Olender et al. BMC Genomics 2012, 13:414
http://www.biomedcentral.com/1471-2164/13/414

BMC Genomics

RESEARCH ARTICLE Open Access

Personal receptor repertoires: olfaction as a model

Tsviya Olender[1*], Sebastian M Waszak[2], Maya Viavant[1], Miriam Khen[1], Edna Ben-Asher[1], Alejandro Reyes[3], Noam Nativ[1], Charles J Wysocki[4], Dongliang Ge[5] and Doron Lancet[1*]

Abstract

Background: Information on nucleotide diversity along completely sequenced human genomes has increased tremendously over the last few years. This makes it possible to reassess the diversity status of distinct receptor proteins in different human individuals. To this end, we focused on the complete inventory of human olfactory receptor coding regions as a model for personal receptor repertoires.

Results: By performing data-mining from public and private sources we scored genetic variations in 413 intact OR loci, for which one or more individuals had an intact open reading frame. Using 1000 Genomes Project haplotypes, we identified a total of 4069 full-length polypeptide variants encoded by these OR loci, average of ~10 per locus, constituting a lower limit for the effective human OR repertoire. Each individual is found to harbor as many as 600 OR allelic variants, ~50% higher than the locus count. Because OR neuronal expression is allelically excluded, this has direct effect on smell perception diversity of the species. We further identified 244 OR segregating pseudogenes (SPGs), loci showing both intact and pseudogene forms in the population, twenty-six of which are annotatively "resurrected" from a pseudogene status in the reference genome. Using a custom SNP microarray we validated 150 SPGs in a cohort of 468 individuals, with every individual genome averaging 36 disrupted sequence variations, 15 in homozygote form. Finally, we generated a multi-source compendium of 63 OR loci harboring deletion Copy Number Variations (CNVs). Our combined data suggest that 271 of the 413 intact OR loci (66%) are affected by nonfunctional SNPs/indels and/or CNVs.

Conclusions: These results portray a case of unusually high genetic diversity, and suggest that individual humans have a highly personalized inventory of functional olfactory receptors, a conclusion that might apply to other receptor multigene families.

Keywords: Olfactory receptor, Genetic polymorphism, Haplotypes, Single nucleotide polymorphism, Copy number variation, Olfaction, Gene family

Background

Olfaction, the sense of smell, is a versatile and sensitive mechanism for detecting and discriminating thousands of volatile odorants. Olfactory recognition is mediated by large repertoires of olfactory receptors (ORs), which activate a G-protein-mediated transduction cascade, located in the cilia of olfactory sensory neurons [1,2]. The human OR repertoire has 851 loci, encompassing 78 genomic clusters and 57 singleton loci, residing on all but two human chromosomes [3-6]. Each sensory cell expresses a single allele of a single OR locus, thus transmitting a molecularly defined signal to the brain [7-10]. A single OR gene may recognize more than a single odorant molecule [11-15]. A widely accepted working hypothesis is that allelic variants of OR genes may harbor different functional characteristics and hence, may generate different odorant sensitivity phenotypes in different members of the human population [16-18].

Human ORs encompass a high number of pseudogenes, whereby more than 50% of the loci annotated as nonfunctional due to frame-disrupting mutations [3,5,6,19]. Primates are less dependent than mouse and dog on olfactory cues, which appears to have resulted in a gradual gene loss process along this lineage [20-22]. Similar OR repertoire diminutions have been reported in other mammals [23]. In higher apes, the gene loss has remarkably accelerated in humans [24]. Such diminution of the functional OR repertoire in humans is an ongoing evolutionary

* Correspondence: tsviya.olender@weizmann.ac.il; doron.lancet@weizmann.ac.il
[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel
Full list of author information is available at the end of the article

process, as demonstrated by the past identification of OR genes that segregate between intact and pseudogene forms [25,26], and by more recent surveys showing an enrichment of loss-of-function OR alleles [27,28]. It was shown that every human individual is characterized by a different combination of such segregating pseudogenes (SPGs), constituting a pronounced genotypic diversity in the population, including ethnogeographic differences [26]. More recently, using a high-resolution microarray applied to 20 individuals [29], and a read-depth-based Copy Number Variation (CNV) genotyping algorithm [30], we showed a wide range of copy-number values across individuals, ranging from zero to nine copies. These results are in-line with other surveys which found a significant enrichment of ORs in CNV regions [31,32]. CNVs involving deletions (copy numbers of 0 or 1) were shown to affect 56 intact OR loci, 14% of the human OR gene repertoire [30].

Cell-surface receptors are often characterized by several haplotypic alleles in the population, sometimes with different functional properties. A prominent example is the group of the major histocompatibility proteins with varying specificities towards antigenic peptides [33,34]. Other examples include the taste receptor TAS38, underlying responsiveness to the bitter compound phenylthiocarbamide (PTC) [35,36], the melanocortin 1 receptor (MC1R), affecting human skin and hair pigmentation [37], and the green opsin OPN1MW, mediating red-green color vision discrimination [38]. Likewise, in the olfactory system, two protein haplotypes of the olfactory receptor OR7D4 were shown to manifest large difference in sensing the steroid odorant androstenone [39,40].

Some missense haplotypic alleles can be nonfunctional, due to a substitution of key amino acids governing protein folding or interaction with signal transduction components. A continuous spectrum of functionality among missense haplotypes may be quantified by algorithms such SIFT [41] or PolyPhen [42]. An analogous algorithm, Classifier for Olfactory Receptor Pseudogenes (CORP) [43], was previously used to identify 30 SNP variations for which one of the alleles is likely inactive [26], with a broader estimate of as many as 135 functionally inactive missense alleles in the reference genome [43].

Here, we performed scrutiny of publicly available data to create a comprehensive catalog of genetic variability in the human OR repertoire. This includes a compendium of all available missense haplotypes of OR proteins and a dramatically expanded list of OR segregating pseudogenes. Our work creates a framework for understanding the evolution and function of OR genes, and a necessary infrastructure for genotype-phenotype association studies for smell deficits. It further highlights the utility of the olfactory system as a model for personalized gene repertoires.

Results

Numerous allelic variants in intact ORs

We performed *in-silico* data mining of genomic variations in OR genes and segregating pseudogenes, including single nucleotide polymorphisms, small indels (< 100 bp) and structural variations. These were obtained from 651 individuals of the 1000 Genomes Project, including three major ethnic groups, as well as from 11 additional resources (Additional file 1: Table S1). Our compendium contains 5,958 polymorphic events (variations) within coding regions of 413 functional gene loci, the latter selected as having an intact open reading frame in at least one of the individual human chromosomes analyzed (including 26 "resurrected" loci, see below). The breakdown of these variations to seven categories is shown in Figure 1. Additional file 2 lists all duplications and inversion structural variations, not further discussed herein. Altogether, we observed an average of 14.4 ± 6.8 polymorphic variations of all types per ~930 bp open reading frame, similar to what we found in OR pseudogenes (14.9 ± 6.7, p = 0.0881 using Kolmogorov-Smirnov test). The combinations of

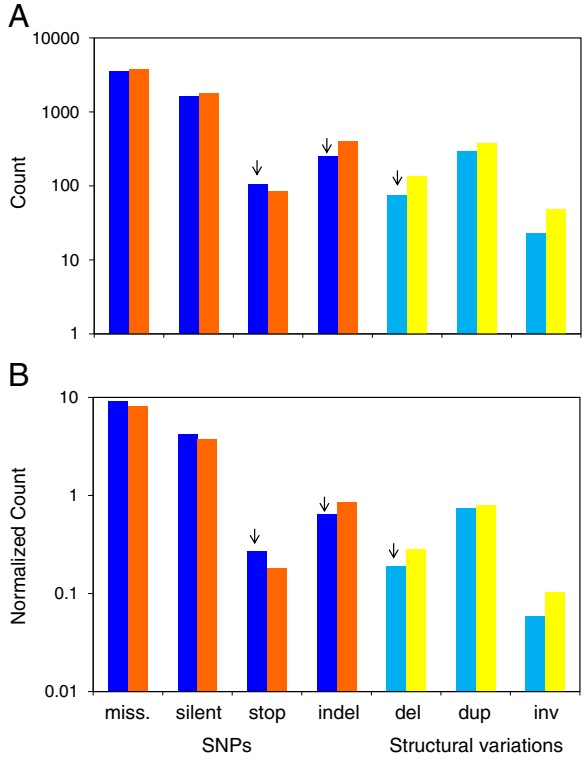

Figure 1 A summary of the genomic variation counts in intact OR coding regions and in OR pseudogenes. A, The absolute count. **B**, Count normalized per gene. Intact genes, blue and light blue; pseudogenes, orange and yellow. Nonfunctional variations are indicated by arrows. Abbreviations: miss, missense SNP; indel, small insertion/deletion up to 100 bp; del, CNV deletion; dup, CNV duplication; inv, CNV inversion; stop, stop gain/stop loss/ loss of the initiating methionine.

polymorphic variations within each OR open reading frame are subsequently used to define haplotypic OR alleles at the DNA and protein levels (see below). All variations are available at the Human Olfactory Receptor Data Explorer database (HORDE database, http://genome.weizmann.ac.il/horde/) [6,44,45].

We subsequently analyzed 2610 missense variants found in the imputed and haplotype-phased data of the 1000 Genomes Project for 651 individuals, to obtain 4069 putative haplotypic OR alleles. Of these, 2682 alleles are present in 3 or more individuals, and hence are less likely to be false positives (Additional file 3). A display of allelic diversity for 30 typical OR loci indicate as many as 35 haplotypic proteins per locus, with an average of 10.4 ± 6.7 (Additional file 1: Figure S1). Every one of these allelic DNA sequence variants ostensibly represents a distinct functional protein, portrayed by a color-coded functional score based on the previously published CORP algorithm [43], including indications for probable non-functionality (CORP>0.9). Figure 2 shows three OR genes with maximal CORP score inter-allele diversity. We also portray three genes with reported odorant specificity [15,39,46]. For the androstenone-binding OR4D7, all 8 haplotypic alleles have similarly high degree of predicted functionality. For the aliphatic thiol-specific OR2C1 the 11 alleles have similar intermediate-level functionality prediction. In contrast, for the amyl butyrate-binding OR2AG1 a bimodal distribution of predicted functionalities is seen, pointing to the possibility of modified odorant responses (Additional file 1: Figure S2).

Figure 3 shows a variation matrix for the 30 OR loci, selected for showing maximal diversity of CORP score values, as viewed in a subset of 30 representative individuals carrying such genotypes. A summary of such patterns for all 413 intact ORs and in 145 individuals of the three major ethnic origins (Figure 4) highlight the vast inter-individual variation in this chemosensory receptor system.

The foregoing analysis embodies a significant enhancement of the OR repertoire in every human individual via haplotypic diversity. Thus, a large majority of human individuals analyzed harbor 490–570 different haplotypes at the 413 loci, i.e. 85–165 loci in a heterozygous state (Figure 5A). This amounts to a repertoire augmentation of 20–40%. The three ethnic groups have pronouncedly different allele count distributions, with Africans having an especially high average of 557 ± 13 different OR sequence variants per individual (Figure 5A). Different ORs often have dissimilar variant distribution in the three populations as exemplified in Figure 5B. These results are consistent with the idea of African origin of modern humans [47,48].

Nonfunctional variations

We next focused on the analysis of nonfunctional variations that eliminate specific members of the OR allele repertoire in a given person, hence are excellent candidates for underlying inter-individual odorant threshold differences [18,49]. First, we analyzed small events, i.e. stop SNPs and indels (up to 84 bases) that result in frame disruption, as derived from 6 different data sources (Additional file 1: Table S1 and Figure S3). Among the 387 OR loci annotated as intact genes in the reference genome we identified 218 cases for which at least one nonfunctional allele was seen. In addition, among the 464 ORs defined as pseudogene in the reference genome, we identified 26 ORs that harbor an intact allele in at least one person, and may be considered as "resurrected" from fixed pseudogene status (Additional file 4). Thus, among 413 thus defined intact loci, a total of 244 loci (59%) show segregation between intact and nonfunctional alleles (segregating pseudogenes, Figure 6). This provides a major enhancement relative to our previously published set of 31 segregating pseudogenes [25]. When analyzing 145 subjects from the 1000 Genome Project for which both SNPs and indels are available, we found that every human individual has 21 ± 4 deletion heterozygotes and 11 ± 2 loci that are homozygously disrupted.

We performed experimental validation for 68 nonfunctional SNPs (stop gain, stop loss, and loss of initiator methionine) and 200 frame-disrupting indels (Additional file 4). For this we designed a custom SNP array (Illumina GoldenGate) that included the total of 268 nonfunctional variations. These were genotyped in a cohort of 468 individuals of two ethnicities, providing validation for 184 of the variations, as compared to a most probable value of validation of 197 ± 2 based on the cohort size and specific minor allele frequencies (validation rate of 93.4%). The number of nonfunctional SNPs per individual (heterozygous and homozygous) thus discovered is shown in Additional file 1: Figure S4. A significant correlation was seen between the allele frequencies in the 1000 Genomes Project data and our validation sets (Additional file 1: Figure S5).

Deletion CNVs

We further performed integration of biallelic deletion CNVs for all OR loci, utilizing five different data sources (Additional file 1: Table S1). This revealed 63 such CNV events (Figure 7A, Additional file 5). This brings the total number of loci that harbor a nonfunctional allele in the examined populations to 271 (Figure 6). As previously seen for segregating pseudogenes [26], here too we observe a great inter-individual variation in the combinations of OR loci affected by deletion CNVs (Figure 7B).

The combined variation results of the deletion CNVs with the SPG genotypes strongly reinforce the notion that practically every individual in the human population

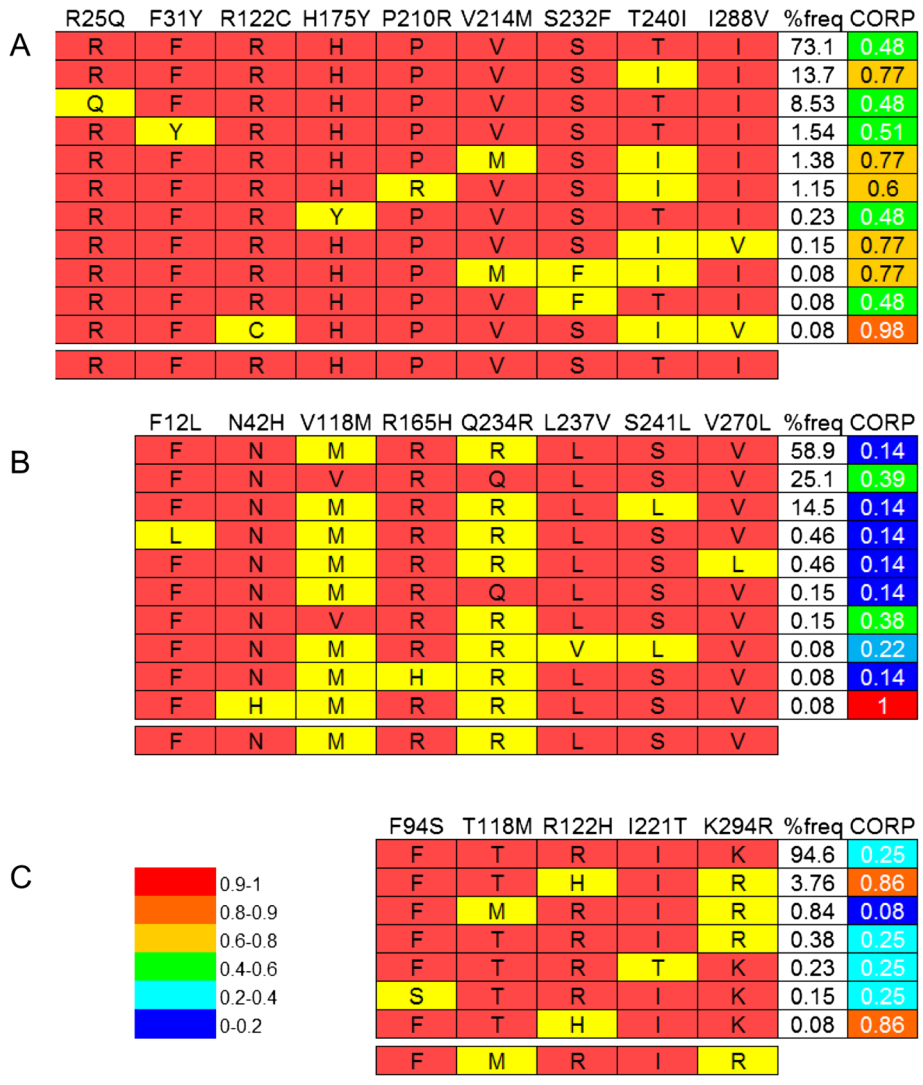

Figure 2 OR protein haplotype alleles for selected ORs. This is shown for OR1D2 (**A**), OR4E2 (**B**) and OR7C2 (**C**), typifying genes with high inter-allele diversity of CORP-predicted functionality. Segregating protein positions (indicated on top) are shown for each haplotype sequence, with yellow indicating non-reference SNP allele. The ancestral chimpanzee allele is shown in the lower row of each panel. The frequency of each allele in the population (%freq) and the CORP pseudogene probability score [43] are indicated in the two right columns. A high CORP score predicts a high pseudogene probability.

has a different combination of intact and inactive alleles (Figure 8). Using a phasing procedure (see methods), we could assign deletion locus haplotypes to 177 ORs, which in some cases harbor more than one event on a given chromosome, and in others create compound heterozygosity for two deletion types (Figure 9 and Additional file 1: Figure S6). Using this combined view we find that, on average, every individual genome carries a disrupted allele at 35 ± 4 loci, of which 11 ± 3 are homozygously affected (Additional file 1: Figure S7). Because every olfactory sensory neuron expresses a single allele at an OR locus, heterozygosly deleted SPGs might have a phenotypic outcome. The personalized repertoire of intact and inactivated ORs significantly differs among ethnic groups (Figure 10A), and such differences are dominated by a subset of OR loci, representing both class I and class II ORs, that manifest a relatively large inter-group variation (Figure 10B, Additional file 1: Table S2). There is however no significant difference in homozygous deletion alleles among the different populations (Additional file 1: Figure S6).

OR Evolution

We asked whether OR genes harbor an unusually high frequency of missense variations. For this, we compared the number of non-synonymous SNPs in two gene sets. The first was 387 OR genes defined as intact in the reference genome, and the second control set constituted 581

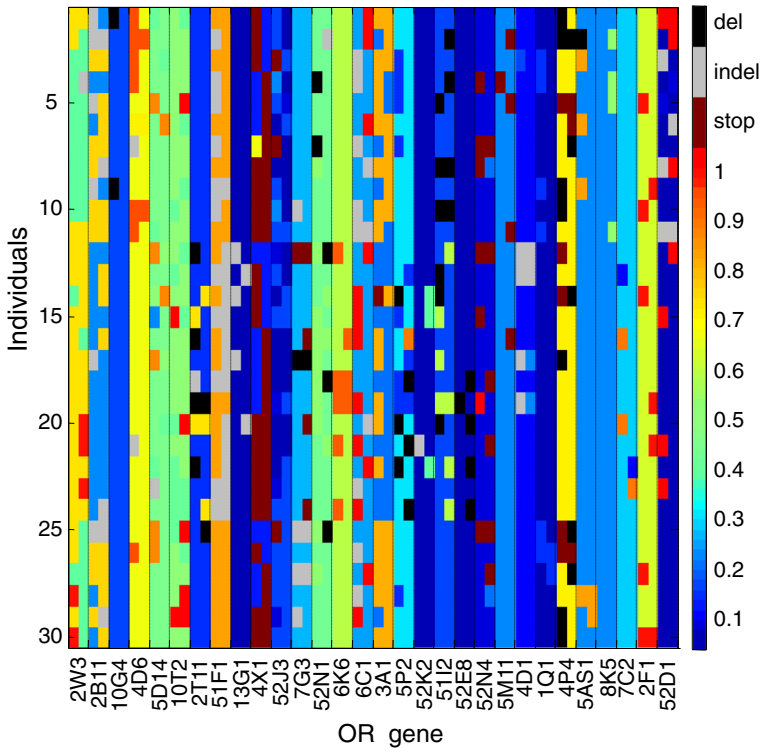

Figure 3 Protein allele genotype for 30 selected OR genes in 30 individuals. The ORs and individuals were selected to show maximal inter-allele diversity of CORP-predicted functionality. The two allelic protein sequences at each locus are shown, color-coded by their CORP scores for missense, and as indicated by the abbreviations (see Figure 1) for nonfunctional, and as depicted by the color scale on right, Ethnicities: 1–11 Europeans, 12–26 Africans, and 27–30 Asians.

protein-coding genes that (like ORs) have a single coding exon. The latter included non-OR G-protein coupled receptors, keratin associated proteins, protocadherins and histones. ORs were found to have 7.7 ± 4.3 missense SNPs per open reading frame, while the controls had 2.2 times less such SNPs (3.5 ± 4.3, $p<2.2X10^{-16}$ Wilcoxon rank sum test with continuity correction, Figure 11A). This was confirmed in a second test set of 15,425 protein coding genes (all GeneCards coding SNPs [50,51] Figure 11C, $p<2.2X10^{-16}$). Synonymous SNP counts showed a much smaller, though significant, difference between ORs and controls (Figure 11B, $p = 1.465X10^{-13}$ and Figure 11D, $p = 0.00789$). We note that OR genes and pseudogenes show a similar propensity of non-synonymous SNPs (Figure 11E), with a slight, statistically significant excess in intact ORs ($p = 0.001149$). The simplest interpretation is that on average ORs may neutrally accumulate genetic variations, mainly due to less stringent purifying selection as compared to non-ORs [31,32].

We asked whether some of the OR genes accrue variations in a non-neutral fashion by examining the ratio of polymorphic non-synonymous substitutions per non-synonymous site to polymorphic synonymous substitutions per synonymous site (pN/pS) [52,53], whereby a value near one would suggest neutrality. While for most ORs the results are consistent with neutrality, there is significant enrichment in the high pN/pS region of the distribution in ORs compared to controls, consistent with selection (Figure 12 and Additional file 1: Figure S8). A subclass of the ORs with pN/pS>1.5 also have a positive value of Tajima's D (Figure 12A) suggesting balancing selection. We asked whether the subgroup of fast evolving ORs (with pN/pS>1.5) is enriched with "evolutionary young" genes, defined as those lacking one-to-one orthology relationships with the chimpanzee orthologs [29]. We find that no such enrichment occurs, as among 47 fast evolving ORs, the fraction of evolutionary young genes is 12.8%, while for all other ORs the fraction is 17.1%. We further note that a relatively small subgroup of 57 ORs (16.8%) in our dataset (in all three populations) show evidence for strong purifying selection (Tajima's D<−0.5 and pN/pS <0.5, Figure 12). This low count as compared to 40.5% in controls, is likely related to the tendency of ORs to evolve towards higher inter-individual diversity [54]. Thus, for the specific receptors showing this evolutionary pattern (Additional file 1: Table S3), such sequence conservation may indicate functional importance, e.g. recognition of essential odorants essential for the species as a whole.

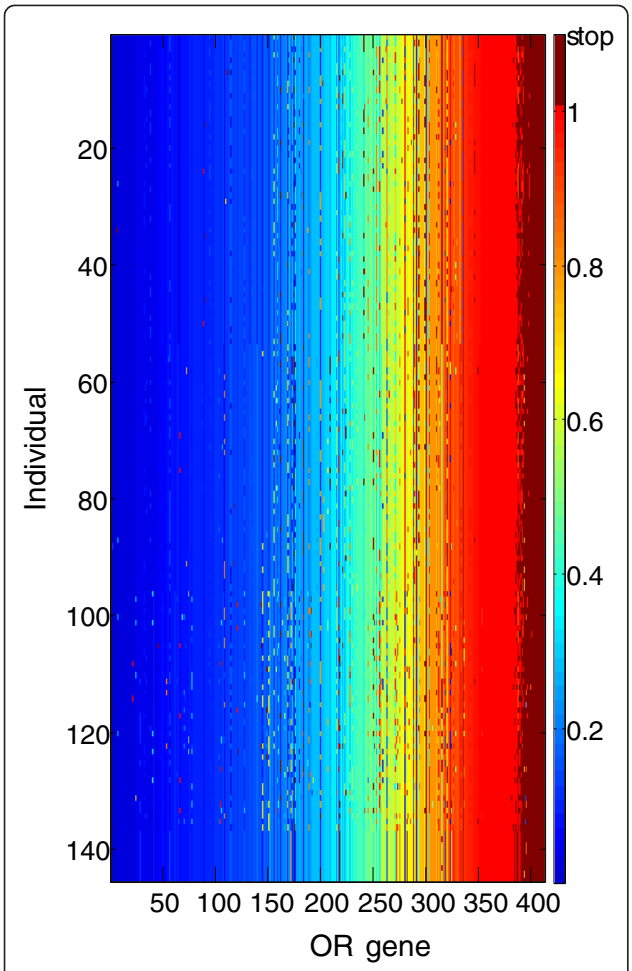

Figure 4 Genomic variation for the entire OR intact haplotype repertoire in 145 individuals. Every individual is represented in every locus by a single randomly selected missense allele, except for Stop loci for which the non-reference allele is preferably shown. Color coding as in Figure 3. Ethnicities: 1–53 - Asians, 54–95 - Europeans and rows 96–145 - Africans. The dataset does not include alleles with concomitant indels and CNV deletions.

Discussion

An OR variation compendium

Using various databases and experimental resources, we have compiled a compendium of synonymous, missense and nonsense SNPs, as well as copy number variations within OR coding regions. A major resource for this work was the 1000 Genomes Project's whole genome sequence data [55], yielding variation and phase information. A significant caveat regarding such data is their low coverage in each sequenced individual and the imputation procedures used in the phasing process [56-58]. This is partly ameliorated by the fact that the main body of our analyses is based on cumulative data from 300–1300 human chromosomes. Another point of concern is that some of the variations were obtained from dbSNP [59], for which population frequencies or validation are sometimes not provided. Indeed, in our experimental validation of 268 OR nonfunctional SNPs, a majority (65%) of the unsupported variations were mined only from dbSNP.

Enormous gene variability

Our results portray an overview of the degree of inter-individual genomic variability harbored in the OR gene inventory. We report on an enormous amount of genomic variation (one variation per 66 bases), 2.5 times larger than in single coding exon control genes. Our analyses suggest that such enhanced variation is largely due to neutral drift, both because the propensity of variations per coding region is similar to that found for OR pseudogenes, and since the average pN/pS value for the intact ORs is 0.9 ± 0.6, consistent with neutrality.

Previous studies reported on positive selection acting in specific OR genes [60-62], potentially related to a recent evolutionary acquisition of a capacity to recognize specific behavior-related odorants [63]. Our results do not provide clear evidence for such selection mode. Other reports suggest that the OR diversity may be maintained to some degree by balancing selection [54,64], similar to that acting upon the major histocompatibility complex alleles [65,66], leading to enhanced ligand recognition success at the population level [67]. While balancing selection for ORs has been disputed [68] our results suggest that a fraction of OR genes may be under such selection mode, a mechanism consistent with the advantage for heterozygosity in a pathway endowed with allelically excluded expression. This is in line with a previous report showing higher than expected count of heterozygotes at OR SNPs in the HapMap populations, which led to the conclusion that the human ORs may have been shaped by balancing selection, stemming from overdominance [54].

Weak purifying selection has also been suggested to affect a subpopulation of human ORs, as seen by human-chimpanzee comparisons [69]. In line with this, we identified nearly 60 ORs in our dataset showing evidence for this evolutionary mechanism. Such evolutionarily conserved OR genes may subserve the recognition of specific odorants important for survival and/or propagation of the species. Interestingly, this group of human genes has a higher fraction of candidate orthologs in mouse, as compared to dog, consistent with a presently accepted phylogeny whereby primates and rodents belong to the same clade, different from that of carnivores [70,71], although a rodent-outside phylogeny was also suggested [72,73].

In sum, it is difficult to negate the possibility that certain modes of selection act on subsets of human OR genes, but it is rather certain that no single mode applies to all ORs. Such heterogeneity of selection modes within the large OR repertoire has also been reported in dog [74,75].

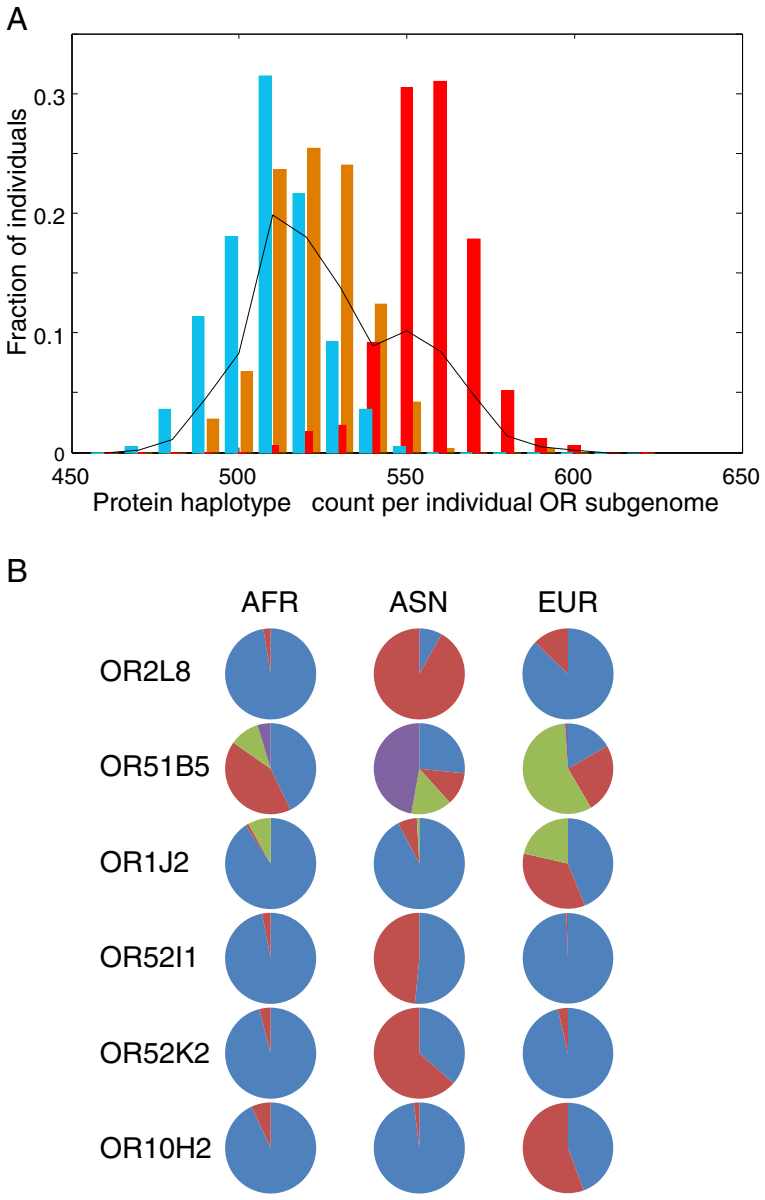

Figure 5 Population differences of personal OR protein allele counts. A) Distribution of the OR missense allele count frequencies in Africans (red), Europeans (brown) and Asians (blue). The black line indicates the average distribution for the whole population. **B**) Haplotype allele frequencies for six OR genes that show the highest inter-population variability. Only alleles with 1000 Genomes frequency > 10% in the entire human population are shown. AFR- Africans, ASN- Asians, EUR- Europeans.

The human allele repertoire

Irrespective of evolutionary path, it is obvious that human ORs show an unusually high variability as compared to other intact protein-coding genes. We report that some human individuals have as many as 600 OR coding regions at their ~400 intact OR loci. Some of these allelic protein variants may have different odorant affinity and/or specificity [39]. Previous reports demonstrate that olfactory sensory neurons express only one of the two alleles at a given locus [2,76,77] with a possibility that allelically excluded neurons report independently to olfactory bulb glomeruli in the brain [78]. This, together with allele plurality, generates a powerful mechanism for augmenting functional variation and enhancing odorant recognition capacities. Furthermore, a higher size of the effective OR repertoire may also signify enhanced average sensitivity to odorants [79,80]. The functional significance of allelic diversity most likely applies to other species as well [75,81].

Loss of function alleles

One of the striking results of the present report is the extremely high prevalence of loss-of-function OR alleles.

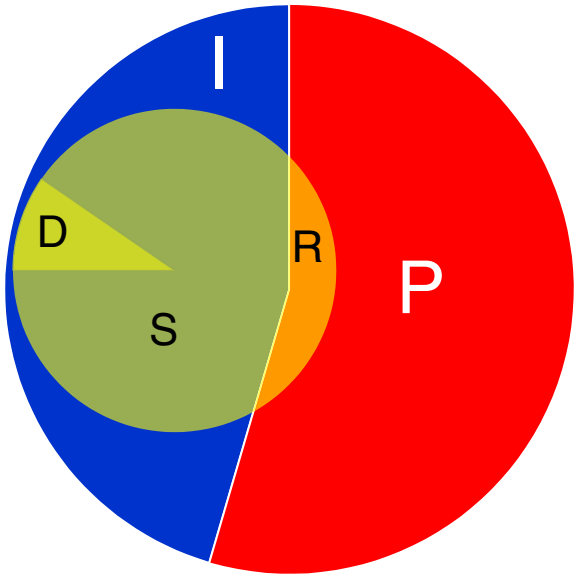

Figure 6 A status diagram of the human OR repertoire. Among the 851 human OR sequences in the reference human genome, 464 were originally annotated as pseudogenes (**P**) and 387 as intact genes (**I**). Our study suggests that 218 (56%) of these intact ORs are segregating pseudogenes (**S**, originating from stop-SNPs and frame-disrupting indels), and 27 (6.9%) have a CNV deletion allele (**D**). Additionally, 26 (5.6%) of the OR pseudogenes are "resurrected (**R**)", by showing an intact allele in some individuals.

Based on the data mining performed, among the 851 human genomic OR loci, 438 have a frame-disrupting pseudogene apparently fixed in the entire population. Of the 413 remaining loci, 271 (66%) have at least one allele lacking an intact open reading frame, including frame disruptions and deletion CNV alleles. The CORP algorithm [43] predicts that an additional 37 loci have missense non-functional alleles, with a CORP score > 0.9, suggesting a probable non-functional OR protein. Thus, as many as 308 OR loci harbor one or more functionally disrupted alleles, and only 105 loci appear to be purely functional in the studied population. This is likely related to the emergence of a large number of OR pseudogenes in higher primate evolution [22,82]. Further, the very high incidence of segregation between intact and nonfunctional alleles attests to a possible highly accelerated gene inactivation in recent human evolution. This potentially took place on a shorter time scale than the previously indicated human-specific acceleration in OR pseudogene accumulation relative to apes [24].

The presently reported number of 308 non-intact loci is fivefold larger than an earlier estimate of ~60 [26]. This number will likely increase even further as many more human genomes become available. Curiously, among the non-intact loci are included 26 that were originally annotated as pseudogenes in the reference genome. Further sequencing would probably show additional such cases of

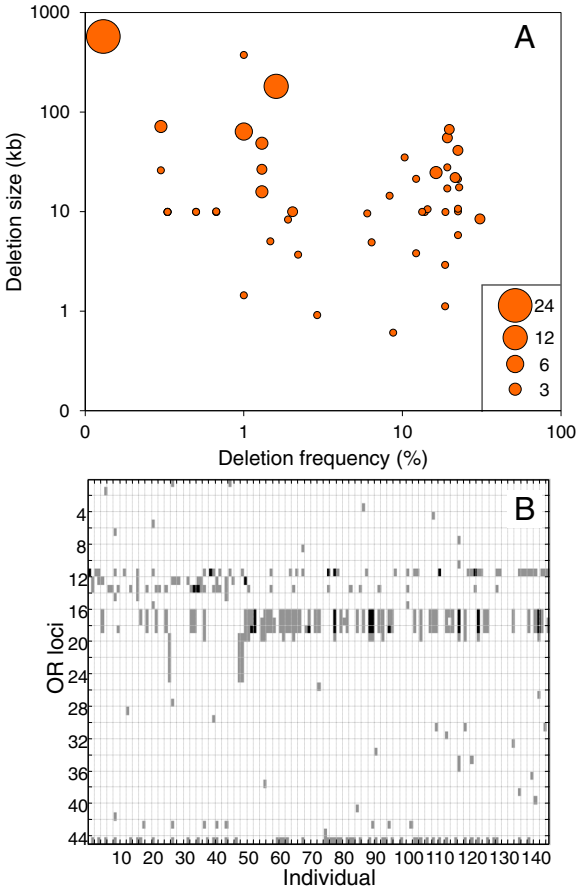

Figure 7 Deletion CNV events in the human OR repertoire. **A**) The deletions size plotted against the deletion frequency in the 145 individuals analyzed. Circle size represents the number of OR genes affected by the deletion. **B**) Genotype calls for the 45 biallelic deletion loci [30] in 145 individuals. Black, homozygotes; grey, heterozygotes.

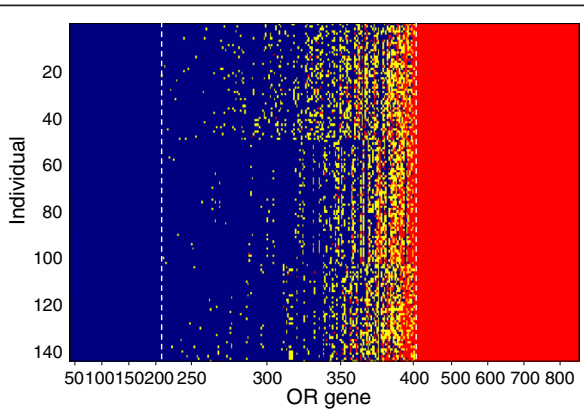

Figure 8 Personalized OR repertoires in 145 individuals. Blue- homozygotes for an intact allele, red- homozygotes for a disrupted allele, yellow- heterozygotes. Nonfunctional allele calls: stop SNPs, indels and deletion alleles.

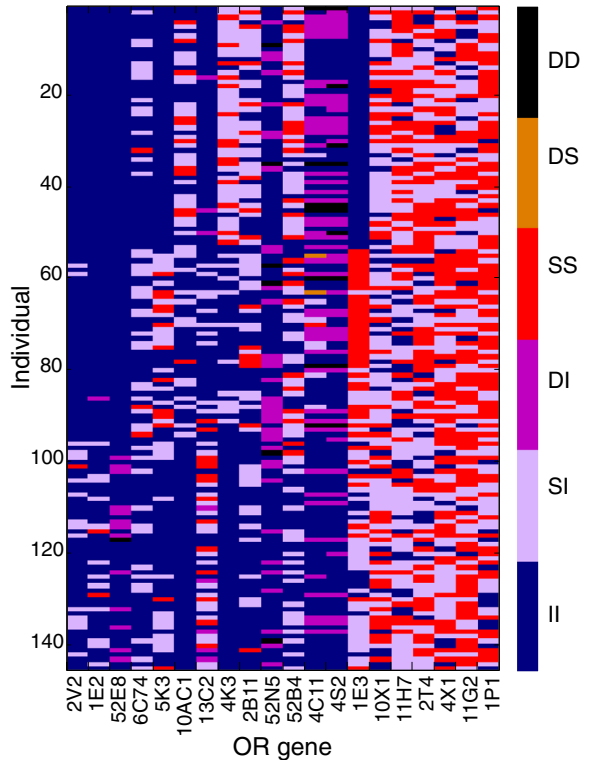

Figure 9 Nonfunctional allele genotypes for 20 OR genes in 145 individuals. The genes have been selected to maximally span the genotype range. Individuals are sorted by ethnicity as in Figure 4. Allele statuses are: intact (I), nonfunctional SNPs/indels (S), bi-allelic deletion CNV (D) [30]. Colors indicate genotypic combinations. The full matrix with all 177 ORs in 145 individuals is shown in Additional file 1: Figure S7.

"resurrected" ORs, most likely from among the 44 fixed OR pseudogenes that have only one frame disruption [6,45]. It should be pointed out that OR pseudogenes are not processed pseudogenes [83], and hence are typically endowed with all features of intact ORs (cis regulatory elements, 5' upstream introns and non-coding exons) and are only different from the intact form by frame-disrupting mutations.

Personal noses

Our comprehensive portrayal of genetic variability in OR genes provides considerably enhanced support for the notion of "different noses for different people" [26]. While for the 145 individuals analyzed from the 1000 Genomes Project data the overall count of homozygous deletion genotypes per individual is not very high (16 ± 3 including missense nonfunctional alleles), the inter-individual variability is vast: there was no case of two individuals having the same deletion pattern across all relevant loci. Furthermore, viewing the broader picture of nonfunctional alleles of all types, as well as protein missense alleles, a randomly selected pair of subjects will on average share only 500 of the alleles, and the remaining 274 (33%) will be different (Figure 13). Importantly, on average 32% of all fully intact OR loci are heterozygously disposed, encoding two different active OR protein variants. A heterozygous deletion event affecting such a locus could have an odorant sensitivity phenotype, as only one of the two different functional alleles would remain active, and the allelically-excluded neuronal pattern could thus be modified.

Analysis of deletion CNVs with high-confidence breakpoints revealed that, for a typical individual, 40% of the deletion CNVs affect more than one (and up to six) intact OR genes, consistent with previous reports [29,30], thus highlighting the large impact of CNVs as opposed to smaller variants. However the contribution of deletion CNVs to the overall number of disrupted alleles per individual is less pronounced.

Receptor diversity and ethnogeography

Our results generally suggest substantial differences among the three major ethnogeographical groups analyzed: Caucasians, Africans and Asians. The most significant result is that Africans have a higher number of OR protein haplotypic variants, with implications to chemosensory diversity. Such findings are in line with the reported higher genetic diversity in this ethnogeographical group [48,84,85]. Some of the protein variants are seen only in one or two of the groups, and others show great disparity of relative allele frequency. The three different human races also have distinct patterns of deletion allele genotypes, which again could affect chemosensory preferences. Previously, we have reported a slightly higher number of intact OR loci in Africans as compared to Caucasians [26]. The results reported here, utilizing a much larger number of deletion loci, shows no statistically significant difference in this realm between ethnic groups.

Conclusions

We used data mining strategies to generate a comprehensive compendium of genomic variations in the inventory of human OR coding regions. Our analyses suggest that the effective size of the functional human OR repertoire is much higher than the number of intact loci, implying considerable enhancement of the potential of human smell perception diversity. Importantly, using both data-mining and experimental verification we show that more than two thirds of human OR loci segregate between an intact and inactivated alleles. These results portray a case of unusually high genetic diversity, and suggest that individual humans have a highly personalized "barcodes" of functional olfactory receptors, a conclusion that likely applies to other receptor multigene families as well.

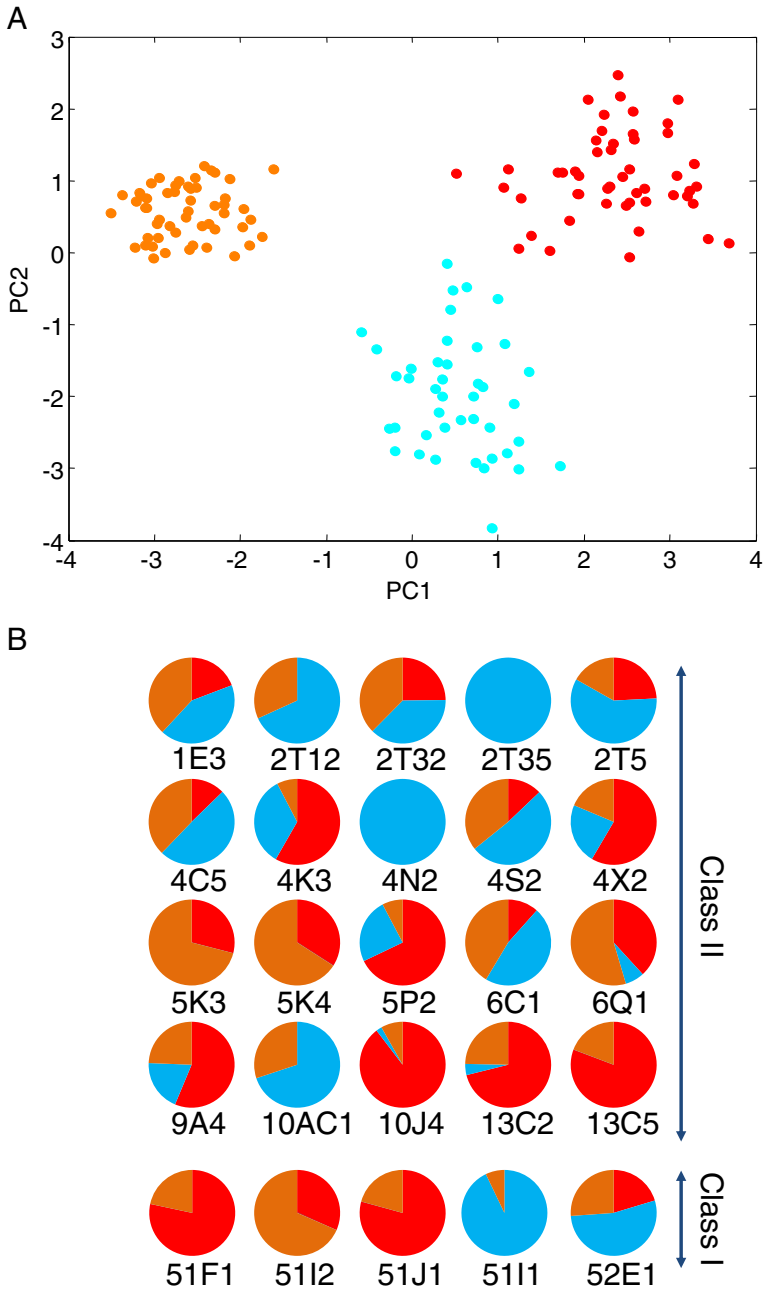

Figure 10 Population differences of OR SPGs. A) Principal component analysis of the nonfunctional SNP genotypes. Each point represents a specific individual, colors as in Figure 5A. **B**) Normalized relative frequencies of the nonfunctional OR allele in the three ethnic populations, color-coded as in (**A**). This is shown for 25 ORs, selected to represent the highest inter-population variability (values are given in Additional file 5). This include 20 ORs belonging to class II ("tetrapod-like"), members of 15 subfamilies (e.g. 1E), and 5 ORs belonging to class I ("fish-like"), represented by members of 5 subfamilies (e.g. 51F). OR classification is as described [3]. Colors as in Figure 5A.

Methods

Genomic variations

Table S1 (Additional file 1) lists the data sources screened for genomic variations in the OR coding regions [26,30,55,59,86-93]. We used the UCSC table browser tool [94] to extract variations from dbSNP, and custom Perl scripts for other databases. We used the GRCh37/hg19 reference genome assembly, and when necessary genomic variations were converted to this version, using the liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Variations that had the same type (SNP or CNV) in the same OR gene symbol with the same start and end locations were considered duplicates and were merged. Indel variations, often located in oligonucleotide repeat loci

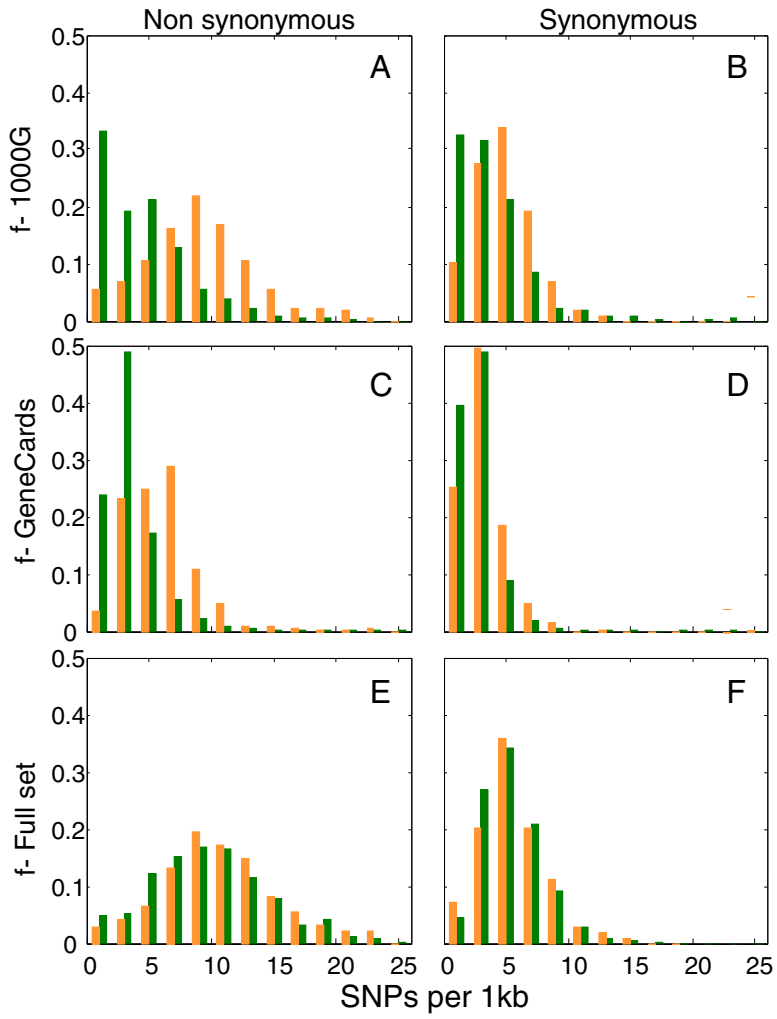

Figure 11 OR genes are enriched with non-synonymous SNPs. Each panel compares the frequency distribution (**f**) for ORs (orange) and control genes (green). The analysis is done using three data sources: 1000 Genomes Project (**A, B**) with 581 genes with a single coding exon as controls; the GeneCards database [50] (**C, D**), with 15,425 protein coding genes as controls; all 10 data sources (Additional file 1: Table S1) (**E, F**), with OR pseudogenes as controls. A, C, E, non-synonymous SNPs, B, D, F, synonymous SNPs.

[95], might have more than a single valid mapping, and were therefore merged manually. Annotation and classification of the variations into the different categories presented in Figure 1 was done by a custom Perl script. Multi-allelic SNPs were removed from the analysis. Unique genomic mapping for dbSNP variations was ascertained by allowing only SNPs with "map weight" equal to 1. SNPs from other sources were analyzed for non-uniqueness by mapping flanking sequences (±50pb) with BLAT [96] and filtering out cases with multiple locations with ≤2 mismatches.

Bi-allelic CNV deletions reported by different sources (Additional file 1: Table S1) were merged by the following procedure: if both beginning and end coordinates of two CNV instances differed by ≤1 kb they were merged into a single entry, and the average genomic coordinates and allele frequencies were used (Additional file 4). From the 1000 Genomes Project data for the first 150 individuals ([93], union.2010_06.deletions.sites.vcf) we kept only deletions with allele frequencies. Multiple overlapped variants from this source were filtered using the following rules (in order):i) When a deletion spanning multiple ORs overlapped with deletions of individual ORs in the same location, the former was preferred; ii) Among overlapping deletions affecting the same OR, the smallest was favored.

OR haplotypes were computed based on phased SNP calling data from the Broad Institute Phase 1 1000 Genomes Project data files (http://www.1000genomes.org/) (AFR.BI_withr2.20100804.genotypes, ASN.BI_withr2.20100804.genotypes, EUR.BI_withr2.20100804.genotypes). Each OR haplotype was defined as a binary vector of non-

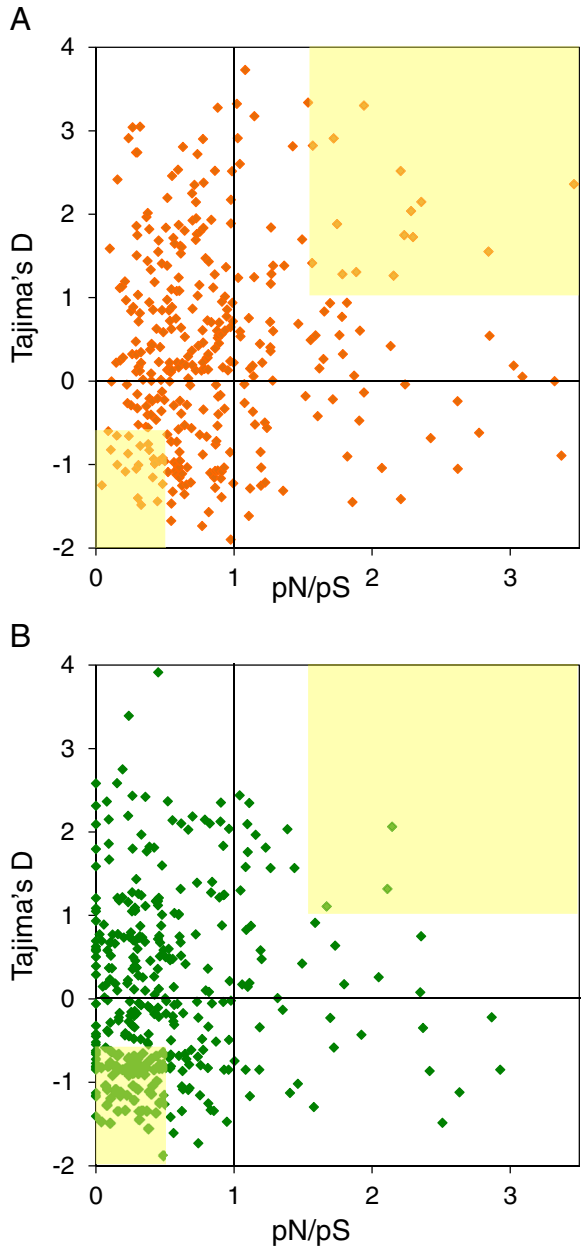

Figure 12 Selection signatures in the OR genes. Correlation of non-synonymous to synonymous substitution rate (pN/pS) with Tajima's D values for A, 364 intact OR genes and B, 439 single coding exon genes. Data are plotted for the European population, other populations in Additional file 1: Figure S4. The difference between the ORs and the control genes was tested using the Kolmogorov-Smirnov test yielding $p = 1.936*10^{-24}$ for pN/pS, and $p = 2.26*10^{-5}$ for Tajima's D. The yellow squares highlight regions which might act under non-neutral selection, top right with D>1, pN/pS>1.5 (balancing selection), and bottom left with Tajima's D<−0.5 and pN/pS<0.5 (purifying selection). Additional file 1: Table S3 lists the 57 ORs found under purifying selection in all populations.

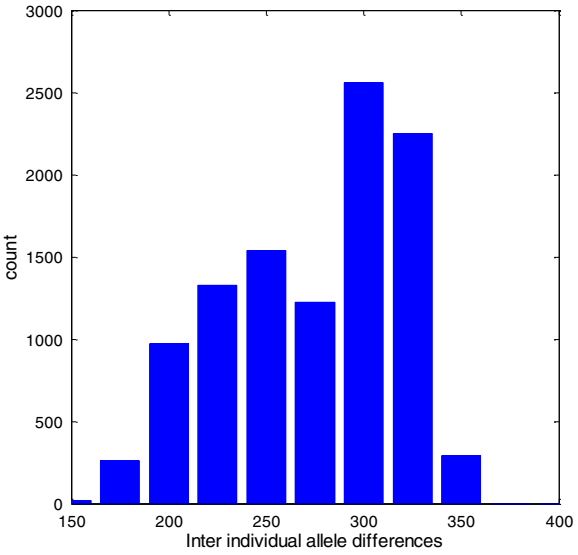

Figure 13 A distribution of the pairwise inter-individual OR missense allele count differences for all 145 individuals shown in the previous figures.

synonymous segregating sites present in all 3 populations, with 1 denoting the non-reference variant. The OR haplotype frequencies for each population were then summarized in Additional file 3.

Haplotype protein functional score

The CORP routine, available in the HORDE database (http://genome.weizmann.ac.il/horde/) [43,45], was used to assign a functional score for each haplotype. CORP examines the amino-acid composition of 60 highly conserved pre-defined positions, where for each site a specific list of present amino-acids is defined. Using a logistic regression model, CORP score (*CS*) is computed using:

$$CS = \frac{1}{1 + \exp(S)}$$

where S is a weighted sum of β coefficients [43]

$$S = -50 + \sum_{i=1}^{60} \alpha_i \times \beta_i$$

and $\alpha_i = -1$ if in the sequence carries an allowed amino-acids in position i, and $\alpha_i = 1$ otherwise.

Variation frequency comparisons

Two control sets were used for variation frequency comparisons. The first was 581 single coding exon genes, retrieved from GeneLoc ([97], http://genecards.weizmann.ac.il/geneloc), further curated with the UCSC table tool [94] to remove non-protein-coding genes. SNPs in these genes were extracted from the 1000 Genomes Project data for the same set of 651 individuals and using the same

computational procedures as applied to the ORs. The SNP count was normalized to gene length using the longest transcript.

The second control gene set was of 15,425 protein coding genes, extracted from GeneCards (http://www.genecards.org/, [50,51]). The same source was also used to obtain SNPs in the 321 intact ORs listed within it. SNPs in OR pseudogenes were classified as "synonymous" or "non-synonymous" based on sequence translation using FASTY [98]. For calling reversion of a pseudogene to an intact status, an open reading frame ≥300 amino-acids was used as a cutoff.

DNA samples

For SNP validation, a cohort of 480 DNA samples was used, collected under ethically-approved protocols as described [91,99]. This panel included 366 individuals of Israeli Jewish origin (271 Ashkenazi, and others of mixed origin) used in a previous study [99], as well as 92 individuals of American origin (57 Caucasians and 22 Afro-Americans) was collected in the framework of a collaborative genotype-phenotype study [91,100].

SNP genotyping

Genomic DNA was extracted from 10 ml of peripheral blood using a DNA Isolation Kit for Mammalian Blood (Roche) [99]. DNA concentration was measured in the Beckman DTX880 Multi-Detection Microplate Reader using PicoGreen (Invitrogen). Genotyping of SNPs was carried out at the Rappaport Research Institute, Technion, Israel, using the Illumina GoldenGate assay according to the manufacturer's instructions (Illumina Inc., SanDiego, CA, USA) [http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn].

The Illumina oligonucleotide pool assay (OPA) was designed using the Illumina Assay Design Tool (ADT) software, with inclusion of all OR nonfunctional variations showing an ADT designability score > 0.4. Inter-variation distances were kept at >60 bp, choosing the variants with highest designability score. The final design included 285 nonfunctional OR variations, of which 268 were successfully genotyped.

For computing the most probable value of validation, we used the minor allele frequencies for the genotyped SNPs, as shown in Additional file 1: Figure S9. We simulated 1000 cohorts of 445 individuals (to account for averaged null calls of 22 individuals per SNP) and obtained a mean and standard deviation for the rate of validation for each variant.

Resolving genotype ambiguities

We developed procedures to obtain unambiguous personal genotypes based on the mining of three independent genotype datasets: 1) The 1000 Genome Project imputed phased SNPs (Broad Institute, version 20100804); 2) The 1000 Genome Project imputed phased indels (Broad Institute, version 2010_07); 3) Bi-allelic CNV calls as described [30]. Ambiguities arise when more than one of these sources reports heterozygosity in the same person and in the same gene. Regarding the merger of nonfunctional SNPs with indels, only 3 genes (OR1B1, OR4C5, OR7G3) showed such an ambiguity, and it was resolved by re-phasing using the PHASE program [101]. The merger of CNV deletions with SNPs/indels was done by the following rules: a. for homozygous CNV deletion concomitant with nonfunctional SNP/indel, the latter was considered as imputation artifact and was ignored; b. heterozygous CNV deletion concomitant with apparently homozygous nonfunctional SNP/indel, was scored as compound heterozygosity; c. Heterozygous SNP/indels along with heterozygous CNV remained unsolved (3 cases). For Figure 3, in cases of unresolvable heterozygous indel/deletion along with claimed missense heterozygosity, one missense allele was selected randomly.

Analyses of selection signatures

The ratio of the number of polymorphic non-synonymous substitutions per non-synonymous sites to the number of polymorphic synonymous substitutions per synonymous sites (pN/pS) was calculated for ORs and control genes following published procedure [102] and using SNPs of the 1000 Genomes Project. This procedure was demonstrated to be correlated with Ka/Ks for divergence [102]. Tajima's D Neutrality test was computed with the DnaSP program [103].

Additional files

Additional file 1: Figures S1-S9, Table S1, Table S2, Table S3.

Additional file 2: A List of duplications and inversions in the OR genes.

Additional file 3: OR protein haplotypes. Haplotypes are represented by their segregating positions (fourth column) where 0 is reference-genome allele and 1 is non-reference allele. Segregating position names are composed from the chromosome name, genomic coordinate, reference amino-acid, protein position and non-reference amino-acid.

Additional file 4: A list of nonfunctional variations in the OR genes.

Additional file 5: OR intact loci for with bi-allelic deletion allele.

Additional file 6: The number of intact and disrupted alleles in OR nonfunctional SNP loci, when using the 1000 Genomes Project, Illumina GoldenGate experiment and Exome sequencing data. Data in this table were used to plot Additional file 1: Figure S8.

Abbreviations

OR: Olfactory receptor; SPG: Segregating pseudogene; CNV: Copy number variation; SNP: Single nucleotide polymorphism; CORP: Classifier for Olfactory Receptor Pseudogenes; pN/pS: The ratio of polymorphic non-synonymous substitutions per non-synonymous site to polymorphic synonymous substitutions per synonymous site.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TO, SW and DG performed the computational mining of genomic variations. TO, SW, MV and AR analyzed the data. TO, DL and SW wrote the paper. CJW, MK and EBA did the experimental validation. NN worked on the database design. All authors read and approved the final manuscript.

Acknowledgments

We are grateful to E.E. Eichler, J.M. Kidd and M. Malig (University of Washington, Seatle, WA, USA) for providing access to fosmid clone reagents under the auspices of the Structural Variation Project; H. Lehrach (Max Plank Institute for Molecular Genetics, Berlin, Germany) for sequencing of OR genes; R. Radtke, A. Husain, S. Sinha, M. Mikati, W. Gallentine, D. Attix, J. McEvoy, E. Cirulli, V. Dixon, N. Walley, K. Linney, E. Heinzen, A. Need, J.P. McEvoy, J. Silver, M. Silver and D. Goldstein (Duke University, Durham NC, USA) for their role in collecting samples used in this study; D. Reed and A. Knaapila (Monell Chemical Senses Center, Philadelphia PA, USA) for collecting some of the samples studied in the work; Y. Hasin-Brumshtein for validating some of the nonfunctional variations; J. Korbel (EMBL, Heidelberg, Germany) for preferred access to the 1000 Genomes Project data and for fruitful discussions.

The work on the fosmid clone reagent (E.E. Eichler's group) was supported by National Institutes of Health Grant HG004120 to E.E.E. Sample collection in D. Goldstein's group was funded in part by NIMH Grant RC2MH089915. Support to DL was from NIDCD/NIH grant 5-R01-DC000298-18 and the Crown Human Genome Center at the Weizmann Institute of Science.

Author details

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland. [3]Genome Biology Unit. European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117, Heidelberg, Germany. [4]Monell Chemical Senses Center, 3500 Market Street, Philadelphia, PA 19104, USA. [5]Center for Human Genome Variation, Duke University School of Medicine, Durham, NC, United States of America.

Received: 3 May 2012 Accepted: 26 July 2012
Published: 21 August 2012

References

1. Kato A, Touhara K: **Mammalian olfactory receptors: pharmacology, G protein coupling and desensitization.** *Cell Mol Life Sci* 2009, **66:**3743–3753.
2. DeMaria S, Ngai J: **The cell biology of smell.** *J Cell Biol* 2011, **191:**443–452.
3. Glusman G, Yanai I, Rubin I, Lancet D: **The complete human olfactory subgenome.** *Genome Res* 2001, **11:**685–702.
4. Zozulya S, Echeverri F, Nguyen T: **The human olfactory receptor repertoire.** *Genome Biol* 2001, **2:**RESEARCH0018.
5. Niimura Y, Nei M: **Evolution of olfactory receptor genes in the human genome.** *Proc Natl Acad Sci U S A* 2003, **100:**12235–12240.
6. Olender T, Lancet D, Nebert DW: **Update on the olfactory receptor (OR) gene superfamily.** *Hum Genomics* 2008, **3:**87–97.
7. Chess A, Simon I, Cedar H, Axel R: **Allelic inactivation regulates olfactory receptor gene expression.** *Cell* 1994, **78:**823–834.
8. Serizawa S, Miyamichi K, Sakano H: **One neuron-one receptor rule in the mouse olfactory system.** *Trends Genet* 2004, **20:**648–653.
9. Kambere MB, Lane RP: **Co-regulation of a large and rapidly evolving repertoire of odorant receptor genes.** *BMC Neurosci* 2007, **8**(Suppl 3):S2.
10. Pathak N, Johnson P, Getman M, Lane RP: **Odorant receptor (OR) gene choice is biased and non-clonal in two olfactory placode cell lines, and OR RNA is nuclear prior to differentiation of these lines.** *J Neurochem* 2009, **108:**486–497.
11. Malnic B, Hirono J, Sato T, Buck LB: **Combinatorial receptor codes for odors.** *Cell* 1999, **96:**713–723.
12. Kajiya K, Inaki K, Tanaka M, Haga T, Kataoka H, Touhara K: **Molecular bases of odor discrimination: Reconstitution of olfactory receptors that recognize overlapping sets of odorants.** *J Neurosci* 2001, **21:**6018–6025.
13. Firestein S: **A code in the nose.** *Sci STKE* 2004, **2004:**pe15.
14. Hatt H: **Molecular and cellular basis of human olfaction.** *Chem Biodivers* 2004, **1:**1857–1869.
15. Saito H, Chi Q, Zhuang H, Matsunami H, Mainland JD: **Odor coding by a Mammalian receptor repertoire.** *Sci Signal* 2009, **2:**ra9.
16. Young JM, Trask BJ: **The sense of smell: genomics of vertebrate odorant receptors.** *Hum Mol Genet* 2002, **11:**1153–1160.
17. Knape K, Beyer A, Stary A, Buchbauer G, Wolschann P: **Genomics of selected human odorant receptors.** *Monatshefte für Chemie* 2008, **139:**1537–1544.
18. Hasin-Brumshtein Y, Lancet D, Olender T: **Human olfaction: from genomic variation to phenotypic diversity.** *Trends Genet* 2009, **25:**178–184.
19. Niimura Y, Nei M: **Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates.** *J Hum Genet* 2006, **51:**505–517.
20. Sharon D, Glusman G, Pilpel Y, Khen M, Gruetzner F, Haaf T, Lancet D: **Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes.** *Genomics* 1999, **61:**24–36.
21. Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, Sharon D: **Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes.** *Nat Genet* 2000, **26:**221–224.
22. Matsui A, Go Y, Niimura Y: **Degeneration of olfactory receptor gene repertories in primates: no direct link to full trichromatic vision.** *Mol Biol Evol* 2010, **27:**1192–1200.
23. Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC: **Ecological adaptation determines functional mammalian olfactory subgenomes.** *Genome Res* 2010, **20:**1–9.
24. Gilad Y, Man O, Paabo S, Lancet D: **Human specific loss of olfactory receptor genes.** *Proc Natl Acad Sci U S A* 2003, **100:**3324–3327.
25. Menashe I, Man O, Lancet D, Gilad Y: **Population differences in haplotype structure within a human olfactory receptor gene cluster.** *Hum Mol Genet* 2002, **11:**1381–1390.
26. Menashe I, Man O, Lancet D, Gilad Y: **Different noses for different people.** *Nat Genet* 2003, **34:**143–144.
27. Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C: **A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs.** *Am J Hum Genet* 2009, **84:**224–234.
28. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, *et al*: **A systematic survey of loss-of-function variants in human protein-coding genes.** *Science* 2012, **335:**823–828.
29. Hasin Y, Olender T, Khen M, Gonzaga-Jauregui C, Kim PM, Urban AE, Snyder M, Gerstein MB, Lancet D, Korbel JO: **High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution.** *PLoS Genet* 2008, **4:**e1000249.
30. Waszak SM, Hasin Y, Zichner T, Olender T, Keydar I, Khen M, Stutz AM, Schlattl A, Lancet D, Korbel JO: **Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity.** *PLoS Comput Biol* 2010, **6:**e1000988.
31. Nozawa M, Kawahara Y, Nei M: **Genomic drift and copy number variation of sensory receptor genes in humans.** *Proc Natl Acad Sci U S A* 2007, **104:**20421–20426.
32. Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ: **Extensive copy-number variation of the human olfactory receptor gene family.** *Am J Hum Genet* 2008, **83:**228–242.
33. The MHC sequencing consortium: **Complete sequence and gene map of a human major histocompatibility complex.** *Nature* 1999, **401:**921–923.
34. Vandiedonck C, Knight JC: **The human Major Histocompatibility Complex as a paradigm in genomics research.** *Brief Funct Genomic Proteomic* 2009, **8:**379–394.
35. Montmayeur JP, Matsunami H: **Receptors for bitter and sweet taste.** *Curr Opin Neurobiol* 2002, **12:**366–371.
36. Feeney E, O'Brien S, Scannell A, Markey A, Gibney ER: **Genetic variation in taste perception: does it have a role in healthy eating?** *Proc Nutr Soc* 2011, **70:**135–143.
37. Dessinioti C, Antoniou C, Katsambas A, Stratigos AJ: **Melanocortin 1 receptor variants: functional role and pigmentary associations.** *Photochem Photobiol* 2011, **87:**978–987.
38. Deeb SS: **Genetics of variation in human color vision and the retinal cone mosaic.** *Curr Opin Genet Dev* 2006, **16:**301–307.
39. Keller A, Zhuang H, Chi Q, Vosshall LB, Matsunami H: **Genetic variation in a human odorant receptor alters odour perception.** *Nature* 2007, **449:**468–472.

40. Knaapila A, Zhu G, Medland SE, Wysocki CJ, Montgomery GW, Martin NG, Wright MJ, Reed DR: **A Genome-Wide Study on the Perception of the Odorants Androstenone and Galaxolide.** *Chem Senses* 2012, **37:**541–552.
41. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073–1081.
42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7:**248–249.
43. Menashe I, Aloni R, Lancet D: **A probabilistic classifier for olfactory receptor pseudogenes.** *BMC Bioinforma* 2006, **7:**393.
44. Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, *et al*: **Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE.** *Nucleic Acids Res* 2003, **31:**142–146.
45. Olender T, Feldmesser E, Atarot T, Eisenstein M, Lancet D: **The olfactory receptor universe–from whole genome analysis to structure and evolution.** *Genet Mol Res* 2004, **3:**545–553.
46. Gelis L, Wolf S, Hatt H, Neuhaus EM, 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 Engl* 2012, **51:**1274–1278.
47. Campbell MC, Tishkoff SA: **The evolution of human genetic and phenotypic variation in Africa.** *Curr Biol* 2010, **20:**R166–R173.
48. Lambert CA, Tishkoff SA: **Genetic structure in African populations: implications for human demographic history.** *Cold Spring Harb Symp Quant Biol* 2009, **74:**395–402.
49. Menashe I, Lancet D: **Variations in the human olfactory receptor pathway.** *Cell Mol Life Sci* 2006, **63:**1485–1493.
50. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, *et al*: **GeneCards Version 3: the human gene integrator.** *Database (Oxford)* 2010, **2010:**baq020.
51. Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N, Oz-Levi D, Olender T, Belinky F, Bahir I, *et al*: **In-silico human genomics with GeneCards.** *Hum Genomics* 2011, **5:**709–717.
52. Ronald J, Akey JM: **Genome-wide scans for loci under selection in humans.** *Hum Genomics* 2005, **2:**113–125.
53. Nei M, Suzuki Y, Nozawa M: **The neutral theory of molecular evolution in the genomic era.** *Annu Rev Genomics Hum Genet* 2010, **11:**265–289.
54. Alonso S, Lopez S, Izagirre N, de la Rua C: **Overdominance in the human genome and olfactory receptor activity.** *Mol Biol Evol* 2008, **25:**997–1001.
55. Consortium. GP: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061–1073.
56. Hao K, Chudin E, McElwee J, Schadt EE: **Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies.** *BMC Genet* 2009, **10:**27.
57. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: **A comprehensive evaluation of SNP genotype imputation.** *Hum Genet* 2009, **125:**163–171.
58. Marchini J, Howie B: **Genotype imputation for genome-wide association studies.** *Nat Rev Genet* 2010, **11:**499–511.
59. Day IN: **dbSNP in the detail and copy number complexities.** *Hum Mutat* 2010, **31:**2–4.
60. Gilad Y, Lancet D: **Population differences in the human functional olfactory repertoire.** *Mol Biol Evol* 2003, **20:**307–314.
61. Gilad Y, Bustamante CD, Lancet D, Paabo S: **Natural selection on the olfactory receptor gene family in humans and chimpanzees.** *Am J Hum Genet* 2003, **73:**489–501.
62. Moreno-Estrada A, Casals F, Ramirez-Soriano A, Oliva B, Calafell F, Bertranpetit J, Bosch E: **Signatures of selection in the human olfactory receptor OR5I1 gene.** *Mol Biol Evol* 2008, **25:**144–154.
63. Zhuang H, Chien MS, Matsunami H: **Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates.** *Proc Natl Acad Sci U S A* 2009, **106:**21247–21251.
64. Tong P, Prendergast JG, Lohan AJ, Farrington SM, Cronin S, Friel N, Bradley DG, Hardiman O, Evans A, Wilson JF, Loftus B: **Sequencing and analysis of an Irish human genome.** *Genome Biol* 2010, **11:**R91.
65. Hedrick PW: **Balancing selection and MHC.** *Genetica* 1998, **104:**207–214.
66. Meyer D, Thomson G: **How selection shapes variation of the human major histocompatibility complex: a review.** *Ann Hum Genet* 2001, **65:**1–26.
67. Sommer S: **The importance of immune gene variability (MHC) in evolutionary ecology and conservation.** *Front Zool* 2005, **2:**16.
68. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, *et al*: **Targets of balancing selection in the human genome.** *Mol Biol Evol* 2009, **26:**2755–2764.
69. Gimelbrant AA, Skaletsky H, Chess A: **Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence.** *Proc Natl Acad Sci U S A* 2004, **101:**9019–9022.
70. Murphy WJ, Pevzner PA, O'Brien SJ: **Mammalian phylogenomics comes of age.** *Trends Genet* 2004, **20:**631–639.
71. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J: **Retroposed elements as archives for the evolutionary history of placental mammals.** *PLoS Biol* 2006, **4:**e91.
72. Niimura Y, Nei M: **Extensive gains and losses of olfactory receptor genes in mammalian evolution.** *PLoS One* 2007, **2:**e708.
73. Cannarozzi G, Schneider A, Gonnet G: **A phylogenomic study of human, dog, and mouse.** *PLoS Comput Biol* 2007, **3:**e2.
74. Tacher S, Quignon P, Rimbault M, Dreano S, Andre C, Galibert F: **Olfactory receptor sequence polymorphism within and between breeds of dogs.** *J Hered* 2005, **96:**812–816.
75. Robin S, Tacher S, Rimbault M, Vaysse A, Dreano S, Andre C, Hitte C, Galibert F: **Genetic diversity of canine olfactory receptors.** *BMC Genomics* 2009, **10:**21.
76. Mombaerts P: **Axonal wiring in the mouse olfactory system.** *Annu Rev Cell Dev Biol* 2006, **22:**713–737.
77. Mori K, Sakano H: **How is the olfactory map formed and interpreted in the mammalian brain?** *Annu Rev Neurosci* 2011, **34:**467–499.
78. Feinstein P, Mombaerts P: **A contextual model for axonal sorting into glomeruli in the mouse olfactory system.** *Cell* 2004, **117:**817–831.
79. Lancet D, Sadovsky E, Seidemann E: **Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system.** *Proc Natl Acad Sci U S A* 1993, **90:**3715–3719.
80. Rosenwald S, Kafri R, Lancet D: **Test of a statistical model for molecular recognition in biological repertoires.** *J Theor Biol* 2002, **216:**327–336.
81. Richgels PK, Rollmann SM: **Genetic Variation in Odorant Receptors Contributes to Variation in Olfactory Behavior in a Natural Population of Drosophila melanogaster.** *Chem Senses* 2012, **37:**229–240.
82. Gilad Y, Przeworski M, Lancet D: **Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates.** *PLoS Biol* 2004, **2:**E5.
83. Vanin EF: **Processed pseudogenes: characteristics and evolution.** *Annu Rev Genet* 1985, **19:**253–272.
84. Long JC, Kittles RA: **Human genetic diversity and the nonexistence of biological races.** *Hum Biol* 2003, **75:**449–471.
85. Holsinger KE, Weir BS: **Genetics in geographically structured populations: defining, estimating and interpreting F(ST).** *Nat Rev Genet* 2009, **10:**639–650.
86. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, *et al*: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5:**e254.
87. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, *et al*: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872–876.
88. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53–59.
89. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, *et al*: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453:**56–64.
90. Lee S, Hormozdiari F, Alkan C, Brudno M: **MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions.** *Nat Methods* 2009, **6:**473–474.
91. Hasin-Brumshtein Y: *Genetic variation in human olfactory receptors: from evolution to olfactory sensitivity*. The Weizmann Institute of Science, Molecular Genetics: PhD thesis; 2010.
92. Bhagwat M: **Searching NCBI's dbSNP database.** *Curr Protoc Bioinformatics* 2010, **Chapter 1:Unit 1:**19.

93. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, *et al*: **Mapping copy number variation by population-scale genome sequencing.** *Nature* 2011, **470:**59–65.
94. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res* 2004, **32:**D493–D496.
95. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE: **An initial map of insertion and deletion (INDEL) variation in the human genome.** *Genome Res* 2006, **16:**1182–1190.
96. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656–664.
97. Rosen N, Chalifa-Caspi V, Shmueli O, Adato A, Lapidot M, Stampnitzky J, Safran M, Lancet D: **GeneLoc: exon-based integration of human genome maps.** *Bioinformatics* 2003, **19(Suppl 1):**i222–i224.
98. Mackey AJ, Haystead TA, Pearson WR: **Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences.** *Mol Cell Proteomics* 2002, **1:**139–147.
99. 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:**e284.
100. Wysocki CJ, Reed DR, Lancet D, Hasin Y, Knaapila A, Louie J, Duke F, Lisa O: **Phenotype/Genotype Associations in Human Olfaction.** *Chem Senses* 2010, **35:**627.
101. Stephens M, Donnelly P: **A comparison of bayesian methods for haplotype reconstruction from population genotype data.** *Am J Hum Genet* 2003, **73:**1162–1169.
102. Liu J, Zhang Y, Lei X, Zhang Z: **Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective.** *Genome Biol* 2008, **9:**R69.
103. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25:**1451–1452.

doi:10.1186/1471-2164-13-414
Cite this article as: Olender *et al.*: Personal receptor repertoires: olfaction as a model. *BMC Genomics* 2012 **13**:414.