The mutational landscape of human olfactory G protein-coupled receptors

Ramón Cierco Jimenez[1,2], Nil Casajuana-Martin[1], Adrián García-Recio[1], Lidia Alcántara[1], Leonardo Pardo[1], Mercedes Campillo[1] and Angel Gonzalez[1*] (iD)

Abstract

Background: Olfactory receptors (ORs) constitute a large family of sensory proteins that enable us to recognize a wide range of chemical volatiles in the environment. By contrast to the extensive information about human olfactory thresholds for thousands of odorants, studies of the genetic influence on olfaction are limited to a few examples. To annotate on a broad scale the impact of mutations at the structural level, here we analyzed a compendium of 119,069 natural variants in human ORs collected from the public domain.

Results: OR mutations were categorized depending on their genomic and protein contexts, as well as their frequency of occurrence in several human populations. Functional interpretation of the natural changes was estimated from the increasing knowledge of the structure and function of the G protein-coupled receptor (GPCR) family, to which ORs belong. Our analysis reveals an extraordinary diversity of natural variations in the olfactory gene repertoire between individuals and populations, with a significant number of changes occurring at the structurally conserved regions. A particular attention is paid to mutations in positions linked to the conserved GPCR activation mechanism that could imply phenotypic variation in the olfactory perception. An interactive web application (hORMdb, Human Olfactory Receptor Mutation Database) was developed for the management and visualization of this mutational dataset.

Conclusion: We performed topological annotations and population analysis of natural variants of human olfactory receptors and provide an interactive application to explore human OR mutation data. We envisage that the utility of this information will increase as the amount of available pharmacological data for these receptors grow. This effort, together with ongoing research in the study of genetic changes in other sensory receptors could shape an emerging sensegenomics field of knowledge, which should be considered by food and cosmetic consumer product manufacturers for the benefit of the general population.

Keywords: Olfactory receptors, OR, Natural variants, Mutations, 7-TM receptors, G protein-coupled receptor, GPCR, Sensegenomics, Database

* Correspondence: Angel.Gonzalez@uab.es
[1]Laboratori de Medicina Computacional, Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain
Full list of author information is available at the end of the article

Jimenez *et al. BMC Biology*   (2021) 19:21

Page 2 of 15

Background

Vertebrate olfactory systems have evolved to sense volatile substances through their recognition by olfactory receptors (ORs) located on the membrane of olfactory sensory neurons in the olfactory epithelium [1] and consequent initiation of signaling cascades that transform odorant-receptor chemical interactions into electrochemical signals [2, 3]. These receptors belong to the class A G protein-coupled receptors (GPCRs), a major drug target protein family [4] involved in the transduction of extracellular signals through second messenger cascades controlled by different heterotrimeric guanine nucleotide-binding proteins (G_{olf} in the case of ORs) coupled at their intracellular regions [5, 6].

ORs are characterized by intronless coding regions of an average length of 310 codons (~ 1 kb) and constitute the largest multigene family in humans, with around 400 intact (functional) loci, divided into two main classes, 18 families and more than 150 subfamilies [7, 8]. This broad array of receptors, like in other terrestrial mammals, is shared with tetrapods (families 1–14) and marine vertebrates (families 51–56) [9] and seems necessary to respond efficiently to the extraordinary chemical diversity of odorants in Earth's ecosystems [10]. However, there is growing evidence that their functional roles are beyond olfactory tissues [11, 12].

Human genomic data reveal that OR loci harbor a considerable number of genetic variants and a high proportion of pseudogenes [13, 14]. Many of these changes may interfere with the receptor expression, interaction with odorants, or signal transduction and consequently could modify the physiological response to a determinate olfactory stimulus. In this regard, it has been long established a considerable variation in the perception of odorants among individuals [10, 15] and populations [16, 17], which in some cases has been associated to genetic changes in OR genes [18–20]. To further study this issue, we used publicly available human sequencing data to conduct in silico data mining and analysis of OR natural variants in 141,456 human exomes and genomes from more than one hundred thousand unrelated individuals [21].

Information of chromosomal localization, type of substitutions, and allele frequencies in several subcontinental populations was obtained for close to a hundred and twenty thousand natural variants identified in 378 human ORs. A detailed topological localization system was developed to assign each mutation to a region within the seven alpha-helical bundle molecular architecture characteristic of GPCRs (i.e., extracellular and intracellular N- and C-terminal sequences, seven transmembrane α-helices [TM 1 to 7], and three extracellular [ECL 1 to 3] and three cytoplasmic [ICL 1 to 3] loops) [22, 23]. This system also includes the assignment of unambiguously positions to all mutations occurring in the TM helices according to the numbering systems developed by Ballesteros-Weinstein (BW) and others for this family of proteins [24, 25].

The analysis of the collected data revealed numerous differences among individuals and populations, with an allele frequency spectrum dominated by low-frequency variants. A significant number of natural changes were identified at GPCR functional regions [26–30] or forming part of ligand-binding cavities [31, 32]. These and the rest of the coding sequence mutations were evaluated according to an amino acid substitution score weighting developed for this family of receptors [33]. The utility of this topological annotation approach is illustrated with selected examples of natural OR variations that could imply phenotypic changes in the odorant perception for a substantial group of individuals. These results are accompanied by a computational application developed to facilitate the public access and analysis of this data. The human Olfactory Receptor Mutation Database (hORMdb) is an interactive database that allows the selection and filtering of human OR natural variants and the analysis of specific dbSNP entries, individual genes or complete families according to their topological localization, population frequencies, and substitution scores, among other features.

Results

Natural variations in human ORs were mined from nucleotide sequence data of 141,456 unrelated individuals in the Genome Aggregation Database (gnomAD) (Additional file 1: Table S1) [21] and annotated at structural level with information of the class A GPCR family as resumed in Fig. 1. This curated dataset comprises 119,069 nucleotide changes in 378 functional OR genes, which belong to 17 OR families (Fig. 2). The overall average number of mutations per receptor was 315, with a prominent variation rate in the OR52 family (average of 343) and five members of the OR4 family (OR4A5, OR4A15, OR4A16, OR4C16, and OR4C46) with more than 500 mutations/receptor. On the other hand, the lowest variation rates correspond to the OR14 family (average of 265) and few more than a dozen receptors with less than 100 mutation counts (Additional file 1: Tables S2-S3). This staggered mutational distribution supports a heterogeneous selective pressure in OR genes, as indicated by other studies [7, 34].

The most common variation types in the collected dataset correspond to missense (~ 64%) and synonymous substitutions (~ 25%), followed by frameshifts, non-coding (3′ UTR and 5′ UTR), stop gained, and a reduced number of other minor mutations events (Fig. 3a). Regarding the nature of the changes, transitions and transversions are the most likely mutational events, representing > 95% of
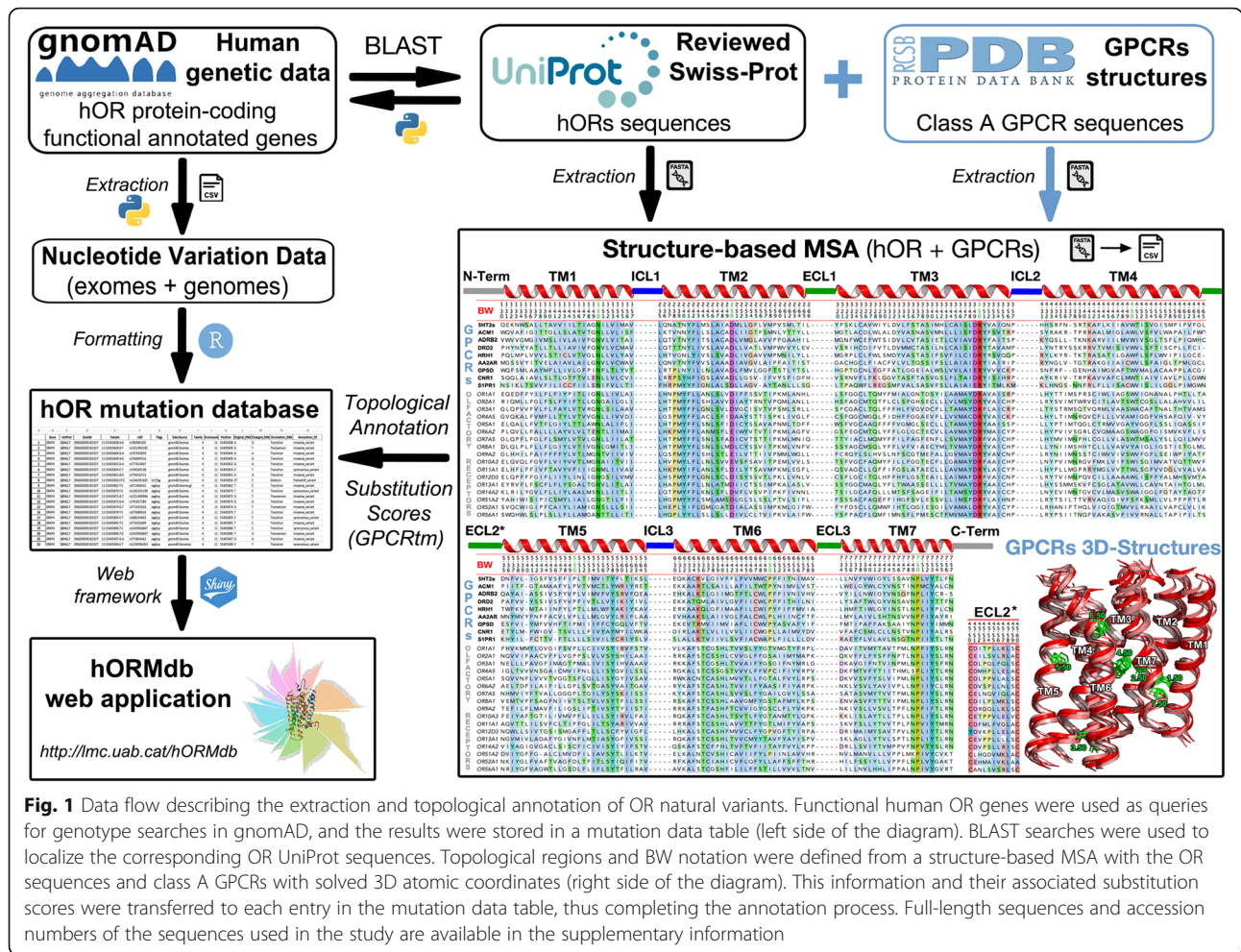
Fig. 1 Data flow describing the extraction and topological annotation of OR natural variants. Functional human OR genes were used as queries for genotype searches in gnomAD, and the results were stored in a mutation data table (left side of the diagram). BLAST searches were used to localize the corresponding OR UniProt sequences. Topological regions and BW notation were defined from a structure-based MSA with the OR sequences and class A GPCRs with solved 3D atomic coordinates (right side of the diagram). This information and their associated substitution scores were transferred to each entry in the mutation data table, thus completing the annotation process. Full-length sequences and accession numbers of the sequences used in the study are available in the supplementary information

the entire dataset, while the remaining correspond to deletions and insertions (inset on Fig. 3a). The large OR multigene family occupies vast amounts of genomic territory. As expected, the number of mutations per chromosome is linked to the genome distribution of the OR genes (Fig. 3b). Chromosome 11, which contains the largest number of receptors, displays the highest number of variants, followed by chromosome 1. For the rest of chromosomes hosting OR genes, the number of variants ranges from ~ 7000 to less than 100, and no data was recorded for chromosomes 4, 13, 18, 20, 21, and Y. A graphical display of the unevenly chromosomal distribution of the mutations within the OR families is available in Additional file 2: Figure S1.

Allele frequencies and population distribution of the variant dataset

Analysis of the frequency values from gnomAD discloses only 2182 natural OR variants with global allele frequency above 1% in the collected dataset. By contrast, > 95% correspond to low-frequency variants (60,312 of which are singletons), exposing an extraordinary interindividual

variation in the human OR gene repertoire. Taking into account that differences in olfactory sensitivity could be at least partly explained by the prevalence of particular mutated OR alleles in individuals within populations [35], independent frequency ranges were analyzed on each of the seven sub-continental populations in the database (Fig. 4a–g, Additional file 1: Table S1) [21]. This analysis shows a similar trend of frequency distribution among ethnic groups, characterized by an elevated number of mutations with allele frequencies below 0.1%. From these changes, 37,013 were exclusively found in the European (non-Finnish), 14,763 in South Asian, 11,178 in African, 10,579 in Latino, 9935 in East Asian, 1784 in Finnish, and 819 in Ashkenazi Jewish populations.

On another note, assessment of the concurrence of mutations reveals 2130 genetic variants common to all populations, of which 1844 display allele frequencies > 1%. Notwithstanding, 29,230 variants were identified in two or more ethnic groups. Pair-wise comparisons of shared mutations between sub-continental populations are summarized in the circos plot of Fig. 4h. As observed in the graph, the largest European (non-Finnish) population
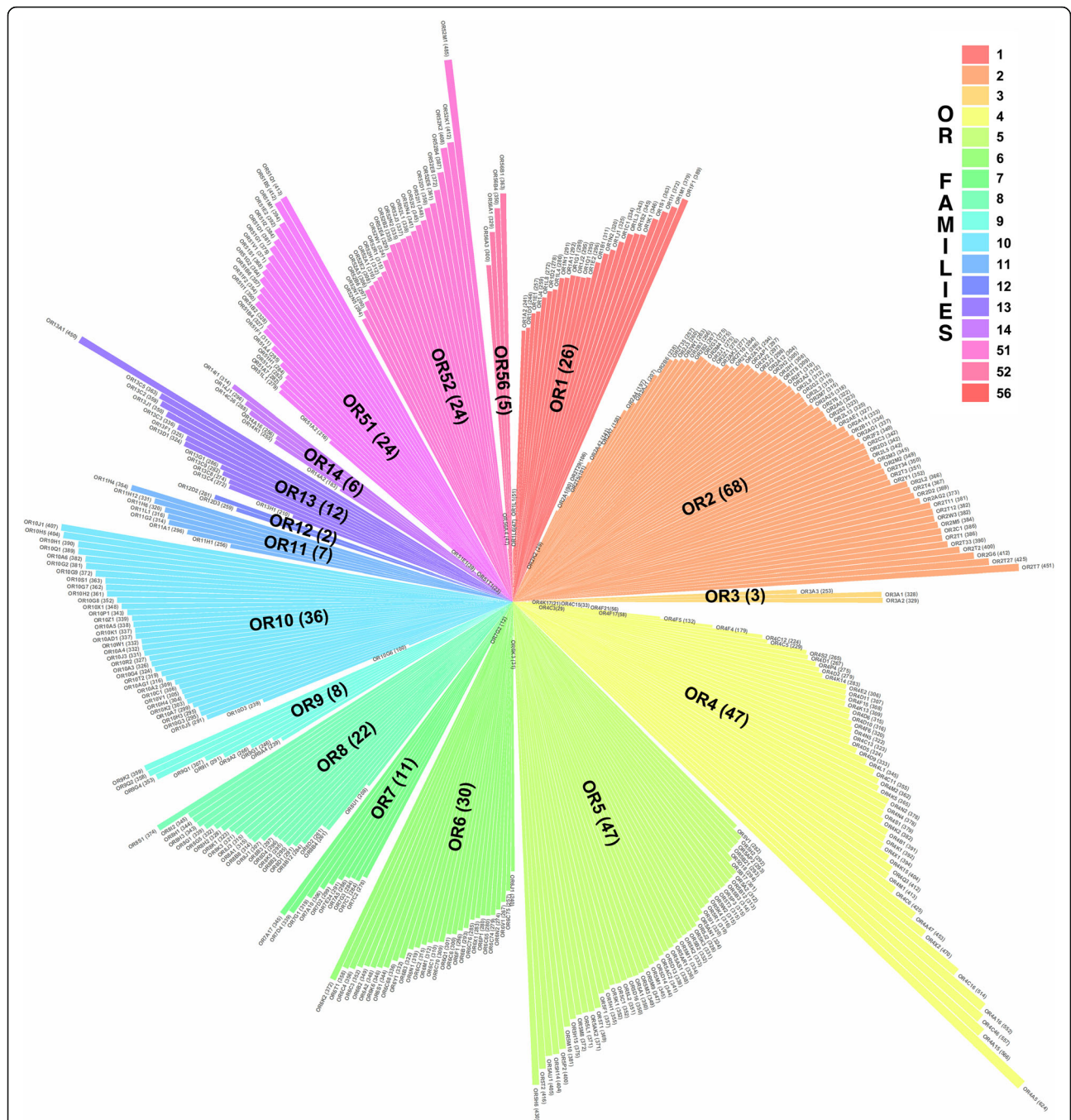
Fig. 2 Wind rose plot representing the mutational landscape of human ORs. The plot shows the distribution of 119,069 nucleotide variants in 378 functional OR genes (color bars) clustered in 17 families (top right legend, ordered clockwise). Numbers in parenthesis correspond to the total number of receptors analyzed at each family. Gene names and family assignation correspond to the recommended terms by the HUGO Gene Nomenclature Committee (HGNC). Names of individual OR family members are in the format "ORnXm": a root name "OR", followed by family numeral (n), subfamily letter (X), and a numeral (m) representing the particular gene within the subfamily. For example, OR2A1 is the first OR gene in the family 2, subfamily A

shares more variants with the rest of the ethnicities. This data, expressed as a percentage of the total number of mutations at each population indicates that approximately 82% of the Ashkenazi Jewish, 72% of Finnish, 48% of Latino, 46% of African, 38% of South Asian, and 36% of the East

Asian natural variants are shared with the European (non-Finnish) population. Likewise, African and Latino share ~ 38% of mutations, whereas the South Asian population shares ~ 28% of mutations with Latino, ~ 26% with African, and less than 20% with the East Asian population.
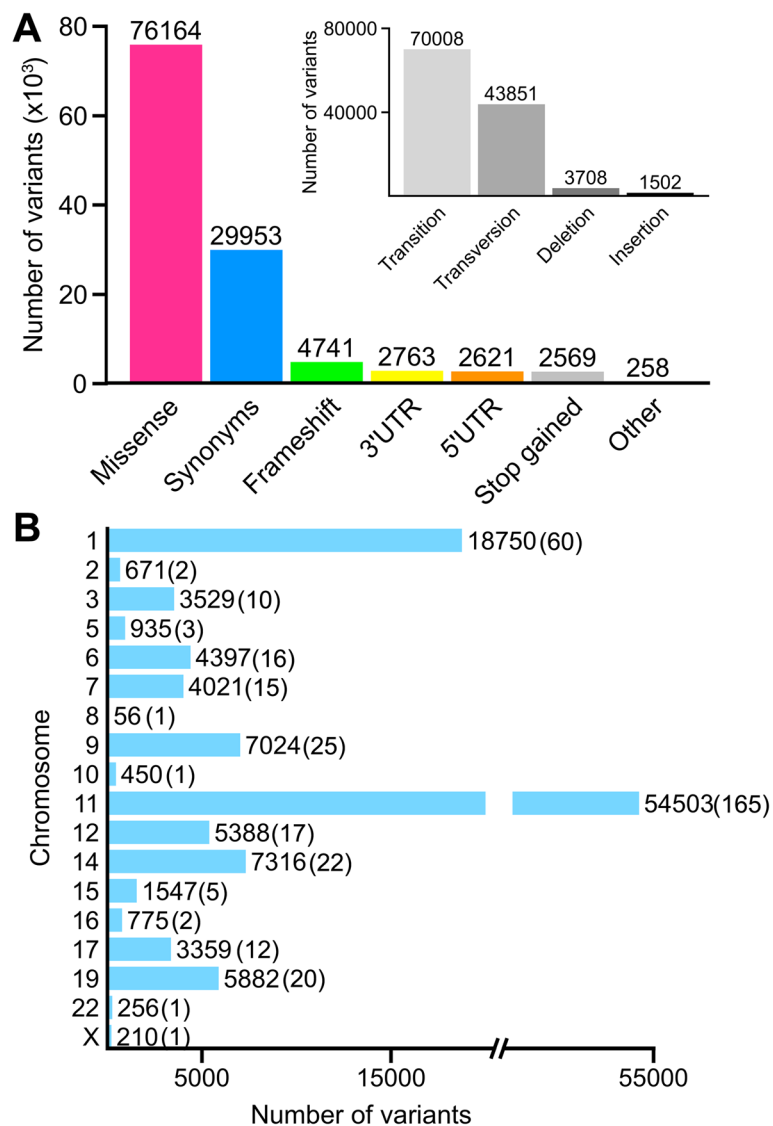
Jimenez *et al. BMC Biology*   (2021) 19:21

Page 5 of 15

Fig. 3 Categorization and chromosomal distribution of human OR natural variants. **a** Functional categories and type of changes (inset) of the sequence variants identified in the 378 investigated OR genes. **b** Number and distribution of the OR variants per chromosome (the number of OR analyzed per chromosome are shown in parenthesis)

Topological assignment to sequence variants

Topological domain assignation of coding sequence variants according to the conserved class A GPCR molecular architecture (i.e., N-term, 7-TMs, 3-ECLs, 3-ICLs, and C-term defined in the structure-based multiple sequence alignment (MSA) in Fig. 1), revealed that ∼ 66% of the mutations were located in the TM regions (53,533 missense, 21,273 synonymous, 3237 frameshifts, and 1656 stop gained variants) (Fig. 5). TM6 accumulates more changes (13,232 variants), followed by TM3 (12,919), TM5 (11,322), TM2 (11,319), and ECL2 (11,151). On the other hand, a lower number of mutations were found in intracellular and extracellular loops, N- and C-terminal domains, and non-coding regions (NCRs). This trend is

observed in all OR families, with most changes occurring in the TMs and ECL2, and major inter-family differences in the NCRs and N- and C-terminal domains because of their variable lengths (Additional file 2: Figure S2).

The analysis of individual positions within the conserved GPCR topological domains, using the BW nomenclature, reveals that, overall, the occurrence of natural variants is not restricted to a specific TM region or particular location, with an average of 361 changes per site (Fig. 6). However, position 3.50 (967 total variations, 826 missense) stands out from the rest of the sites (Fig. 6c). This conserved position constitutes a switch for the signal transmission mechanism, which involves the structural rearrangement of the TM regions, opening
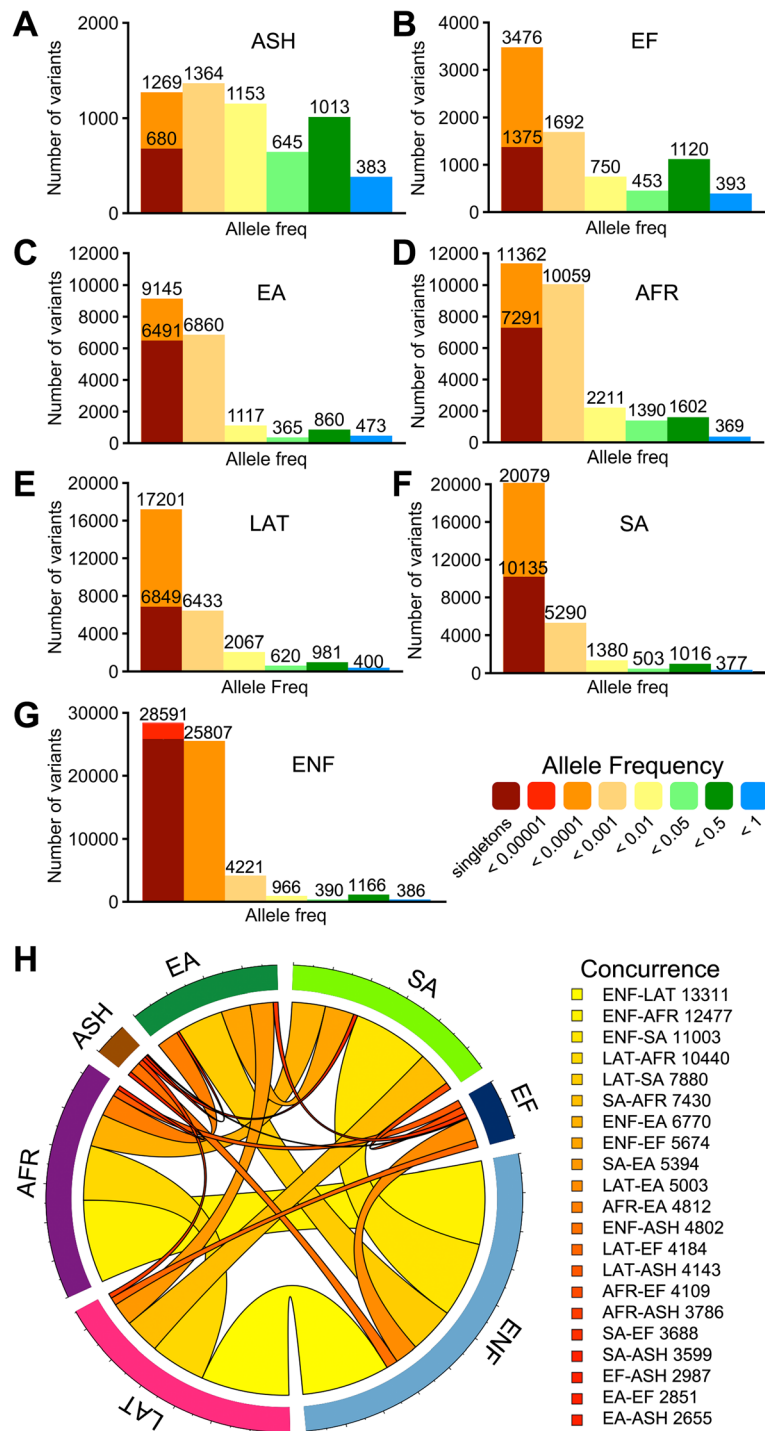
Fig. 4 Allele frequencies and concurrence of OR mutations within human populations. The number and frequency distribution of natural variants from seven sub-continental populations obtained from gnomAD and corresponding to **a** Ashkenazi Jewish (ASH), **b** European Finnish (EF), **c** East Asian (EA), **d** African (AFR), **e** Latino (LAT), **f** South Asian (SA), and **g** European non-Finnish (ENF). Bars are colored according to the allele frequency scale on the right. **h** Circos plot of the concurrence of mutations between populations. The width of the link between two populations corresponds to the number of shared mutations between them, also represented by a color gradient (numerical values are displayed on the right)

the intracellular cavity for G protein binding, through changes in the DR$^{3.50}$Y interaction environment [36, 37]. Consequently, this position is very sensitive to natural

sequence variations linked to pathological outcomes in several GPCRs [38–42]. This high variant enrichment has been noted earlier, and although there is no

Jimenez *et al. BMC Biology*   (2021) 19:21

Page 7 of 15

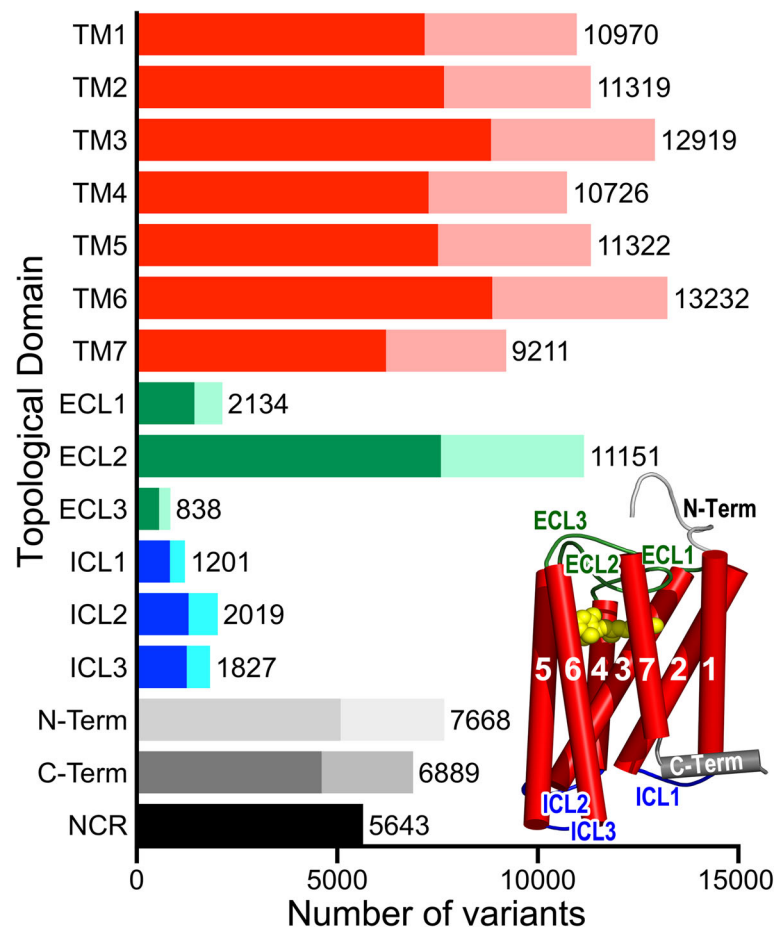**Fig. 5** Number and distribution of mutations within topological domains. The number of variants (*x*-axis) per topological domain (*y*-axis) as defined by the conserved class A GPCR molecular architecture of seven TM helices (TM1 to 7 in red), three extracellular loops (ECL 1 to 3 in green), three intracellular loops (ICL 1 to 3 in blue), and N- and C-terminal regions (in gray). NCR, non-coding regions. The darkest regions on the *y*-axis bars correspond to missense substitutions. The cartoon model in the lower right exemplifies the extent and arrangement of the topological regions according to the crystallographic structure of rhodopsin (PDBid: 4J4Q, retinal is shown in yellow vdW spheres)
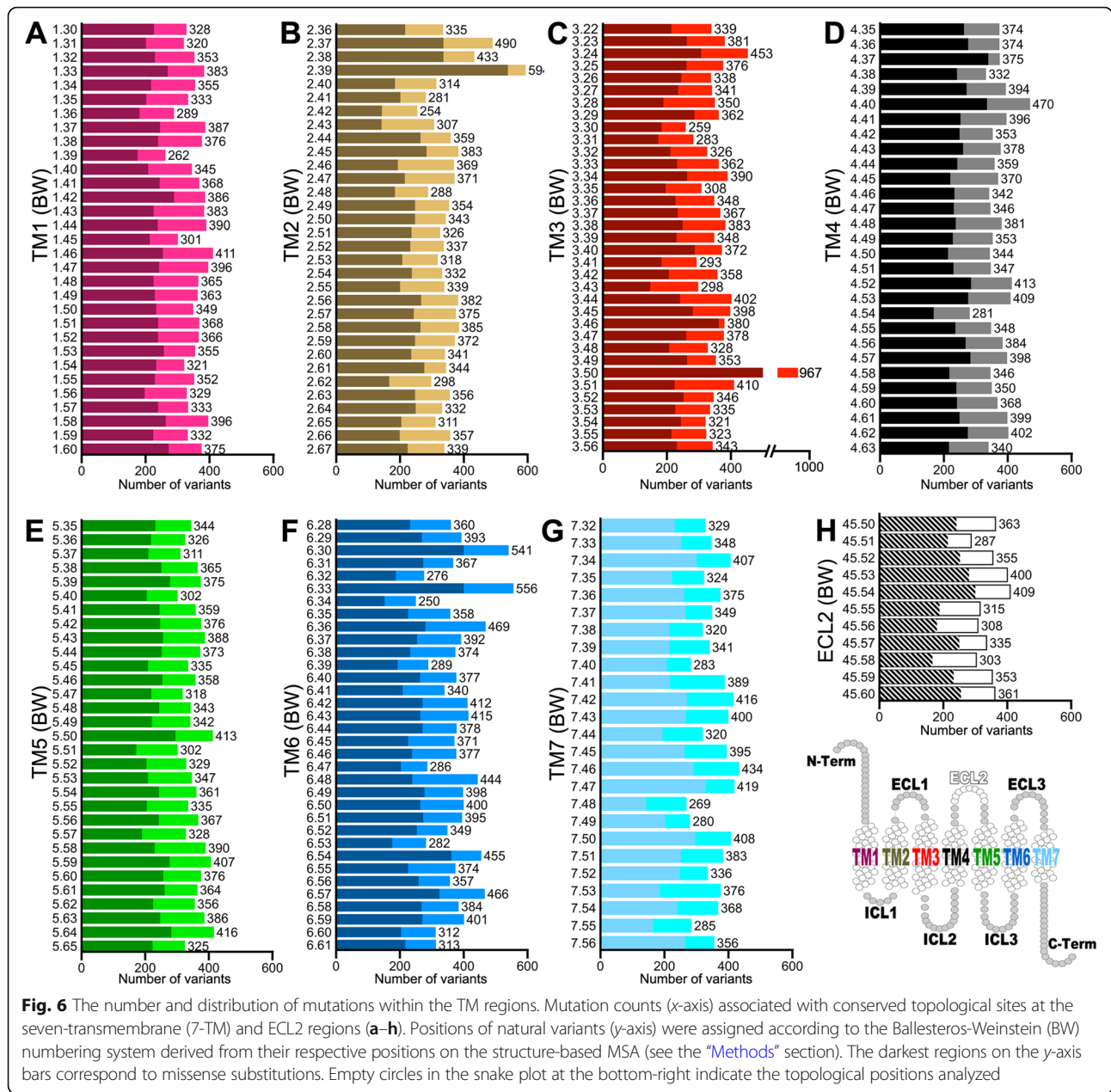
conclusive evidence, it has been suggested a positive selection at this position [43]. Interestingly, most frequent substitutions of the conserved Arg[3.50] (96% conservation in GPCRs, 92% in ORs) involved the amino acids His (195 occurrences) and Cys (188 occurrences), which is consistent with a previous study conducted on non-olfactory GPCRs [44].

Mutability landscapes of amino acids changes

Single amino acid variants in human ORs can alter the resulting phenotype, for example, by altering the odorant perception [45]. Thus, we investigate the type and magnitude of the amino acid changes in missense substitutions (76,164 variants in the dataset) as the first approximation to evaluate their functional consequences at the molecular level. As displayed in Fig. 7a, hydrophobic residues (Leu, Ile, Val, and Ala) exhibited the highest levels of mutability, followed by Ser and Thr in

agreement with their stabilization roles on the structure of TM helices [46, 47]. Conversely, substitutions of Trp or polar/charged Gln, Glu, Lys, Asp, His, and Asn (often associated with protein malfunction in TM proteins) were less frequent [48, 49].

The evaluation of the magnitude of changes was conducted using amino acid substitution scores derived from more than one thousand class A GPCR sequences (including ORs) and thus reflecting the compositional bias distinctive of this particular family of proteins (Fig. 7b, Additional file 2: Figure S3) [33]. From this analysis, ~ 68% of the missense substitutions were associated with zero or positive substitution scores (52,048 variants), indicating a preservation of physico-chemical properties of the original residue. Nonetheless, 24,116 changes compute negative scores, reflecting significant differences between the original and substituted amino acid, with possible impact on the receptor structural integrity and/or the binding of odorant molecules.

Fig. 6 The number and distribution of mutations within the TM regions. Mutation counts (x-axis) associated with conserved topological sites at the seven-transmembrane (7-TM) and ECL2 regions (**a**–**h**). Positions of natural variants (y-axis) were assigned according to the Ballesteros-Weinstein (BW) numbering system derived from their respective positions on the structure-based MSA (see the "Methods" section). The darkest regions on the y-axis bars correspond to missense substitutions. Empty circles in the snake plot at the bottom-right indicate the topological positions analyzed

Use of topological annotation, substitution metrics, and allele frequencies in the impact evaluation of the mutations

Topological mapping of natural variations and their associated substitution scores were used in the functional imputation of missense substitutions. These features were analyzed in two subsets of topological positions within the conserved TMs and ECL2, which could either be involved in the receptor integrity and functional mechanism (functional core, FC) or in odorant-receptor interactions (binding cavity, BC) (Additional file 2: Figures S4-S5). FC and BC topological

subsets comprise 60 BW annotated positions and accumulate 8049 and 7394 missense variants counts, respectively, of which 5554 computed negative substitution scores. From this data, we identify 80 changes with allele frequencies > 1% in at least one of the subcontinental populations that could implicate distinctive odorant sensitivities for a considerable group of carriers (Additional file 2: Figure S6). At the moment, based on the limited published information of known ligands for human ORs, we can only hypothesize about the impact of such changes through a few concrete examples described below:
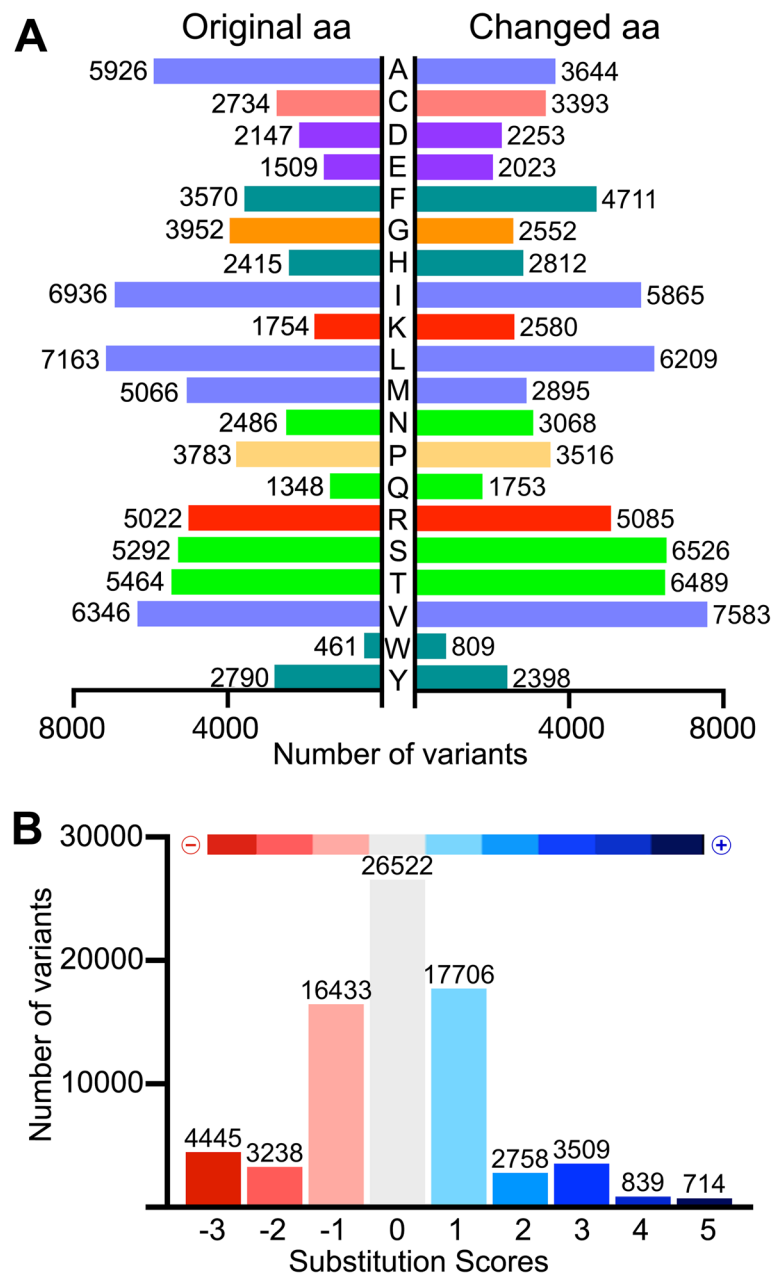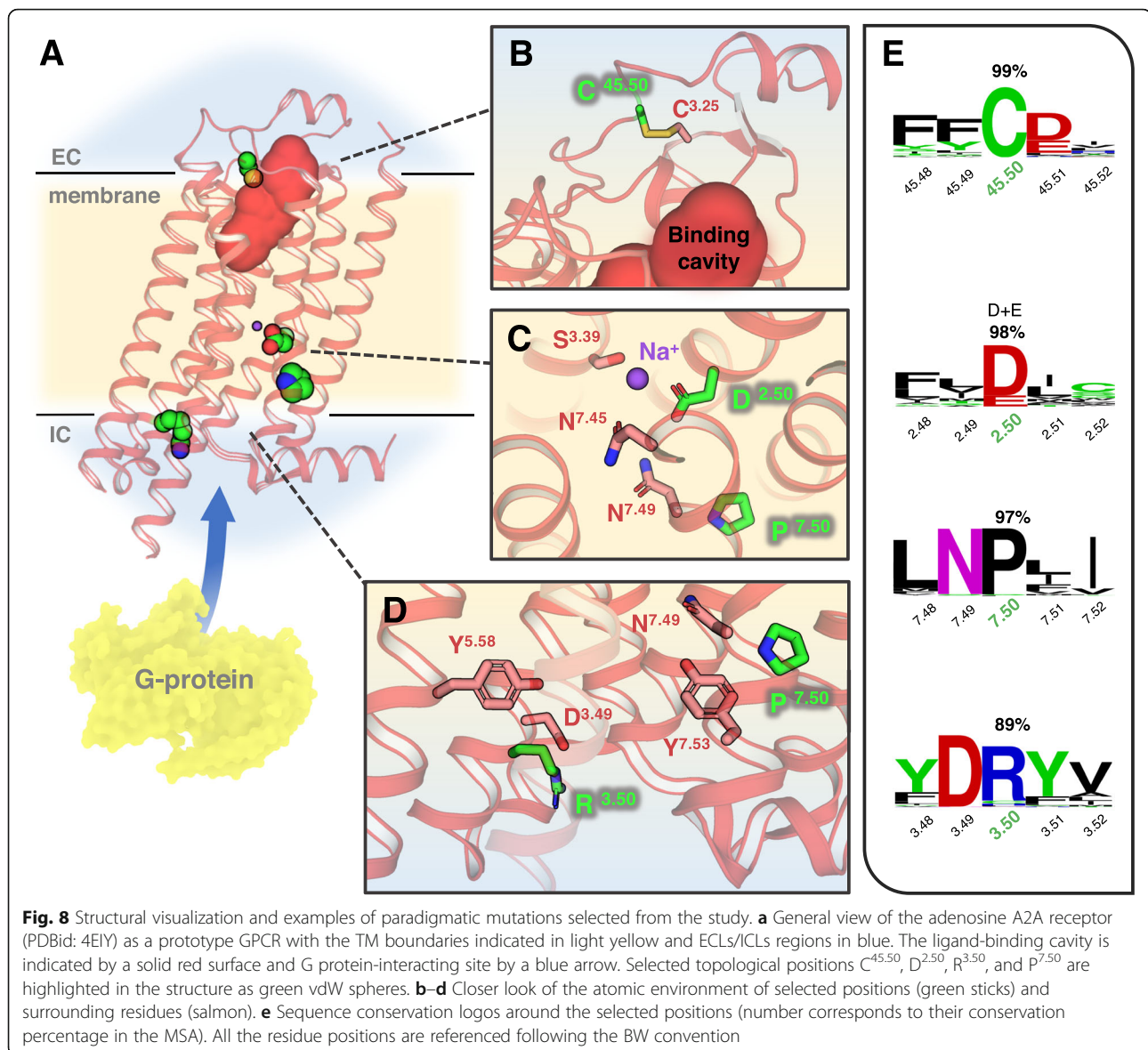
Fig. 7 Mutability landscapes of amino acid substitutions in human ORs. **a** The number of original (left) vs. changed (right) amino acids due to missense substitutions. Amino acid bars are colored by physico-chemical properties of the residues (blue = hydrophobic, green = polar, dark cyan = aromatic, purple = negatively charged, red = positively charged, salmon, yellow and orange = special residues Cys, Pro, and Gly). **b** Categorization of all amino acid replacements according to substitution scores extracted from GPCRtm. The color scale bar on top indicates the range of score values for all computed changes (negative = red, positive = blue)

Extracellular loop 2 at the conserved Cys45.50

A conserved cysteine residue in this position is involved in a disulfide bridge between ECL2 and TM3 in > 80% of class A GPCRs, and its substitution is related to a loss of function [30, 50, 51] (Fig. 8a, b, e). An example of this type of mutation is found in the OR8B4, a recently deorphanized receptor for anisic aldehyde and muguet alcohol [52]. Variation rs4057749 (c.532 T>C, p.Cys178Arg) in the OR8B4 may lead to impairment in the ability to perceive these aromatic cosmetic substances in a considerable proportion of the population (Additional file 2: Figure S6).

Fig. 8 Structural visualization and examples of paradigmatic mutations selected from the study. **a** General view of the adenosine A2A receptor (PDBid: 4EIY) as a prototype GPCR with the TM boundaries indicated in light yellow and ECLs/ICLs regions in blue. The ligand-binding cavity is indicated by a solid red surface and G protein-interacting site by a blue arrow. Selected topological positions $C^{45.50}$, $D^{2.50}$, $R^{3.50}$, and $P^{7.50}$ are highlighted in the structure as green vdW spheres. **b**–**d** Closer look of the atomic environment of selected positions (green sticks) and surrounding residues (salmon). **e** Sequence conservation logos around the selected positions (number corresponds to their conservation percentage in the MSA). All the residue positions are referenced following the BW convention

Transmembrane helix 2 at the conserved Asp2.50

It is characterized by the presence of a negative ionizable residue in the conserved $(N/S)LxxxD^{2.50}$ motif, which is involved in the GPCR activation mechanism through allosteric modulation mediated by ionic species [53] (Fig. 8a, c, e). Replacement of the conserved $D^{2.50}$ would impair the coordination of modulating ions due to the loss of the negatively ionizable center [54]. Carriers of mutations on this site, such as the rs4501959 (c.262G>A, p.Asp88Asn) in the OR52L1, might have different abilities to perceive carboxylic acids present in human sweat [55], and some of the components from the butter smell like butanoic acid and gamma decalactone that interact with this receptor [56].

Transmembrane helix 7 at the conserved Pro7.50

A conserved Pro in this position forms part of the $NP^{7.50}xxY$ motif involved in the transition from the ground state to the active forms of the GPCRs and internalization [57] (Fig. 8a, c–e). Substitution of the $P^{7.50}$ would modify the TM7 conformation producing a change of signalization patterns as observed in rhodopsin [29]. An example of mutation on this site is found in the OR1A1, rs769427 (c.853C>T, p.Pro285Ser), which probably would affect their carriers for the detection of citronellic terpenoid substances identified as ligands for this receptor [58].

Transmembrane helix 3 at the conserved Arg3.50

A conserved Arg is the central component in the $DR^{3.50}Y$ motif directly implicated in the general

Fig. 9 Overview of the hORMdb web interface. The hORMdb is an online resource to study natural variations in human ORs, and it is structured in three main panels. **a** A mutation data panel containing a downloadable table of the natural variants with multiple columns of information. **b** A filtering variable panel allows the selection, concatenation, and filtering of the data. **c** A graphical panel for the visualization of the mutation data through several interactive graphs. Further details of the contents and organization of the database are provided in the HELP panel

activation mechanism of the class A GPCRs and its substitution generally modifies the transduction capacity of the receptor [26, 38–42] (Fig. 8a, d, e). Natural variations at this position are found in most ORs, some of them at moderate to high frequencies in the populations investigated; examples include rs2072164 in OR2F1, rs3751484 in OR6J1, rs10176036 in OR6B2, rs12224086 in OR5AS1, rs2512219 in OR8D2, rs16930982 in OR51I1, and rs11230983 in OR5D13.

Development of an interactive application to explore the human OR mutation data

It is expected that progress on OR genome association studies will continue to be made in the future. Thus, an

interactive computational application was developed for the free access and analysis of this data by academics and industry professionals. The human Olfactory Receptor Mutation Database (hORMdb) provides a curated and downloadable repository of natural variations in human ORs and several interactive tools for the selection, filtering, and analysis of its contents (Fig. 9).

The hORMdb is structured as a data table (Fig. 9a), containing information about individual dbSNP entries, particular genes, or entire OR families, including the types of nucleotide and amino acid changes, allele frequencies in several sub-continental populations, and topological location in the receptor structure. All the mutation data can be selectively accessed through a filtering variable panel (Fig. 9b) that allows the possibility of concatenate multiple selection choices (including numerical ranges for allele frequencies) or predefined topological subsets to analyze (e.g., BC, FC). Finally, a graphical panel interface (Fig. 9c) allows to interactively display the selected content according to receptor types, chromosomal location, mutation impact, original/changed amino acids, substitution score, topological domain, BW position, allele frequencies, and concurrence within populations. Altogether, this tool is intended to be used for the functional assessment of natural variations, rationalization of mutation data experiments, or comparative population studies.

Discussion

It is common ground that olfactory sensitivity differs across individuals, and in some cases, this feature has been related to genetic variations. Thus, the contribution of the genotype in the perception of odorants and volatile chemical mixtures seems particularly relevant. The highly diverse ORs, at the membrane of the olfactory neurons, trigger the first input of the olfactory signal. Thus, genomic studies of this family of receptors represent an important source of knowledge for academics and industry professionals who study human olfaction. To this end, we can take advantage of the vast amount of information on natural genetic variations coming from the genome-data community shared initiatives freely available in the public domain.

Using data mining tools, close to 120,000 nucleotide variations in human ORs were obtained from the large-scale sequencing data repository gnomAD, which provides well-structured information of sequencing data from a wide variety of sequencing projects all over the world [21]. The curation and computer analysis of this variation data revealed an uneven distribution of mutations in OR genes, reflecting the active role of natural selection in this family of receptors. Moreover, a considerable proportion of the identified mutations occur at very low frequencies, many of them uniquely identified

at definite ethnic groups or individuals. This extraordinary genotypic variation has been earlier described [59] and suggests a great phenotypic diversity in the olfactory perception between humans.

The striking variation in the OR gene repertoire has motivated their study and characterization by computational methods for several years [60]. These tools have been fundamental in the identification of inactive members of the family (e.g., the Classifier for Olfactory Receptor Pseudogenes (CORP) algorithm [14]), as well as for exploring the olfactory repertoires (e.g., the Olfactory Receptors Database (ORDB) [61] and the Human Olfactory Data Explorer (HORDE) [62]). Nevertheless, more progress is required in the development of new data analysis interfaces that facilitate the integration of OR information with structural knowledge. Taking into account the increasing need for tools providing accurate predictions of functional consequences of natural variants identified in genomic studies [63]; evolutionary conservation and structural context were considered as key elements in the estimation of the functional role of the natural variations identified. It is worth stressing that, in many cases, the structural framework of the mutated sites (intimately linked to the stability, function, and interactions) is often overlooked due to a limited structural knowledge [64]. ORs are not an exception to this reality, with no molecular structure reported to date. However, the highly conserved molecular architecture and sequence motifs that characterize the class A GPCR family make it possible to reliably predict the topological positions of the identified mutations from structure-informed sequence alignments. Using this approach, we provide a 3D context for the many variants occurring in ORs facilitating the functional interpretation of the changes attending to their structural location, biochemical associated data, and substitution score weightings. This method is exemplified through the identification of several natural OR variants located at conserved topological sites (e.g., BW 2.50, 3.50, 7.50, 45.50 at ECL2), either involved in the structural stability or in the functional mechanism of the receptors, and which might induce changes in the odorant sensitivity.

We believe the integration of high-throughput sequencing data with structural information is crucial for the interpretation of the complex genotype-phenotype associations occurring not only in human olfaction, but also in any other biological process. These would require in many cases the development of automatic interfaces to facilitate the management and organization of large quantities of data. Hence, we developed an interactive computational application that integrates both genomic and structural knowledge with analytical graphical tools for the study of the OR mutational landscape. The human Olfactory Receptor Mutation Database (hORMdb)

Jimenez *et al. BMC Biology*   (2021) 19:21

Page 13 of 15

allows the comparison, topological localization, and evaluation of natural variations occurring in human ORs, and represents to our knowledge, one of the largest collections of variation data of human sensory proteins annotated at the structural level.

Conclusions

We performed topological annotations and population analysis of natural variants of human olfactory receptors, and provide an interactive application to explore human OR mutation data. We envisage that the utility of this information will increase as the amount of available pharmacological data for these receptors grow. This effort, together with ongoing research in the study of genetic changes in other sensory receptors [65], could shape an emerging *sensegenomics* field of knowledge, which should be considered by food and cosmetic consumer product manufacturers for the benefit of the general population.

Methods

Data acquisition and filtering

Natural sequence variations from functionally annotated human ORs [62, 66] were obtained from the Genome Aggregation Database (gnomAD v2, http://gnomad.broadinstitute.org/) using Python (v.3.7.6) data mining scripts. Variant tables for each OR were imported to R (v.3.6.2), including information of chromosome location, transcript consequence, and allele frequencies in seven sub-continental populations (Additional file 1: Table S1) [21]. Basic Local Alignment Search Tool (BLAST, v.2.10.0) and Python scripts were used to compare the collected sequence information with UniProt database (release 2019_11, https://www.uniprot.org/). The collected data was then filtered to remove null values, duplicates, missing rsIDs, and sequence conflicts with reference Swiss-Prot entries, resulting in a curated dataset of 119,069 nucleotide variants from 378 human OR genes (Additional file 1: Tables S2-S3).

Topological mapping and BW annotation

Python data mining scripts were used to assign each coding-sequence mutation a topological location according to a structure-based multiple sequence alignment (MSA) of 378 ORs Swiss-Prot reference sequences and class A GPCRs of known three-dimensional structure (Additional file 3). Natural variants at the TM regions were further annotated with the generic two number system developed by BW consisting of two digits: the first (1 through 7) corresponds to the helix in which the change is located, and the second indicates its position relative to the most conserved residue in the helix (arbitrarily assigned to 50) [24]. This nomenclature was also applied to a 10 residue stretch located between two highly conserved cysteines at the ECL2 (indicated by 45 as the first number attending to its location between the TMs 4 and 5) (Additional file 2: Figure S7) [25].

Impact evaluation of coding sequence variants

The impact of non-synonymous changes was estimated from the amino acid substitution scores derived from the GPCRtm matrix (Additional file 2: Figure S3) [33]. In addition, two subsets of BW topological sites were outlined: (i) a functional core (FC) subset of 30 topological positions with a high degree of conservation and likely involved in the receptor activation, G protein binding, or disulfide bond formation (Additional file 1: Table S4, Additional file 2: Figure S4) and (ii) a binding cavity (BC) subset of 30 amino acid positions within a distance of $\leq$ 4.0 Å to bound ligands in 39 reference class A GPCR 3D structures (Additional file 1: Table S5, Additional file 2: Figure S5). This selection exhibited a high degree of correspondence with positions identified in a reference study conducted on orthosteric and allosteric GPCR ligand interactions sites [32], including the 45.52 at ECL2.

Development of an interactive database with the annotated variation data

Substitution scores and topological annotation (including BC/FC and BW numbering) were transferred to the mutation data table using Python data mining scripts, completing the annotation process (Fig. 1, Additional file 4). A standalone application was programmed with the open-source RStudio (v.1.2.5003) to manage and visualize this curated mutation dataset (https://github.com/lmc-uab/hORMdb). This database resource is also made available online as an interactive web server programmed with the Shiny Server package (v.1.5.12.933) (http://lmc.uab.cat/hORMdb).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-021-00962-0.

Additional file 1: Table S1. Nucleotide sequencing data sources used in the study. **Table S2.** Number of mutations in human ORs collected in the study. **Table S3.** Human ORs genes with functional annotation excluded from the study. **Table S4.** Conserved topological sites with functional implication in the GPCR activity. **Table S5.** Non-olfactory class A GPCRs used in topological annotation.

Additional file 2: Figure S1. Chromosomal distribution of natural variants within OR families. **Figure S2.** Topological distribution of natural variants within OR families. **Figure S3.** The GPCRtm amino acid substitution scores. **Figure S4.** Functional core (FC) topological positions in class A GPCRs. **Figure S5.** Binding cavity (BC) topological positions in class A GPCRs. **Figure S6.** Human OR mutations with potential functional effects. **Figure S7.** Structure-based sequence alignment used in topological annotation.

Additional file 3. Multiple sequence alignment (MSA) of human ORs. MSA of the 378 human OR UniProt sequences used in the topological annotation of protein-coding mutations. Receptor sequences were

Jimenez et al. BMC Biology   (2021) 19:21

Page 14 of 15

aligned with ClustalW (v2.1) using a customized GPCR substitution score matrix. The resulted MSA was manually adjusted to fulfill the structural information derived from non-olfactory class A GPCRs (Additional file 2 Fig. S7). Topological regions (N- and C-terminal sequences, transmembrane α-helices TM 1 to 7, extracellular ECL 1 to 3 and cytoplasmic loops ICL 1 to 3), as well as Ballesteros-Weinstein (BW), Functional Core (FC) and ligand Binding Cavity (BC) topological positions are indicated on top of the alignment. The alignment file in FASTA format is available at the human Olfactory Receptor Mutation database (hORMdb) website.

Additional file 4. The human OR mutation database table. The mutation data table is available at the human Olfactory Receptor Mutation database (hORMdb) website and contains information of 119,069 natural human OR nucleotide variants extracted from gnomAD v2 and annotated with genomic and structural information as described in the main text (resumed in the diagram of Fig. 1). A total of 78 descriptors were associated with each of the natural variants generating a total of 9,287,382 data points. More information about the data types at each column can be found at the HELP panel on the hORMdb web application.

Abbreviations

OR: Olfactory receptor; GPCR: G protein-coupled receptor; G$_{olf}$: Olfactory-specific guanosine triphosphate (GTP)-binding protein alpha subunit; TM: Transmembrane; ECL: Extracellular loop; ICL: Intracellular loop; FC: Functional core; BC: Binding cavity; BW: Ballesteros-Weinstein; BLAST: Basic Local Alignment Search Tool; MSA: Multiple sequence alignment; CORP: Classifier for Olfactory Receptor Pseudogenes; gnomAD: Genome Aggregation Database; ORDB: Olfactory Receptors Database; HORDE: Human Olfactory Data Explorer; hORMdb: Human Olfactory Receptor Mutation Database; 5′ UTR: Five prime untranslated region; 3′ UTR: Three prime untranslated region; NCR: Non-coding region; AFR: African; LAT: Latino; ASH: Ashkenazi Jewish; EA: East Asian; EF: European Finnish; ENF: European Non-Finnish; SA: South Asian

Authors' contributions

A. G., L. P., and M. C. conceived and designed the research. R. C., L. A., N. C.. and A. G. collected and analyzed the data. R. C. and A. GR. developed the database application. A. G., and M. C. interpreted the data and wrote the paper. L. P. provided funding and computational resources.
All authors read and approved the final manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. OR human genes and class A GPCRs used in topological annotation are provided in Additional file 1. Human OR protein sequences and MSA are available in Additional file 3. The mutation data table is provided in Additional file 4. The hORMdb application code is freely accessible at the GitHub repository (https://github.com/lmc-uab/hORMdb). An interactive web browser with filtering functionality and graphical display options is publicly available at http://lmc.uab.cat/hORMdb.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests to declare. All raw data is available upon request.

Author details

¹Laboratori de Medicina Computacional, Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain. ²Present Address: International Agency for Research on Cancer, Evidence Synthesis and Classification Section, WHO Classification of Tumours Group, 150 Cours Albert Thomas, 69008 Lyon, France.

References

1. Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell. 1991;65(1):175–87.
2. Firestein S. How the olfactory system makes sense of scents. Nature. 2001; 413(6852):211–8.
3. Su CY, Menuz K, Carlson JR. Olfactory perception: receptors, cells, and circuits. Cell. 2009;139(1):45–59.
4. Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov. 2017;16(12):829–42.
5. Jones DT, Reed RR. Golf: an olfactory neuron specific-G protein involved in odorant signal transduction. Science. 1989;244(4906):790–5.
6. Du Y, Duc NM, Rasmussen SGF, Hilger D, Kubiak X, Wang L, Bohon J, Kim HR, Wegrecki M, Asuru A, et al. Assembly of a GPCR-G protein complex. Cell. 2019;177(5):1232–42 e1211.
7. Olender T, Waszak SM, Viavant M, Khen M, Ben-Asher E, Reyes A, Nativ N, Wysocki CJ, Ge D, Lancet D. Personal receptor repertoires: olfaction as a model. BMC Genomics. 2012;13:414.
8. Malnic B, Godfrey PA, Buck LB. The human olfactory receptor gene family. Proc Natl Acad Sci U S A. 2004;101(8):2584–9.
9. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome Res. 2001;11(5):685–702.
10. Bushdid C, Magnasco MO, Vosshall LB, Keller A. Humans can discriminate more than 1 trillion olfactory stimuli. Science. 2014;343(6177):1370–2.
11. Chen Z, Zhao H, Fu N, Chen L. The diversified function and potential therapy of ectopic olfactory receptors in non-olfactory tissues. J Cell Physiol. 2018;233(3):2104–15.
12. Massberg D, Hatt H. Human olfactory receptors: novel cellular functions outside of the nose. Physiol Rev. 2018;98(3):1739–63.
13. Hasin-Brumshtein Y, Lancet D, Olender T. Human olfaction: from genomic variation to phenotypic diversity. Trends Genet. 2009;25(4):178–84.
14. Menashe I, Aloni R, Lancet D. A probabilistic classifier for olfactory receptor pseudogenes. BMC Bioinformatics. 2006;7:393.
15. Shepherd GM. The human sense of smell: are we better than we think? PLoS Biol. 2004;2(5):E146.
16. Ayabe-Kanamura S, Schicker I, Laska M, Hudson R, Distel H, Kobayakawa T, Saito S. Differences in perception of everyday odors: a Japanese-German cross-cultural study. Chem Senses. 1998;23(1):31–8.
17. Sorokowska A, Sorokowski P, Hummel T, Huanca T. Olfaction and environment: Tsimane' of Bolivian rainforest have lower threshold of odor detection than industrialized German people. PLoS One. 2013;8(7):e69203.
18. Keller A, Zhuang H, Chi Q, Vosshall LB, Matsunami H. Genetic variation in a human odorant receptor alters odour perception. Nature. 2007;449(7161): 468–72.
19. Menashe I, Abaffy T, Hasin Y, Goshen S, Yahalom V, Luetje CW, Lancet D. Genetic elucidation of human hyperosmia to isovaleric acid. PLoS Biol. 2007; 5(11):e284.
20. McRae JF, Mainland JD, Jaeger SR, Adipietro KA, Matsunami H, Newcomb RD. Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the "grassy" smelling odor, cis-3-hexen-1-ol. Chem Senses. 2012;37(7):585–93.
21. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434–43.
22. Gonzalez A, Cordomi A, Caltabiano G, Pardo L. Impact of helix irregularities on sequence alignment and homology modeling of G protein-coupled receptors. ChemBioChem. 2012;13(10):1393–9.
23. Munk C, Mutt E, Isberg V, Nikolajsen LF, Bibbe JM, Flock T, Hanson MA, Stevens RC, Deupi X, Gloriam DE. An online resource for GPCR structure determination and analysis. Nat Methods. 2019;16(2):151–62.

24. Ballesteros JA, Weinstein H. Integrated methods for the construction of three dimensional models and computational probing of structure-function relations in G-protein coupled receptors. Methods Neurosci. 1995;25:366–428.

25. Isberg V, de Graaf C, Bortolato A, Cherezov V, Katritch V, Marshall FH, Mordalski S, Pin JP, Stevens RC, Vriend G, et al. Generic GPCR residue numbers - aligning topology maps while minding the gaps. Trends Pharmacol Sci. 2015;36(1):22–31.

26. Rovati GE, Capra V, Neubig RR. The highly conserved DRY motif of class A G protein-coupled receptors: beyond the ground state. Mol Pharmacol. 2007; 71(4):959–64.

27. Urizar E, Claeysen S, Deupi X, Govaerts C, Costagliola S, Vassart G, Pardo L. An activation switch in the rhodopsin family of G protein-coupled receptors: the thyrotropin receptor. J Biol Chem. 2005;280(17):17135–41.

28. Garcia-Nafria J, Tate CG. Cryo-EM structures of GPCRs coupled to Gs, Gi and Go. Mol Cell Endocrinol. 2019;488:1–13.

29. Fritze O, Filipek S, Kuksa V, Palczewski K, Hofmann KP, Ernst OP. Role of the conserved NPxxY(x)5,6F motif in the rhodopsin ground state and during activation. Proc Natl Acad Sci U S A. 2003;100(5):2290–5.

30. Woolley MJ, Conner AC. Understanding the common themes and diverse roles of the second extracellular loop (ECL2) of the GPCR super-family. Mol Cell Endocrinol. 2017;449:3–11.

31. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. Nature. 2013; 494(7436):185–94.

32. Chan HCS, Li Y, Dahoun T, Vogel H, Yuan S. New binding sites, new opportunities for GPCR drug discovery. Trends Biochem Sci. 2019;44(4):312–30.

33. Rios S, Fernandez MF, Caltabiano G, Campillo M, Pardo L, Gonzalez A. GPCRtm: an amino acid substitution matrix for the transmembrane region of class A G protein-coupled receptors. BMC Bioinformatics. 2015;16:206.

34. Gilad Y, Lancet D. Population differences in the human functional olfactory repertoire. Mol Biol Evol. 2003;20(3):307–14.

35. Trimmer C, Keller A, Murphy NR, Snyder LL, Willer JR, Nagai MH, Katsanis N, Vosshall LB, Matsunami H, Mainland JD. Genetic variation across the human olfactory receptor repertoire alters odor perception. Proc Natl Acad Sci U S A. 2019;116(19):9475–80.

36. Rosenbaum DM, Rasmussen SG, Kobilka BK. The structure and function of G-protein-coupled receptors. Nature. 2009;459(7245):356–63.

37. Weis WI, Kobilka BK. Structural insights into G-protein-coupled receptor activation. Curr Opin Struct Biol. 2008;18(6):734–40.

38. Alewijnse AE, Timmerman H, Jacobs EH, Smit MJ, Roovers E, Cotecchia S, Leurs R. The effect of mutations in the DRY motif on the constitutive activity and structural instability of the histamine H(2) receptor. Mol Pharmacol. 2000;57(5):890–8.

39. Moore SA, Patel AS, Huang N, Lavin BC, Grammatopoulos TN, Andres RD, Weyhenmeyer JA. Effects of mutations in the highly conserved DRY motif on binding affinity, expression, and G-protein recruitment of the human angiotensin II type-2 receptor. Brain Res Mol Brain Res. 2002; 109(1–2):161–7.

40. Rompler H, Yu HT, Arnold A, Orth A, Schoneberg T. Functional consequences of naturally occurring DRY motif variants in the mammalian chemoattractant receptor GPR33. Genomics. 2006;87(6):724–32.

41. Chung DA, Wade SM, Fowler CB, Woods DD, Abada PB, Mosberg HI, Neubig RR. Mutagenesis and peptide analysis of the DRY motif in the alpha2A adrenergic receptor: evidence for alternate mechanisms in G protein-coupled receptors. Biochem Biophys Res Commun. 2002;293(4):1233–41.

42. D'Antona AM, Ahn KH, Wang L, Mierke DF, Lucas-Lenard J, Kendall DA. A cannabinoid receptor 1 mutation proximal to the DRY motif results in constitutive activity and reveals intramolecular interactions involved in receptor activation. Brain Res. 2006;1108(1):1–11.

43. Raimondi F, Betts MJ, Lu Q, Inoue A, Gutkind JS, Russell RB. Genetic variants affecting equivalent protein family positions reflect human diversity. Sci Rep. 2017;7(1):12771.

44. Kim HR, Duc NM, Chung KY. Comprehensive analysis of non-synonymous natural variants of G protein-coupled receptors. Biomol Ther (Seoul). 2018; 26(2):101–8.

45. Jaeger SR, McRae JF, Bava CM, Beresford MK, Hunter D, Jia Y, Chheang SL, Jin D, Peng M, Gamble JC, et al. A Mendelian trait for olfactory sensitivity affects odor experience and food selection. Curr Biol. 2013;23(16):1601–5.

46. Dawson JP, Weinger JS, Engelman DM. Motifs of serine and threonine can drive association of transmembrane helices. J Mol Biol. 2002;316(3): 799–805.

47. Deupi X, Olivella M, Sanz A, Dolker N, Campillo M, Pardo L. Influence of the g- conformation of Ser and Thr on the structure of transmembrane helices. J Struct Biol. 2010;169(1):116–23.

48. Ridder A, Skupjen P, Unterreitmeier S, Langosch D. Tryptophan supports interaction of transmembrane helices. J Mol Biol. 2005;354(4):894–902.

49. Partridge AW, Therien AG, Deber CM. Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. Proteins. 2004;54(4):648–56.

50. Mirzadegan T, Benko G, Filipek S, Palczewski K. Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin. Biochemistry. 2003; 42(10):2759–67.

51. Wheatley M, Wootten D, Conner MT, Simms J, Kendrick R, Logan RT, Poyner DR, Barwell J. Lifting the lid on GPCRs: the role of extracellular loops. Br J Pharmacol. 2012;165(6):1688–703.

52. Ashtibaghaei K, Gisselmann G, Hatt H, Panten J. Method for evaluating the scent performance of perfumes or perfume mixtures. EP2884280. 2018. https://patentscope.wipo.int/search/en/detail.jsf?docId=EP134004539.

53. White KL, Eddy MT, Gao ZG, Han GW, Lian T, Deary A, Patel N, Jacobson KA, Katritch V, Stevens RC. Structural connection between activation microswitch and allosteric sodium site in GPCR signaling. Structure. 2018; 26(2):259–69 e255.

54. Liu W, Chun E, Thompson AA, Chubukov P, Xu F, Katritch V, Han GW, Roth CB, Heitman LH, IJzerman AP, et al. Structural basis for allosteric regulation of GPCRs by sodium ions. Science. 2012;337(6091):232–6.

55. Chatelain P, Veithen A: Olfactory receptors involved in the perception of sweat carboxylic acids and the use thereof. PCT/EP2013/061243. 2013.

56. Geithe C, Andersen G, Malki A, Krautwurst D. A butter aroma recombinate activates human class-I odorant receptors. J Agric Food Chem. 2015;63(43):9410–20.

57. Bouley R, Sun TX, Chenard M, McLaughlin M, McKee M, Lin HY, Brown D, Ausiello DA. Functional role of the NPxxY motif in internalization of the type 2 vasopressin receptor in LLC-PK1 cells. Am J Physiol Cell Physiol. 2003; 285(4):C750–62.

58. Schmiedeberg K, Shirokova E, Weber HP, Schilling B, Meyerhof W, Krautwurst D. Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. J Struct Biol. 2007;159(3):400–12.

59. Mainland JD, Keller A, Li YR, Zhou T, Trimmer C, Snyder LL, Moberly AH, Adipietro KA, Liu WL, Zhuang H, et al. The missense of smell: functional variability in the human odorant receptor repertoire. Nat Neurosci. 2014; 17(1):114–20.

60. Marenco L, Wang R, McDougal R, Olender T, Twik M, Bruford E, Liu X, Zhang J, Lancet D, Shepherd G, et al. ORDB, HORDE, ODORactor and other on-line knowledge resources of olfactory receptor-odorant interactions. Database (Oxford). 2016;2016:baw132. https://academic.oup.com/database/article/doi/10.1093/database/baw132/2630523.

61. Crasto C, Marenco L, Miller P, Shepherd G. Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Res. 2002;30(1):354–60.

62. Olender T, Nativ N, Lancet D. HORDE: comprehensive resource for olfactory receptor genomics. Methods Mol Biol. 2013;1003:23–38.

63. Slodkowicz G, Babu MM. From prioritisation to understanding: mechanistic predictions of variant effects. Mol Syst Biol. 2018;14(12):e8741.

64. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? J Mol Biol. 2019;431(11):2197–212.

65. Chamoun E, Mutch DM, Allen-Vercoe E, Buchholz AC, Duncan AM, Spriet LL, Haines J, Ma DWL, Guelph Family Health S: A review of the associations between single nucleotide polymorphisms in taste receptors, eating behaviors, and health. Crit Rev Food Sci Nutr 2018, 58(2):194–207.

66. Olender T, Lancet D, Nebert DW. Update on the olfactory receptor (OR) gene superfamily. Hum Genomics. 2008;3(1):87–97.

Publisher's Note