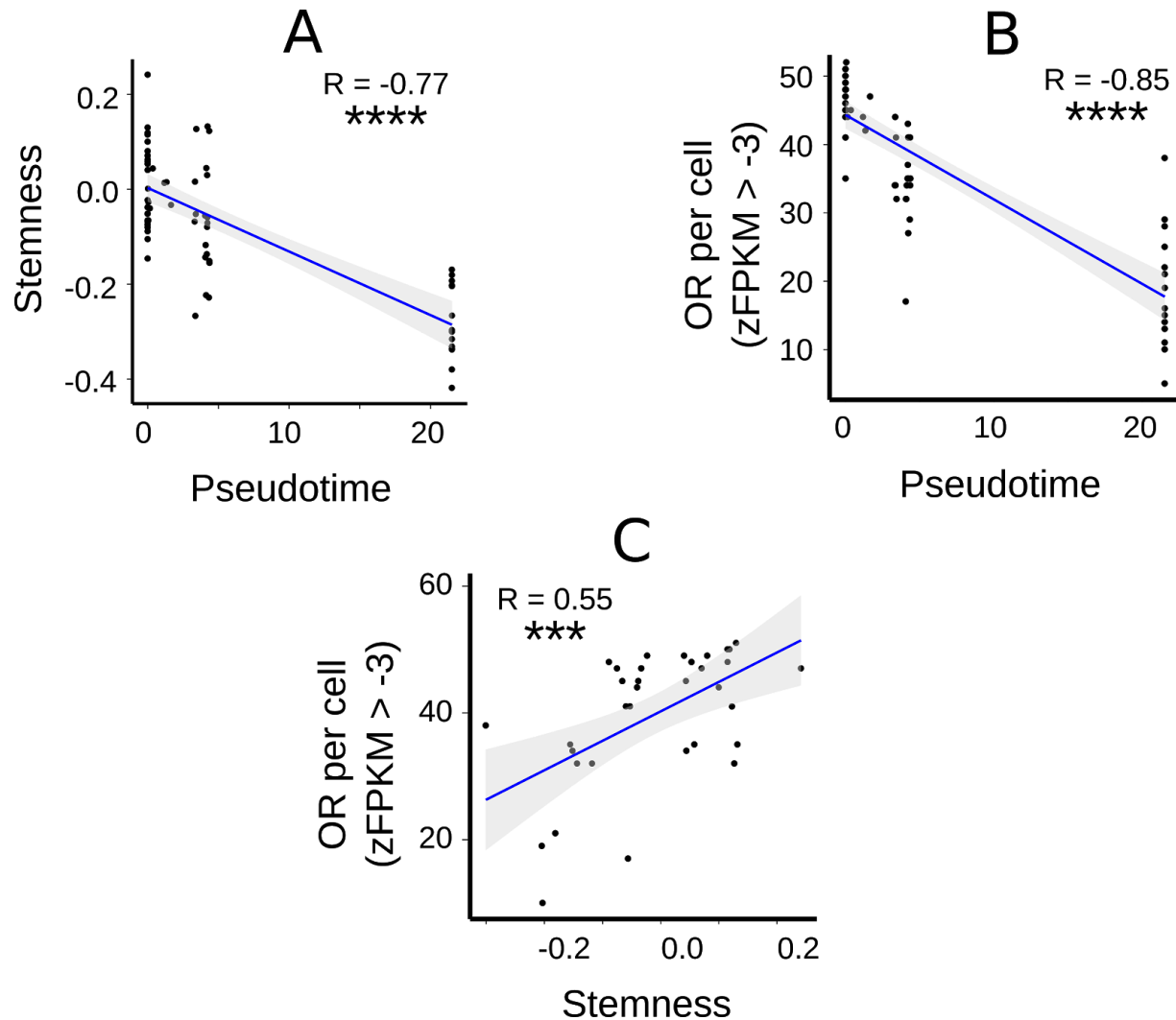

Figure 26: A) Scatter plot depicting the negative correlation between cellular stemness and cell differentiation time points (pseudotime). B) Scatter plot depicting the negative correlation between cellular frequency of expressed ORs and cancer cell differentiation time points (pseudotime) in breast carcinoma. C) Scatter plot depicting the positive correlation between cellular frequencies of expressed ORs in breast carcinoma and cellular stemness.

The exact values for Luminal A dataset is summarized in **Table 4**.

Table 4: The table contains information about the pseudo time, OR frequency, OR expression, Stemness, and Differentiation for Luminal A population of breast carcinoma dataset.

ID	Pseudotime	OR count	OR expression	Stemness	Differentiation
SRR2973279	21.52420823	38	1.262755604	-0.3006978196	-0.2718085183
SRR2973280	21.52466802	15	2.69496117	-0.4189379134	-0.208272138
SRR2973284	21.52512781	22	1.682205434	-0.2970746378	-0.0373324635
SRR2973287	21.5255876	5	0.4153102075	-0.1697904766	-0.0874410006
SRR2973288	21.5260474	29	1.781044302	-0.3316801538	-0.2185324096
SRR2973289	21.52650719	25	2.01078783	-0.3801465988	-0.1930237287
SRR2973290	21.52696698	28	6.054491698	-0.1933474445	-0.1413975152
SRR2973292	21.52742677	11	1.389538491	-0.3161421442	-0.2542201246
SRR2973293	21.52788657	10	2.197699755	-0.2030661657	-0.1832278115
SRR2973294	21.52834636	14	0.7822740943	-0.2665449842	0.007080877
SRR2973295	21.52880615	16	1.345335321	-0.338111099	-0.2097645005
SRR2973297	21.52926594	21	1.062580792	-0.1810156684	-0.1784262246
SRR2973298	21.52972574	13	1.06459617	-0.3365720005	-0.1993824679
SRR2973299	21.53018553	19	2.144369057	-0.204406583	-0.159057308
SRR5023386	0.0032185463	47	4.156377264	0.2411531233	0.132067811
SRR5023387	3.356580927	44	4.154437736	0.0156313517	0.2768252767
SRR5023388	3.4106078	41	2.145221472	-0.0527415046	-0.0403392913
SRR5023389	0.002758754	50	3.032395396	0.1184395138	0.0757643182
SRR5023390	1.662364479	47	2.105459	-0.0333220896	-0.0763218935
SRR5023391	4.213850242	43	1.353116472	0.029417827	0.1134154646
SRR5023392	0.0022989617	45	2.524795226	-0.0807995063	0.1172134742
SRR5023393	0.0018391693	48	3.535033377	0.0530772053	-0.1354238765
SRR5023394	4.237871432	27	0.4634990566	-0.0713338064	0.0673235725
SRR5023395	0.001379377	50	3.902730887	0.1149667417	0.0917927866
SRR5023396	0.0009195847	44	5.338857358	0.0999417997	0.0206250219
SRR5023397	4.37780286	35	9.936629943	-0.1554943662	-0.2158563866

SRR5023398	0.0004597923	48	5.192682906	0.1150384687	0.0749376193
SRR5023399	0	49	5.141005585	-0.0659107522	0.0623140501
SRR5023400	0.0004597923	47	2.220803962	-0.0750384848	-0.0593318122
SRR5023402	0.0009195847	48	4.558048453	-0.0520571023	0.2115152413
SRR5023403	4.086356501	32	1.639626075	-0.1437164589	-0.1393351317
SRR5023405	0.0334593171	48	3.406396604	0.0010347729	0.2088392925
SRR5023406	0.1582670759	44	2.642164811	-0.0404887779	-0.1521410486
SRR5023407	4.069696868	17	1.687739566	-0.0561387651	-0.0485229059
SRR5023408	0.001379377	45	1.918187472	-0.0383993142	-0.1766806644
SRR5023409	0.0514998212	45	2.268365906	-0.0658691939	-0.1206590429
SRR5023411	0.0018391693	46	1.874967943	-0.1463098019	0.0124698787
SRR5023412	1.344039539	42	4.897772226	0.0146105035	0.180211142
SRR5023414	4.23622963	41	3.154832038	-0.0595766464	-0.1188202918
SRR5023415	4.338281191	29	1.564337943	-0.2284859606	0.1544628259
SRR5023416	3.446066188	32	1.982048396	0.1266077948	0.0488244169
SRR5023417	4.105811226	32	2.203877698	-0.1179450225	-0.1129560538
SRR5023418	4.212930658	37	5.58864483	-0.0797662411	-0.0061928967
SRR5023419	4.139446157	34	1.177685151	-0.2235924537	-0.0945600916
SRR5023420	0.0022989617	49	4.228366358	0.080074827	0.0614290192
SRR5023421	0.002758754	51	3.106094755	0.1297893918	0.0901014721
SRR5023422	0.0032185463	47	2.484076547	0.0705420991	-0.0965974244
SRR5023423	4.24687931	35	1.023866962	0.1320436722	-0.0259302083
SRR5023424	0.3679384298	45	1.800391491	0.0435302512	0.0853074303
SRR5023425	0.0036783387	48	2.046605566	-0.0889389722	-0.1009217412
SRR5023426	0.004138131	49	2.459066528	0.040255781	-0.0216898736
SRR5023427	0.0573363477	52	4.013252302	-0.0274742433	0.052938118
SRR5023428	4.357820759	41	2.698842434	0.1225489776	-0.0923954202
SRR5023429	4.156033629	34	1.12003417	0.0442778502	0.0910836368
SRR5023431	0.0045979233	41	2.72307217	-0.1053835443	0.2274358539
SRR5023432	0.0050577157	51	7.876948943	-0.051784717	0.0176206823
SRR5023433	0.005517508	49	4.860666075	-0.0231976966	-0.0038663848

SRR5023434	0.0059773003	35	3.015584887	0.0580084557	-0.0219484917
SRR5023435	3.377270725	34	0.9536371321	-0.267236402	-0.027047518
SRR5023436	4.368055683	34	1.147200981	-0.1515201875	-0.2011232742
SRR5023437	0.0064370927	48	3.536559811	0.0621700036	0.137766066
SRR5023438	1.185386416	44	5.243386962	0.0129800824	0.1091095463
SRR5023439	4.19117707	35	1.355108642	-0.1371597088	0.0681697722
SRR5023440	0.006896885	45	1.918120566	-0.0691338279	0.2507971825
SRR5023441	3.333083023	34	2.054601264	-0.0682415866	0.1090918678

In order to check that such decline is also applicable to healthy individuals or not, the same pseudo temporal analysis was performed on a healthy dataset. The healthy luminal cells exhibit no such correlations **(Figure 27, 28)**. No significant relation was found in healthy luminal cells for active ORs per cell and pseudotime. (R=0.048 , p value= 0.4). Neither any significant relation was observed between the active ORs per cell and stemness. (R=0.026, p value= 0.65) **(Figure 29)**.
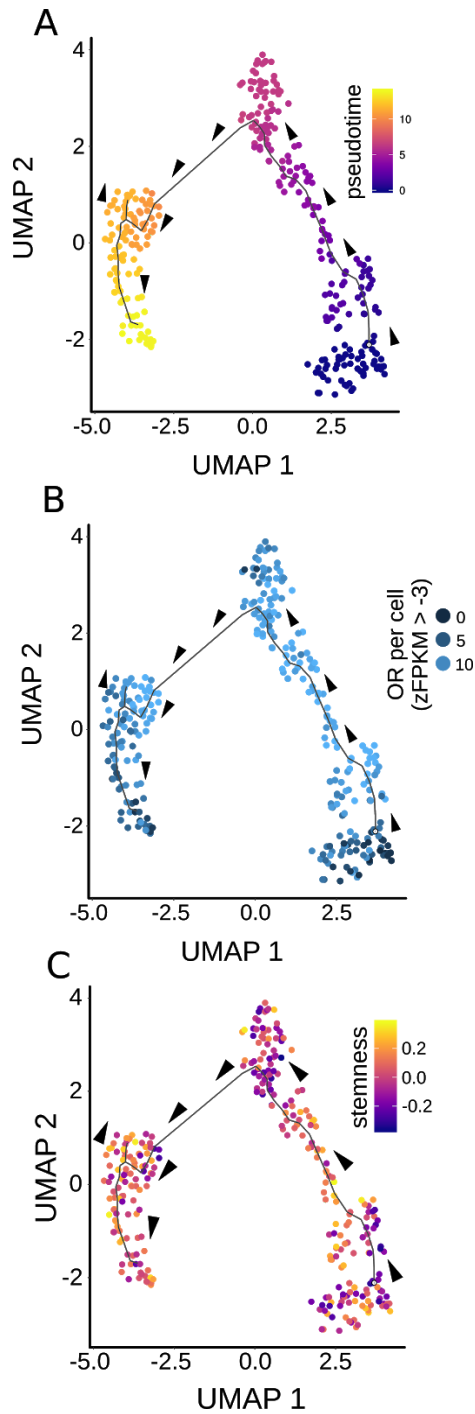
Figure 27: A) Pseudo Temporal trajectory of healthy luminal breast epithelial cells. Arrowheads indicating the direction of cellular differentiation. B) UMAP represents the number of active OR genes in healthy luminal breast epithelial cells during pseudo temporal trajectory. C) UMAP depicts cellular stemness along the pesudotemporal trajectory in the healthy luminal breast epithelial cells.
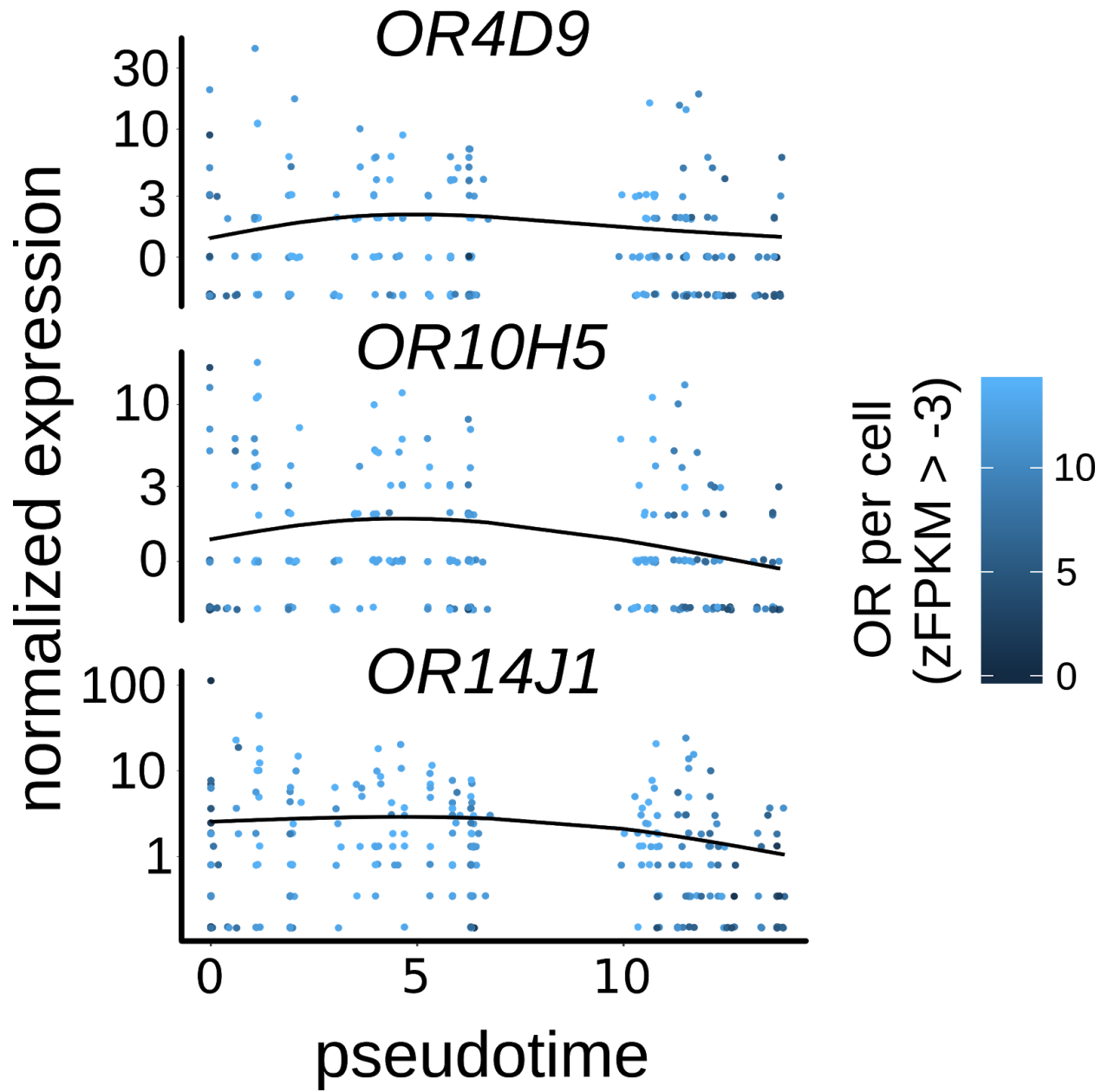
Figure 28: Scatter plots depicting expression dynamics of representative ORs of healthy luminal cells along the pseudo temporal trajectory in the indicated conditions.
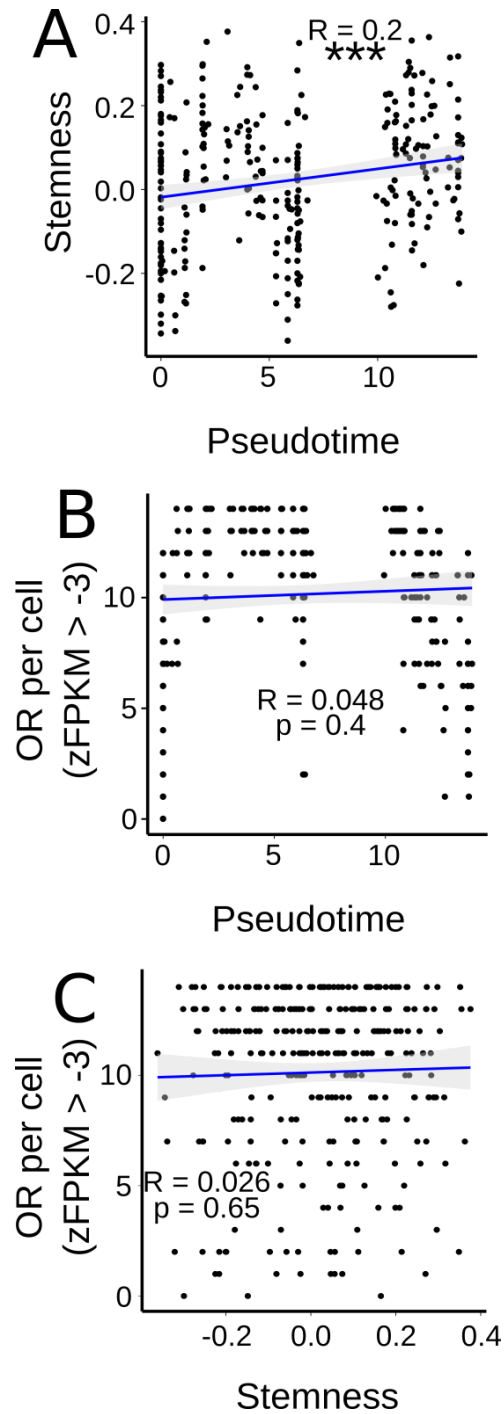
Figure 29: A) Scatter plot depicting the correlation between cellular stemness and pseudo time in healthy breast epithelial dataset. B) Scatter plot depicting the correlation between cellular frequencies of expressed ORs and pseudo time in healthy breast epithelial dataset. C) Scatter plot depicting the correlation between cellular frequencies of the expressed ORs and cellular stemness in breast epithelial dataset.

Cellular OR frequency stratify BRCA tumors into subgroups with divergent patient survival.

The tumor environment harbors multiple cell types, thereby obscuring the detectability of cancer specific ORs through bulk transcriptomics. In order to check the overall decline in the cellular OR number/expression during cancer cell differentiation, we asked if multivariate OR signatures can be used to deconvolute the bulk expression profiles and could segregate patients with distinct survival. First, the Breast cancer bulk dataset was obtained from TCGA and it's corresponding zFPKM score was calculated. This zFPKM matrix was converted into a binary matrix with -3 as cutoff, where =< -3 is 0 and otherwise is 1. This binary matrix was subjected to clustering **(Figure 30-32)**.

Figure 30: Heatmap, along with the cellular clustering (columns) depicting the segregation of cells based on OR expression status in breast carcinoma (tissue derived single cell dataset).

Figure 31: Heatmap, along with the cellular clustering (columns) depicting the segregation of cells based on OR expression status in breast carcinoma single cell dataset of circulating tumor cells of breast carcinoma.
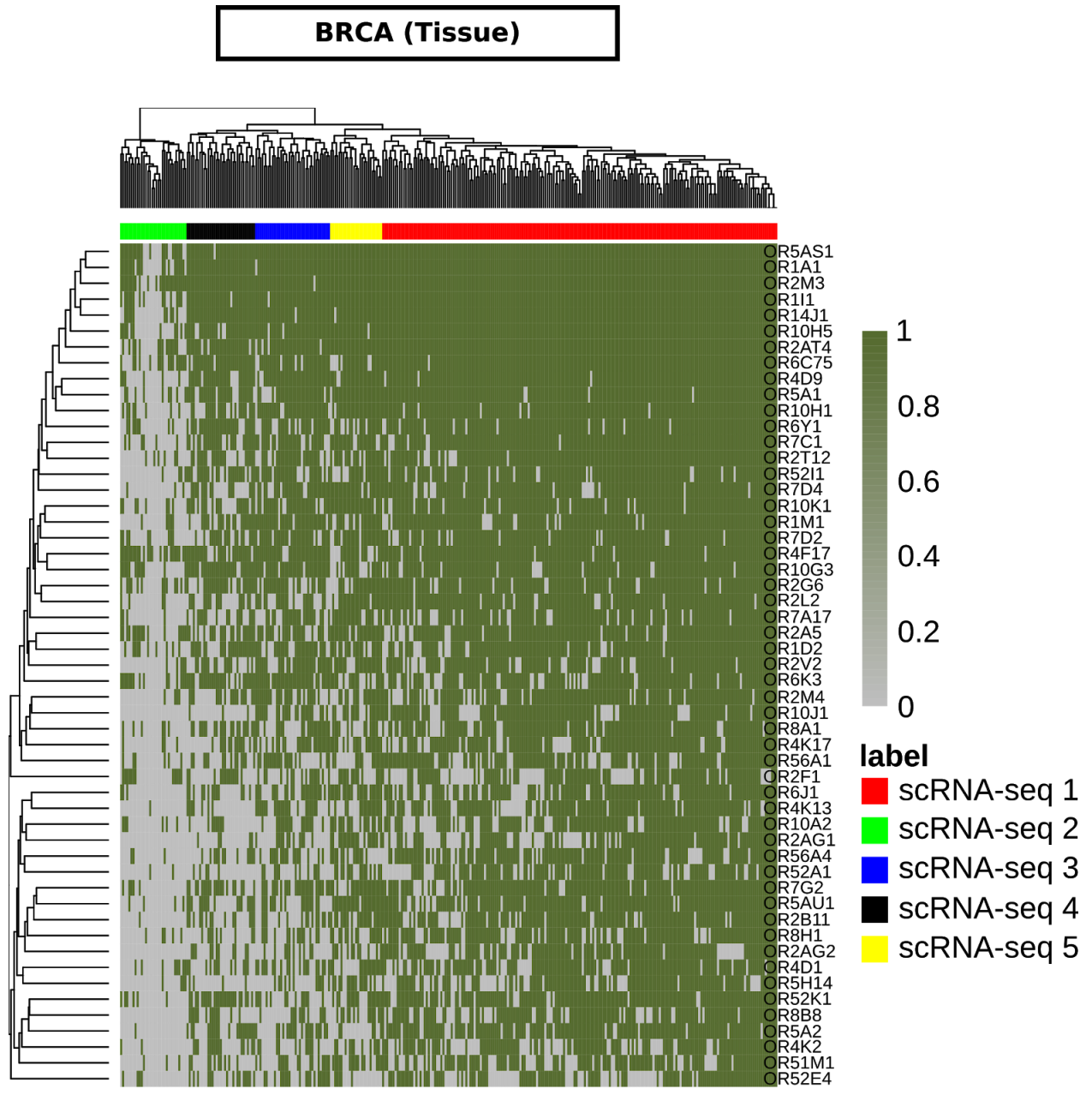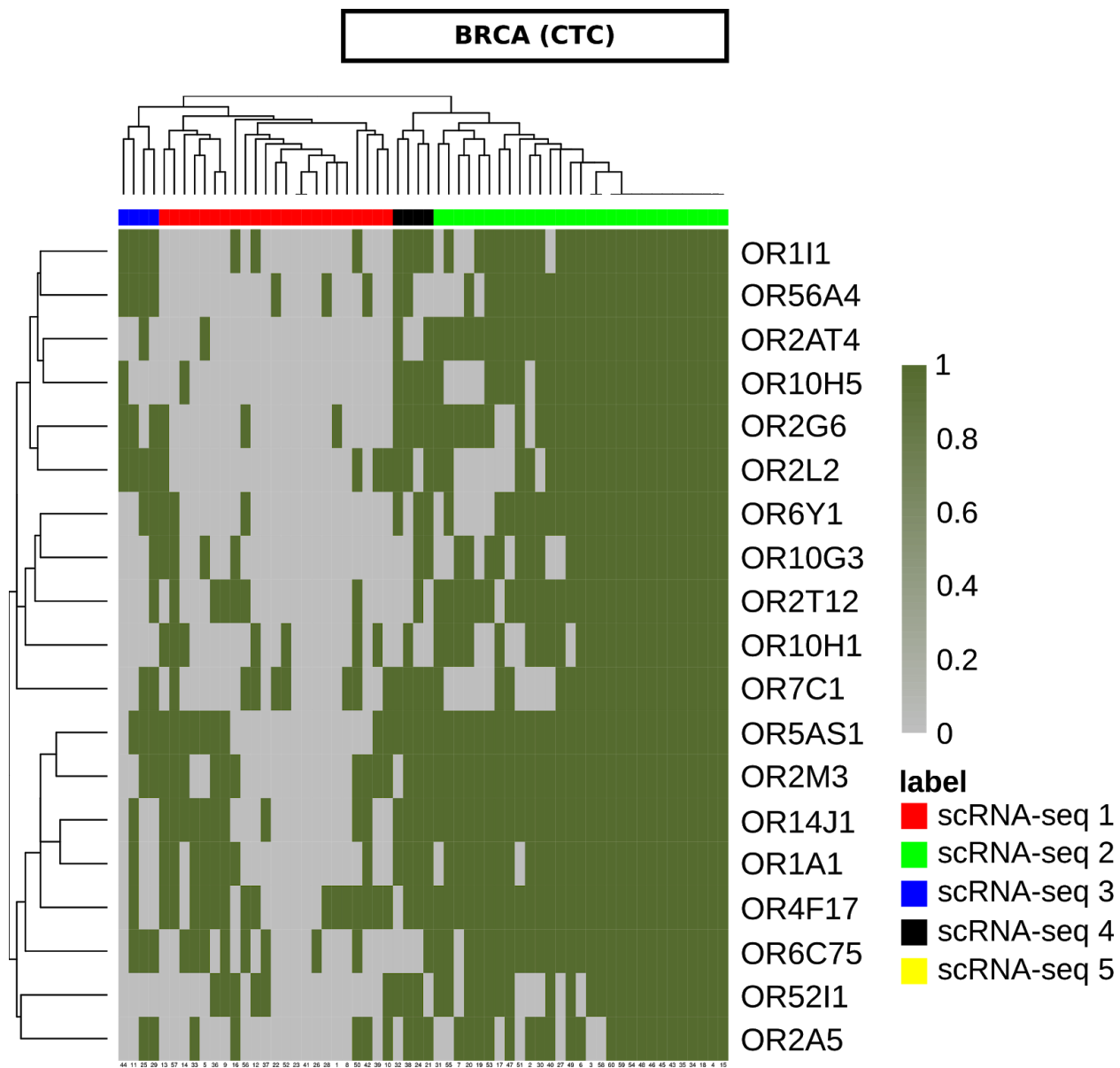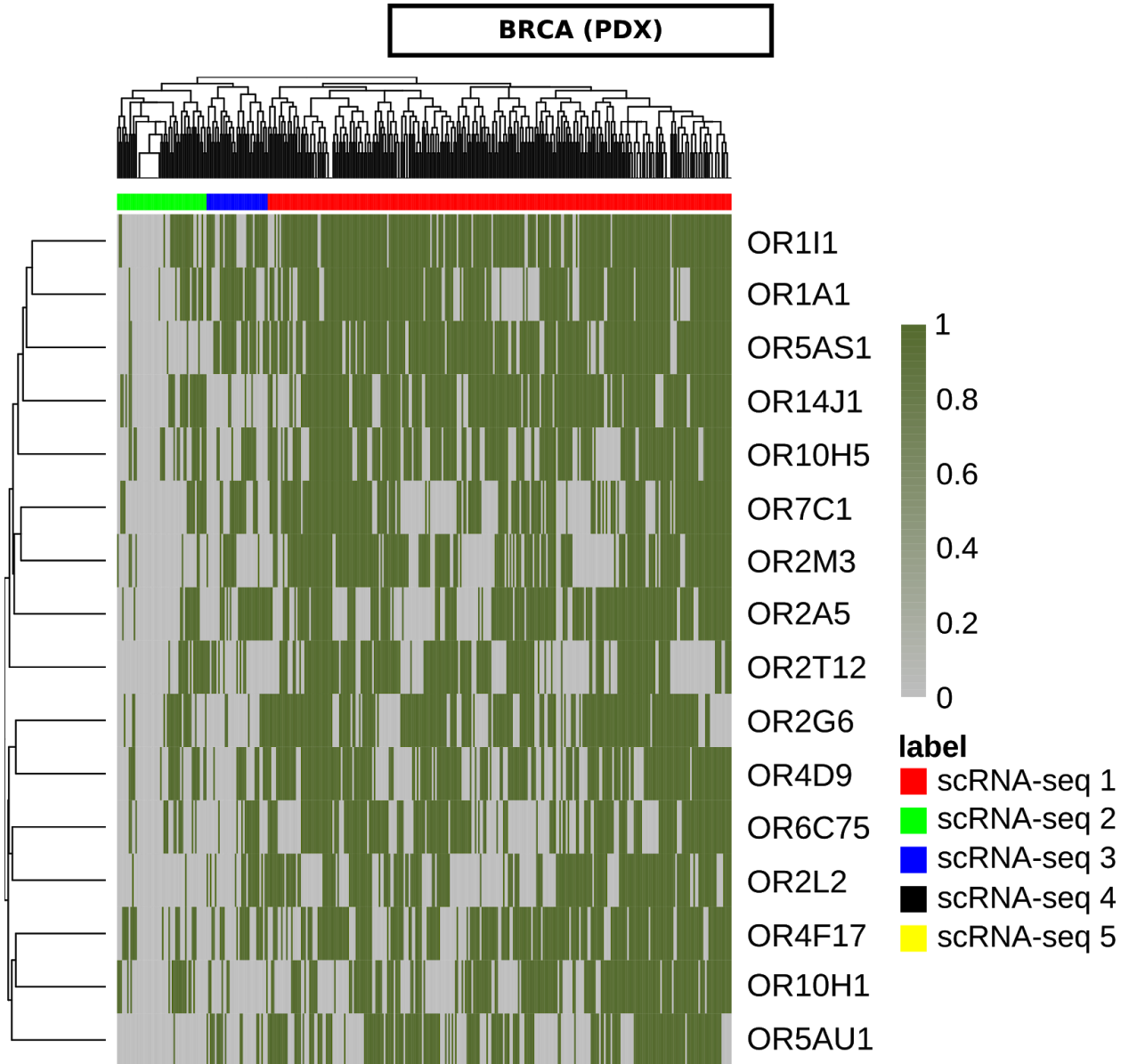
Figure 32: Heatmap, along with the cellular clustering (columns) depicting the segregation of cells based on OR expression status in breast carcinoma single cell dataset of patient derived xenografts of breast carcinoma.

Once this clustering was done, the OR centric signatures were created by taking the average of each cluster sample. These signatures were deconvoluted on the TCGA data **(Figure 33-35)**.

Figure 33: Heatmap depicting the average OR five different signatures representing distinct enrichment of tumor related ORs in breast carcinoma (tissue derived single cell dataset).
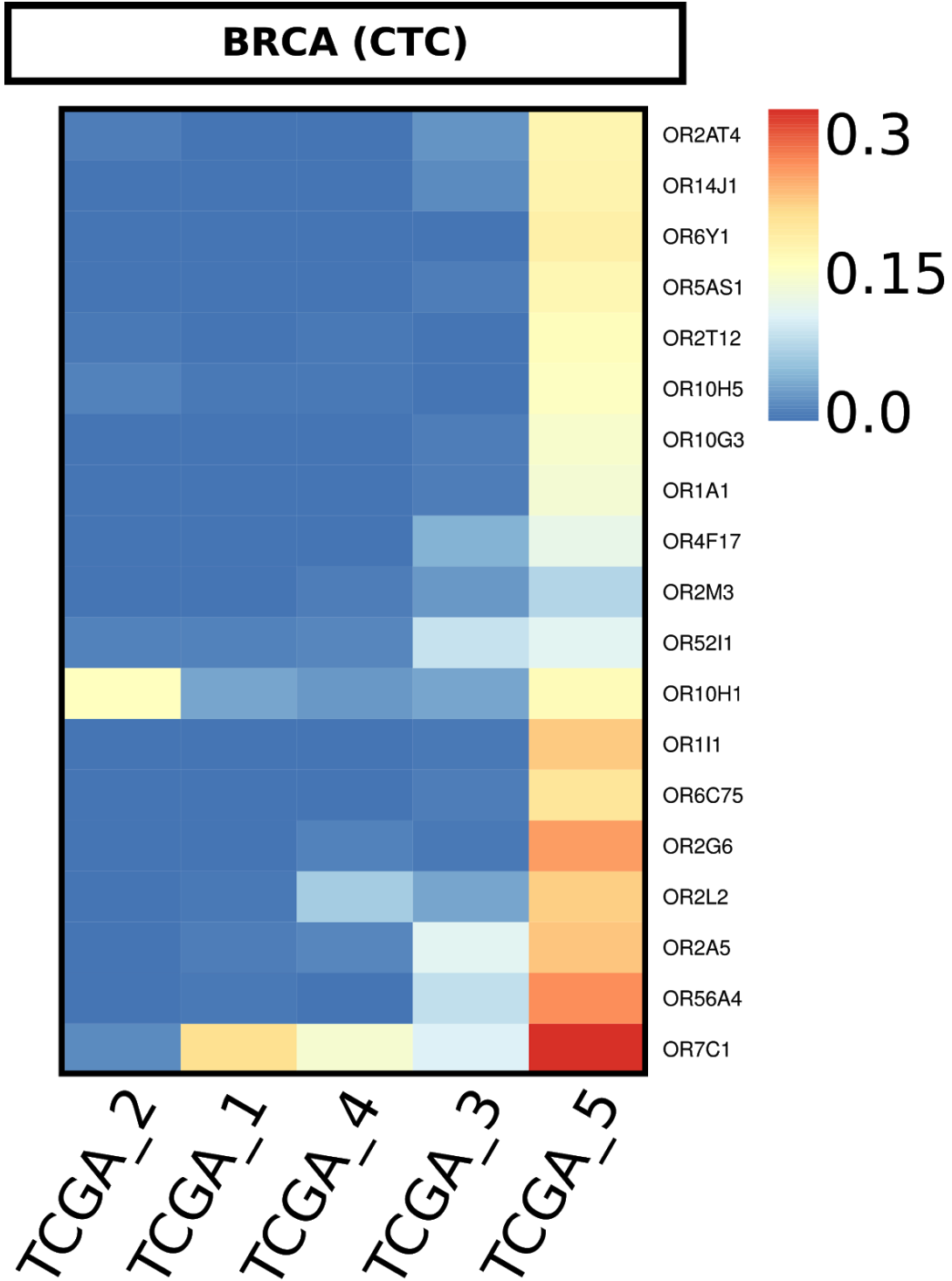
Figure 34: Heatmap depicting the average OR five different signatures representing distinct enrichment of tumor related ORs in breast carcinoma (circulating tumor cells of breast carcinoma).
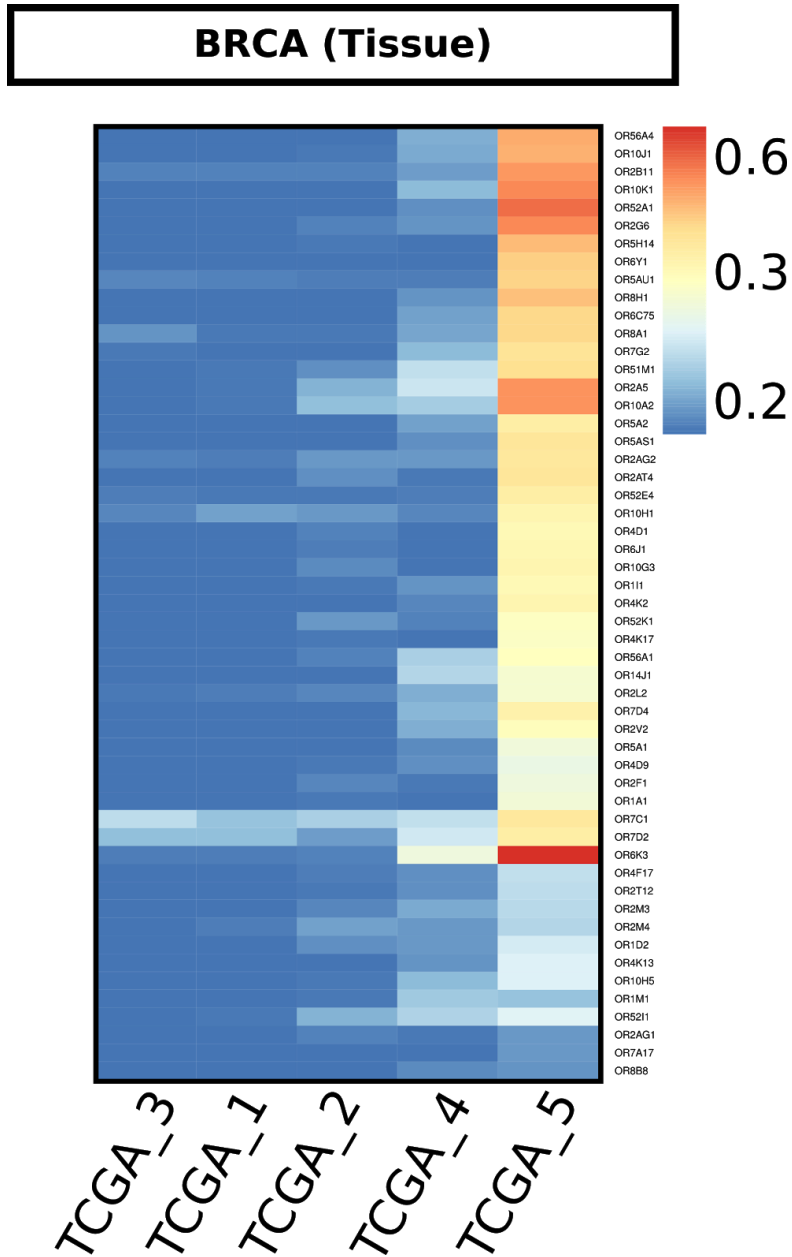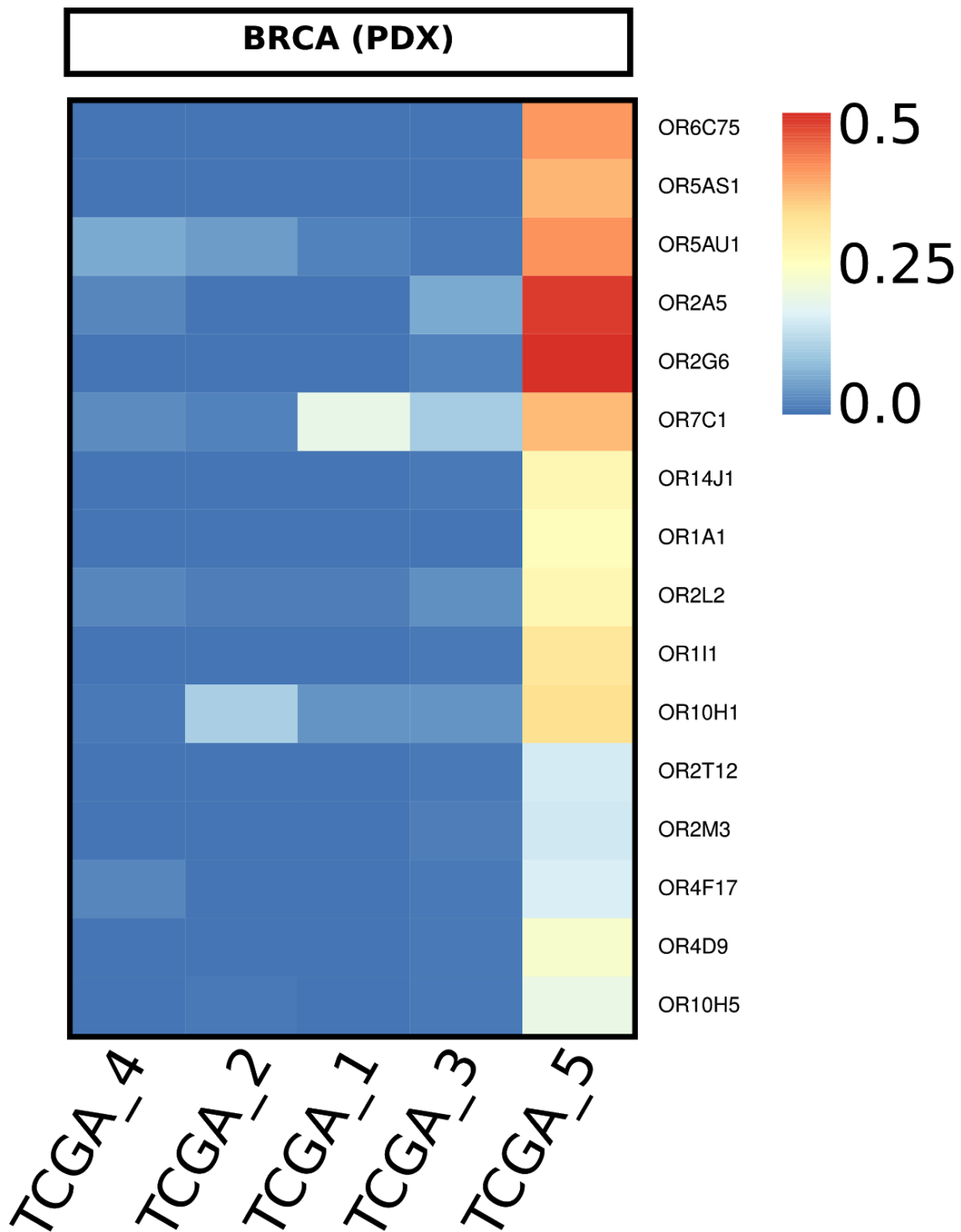
Figure 35: Heatmap depicting the average OR five different signatures representing distinct enrichment of tumor related ORs in breast carcinoma (patient derived xenografts of breast carcinoma).

The cosine distance with the OR-centric signatures representing the low number of expressed OR genes per cell harbors a good prognosis and better survival **(Figure 36)**.
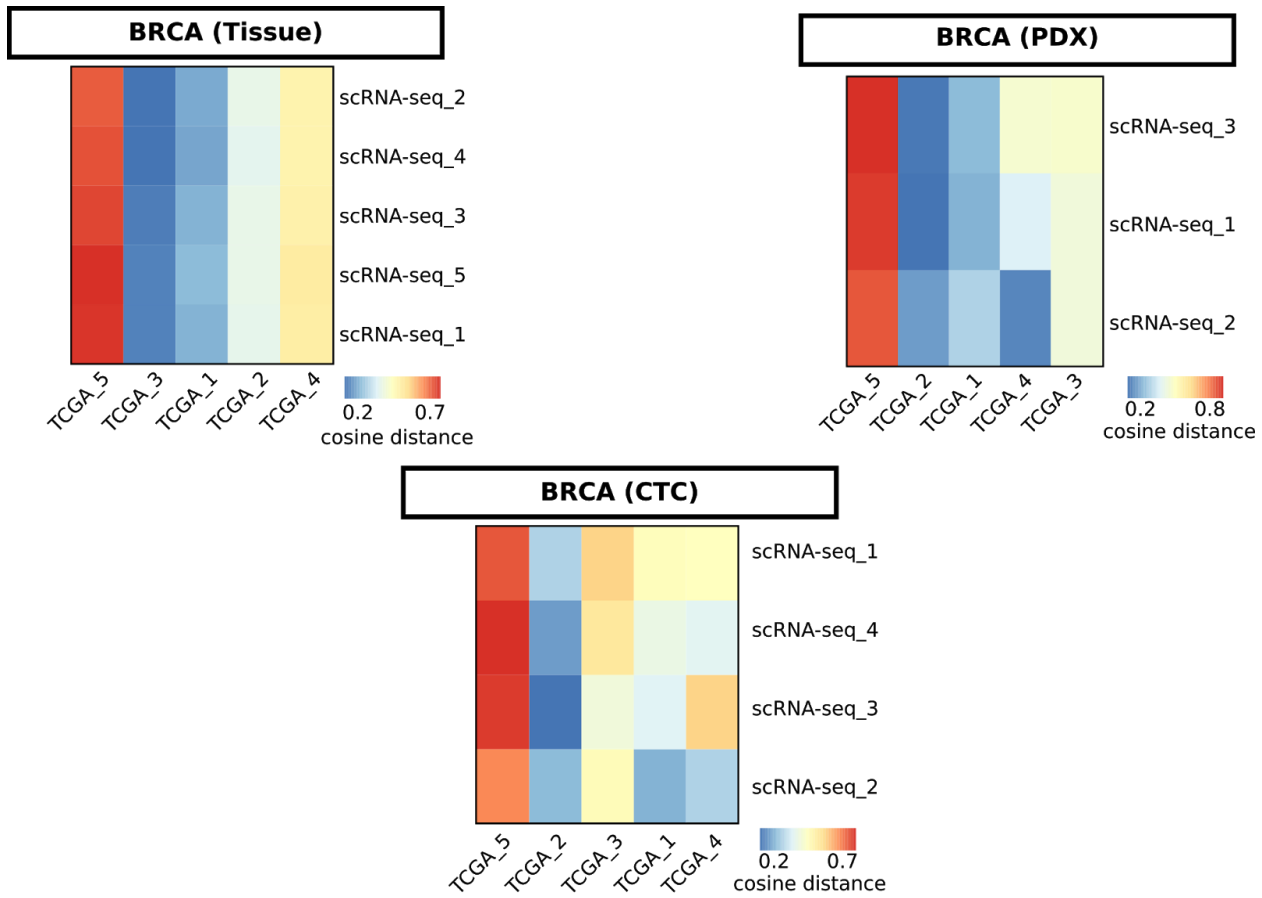
Figure 36: Heatmap depicting the cosine similarity between the scRNA sequencing datasets derived OR centric signatures and stratified TCGA patients cohorts.

The OR centric signatures were projected on TCGA bulk transcriptomics breast carcinoma data, annotated with survival information and obtained multiple patient groups, with at least two major groups having significantly distinct survival probabilities **(Figure 37-39)**.

Figure 37: (Top) Kaplan-Meier plot depicting patient's survival in the indicated group segregated based on cosine similarity towards OR enrichment signatures, inferred from tissue derived scRNA sequencing profiling of malignant cells. (Bottom) Comparison of the tumor stages between two prominent clusters representing TCGA patients cohorts with distinct survival.

BRCA (CTC)

TCGA
1 — (n=702, median survival=131.4)
2 — (n=204, median survival=98.8)
3 — (n=19, median survival=46.3)
4 — (n=150, median survival=131.9)
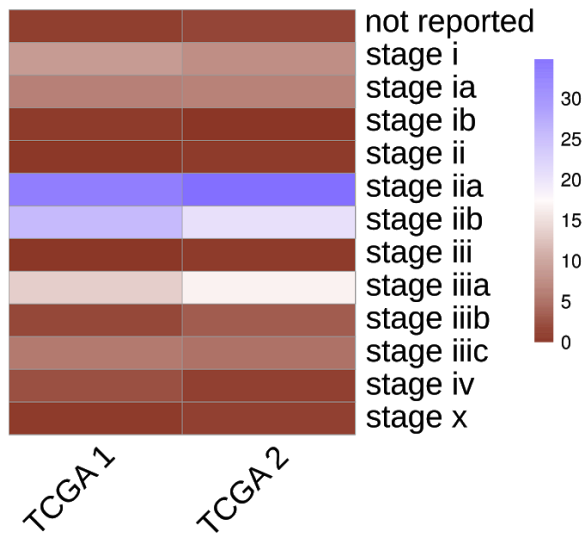
Figure 38: (Top) Kaplan-Meier plot depicting patient's survival in the indicated group segregated based on cosine similarity towards OR enrichment signatures, inferred from single cell transcriptional profiling of circulating tumor cells. (Bottom) Comparison of the tumor stages between two prominent clusters representing TCGA patients cohorts with distinct survival.
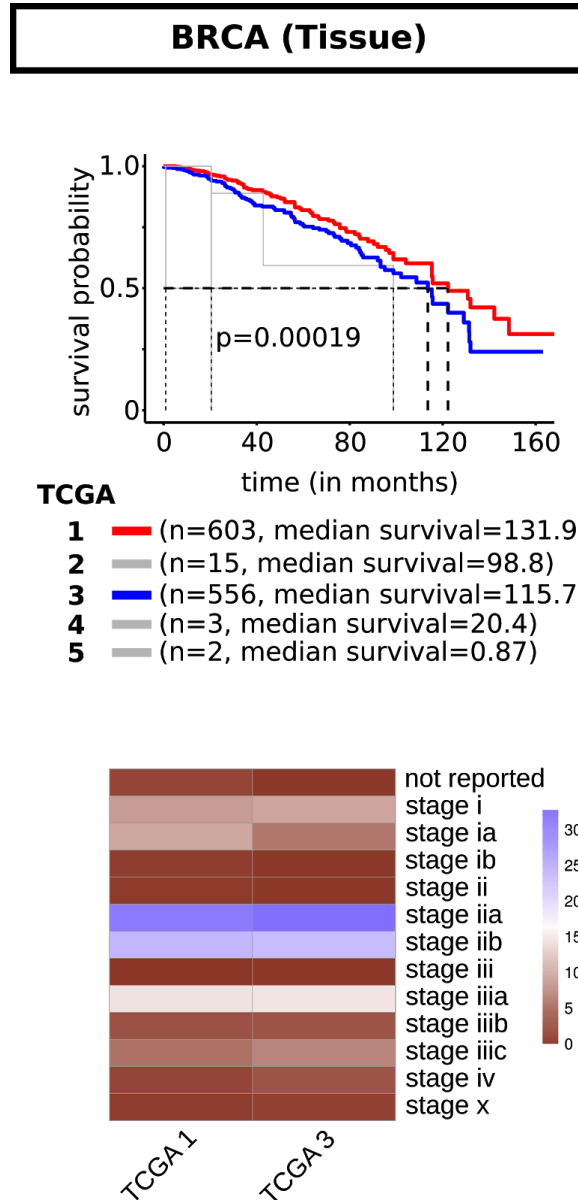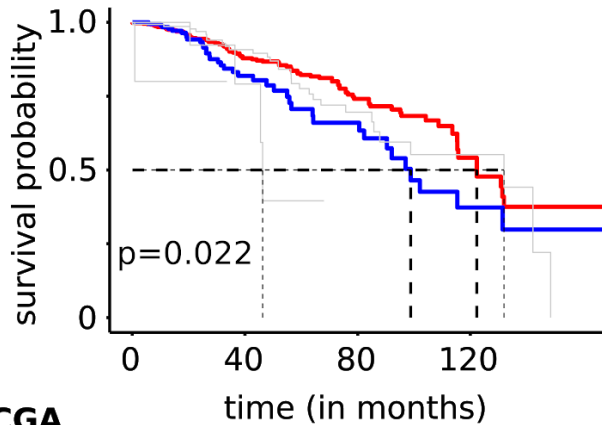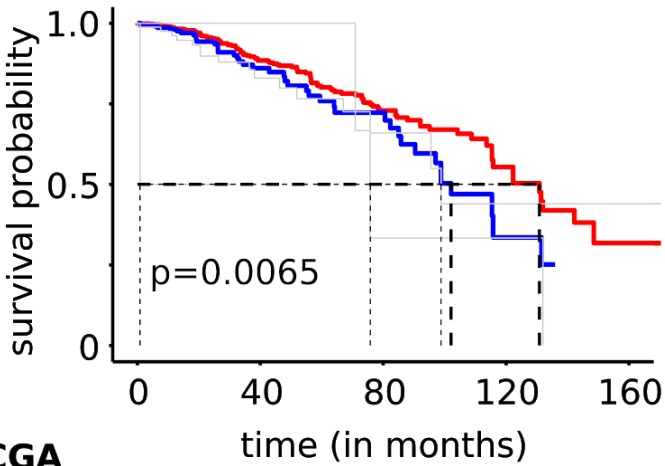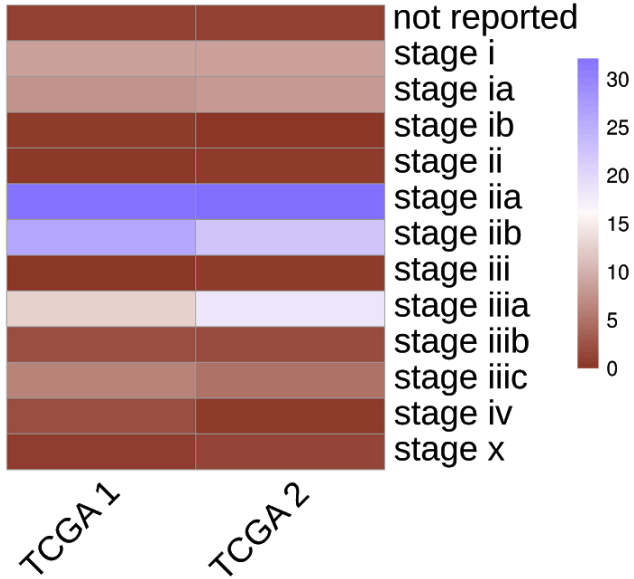
Figure 39: (Top) Kaplan-Meier plot depicting patient's survival in the indicated group segregated based on cosine similarity towards OR enrichment signatures, inferred from transcriptomic profiling of breast carcinoma patient derived xenograft dataset. (Bottom) Comparison of the tumor stages between two prominent clusters representing TCGA patients cohorts with distinct survival.

In clinical practice the Nottingham grading system is used for the assessment of the breast cancer stage. Our results suggest that the patients with the same tumor grading can further be distinguished on the basis of the expression of the ORs. Patient groups possessing the higher cosine distance with the OR centric signatures with lower cellular OR frequencies possess a good prognosis for better survival.

Discussion

Dysregulation in the cancer environment results in the ectopic expression of olfactory receptors. THese receptors have been previously reported to have function in different tissues responsible for wound healing, metabolism, hair growth apart from causing the sense of smell. This study helps to understand the role of these receptors in cancer by analyzing 42529 single malignant cells of 22 different cancer types.

We here report 68 tumor-OR pairs, among which only a mere ~15% have been mentioned in cancer literature. This study is first of its kind that provides a large catalogue of chemosensory receptors across pan tumours at a single cell resolution. Since ORs specifically shows higher expression in tumors, this work clearly states that the tumor specific ORs like OR8H1, expressed in breast cancer can be used for the diagnostic purpose just like OR51E1 is an FDA approved biomarker for prostate cancer. Prognostic value of ORs was reinforced as we obtained the distinct survival groups, simply by tracking the enrichment of key OR-centric molecular signatures in TCGA tumor samples.

This work clearly shows the function of genes away from its native environment and thus suggests that genes can have multiple functions and to understand them is relevant for mechanistic insights. The long term vision is to identify the set of ligands that can either activate or deactivate these sets of receptors as a novel anti-tumor therapy. It is very easy to design drugs against a pathogen which do not share any homology with our body but it is hard to kill a cell that has an identity of our own body. In such cases these ligands when acted on the olfactory receptors can be used for drug therapy.

In this study we make several observations shedding light on the potential relationship between widespread expression of a large spectrum of ORs and expansion/differentiation of cancer clones. Despite their presence in the malignant cells,

so far the experimental validation of their predictive role in tumor-related molecular pathways is still elusive. Moreover, the majority of the identified ORs are orphan receptors, therefore, in order to delineate their contribution in tumor biology, identification of their agonist or antagonist is the first step forward. We assume that the cancer-specific metabolic intermediates could be the potential endogenous agonists for these cancer-specific ectopic ORs. Lastly, since the present results link the cellular count of expressed ORs with tumor cell differentiation, therefore, it is important to functionally validate these findings by the loss-of-function experiments.

BIBLIOGRAPHY

1. Young, J. M. & Trask, B. J. The sense of smell: genomics of vertebrate odorant receptors. *Hum. Mol. Genet.* **11,** 1153–1160 (2002).

2. Ryu, S. E. *et al.* Odorant receptors containing conserved amino acid sequences in transmembrane domain 7 display distinct expression patterns in mammalian tissues. *Mol. Cells* **40,** 954–965 (2017).

3. Bush, C. F. & Hall, R. A. Olfactory receptor trafficking to the plasma membrane. *Cell. Mol. Life Sci.* **65,** 2289–2295 (2008).

4. Hall, R. A. Olfactory receptor interactions with other receptors. *Ann. N. Y. Acad. Sci.* **1170,** 147–149 (2009).

5. Antunes, G. & Simoes de Souza, F. M. Olfactory receptor signaling. *Methods Cell Biol.* **132,** 127–145 (2016).

6. Silva Teixeira, C. S., Cerqueira, N. M. F. S. A. & Silva Ferreira, A. C. Unravelling the olfactory sense: from the gene to odor perception. *Chem. Senses* **41,** 105–121 (2016).

7. Ihara, S., Yoshikawa, K. & Touhara, K. Chemosensory signals and their receptors in the olfactory neural system. *Neuroscience* **254,** 45–60 (2013).

8. Olfactory Receptor - an overview | ScienceDirect Topics. at <https://www.sciencedirect.com/topics/medicine-and-dentistry/olfactory-receptor>

9. Fleischer, J., Breer, H. & Strotmann, J. Mammalian olfactory receptors. *Front. Cell. Neurosci.* **3,** 9 (2009).

10. Jiang, Y. *et al.* Molecular profiling of activated olfactory neurons identifies odorant receptors for odors in vivo. *Nat. Neurosci.* **18,** 1446–1454 (2015).

11. Buck, L. B. Olfactory receptors and odor coding in mammals. *Nutr. Rev.* **62,** S184-8; discussion S224 (2004).

12. Pick, H. *et al.* Dual activities of odorants on olfactory and nuclear hormone receptors. *J. Biol. Chem.* **284,** 30547–30555 (2009).

13. Glusman, G. *et al.* The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mamm. Genome* **11,** 1016–1023 (2000).

14. Krautwurst, D. Human olfactory receptor families and their odorants. *Chem. Biodivers.* **5,** 842–852 (2008).

15. Kurahashi, T., Lowe, G. & Gold, G. H. Suppression of odorant responses by odorants in olfactory receptor cells. *Science* **265,** 118–120 (1994).

16. Olfactory System Anatomy: Overview, Olfactory Epithelium, Olfactory Nerve and the Cribriform Plate. at <https://emedicine.medscape.com/article/835585-overview>

17. Malnic, B., Godfrey, P. A. & Buck, L. B. The human olfactory receptor gene family. *Proc Natl Acad Sci USA* **101,** 2584–2589 (2004).

18. Rivière, S., Challet, L., Fluegge, D., Spehr, M. & Rodriguez, I. Formyl peptide receptor-like proteins are a novel family of vomeronasal chemosensors. *Nature* **459,** 574–577 (2009).

19. He, H.-Q. *et al.* Functional characterization of three mouse formyl peptide receptors. *Mol. Pharmacol.* **83,** 389–398 (2013).

20. Feldmesser, E. *et al.* Widespread ectopic expression of olfactory receptor genes. *BMC Genomics* **7,** 121 (2006).

21. Lee, S.-J., Depoortere, I. & Hatt, H. Therapeutic potential of ectopic olfactory and taste receptors. *Nat. Rev. Drug Discov.* **18,** 116–138 (2019).

22. Manteniotis, S. *et al.* Functional characterization of the ectopically expressed olfactory receptor 2AT4 in human myelogenous leukemia. *Cell Death Discov.* **2,** 15070 (2016).

23. Spehr, J. *et al.* G protein-coupled receptor signaling via Src kinase induces endogenous human transient receptor potential vanilloid type 6 (TRPV6) channel activation. *J. Biol.*

Chem. **286,** 13184–13192 (2011).

24. Weyand, I. *et al.* Cloning and functional expression of a cyclic-nucleotide-gated channel from mammalian sperm. *Nature* **368,** 859–863 (1994).

25. Flegel, C. *et al.* RNA-Seq Analysis of Human Trigeminal and Dorsal Root Ganglia with a Focus on Chemoreceptors. *PLoS ONE* **10,** e0128951 (2015).

26. Veitinger, T. *et al.* Chemosensory Ca2+ dynamics correlate with diverse behavioral phenotypes in human sperm. *J. Biol. Chem.* **286,** 17311–17325 (2011).

27. Hartmann, C. *et al.* Sperm-Activating Odorous Substances in Human Follicular Fluid and Vaginal Secretion: Identification by Gas Chromatography-Olfactometry and Ca2+ Imaging. *Chempluschem* **78,** 695–702 (2013).

28. Neuhaus, E. M., Mashukova, A., Barbour, J., Wolters, D. & Hatt, H. Novel function of beta-arrestin2 in the nucleus of mature spermatozoa. *J. Cell Sci.* **119,** 3047–3056 (2006).

29. Flegel, C. *et al.* Characterization of the Olfactory Receptors Expressed in Human Spermatozoa. *Front. Mol. Biosci.* **2,** 73 (2015).

30. Pluznick, J. L. *et al.* Functional expression of the olfactory signaling system in the kidney. *Proc Natl Acad Sci USA* **106,** 2059–2064 (2009).

31. Kalbe, B. *et al.* Olfactory signaling components and olfactory receptors are expressed in tubule cells of the human kidney. *Arch. Biochem. Biophys.* **610,** 8–15 (2016).

32. Pluznick, J. L. *et al.* Olfactory receptor responding to gut microbiota-derived signals plays a role in renin secretion and blood pressure regulation. *Proc Natl Acad Sci USA* **110,** 4410–4415 (2013).

33. Santillo, M. & Damiano, S. Commentary: Oxygen regulation of breathing through an olfactory receptor activated by lactate. *Front. Neurosci.* **10,** 213 (2016).

34. Zhou, T., Chien, M.-S., Kaleem, S. & Matsunami, H. Single cell transcriptome analysis of

mouse carotid body glomus cells. *J Physiol (Lond)* **594,** 4225–4251 (2016).

35. Aisenberg, W. H. *et al.* Defining an olfactory receptor function in airway smooth muscle cells. *Sci. Rep.* **6,** 38231 (2016).

36. Mariman, E. C. M. *et al.* Olfactory receptor genes cooperate with protocadherin genes in human extreme obesity. *Genes Nutr.* **10,** 465 (2015).

37. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H. & Gisselmann, G. Expression profile of ectopic olfactory receptors determined by deep sequencing. *PLoS ONE* **8,** e55368 (2013).

38. Pichavant, C., Burkholder, T. J. & Pavlath, G. K. Decrease of myofiber branching via muscle-specific expression of the olfactory receptor mOR23 in dystrophic muscle leads to protection against mechanical stress. *Skelet. Muscle* **6,** 2 (2015).

39. Griffin, C. A., Kafadar, K. A. & Pavlath, G. K. MOR23 promotes muscle regeneration and regulates cell adhesion and migration. *Dev. Cell* **17,** 649–661 (2009).

40. Kalbe, B. *et al.* OR2H2 regulates the differentiation of human myoblast cells by its ligand aldehyde 13-13. *Arch. Biochem. Biophys.* **645,** 72–80 (2018).

41. Choquette, A. C. *et al.* Association between olfactory receptor genes, eating behavior traits and adiposity: results from the Quebec Family Study. *Physiol. Behav.* **105,** 772–776 (2012).

42. Jarick, I. *et al.* Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum. Mol. Genet.* **20,** 840–852 (2011).

43. Denda, M. Newly discovered olfactory receptors in epidermal keratinocytes are associated with proliferation, migration, and re-epithelialization of keratinocytes. *J. Invest. Dermatol.* **134,** 2677–2679 (2014).

44. Busse, D. *et al.* A synthetic sandalwood odorant induces wound-healing processes in human keratinocytes via the olfactory receptor OR2AT4. *J. Invest. Dermatol.* **134,**

2823–2832 (2014).

45. Chéret, J. *et al.* Olfactory receptor OR2AT4 regulates human hair growth. *Nat. Commun.* **9,** 3624 (2018).

46. A New Baldness Cure: Smell Receptors in the Human Scalp | Psychology Today. at <https://www.psychologytoday.com/intl/blog/the-fallible-mind/201809/new-baldness-cure-smell-receptors-in-the-human-scalp>

47. Weber, L. *et al.* Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis. *PLoS ONE* **12,** e0172491 (2017).

48. Grison, A. *et al.* Mesencephalic dopaminergic neurons express a repertoire of olfactory receptors and respond to odorant-like molecules. *BMC Genomics* **15,** 729 (2014).

49. Hanchate, N. K. *et al.* Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science* **350,** 1251–1255 (2015).

50. Goff, S. A. & Klee, H. J. Plant volatile compounds: sensory cues for health and nutritional value? *Science* **311,** 815–819 (2006).

51. Dreyer, W. J. The area code hypothesis revisited: olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos. *Proc Natl Acad Sci USA* **95,** 9072–9077 (1998).

52. Kang, N. & Koo, J. Olfactory receptors in non-chemosensory tissues. *BMB Rep.* **45,** 612–622 (2012).

53. Ranzani, M. *et al.* Revisiting olfactory receptors as putative drivers of cancer. [version 1; peer review: 2 approved]. *Wellcome Open Res.* **2,** 9 (2017).

54. Sanz, G. *et al.* Structurally related odorant ligands of the olfactory receptor OR51E2 differentially promote metastasis emergence and tumor growth. *Oncotarget* **8,** 4330–4341 (2017).

55. Abaffy, T. *et al.* A Testosterone Metabolite 19-Hydroxyandrostenedione Induces Neuroendocrine Trans-Differentiation of Prostate Cancer Cells via an Ectopic Olfactory Receptor. *Front. Oncol.* **8,** 162 (2018).

56. Neuhaus, E. M. *et al.* Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *J. Biol. Chem.* **284,** 16218–16225 (2009).

57. Abaffy, T. Human Olfactory Receptors Expression and Their Role in Non-Olfactory Tissues Ã¢Ââ A Mini-Review. *J. Pharmacogenomics Pharmacoproteomics* **06,** (2015).