scientific reports

Check for updates

OPEN

Conserved pattern-based classification of human odorant receptor multigene family

Sang Eun Ryu[1,3,5], Jisub Bae[1,2,4,5], Tammy Shim[1,2], Won-Cheol Kim[1], Kwangsu Kim[1,2] & Cheil Moon[1,2]✉

Conserved protein-coding sequences are critical for maintaining protein function across species. Odorant receptors (ORs), a large poorly understood multigene family responsible for odor detection, lack comprehensive classification methods that reflect their functional diversity. In this study, we propose a new approach called conserved motif-based classification (CMC) for classifying ORs based on amino acid sequence similarities within conserved motifs. Specifically, we focused on three well-conserved motifs: MAYDRYVAIC in TM3, KAFSTCASH in TM6, and PMLNPFIY in TM7. Using an unsupervised clustering technique, we classified human ORs (hORs) into two main clusters with six sub-clusters. CMC partly reflects previously identified subfamilies, revealing altered residue positions among the sub-clusters. These altered positions interacted with specific residues within or adjacent to the transmembrane domain, suggesting functional implications. Furthermore, we found that the CMC correlated with both ligand responses and ectopic expression patterns, highlighting its relevance to OR function. This conserved motif-based classification will help in understanding the functions and features that are not understood by classification based solely on entire amino acid sequence similarity.

Keywords Odorant receptor, Conserved motif, Classification, Olfactory, GPCR

G protein-coupled receptors (GPCRs) are diverse and evolutionarily related. They constitute a large superfamily that is classified into several subfamilies based on sequence similarities, structural features, and the types of ligands and signaling molecules they activate. Odorant receptors (ORs) represent the most diverse family of GPCRs; therefore, the function of each receptor in this system remains unknown.

Despite the diversity of ORs, they exhibit certain conserved amino acid motifs specific to ORs, such as 'GN' in transmembrane domain I, 'PMYF/LFL' in transmembrane domain II (TM2), 'MAYDRYVAIC' in TM3, 'KAFSTCxSH' in TM6, and 'PMLNPF/LIY' in TM7[1–3]. Notably, the motifs in TM3 and TM6 are highly conserved across OR multigene families, thus enabling the efficient amplification of a large fraction of mouse OR genes[4–6]. These motifs are critical for receptor function, ligand binding, and signal transduction[7–9]. For example, the DRY motif is highly conserved in class A GPCRs and, contributes to OR activation and signal transduction[10–16]. The 'S' in KAFST^{TM6} is crucial for receptor conformational dynamics[13], while the NPxxY^{TM7} motif regulates receptor activation, promoting cellular trafficking[17] and imposing structural constraints in class A GPCRs[18], although it has not been extensively studied in ORs. Interestingly, the recent determination of the structure of OR51E2 has verified the impact of these conserved residues R$^{3.50}$ in DRY^{TM3} and H$^{6.40}$ in CxSH^{TM6} in OR-specific hydrogen bonding involved in receptor activation[19]. In addition, conserved C-terminal motifs in ORs influence cell surface expression and cAMP signaling[20].

Variable residues in the conserved motifs appear to enhance the ability of each OR to detect and discriminate between different odorants, thereby enriching smell perception. In an evolutionary study, it was found that a Neanderthal variant of OR1C1 (Y120H$^{3.48}$) had decreased sensitivity compared to human OR1C1[21]. Furthermore, residues near highly conserved regions—such as the M$^{3.46}$ and Y$^{3.48}$ in the MAY motif and the I$^{7.52}$ in the NPxIY motif—exhibit opposite effects on OR function in OR10 and OR52 models. Specifically, the MAY mutation in OR10 increased basal activity and lowered the EC50, while the same mutation in OR52 decreased

[1]Department of Brain Sciences, Graduate School, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Republic of Korea. [2]Convergence Research Advanced Centre for Olfaction, Daegu Gyeungbuk Institute of Science and Technology (DGIST), Daegu, Republic of Korea. [3]Present address: Korea Brain Research Institute (KBRI), 61 Choemdan-Ro, Dong-Gu, Daegu 41062, Republic of Korea. [4]Present address: Center for Cognition and Sociality, Institute for Basic Science (IBS), 55 Expo-Ro, Yuseong-Gu, Daejeon 34126, Republic of Korea. [5]Sang Eun Ryu and Jisub Bae contributed equally. ✉email: cmoon@dgist.ac.kr

efficacy. Additionally, the $I^{7.52}$ mutation enhanced responsiveness in OR10 but reduced sensitivity in OR52[22]. These studies demonstrate that variations in conserved sequences significantly impact odorant-OR interactions. Moreover, our previous studies have shown that variations in conserved motifs cause different patterns of OR expression[23], suggesting that ORs with similar variation patterns may have similar functional characteristics. Given that highly conserved regions within motifs play critical roles in major receptor functions such as activation, stability, and G-protein coupling, it can be inferred that variations in adjacent residues contribute to the diversity of smell detection abilities and perceptual differences. Therefore, classifying ORs based on these variations could be a valuable approach for inferring the functions of multiple ORs.

Various classification approaches, including motif-based methods and analysis of the physicochemical properties of amino acid sequences, have been used to infer the functional relevance across GPCR families[24–26]. Therefore, we hypothesized that similar functions could be inferred by comparing ORs with similar conserved motifs and sought to classify ORs according to variations in variable regions within each conserved motif. Specifically, we focused on the following three well-conserved motifs—MAYDRYVAIC^{TM3}, KAFSTCASH^{TM6}, and PMLNPFIY^{TM7}—each noted for their defined OR functions. These motifs consist of consecutive amino acids and contain variable residues. Using the selected motifs, we classified the ORs into two main clusters, which were further subdivided into six subclusters. Subsequently, we observed the variability of the motif sites and their interactions within OR sequences to elucidate their structural implications. We also verified the potential associations between our classification and functions such as ligand binding or ectopic expression. Our classification showed the possibility of shared characteristics among the classified OR groups, which can offer a framework for further research into their diverse roles and help understand functions and features that are not apparent when classification is based solely on entire amino acid sequence similarity.

Results

MAYDRYVAIC^{TM3}, KAFSTCASH^{TM6}, and PMLNPFIY^{TM7} were highly conserved in the intact hOR

In our study, we focused on three specific conserved motifs within hORs known for their distinct OR functions: MAYDRYVAIC^{TM3}(mof1), KAFSTCASH^{TM6}(mof2), and PMLNPFIY^{TM7} (mof3). Each of these motifs, which consists of sequences of consecutive amino acids, includes a variable residue and is OR-specific. Thus, to determine how hORs are distributed depending on mof1, 2, and 3, we first evaluated the number of hORs based on the matching rate (see Materials and Methods) with three conserved motifs (Supplementary Fig. S2). We found that each motif of the overall hORs scored above 0.5 of the matching rates for each condition. These results indicate that most hORs possess highly conserved mof1, 2, and 3. We further investigated the number of intact, and pseudogenes scored above 0.5 of the matching rates for each condition. We found that mof1, 2, and 3 were highly conserved in 97.9% of the intact hORs and 76.7% of the pseudogenes, with an above 0.5 matching rate (Fig. 1). Among the three, 2.1% of the intact hORs have one of the poorly conserved motifs, and all of them have an arginine (Arg; $R^{3.50}$) residue in mof1, which is a critical residue for mediating receptor signaling, but no serine (Ser; $S^{6.35}$) in mof2, which is known for OR conformation dynamics[13]. However, 23.3% of the pseudo-ORs were poorly conserved at the Arg residue position in mof1 and relatively well-conserved at the Ser position in mof2. Regarding mof3, both intact and pseudo-ORs showed low conservation rates. These results indicate that intact human ORs (known as functional OR) contain more well-conserved motifs. As hypothesized, these results support the hypothesis that mof1 to 3 may be related to hOR function.

Intact hORs are classified into two groups according to their conserved patterns of mof2 and mof3

Figure 1 suggested that conservation rates varied between 0.5 and 1 among the intact hORs. Based on this, we performed unsupervised classification (AHC) to determine whether the conservation rate difference could serve as a criterion for classifying OR. AHC analysis classified hORs into two clusters consisting of six sub-clusters (SC) (Fig. 2A, $p < 0.01$ by SHC). We used Euclidean dissimilarity with Ward's linkage for the AHC. 60% ($n = 232$) of all hORs were included in Cluster1 (C1), showing a high conservation rate (above 0.85) for all three motifs, but 40% of the hORs ($n = 157$) that clustered in Cluster2 (C2) showed a relatively low conservation rate (< 0.75) compared to C1 ORs. We call this the conserved motif-based classification (CMC). Each hOR classified through the CMC was visualized in a scatter plot corresponding to its conservation rate, with a single dot of different colors (Fig. 2B-D). The hORs in SC3 showed the highest conservation rates for all three conserved motifs, and the SC1 ORs showed comparatively low conservation patterns at mof1 among the C1 SCs (Fig. 2C). In contrast, hORs in SC5 showed the lowest conservation rate for all three conserved motifs; specifically, all hORs in C2 had poorly conserved motifs 2 and 3 (Fig. 2D). The conservation rates of each OR are summarized in the table (Fig. 2E).

We observed where these ORs were distributed in the CMC to evaluate the relationship between entire amino acid sequence similarities (Fig. 3). We found that hOR subfamilies 2 and 4, which accounted for a significant proportion of hORs, were distributed across the six SCs. Interestingly, most of the intact fish-like class I hORs (families 51, 52, and 56) clustered as C2. Moreover, OR subfamilies clustered in C2 showed clustered patterns similar to the phylogenetic classification[27,28]. Specifically, OR subfamilies 14, 52, 51, and 56 were phylogenetically close to the other subfamilies. Our results showed that OR subfamilies 1,7 and 8 were primarily distributed in SC3, with a high conservation rate for all three motifs. These three OR subfamilies were also closely clustered in the phylogenetic analysis of intact hOR protein sequences[27]. In contrast, few relationships were observed between SCs and OR subfamilies.

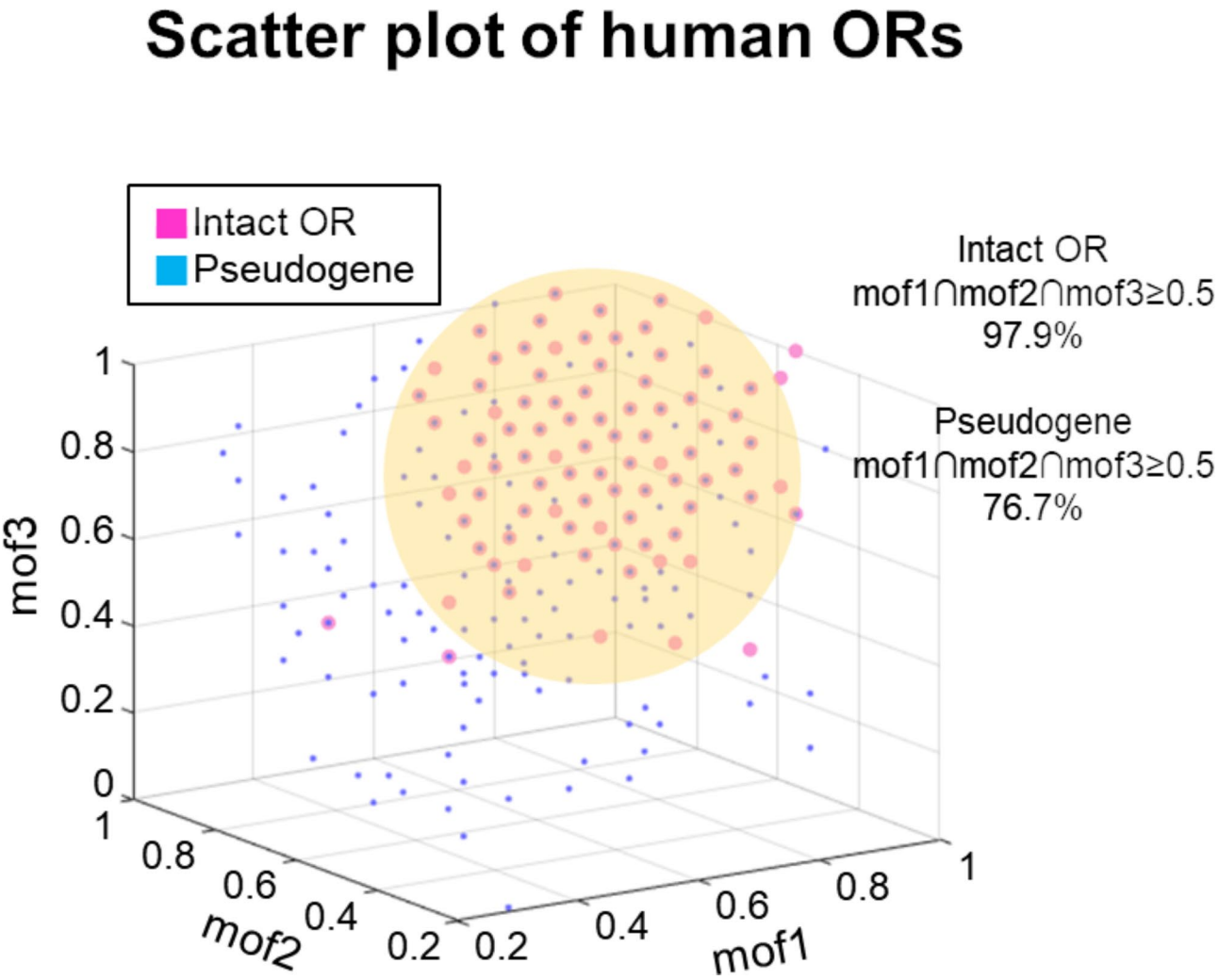
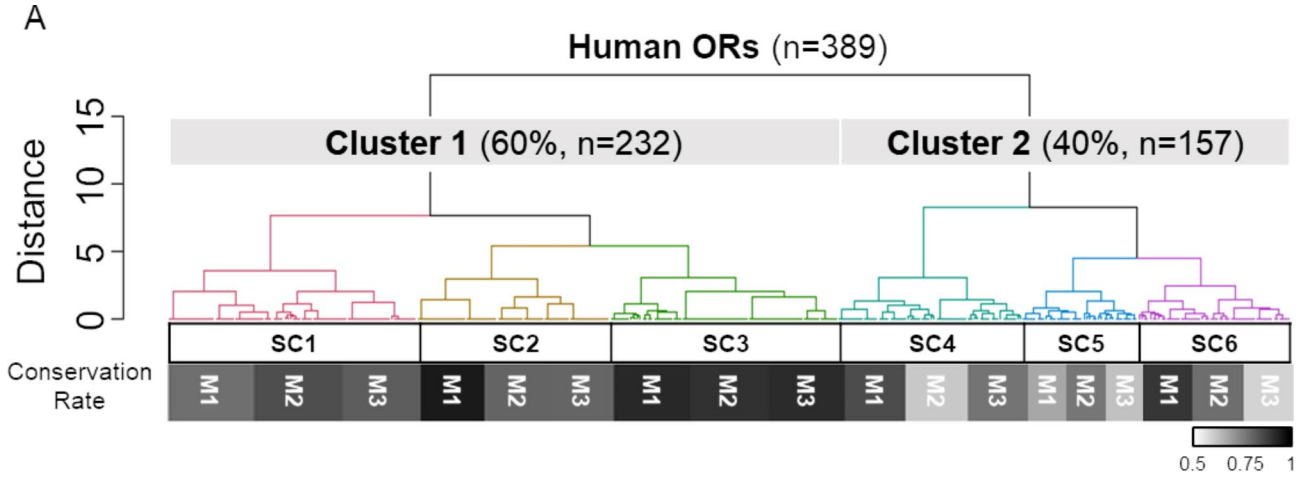


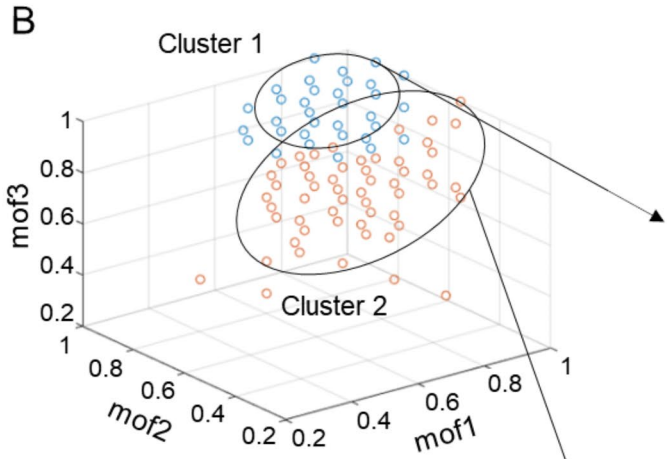
Fig. 1. Scatter plot of intact hORs and pseudogene hORs by the conservation rate of MAYDRYVAIC^{TM3}, KAFSTCASH^{TM6}, and PMLNPFIY^{TM7}. Each intact OR (pink) and pseudogene OR (blue) is represented as a single dot in a scatter plot. The sequence matching rate with each motif was represented as X, Y, and Z-axis (mof1: MAYDRYVAIC, mof2: KAFSTCASH, and mof3: PMLNPFIY, respectively). The yellow-colored sphere represents a matching rate ≥ 0.5 in every three axes. Almost all intact hORs showed a high matching rate with all frame sequences of three motifs (97.9%), and a lower number of pseudogene (76.7%) have a similar conserved pattern with intact hORs.

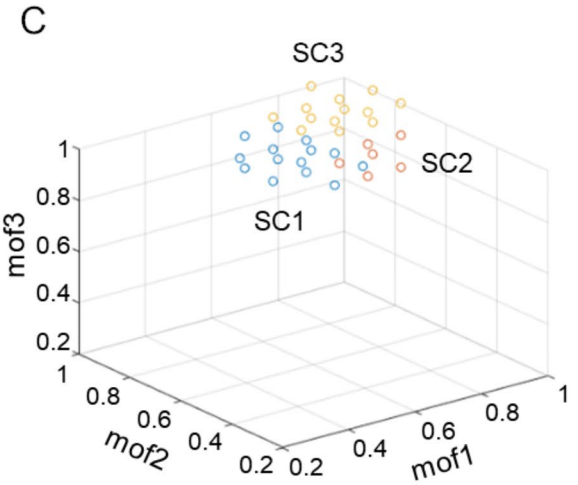
Variant residues in each motif have interacting residues in other transmembrane domain

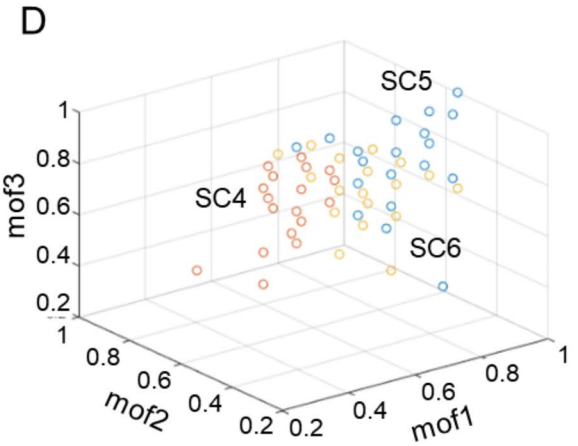
We next generated consensus sequences for each motif in every group using WebLogo[29] to examine which residues were variants within the entire sequence of each motif (Fig. 4). While hORs in C1 (SC1-SC3) showed high sequence conservation at every position, hORs in every subcluster (SC4-SC6) of C2 showed variable patterns. Interestingly, residue S$^{6.35}$ in mof2 and M$^{7.47}$ in mof3 showed low conservation across all C2 subclusters, which are the most pronounced differences that distinguish C2 from C1 (Fig. 4). Moreover, TM3 Y$^{3.48}$ and V$^{3.52}$ beside the 'DRY' in mof1 and TM6 F$^{6.34}$ in 'KAFST' of mof2 are altered to F$^{3.48}$, L$^{3.52}$, mof2: L$^{6.34}$, respectively (Fig. 4A). The conservation scores for each position within each motif are represented in the heat map, and the detailed score values are provided in the supplementary information (Fig. 4B and Table, S3). To examine how these variant residues (residues 3.48, 3.52, 6.34, and 7.47 in Fig. 4) influence the receptor structure, we evaluated the interaction with the rest of the hOR residue positions using residue-residue contact score (RRCS) analysis[15].

To calculate the RRCS, we need to summarize the total contact score from the min-max normalized data (0,1) of the inter-atomic distances among heavy atoms. The maximum contact score was obtained from the minimum inter-atomic distance. Using the RRCS, we calculated the general interactions among the total intact ORs. We summarized the total RRCS for each residue position from the entire OR and set the threshold to a significance level for it. We set one RRCS per OR; therefore, the threshold was set to 384. The RRCS is determined by summing the distances of all possible heavy atom pairs between the target residue pairs. Therefore, when the RRCS equals one, the distances of all heavy-atom pairs between the target residue pairs are identical to the closest distance data observed in the targeted receptor. We found that variant residue positions interacted strongly with









E

	Centroid of clusters		
	M1	M2	M3
Cluster1	0.88	0.85	0.85
Cluster2	0.82	0.71	0.67
SC1	0.78	0.85	0.82
SC2	0.95	0.80	0.80
SC3	0.92	0.90	0.91
SC4	0.85	0.61	0.77
SC5	0.67	0.77	0.62
SC6	0.89	0.78	0.59

Fig. 2. Classification of hOR according to their conservation rate. **(A)** Classified hORs and conservation rate of each cluster (number of hOR: SC1 = 86, SC2 = 67, SC3 = 79, SC4 = 64, SC5 = 40, SC6 = 53, p value < 0.01 by SHC). Distance calculated based on Euclidean dissimilarity with Wards linkage. The intensity of the colored box corresponds to the conservation rate (0.5 to 1). **B-D.** Each intact OR is represented as a single dot in a scatter plot. The sequence matching rate with each motif was represented as X, Y, and Z-axis (mof1: MAYDRYVAIC, mof2: KAFSTCASH, and mof3: PMLNPFIY). **(B)** Scatter plot of intact hOR. Each dot represents each hOR in Cluster1 (blue) and Cluster2 (orange). **(C)** Scatter plot representation of sub-clusters in C1. Every SC is described as different colors; SC1(blue), SC2 (orange), and SC3 (yellow). **(D)** Scatter plot representation of sub-clusters in C2, and each SC is represented as different colors. SC4 is orange, SC5 is blue, and SC6 is yellow. **(E)** Table shows the centroid of each cluster. SC: sub-cluster.

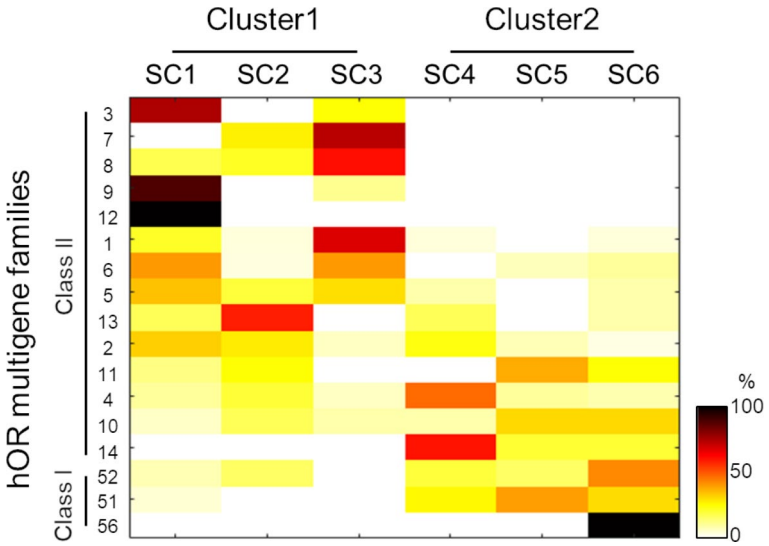


Fig. 3. OR distribution between OR multigene families and CMC. A heatmap of OR distribution between OR multigene families and CMC. The color bar shows the percentage of ORs of each multigene family. SC: sub-cluster.

specific residue positions in the same or different TM (Fig. 5). The residues $Y^{3.48}$ and $V^{3.52}$ in mof1 mainly interact within TM3, but the residues $F^{6.34}$ and $M^{7.47}$ interacted with other TM residues. Specifically, residue $F^{6.34}$ in mof2 interacted with TM5, and residue $M^{7.47}$ in mof3 interacted with TM1. Overall, the interacting residues were highly conserved (> 40%) except $X^{1.42}$ residue positions (Supplementary Table. S1). Residue TM7 $F^{7.51}$ also showed variation patterns; however, this residue was excluded because of its lower conservation rate and low RRCS (Supplementary Fig. S4). As hORs are included in class A GPCRs, these results align with previous class A GPCR study[15]. As we have shown in the results that the R residue in mof1 is well maintained throughout the whole hORs (Figs. 1 and 2), the residue variations besides the DRY^{TM3} motif seem to support its activation of the G-protein, which is critical for olfactory signaling. From a different perspective than mof1, the fact that variant residues in mof2 and 3 interact with other residues of TMs indicates their flexibility in conformational changes. Considering the function of the residue positions in the OR, these different conserved patterns raise the possibility of their roles in interacting with ligands or G-proteins.

CMC may predict ligand responses and ectopic expression of hORs

To understand the questions raised by previous results, we first focused on the effect of variant residues on ligand interactions. We investigated the known ligands of hORs in each SC (Table 1). We sorted the ligands using four criteria: (1) They must be based on experimental data, (2) They must be specific to each SC, (3) There must be at least one OR result for each SC, and (4) They must be available in the physicochemical database. Since our results are based on sequence information, comparing them with experimental data is crucial. SC specificity is also important because some ligands can activate multiple receptors, making it difficult to evaluate SC-specific effects. The third step is to prevent data bias. Table 1 showed that each SC contained a specific ligand-binding list. Supplementary Figure S4 shows the hOR family distribution used to create Table 1. This result shows the effect of the bias from one OR family on the ligand interaction list.

Based on this table, we compared the physicochemical properties of the ligands to examine ligand similarity among the SCs (Fig. 6). We then performed PCA using 24 ligands with 2004 physicochemical features and performed GMM clustering on the PCA data (Fig. 6). We used only 24 ligands for this analysis because the remaining ligands activated the hOR from two or more SCs or lacked information on physicochemical features from our database. The principal component 1 (PC1) axis suggests the maximum covariance axis among the physicochemical features, and the PC2 axis indicates the following maximum covariance axis, which is orthogonal to PC1. Figure 6B shows that the dimension-reduction ligand data are similar to the CMC results. Remarkedly, 71% of the ligand data were classified in the same manner as that in our CMC results. Approximately 37.5% of C1 and 87.5% of C2 ligands were classified into the same group (Fig. 6B). The chance level of each cluster was 33% for C1 and 67% for C2. Although our results were insufficient to identify a distinct pattern in SCs (data not shown), we observed that clusters 1 and 2 interacted with different ligand properties. These results suggest that the CMC may predict which hOR responds to specific ligands based on their physiochemical features.

In addition to our previous study, we observed that hORs with similar conserved motifs display distinct expression patterns in olfactory and non-olfactory tissues[23,30,31]. This suggests that specific conserved motifs facilitating interactions with potential internal ligands may determine the ectopic expression patterns of odorant receptors. In our previous results (Figs. 5 and 6), we demonstrated that hORs classified into the same cluster by CMC are likely to detect similar ligands. If the ectopic expression patterns of hORs are influenced by the type of internal ligand, then the hORs grouped into the same cluster in our CMC might exhibit similar expression patterns. To verify this hypothesis, we collected ectopic expression data for hORs using a literature

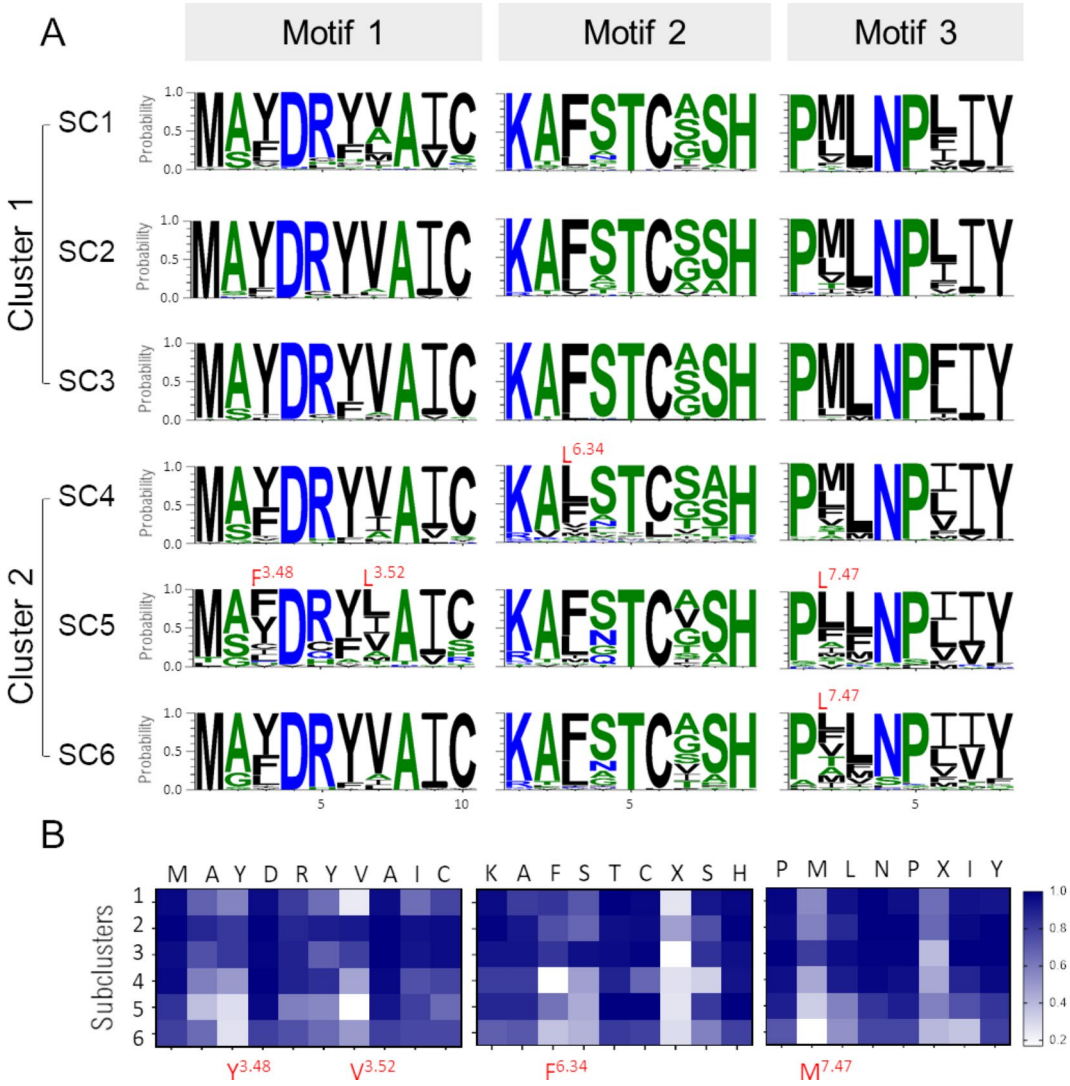


Fig. 4. The consensus sequence of each cluster. The consensus sequences of each SCs for the three conserved motifs are graphically represented. Each color represents the Hydrophobicity of each amino acid. The blue color indicates the hydrophilic (RKDENQ), the green color indicates neutral (SGHTAP), and the black color indicates the hydrophobic (YVMCLFIW) amino acid. The X-axis shows the sequence position. The Y-axis shows residue probabilities within each motif and the sequence size represents the portion of amino acids in each position. The significantly variable positions were highlighted with a red star marker that altered residue position compared to conserved motifs (mof1: $F^{3.48}$, $L^{3.52}$, mof2: $L^{6.34}$, and mof3: $L^{7.47}$). The sixth site of mof3 was not highlighted because of the weak conservation pattern across the SCs. **(B)** Heatmap analysis showing average conservation rate of each residue within each cluster. The conservation score represents the average probability that ORs in each cluster have the consensus residue at a given position, with higher probabilities indicated by darker colors.

survey (Supplementary Data. S2). Previous studies performed microarray analysis on non-olfactory tissues, and we compiled these public data for our analysis. Figure 7A shows the overall ectopic expression pattern of hORs. Then, we analyzed the similarities in ectopic expression patterns between SCs using AHC (Fig. 7B). The hORs with similar expression patterns clustered closely with low distances. Interestingly, we found a close distance of ectopic expression patterns between SC1 and SC2, which showed similar grouping patterns for CMC. This result suggested that ORs in SC1 and 2 may mostly be expressed in the same non-olfactory tissues. However, at comparatively far distances, SC4 and SC5 were clustered in the same class. In contrast, SC6 showed a different pattern from SC4 and SC5. SC6 was closer to the Cluster 1 components (SC1 to 3). SC6 has a very low conservation rate for mof3 at 0.59, while mof1 and mof2 have relatively high rates of 0.89 and 0.78, respectively. These rates are higher than those of other ORs in the same cluster and are similar to the conservation rates of SC1 and SC2 (Fig. 2E). This observation suggests that the conserved patterns of mof1 and mof2 might be more involved in distinguishing the ectopic expression of hORs. Furthermore, it's interesting to note that SC3, which has high conservation rates across all three motifs, is grouped at the furthest distance from SC5, which has low

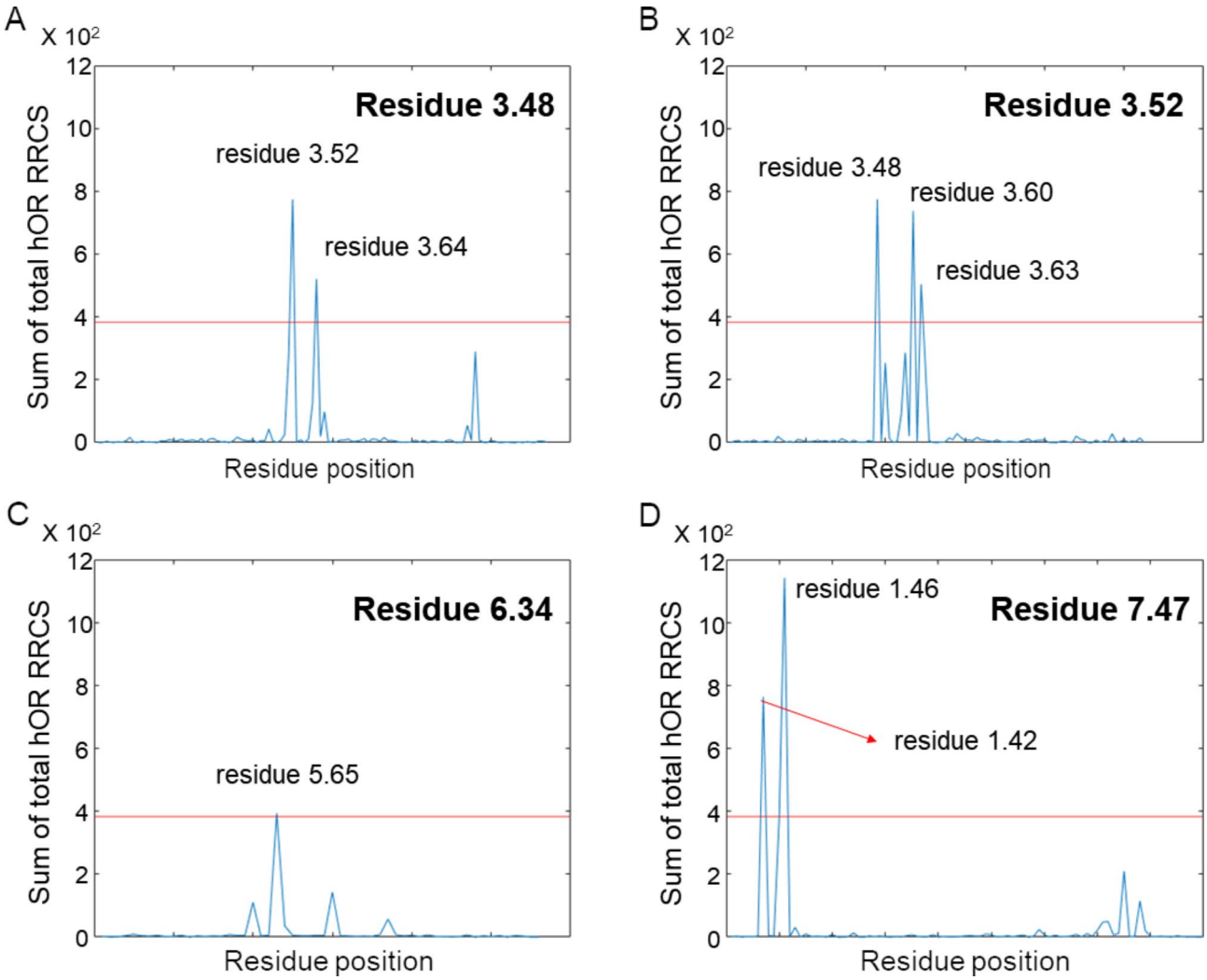


Fig. 5. Interaction patterns by RRCS. In three conserved motifs, the highly variable positions, $F^{3.48}$, $L^{3.52}$, $L^{6.34}$, and $L^{7.47}$, underwent RRCS to observe interactions with other residues in the OR structure. The X-axis is the residue position, and the y-axis is the interaction score represented by RRCS. The red line shows the threshold of a significant interaction. **(A-B)** RRCS pattern of the mof1 residues (Residue 3.48, 3.52). Mof1 mainly interacts with TM3. **(C)** RRCS pattern of the mof2 residue (Residue 6.34). Mof2 showed interaction with TM5 but was comparatively weak from other motifs. **(D)** RRCS pattern of the mof3 residue (Residue 7.47). Mof3 mainly interacts with TM7.

conservation rates in all three motifs. This finding further supports the potential role of motif conservation in the ectopic expression patterns of hORs.

Discussion

We propose a novel approach for classifying hORs based on the variation patterns of conserved motifs, called CMC. This classification system offers a new perspective for understanding the diversity and functionality of hORs.

hORs were classified into two clusters consisting of six SCs, using three conserved motifs ($MAYDRYVAIC^{TM3}$, $KAFSTCASH^{TM6}$, and $PMLNPFIY^{TM7}$). Interestingly, over 87% of the class I (fish-like) OR were classified as Cluster 2, partly representing the hOR family (Fig. 3). However, only a few relationships between the SCs and OR families have been observed. In the classified OR clusters, we found variation patterns in each SC, with some variant residues within each conserved motif (Fig. 4) and interactions between variant residues themselves or with other regions of TMs (Fig. 5). Since the three conserved motifs play crucial roles in OR or GPCR activation and receptor conformational dynamics, our data suggest that each SC may have distinct structural conformation mechanisms due to variant residues in conserved motifs. In addition, we found that CMC was related to ligand binding or ectopic expression (Figs. 6 and 7). Previous studies have suggested that conserved motifs influence ligand responses[10–15]and expression patterns[23]; thereby, these results are in line with those of previous studies. Although we did not find significant differences among the SCs, we observed different ligand-binding patterns between Clusters 1 and 2. Additionally, ectopic expression of SCs is related to CMC. Our classification showed

	Cluster 1			Cluster 2		
	SC1	SC2	SC3	SC4	SC5	SC6
(-)-citronellal	o					
(+)-menthol	o					
1-octanethiol	o					
beta-ionone	o					
octanethiol	o					
caramel furanone		o				
dicyclohexyl disulfide		o				
nonyl aldehyde		o				
p-cymene		o				
quinoline		o				
(-)-carvone			o			
(+)-carvone			o			
1-Nonanol			o			
androstenone			o			
diacetyl			o			
methyl salicylate			o			
muscone			o			
myrac aldehyde			o			
TMT			o			
amyl butyrate				o		
butyric acid				o		
ethyl vanillin				o		
n-amyl acetate				o		
phenyl acetaldehyde				o		
(-)-menthol					o	
2-methoxy-4-methylphenol					o	
2-pentylpyridine					o	
3-phenyl propyl propionate					o	
acetophenone					o	
beta-citronellal					o	
citral					o	
dihydrojasmone					o	
guaiacol					o	
helional					o	
hexyl acetate					o	
2-ethyl fenchol						o
allyl phenyl acetate						o
anisaldehyde						o
decanal						o
fenchone						o
isovaleric acid						o
lyral						o
propionic acid						o
r-limonene						o
undecanal						o
eugenol	o	o	o	o	o	o
eugenol acetate	o	o	o	o		
isoeugenol	o	o		o	o	
eugenol methyl ether	o		o	o	o	
cinnamaldehyde	o	o		o		o
geraniol	o			o	o	o
geranyl acetate		o	o	o		o
coumarin		o	o	o		
cinnamon	o		o			
nutmeg	o				o	

Continued

	Cluster 1			Cluster 2		
	SC1	SC2	SC3	SC4	SC5	SC6
linalool		o		o		
spearmint		o			o	
amyl laurate		o			o	
sandalwood		o				o
butyl anthranilate			o	o		
androstadienone			o		o	
caproic acid			o		o	
jasmine			o			o
bourgeonal			o			o
1-octanol				o	o	
cis-3-hexen-1-ol				o	o	
in				o	o	
ambrette					o	o

Table 1. List of ligands known to activate hOR in each sub-cluster. Ligands known to activate hORs are highlighted in green at each cluster. SC: sub-cluster.

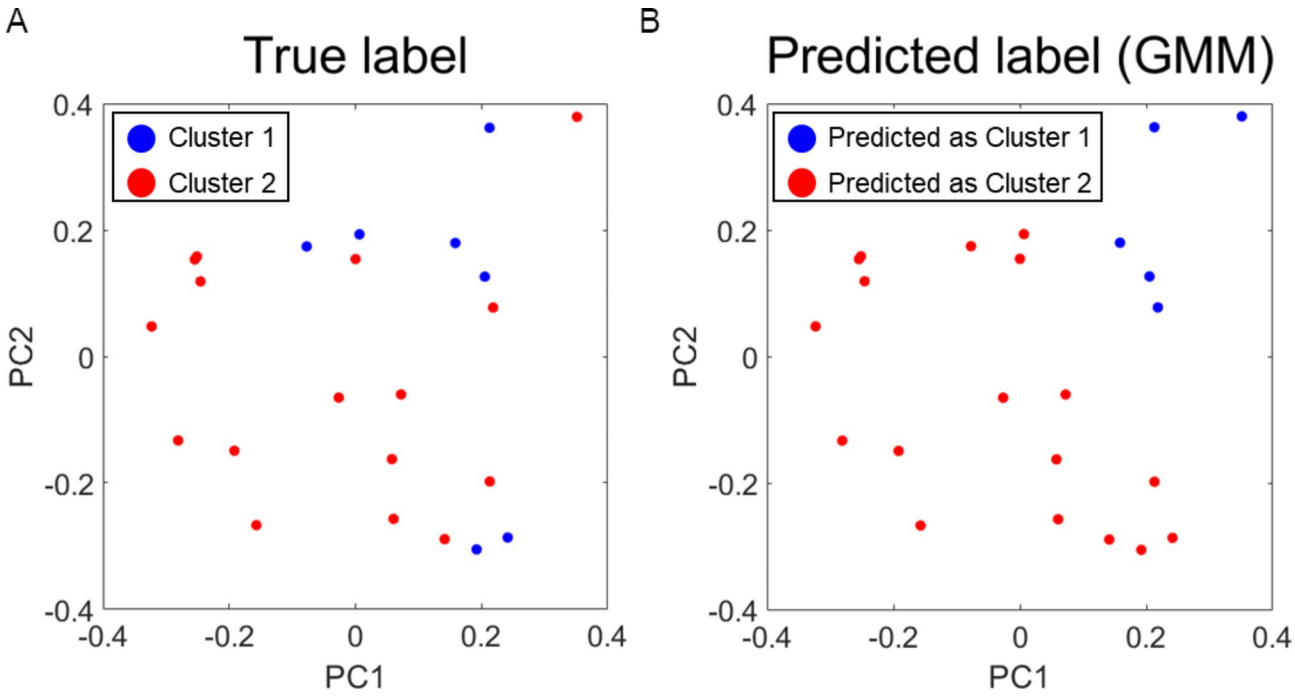


Fig. 6. Similarity space of ligand that respond to a particular hOR. **A-B.** PCA results of physicochemical properties of ligands. The X-axis is PC1 and the Y-axis is PC2. A circle marker depicts each ligand. Blue represents ligands interacting with cluster 1 ORs, while red represents ligands interacting with cluster 2 ORs. **(A)** PCA of true label. **(B)** PCA of predicted label by GMM(k = 2). GMM clustering yielded 71% accuracy (17/24 = 0.71).

the possibility of shared characteristics among the classified hOR groups, which would help to understand functions and features that are not understood in a family classified by full amino acid sequence similarity.

In our previous study[23], highly conserved consensus sequences across species at the C-terminal of ORs were found. We verified that this conserved region influences OR expression patterns in olfactory and non-olfactory systems. We were also interested in how the patterns of other conserved motifs within each conserved area affected the overall structure and accompanying functions, such as ligand detection and receptor activation. For this reason, we presuppose that MAYDRYVAIC^{TM3}, KAFSTCASH^{TM6}, and PMLNPFIY^{TM7}may play crucial roles in the function of OR, according to previous findings[1–3]. Similar to previous studies, our data showed a high level of conservation in the DRY^{TM3}, KAFSTCxSH^{TM6}, and NPxxY^{TM7} motifs.

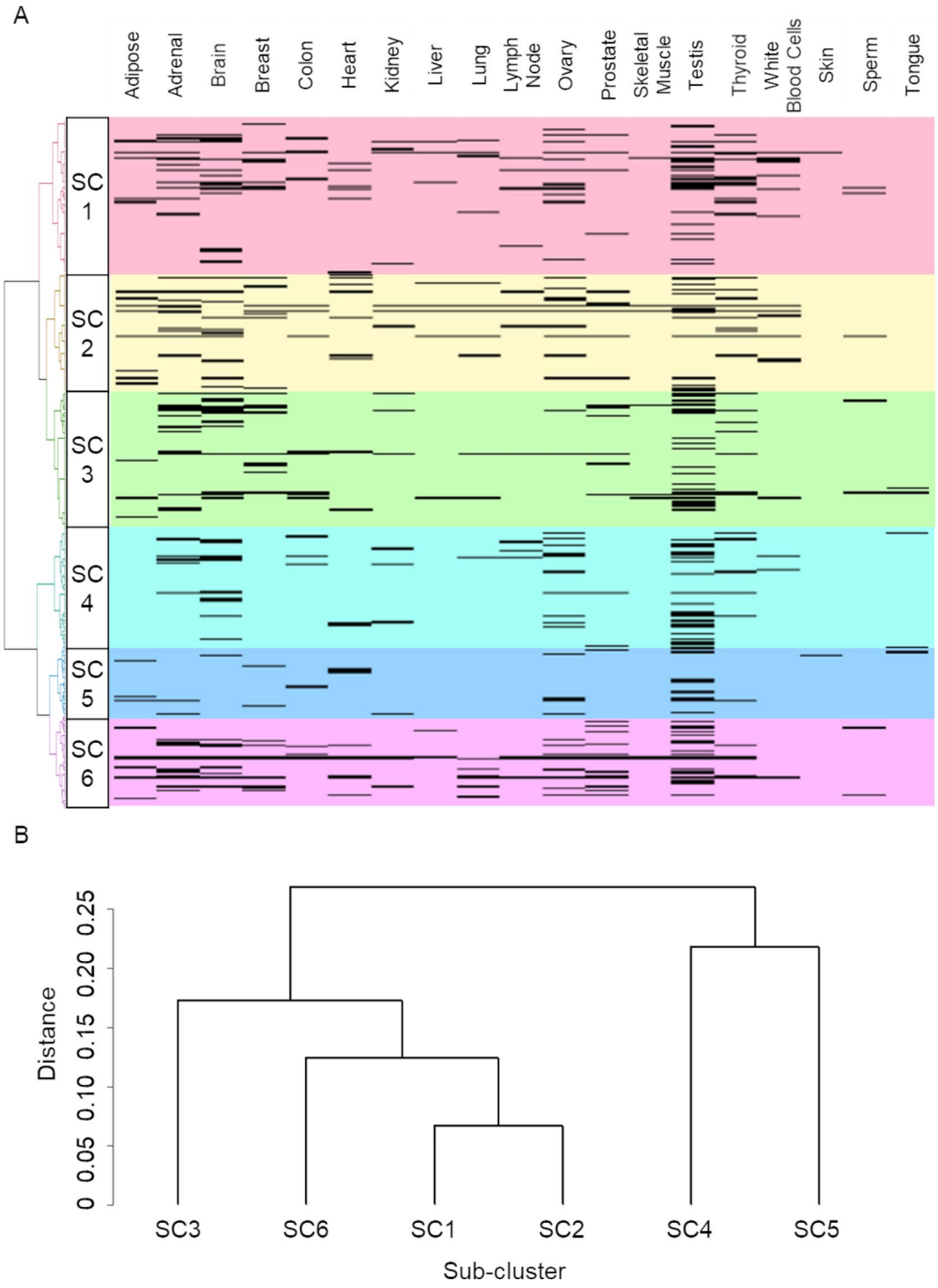


Fig. 7. Ectopic expression pattern of sub-clusters. **(A)** Ectopic expression pattern among the hOR. X-axis shows non-olfactory tissues. Y-axis shows hORs with SCs. Black stripe marker represents specific hOR expressed at specific tissues. **(B)** Similarity of ectopic expression pattern between sub-clusters. Distance calculated based on Euclidean dissimilarity with Wards linkage. Before analyzing data was normalized by the total number of expressions.

Since highly conserved regions within conserved motifs-$R^{3.50}$. $H^{6.40}$-were verified to play a critical role in ligand-OR activation; nearby variable regions may significantly influence ligand responsiveness or sensitivity. The amino acids adjacent to conserved residues exhibit different ligand responses depending on the OR type[22]. Interestingly, we also observed the variant residues were positioned adjacent to DRY^{TM3} or in the middle of the $KAFST^{TM6}$ motifs, except for $NPxxY^{TM7}$. These observations indicated that if different amino acids are present in these variants, they can physically influence the DRY^{TM3}or $KAxSTCxSH^{TM6}$ motifs. Additionally, variant

residues in mof3 are also positioned at 3–4 amino acid distances from NPxxY^{TM7}, which suggests that they may have a chance to interact physically since this motif forms an alpha helix. We also found that these residues interacted with other TM residues. Figure 5 shows that the variant residue positions in mof2 and 3 exhibited higher interactions with other TMs. Moreover, these interacting TM residues showed conserved patterns (except for 1.42 residue), which indicates that changes in significantly variable regions might structurally influence other TMs (Supplementary Table. S1). These results suggest that different patterns of motifs (mof1-3) are linked to ligand interactions or G protein-binding affinity. In addition, these distinct ligand responses may have been affected by evolutionary pressure, thus specifying the ectopic expression patterns of each SC.

Our results indicate that mof3 variation in the CMC might be linked to structural stability or structural constraints rather than ligand or G protein interactions. A study of class A GPCR suggested that NPxxY^{TM7}, located inside the ligand-binding pocket, may interact with ligand[15]. In contrast, we found that M$^{7.47}$residue, which was the most variable residue in our results, was positioned two more amino acids away from NPxxY^{TM7}. This position of the residue is less related to the ligand-binding pocket. According to class A GPCR studies, NPxxY^{TM7} functions as a rotamer switch[32] and provides structural constraints[18]by interacting with other TM residues. Interestingly, we found comparatively higher interaction values in M$^{7.47}$ residue (Fig. 4, red star marked residue position in mof3) than in the other variant residue positions in mof1 or 2 (Fig. 5). Figure 5-D showed that M$^{7.47}$ interacted strongly with TM1, specifically, X$^{1.42}$ and X$^{1.46}$ residues. Variation in the M$^{7.47}$ position can induce different interactions with TM1, which may change the structural stability or constraints, as suggested by class A GPCR studies.

Interestingly, a recent study has also pointed out that conserved motifs are related to ligand interactions. Structural determination of OR confirmed that highly conserved residues are critical for receptor activation. Based on the crystalized structure of OR51E2, they identified an OR-specific hydrogen bonding network among the highly conserved R$^{3.50}$ in DRY^{TM3} and H$^{6.40}$ in KAFSTCxSH^{TM6}, and Y217$^{5.58}$ in TM5[19]. Mutations in these conserved residues significantly decreased receptor activity, while mutations in less conserved residues (such as Y291$^{7.53}$) had a less pronounced effect. This underscores the pivotal role these conserved amino acid residues play in maintaining receptor activation. Notably, OR51E2 is classified into Cluster 2, Subcluster 6 (C2, SC6) in our classification, and it possesses F$^{3.48}$, G$^{6.35}$, V$^{7.47}$ in the three conserved motifs, which aligns with the consensus sequence variation pattern of SC6. This structural insight reinforces our findings that CMC correlates with ligand-OR interactions.

The OR multigene family is phylogenetically classified into Class I and Class II based on sequence homology. Class I OR are fish-like ORs, and Class II ORs are mammalian-like ORs based on their origin of identity[33]. Class I ORs are activated by water-soluble odorants, whereas Class II ORs are activated by volatile odorants[4,34]. Interestingly, most Class I ORs were classified as Cluster 2 in the CMC. This is similar to the division of Classes I and II into phylogenetic classifications. Considering that our CMC classification was based on the variation pattern of conserved motifs rather than the entire sequence, this may reflect the influence of ligands that activate olfactory receptors, causing modifications in the associated amino acid residues. In line with this, the ectopic expression of ORs suggests a wide distribution pattern, owing to the influence of internal ligands that activate these ORs.

The following points should be considered when interpreting the results of this study. Although we observed patterns in ligand response and ectopic expression, there is still a lack of information regarding these aspects. Ligand responses were assessed by the cAMP assay; therefore, these results were combined with various factors such as ligand interaction, G protein interaction, and cell state. Moreover, the number of odorants and ORs tested was limited. Ectopic expression results also relied on mRNA assays and not on the protein levels. Additionally, we hypothesize that evolutionary pressure may have affected distinct internal ligand responses, thereby specifying the ectopic expression patterns of each SC. However, this hypothesis has not yet been verified in this study. Finally, we did not consider the receptor expression rate or other factors that could affect ligand responses or ectopic expression. Further studies are required to provide more evidence on these issues.

We propose a classification method that is more aligned with olfactory receptor function by focusing on conserved regions rather than simply listing them based on overall sequence similarity. The previously predicted aspects of traditional classification methods have not been fully interpreted due to evolving perspectives from subsequent research. In our CMC, it is anticipated that specific variant residue points, which are the basis for OR classification, might play a role in ligand interactions, structural flexibility, and other related functions of olfactory receptors.

Methods

Database

We used the HODRE database (https://genome.weizmann.ac.il/horde/) and excluded five ORs with empty sequences from the database; OR14A16, OR7E162P, OR7E103P, OR4C4P, and OR8G1. For physicochemical properties, we used the Mol-Instincts database (https://www.molinstincts.com/). We used a database from previous studies to analyze ligand responses[31–51] and ectopic expression patterns[35–37,48,52–54]. These data are listed in supplementary information (Supplementary Fig. S2 and S3).

The procedure of classifying human ORs (hORs)

To calculate the conservation rate of the motifs, we evaluated the matching rate (V) between the conserved and target motifs by normalizing the matched number of target motifs compared to the frame sequence (v) with the target motif length (L). We did not consider the matching positions of the residues (Supplementary Fig. S2A).

$$V = \frac{v}{L} \tag{1}$$

Based on this approach, we created a rate-wise conservation rate-wise database for the classification of human ORs (hORs) (Supplementary Data. S1). Agglomerative hierarchical clustering (AHC) was used to classify the hORs. We obtained our results using the Euclidean dissimilarity with Ward's linkage. To determine statistical significance, we applied the statistical Significance of Hierarchical Clustering (SHC)[55]. We used empirical p-values for the cutoff nodes ($p < 0.01$).

Calculation of chance level depending on conservation rate

We calculated the chance level of conservation rates for each target motif (Supplementary Fig. S2), setting the total number of amino acids to twenty for this computation (excluding Selenocysteine and Pyrrolysine).

$$H_{OR} = \frac{1}{20}^{k} \times (n_{OR} + 1 - k) \geq 1 \tag{2}$$

$$\mathrm{N} = \sum_{OR=1}^{852} H_{OR} \tag{3}$$

$k$represents the number of matched sites with a target motif and, n_{OR} is the sequence length of each OR. Notably, k cannot exceed the length of the target motif. The equation $\frac{1}{20}^{k}$ represents chance level depending on k while $n_{OR} + 1 - k$ accounts for the repetitive calculation of the chance level along the entire OR length. Since k represents the number of matched sites, we subtracted k from n_{OR}. For instance, if we consider an OR with a length of 100 and five matched sites, we iteratively compute and aggregated $\frac{1}{20}^{5}$ over 96 times. Note that if we incorporate a window size of five and shift one site at a time within the 100-length OR, we obtain 96 iterations. We obtained the number of true values from these steps, which indicates the number of OR-containing segments in the sequence matching the specified conservation rates for the target motif.

Calculation of interaction score of residue pairs

We evaluated the interaction score using the residue-residue contact score (RRCS)[15]. The following steps were performed to calculate the RRCS.

$$f_{A:B} = [x_1,\ x_2,\ x_3,\ \ldots\ ,\ x_n] \tag{4}$$

where $f_{A:B}$ is distance data from the target residue pairs (e.g., A: V45 and B: Y288), and x is the inter-atomic distance among heavy atoms (Å). If the target residue pairs were located within four amino acids of the protein sequence, only side chain heavy atom pairs were considered; otherwise, all possible heavy atom pairs were used to calculate the RRCS. Next, we normalized f to reduce receptor variation among individuals.

$$x\prime = \frac{x - \min(F)}{\max(F) - \min(F)} \tag{5}$$

$$f_{norm} = [x\prime_1,\ x\prime_2,\ x\prime_3,\ \ldots\ ,\ x\prime_n] \tag{6}$$

where F is the total residue pair distance data from a targeted receptor and, f_{norm} is the normalized $f_{A:B}$. Finally, we sum the total f_{norm} elements to calculate RRCS:

$$RRCS_{A:B} = \sum f_{norm} \tag{7}$$

Using the RRCS, we calculated the general interactions between the total intact ORs. We summarized the total RRCS for each residue position from the entire OR. We set the threshold to a significant level for the RRCS. We set one RRCS per OR, which represented max(F); therefore, the threshold was set at 384.

Multivariate analysis

We employed Principal Component Analysis (PCA) and a Gaussian Mixture Model (GMM) to cluster the ligands responsive to specific hORs (Fig. 6). Prior to analysis, the input data underwent z-score normalization. Subsequently, after dimensionality reduction, we applied the GMM with 1000 iterations for the expectation-maximization algorithm and k was fixed at two. A full covariance matrix is used to construct the Gaussian model. To classify hORs (Fig. 2) and assess the similarity of ectopic expression patterns (Fig. 3), we employed Agglomerative Hierarchical Clustering (AHC) using Euclidean dissimilarity with Ward's linkage. Detailed procedures for classifying hORs are outlined in the "Procedure of classifying hORs" section of the methods. The statistical analysis was performed using MATLAB 2018b and R.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

Received: 16 August 2024; Accepted: 6 November 2024
Published online: 08 November 2024

References

1. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*. **65**, 175–187 (1991).
2. Pilpel, Y. & Lancet, D. The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8**, 969–977 (1999).
3. Probst, W. C., Snyder, L. A., Schuster, D. I., Brosius, J. & Sealfon, S. C. Sequence alignment of the G-protein coupled receptor superfamily. *DNA Cell. Biol.* **11**, 1–20 (1992).
4. Malnic, B., Hirono, J. & Sato, T. Buck, L. B. Combinatorial receptor codes for odors. *Cell*. **96**, 713–723 (1999).
5. Michaloski, J. S., Galante, P. A. F. & Malnic, B. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. *Genome Res.* **16**, 1091–1098 (2006).
6. Ressler, K. J., Sullivan, S. L. & Buck, L. B. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell*. **73**, 597–609 (1993).
7. Bushdid, C. et al. Mammalian class I odorant receptors exhibit a conserved vestibular-binding pocket. *Cell. Mol. Life Sci.* **76**, 995–1004 (2019).
8. Gelis, L., Wolf, S., Hatt, H., Neuhaus, E. M. & Gerwert, K. Prediction of a ligand-binding niche within a human olfactory receptor by combining site-directed mutagenesis with dynamic homology modeling. *Angew Chem. Int. Ed.* **51**, 1274–1278 (2012).
9. Geithe, C., Protze, J., Kreuchwig, F., Krause, G. & Krautwurst, D. Structural determinants of a conserved enantiomer-selective carvone binding pocket in the human odorant receptor OR1A1. *Cell. Mol. Life Sci.* **74**, 4209–4229 (2017).
10. de March, C. A., Kim, S. K., Antonczak, S., Goddard, W. A. & Golebiowski, J. G protein-coupled odorant receptors: from sequence to structure. *Protein Sci.* **24**, 1543–1548 (2015).
11. Gaillard, I., Rouquier, S., Chavanieu, A., Mollard, P. & Giorgi, D. Amino-acid changes acquired during evolution by olfactory receptor 912 – 93 modify the specificity of odorant recognition. *Hum. Mol. Genet.* **13**, 771–780 (2004).
12. Imai, T., Suzuki, M. & Sakano, H. Odorant receptor-derived cAMP signals direct axonal targeting. *Science*. **314**, 657–661 (2006).
13. Kato, A., Katada, S. & Touhara, K. Amino acids involved in conformational dynamics and G protein coupling of an odorant receptor: targeting gain-of-function mutation. *J. Neurochem*. **107**, 1261–1270 (2008).
14. Nguyen, M. Q., Zhou, Z., Marks, C. A., Ryba, N. J. P. & Belluscio, L. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell*. **131**, 1009–1017 (2007).
15. Zhou, Q. et al. Common activation mechanism of class A GPCRs. *eLife* **8**, (2019).
16. Choi, C. et al. Understanding the molecular mechanisms of odorant binding and activation of the human OR52 family. *Nat. Commun.* **14**, 8105 (2023).
17. Ikegami, K. et al. Structural instability and divergence from conserved residues underlie intracellular retention of mammalian odorant receptors. *Proc. Natl. Acad. Sci. USA*. **117**, 2957–2967 (2020).
18. Fritze, O. et al. Role of the conserved NPxxY(x)5,6F motif in the rhodopsin ground state and during activation. *Proc. Natl. Acad. Sci. USA*. **100**, 2290–2295 (2003).
19. Billesbølle, C. B. et al. Structural basis of odorant recognition by a human odorant receptor. *Nature*. **615**, 742–749 (2023).
20. Kotthoff, M., Bauer, J., Haag, F. & Krautwurst, D. Conserved C-terminal motifs in odorant receptors instruct their cell surface expression and cAMP signaling. *FASEB J.* **35**, e21274 (2021).
21. de March, C. A., Matsunami, H., Abe, M., Cobb, M. & Hoover, K. C. Genetic and functional odorant receptor variation in the Homo lineage. *iScience* 26, 105908 (2023).
22. Ghosh, S. et al. Sequence coevolution and structure stabilization modulate olfactory receptor expression. *Biophys. J.* **121**, 830–840 (2022).
23. 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).
24. Lapinsh, M. et al. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.* **11**, 795–805 (2002).
25. Davies, M. N. et al. Proteomic applications of automated GPCR classification. *Proteomics*. **7**, 2800–2814 (2007).
26. Nagarathnam, B. et al. TM-MOTIF: an alignment viewer to annotate predicted transmembrane helices and conserved motifs in aligned set of sequences. *Bioinformation*. **7**, 214–221 (2011).
27. Zozulya, S., Echeverri, F. & Nguyen, T. The human olfactory receptor repertoire. *Genome Biol.* **2**, RESEARCH0018 (2001).
28. Glusman, G. & Lancet, D. Visualizing large-scale genomic sequences. *IEEE Eng. Med. Biol. Mag.* **20**, 49–54 (2001).
29. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
30. Jundi, D. et al. Expression of olfactory receptor genes in non-olfactory tissues in the developing and adult zebrafish. *Sci. Rep.* **13**, 4651 (2023).
31. Ferrer, I. et al. Olfactory receptors in Non-chemosensory organs: the nervous system in Health and Disease. *Front. Aging Neurosci.* **8**, 163 (2016).
32. Nygaard, R., Frimurer, T. M., Holst, B., Rosenkilde, M. M. & Schwartz, T. W. Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol. Sci.* **30**, 249–259 (2009).
33. Niimura, Y. & Nei, M. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene*. **346**, 13–21 (2005).
34. Saito, H., Chi, Q., Zhuang, H., Matsunami, H. & Mainland, J. D. odor coding by a mammalian receptor repertoire. *Sci. Signal.* **2**, ra9 (2009).
35. Adipietro, K. A., Mainland, J. D. & Matsunami, H. Functional evolution of mammalian odorant receptors. *PLoS Genet.* **8**, e1002821 (2012).
36. Ahmed, L. et al. Molecular mechanism of activation of human musk receptors OR5AN1 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. *Proc. Natl. Acad. Sci. USA*. **115**, E3950–E3958 (2018).
37. Audouze, K. et al. Identification of odorant-receptor interactions by global mapping of the human odorome. *PLoS ONE.* **9**, e93037 (2014).
38. Geithe, C., Noe, F., Kreissl, J. & Krautwurst, D. The broadly tuned odorant receptor OR1A1 is highly selective for 3-Methyl-2,4-nonanedione, a key food odorant in aged wines, tea, and other Foods. *Chem. Senses*. **42**, 181–193 (2017).
39. Jaeger, S. R. et al. A mendelian trait for olfactory sensitivity affects odor experience and food selection. *Curr. Biol.* **23**, 1601–1605 (2013).
40. Mainland, J. D. et al. The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **17**, 114–120 (2014).
41. Mainland, J. D., Li, Y. R., Zhou, T., Liu, W. L. L. & Matsunami, H. Human olfactory receptor responses to odorants. *Sci. Data*. **2**, 150002 (2015).
42. Mashukova, A., Spehr, M., Hatt, H. & Neuhaus, E. M. Beta-arrestin2-mediated internalization of mammalian odorant receptors. *J. Neurosci.* **26**, 9902–9912 (2006).
43. Matarazzo, V. et al. Functional characterization of two human olfactory receptors expressed in the baculovirus Sf9 insect cell system. *Chem. Senses*. **30**, 195–207 (2005).
44. Menashe, I. et al. Genetic elucidation of human hyperosmia to isovaleric acid. *PLoS Biol.* **5**, e284 (2007).
45. Sanz, G., Schlegel, C., Pernollet, J. C. & Briand, L. Comparison of odorant specificity of two human olfactory receptors from different phylogenetic classes and evidence for antagonism. *Chem. Senses*. **30**, 69–80 (2005).

46. Sato-Akuhara, N. et al. Ligand specificity and evolution of mammalian musk odor receptors: effect of single receptor deletion on odor detection. *J. Neurosci.* **36**, 4482–4491 (2016).
47. Schmiedeberg, K. et al. Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. *J. Struct. Biol.* **159**, 400–412 (2007).
48. Spehr, M. et al. Identification of a testicular odorant receptor mediating human sperm chemotaxis. *Science*. **299**, 2054–2058 (2003).
49. Trimmer, C. et al. Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proc. Natl. Acad. Sci. USA*. **116**, 9475–9480 (2019).
50. Veitinger, T. et al. Chemosensory Ca2 + dynamics correlate with diverse behavioral phenotypes in human sperm. *J. Biol. Chem.* **286**, 17311–17325 (2011).
51. Lalis, M. et al. M2OR: a database of olfactory receptor-odorant pairs for understanding the molecular mechanisms of olfaction. *Nucleic Acids Res.* **52**, D1370–D1379 (2024).
52. 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).
53. Hasin, Y. et al. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet.* **4**, e1000249 (2008).
54. Zhang, X. et al. Characterizing the expression of the human olfactory receptor gene family using a novel DNA microarray. *Genome Biol.* **8**, R86 (2007).
55. Kimes, P. K., Liu, Y., Hayes, N., Marron, J. S. & D. & Statistical significance for hierarchical clustering. *Biometrics*. **73**, 811–821 (2017).

Acknowledgements

This work was supported by the Ministry of Education (2020R1A6A1A0304051621; RS-2023-00239274) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020R1A2C2005174).

Author contributions

S.E.R., J.B., and C.M. conceptualized this study. J.B. and K.K. performed the experiments and analyzed the data. T.S. and W.C.K. collected the data resources. S.E.R. and J.B. interpreted the results and wrote the manuscript. S.E.R., J.B., T.S., and C.M. edited and approved the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-79183-8.

Correspondence and requests for materials should be addressed to C.M.