Genetic variation of olfactory receptor gene family in a Japanese population

Muhammad Shoaib AKHTAR¹, Ryuichi ASHINO¹, Hiroki OOTA², Hajime ISHIDA³, Yoshihito NIIMURA^{4,5}, Kazushige TOUHARA⁵, Amanda D. MELIN⁶, Shoji KAWAMURA^{1*}

¹Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8562, Japan

²Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan

³Department of Human Biology and Anatomy, Faculty of Medicine, University of the Ryukyus, Nishihara, Okinawa 903-0215, Japan

⁴Department of Veterinary Sciences, Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan

⁵Department of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan

⁶Department of Anthropology and Archaeology & Department of Medical Genetics, University of Calgary, Calgary T2N 1N4, Canada

Received 16 September 2021; accepted 24 October 2021

Abstract The olfactory receptor (OR) gene family is comprised of hundreds of intact and disrupted genes in humans. The compositions and copy number variation (CNV) of disrupted and intact OR genes among individuals is expected to cause variation in olfactory perception. However, little is known about OR genetic variation in many human populations. In this study, we used targeted capture enrichment and massive parallel short-read sequencing methods to examine genetic variation of OR genes, as well as of neutral genome regions as references, for 69 anonymized unrelated Japanese individuals. The capture probes were designed for 398 intact OR genes in the human reference genome hg38, and 85 neutral references. Probes were also designed for four unannotated and 99 ‘nearly-intact’ (hg38-pseudo) OR genes in hg38 and 53 chimpanzee OR genes in the Pantro3.0 genome database with no orthologs in hg38. All the hg38 OR genes and one Pantro 3.0 OR gene were retrieved. The mean sequencing depth was significantly higher than that of the 1000 Genomes Project. A total of 30 OR genes from hg38-intact and hg38-pseudo categories were newly found to be segregating pseudogenes. One hg38-pseudo OR gene was intact in all individuals. CNV was detected in 63 OR genes. Tajima’s *D* analysis for OR genes and neutral references was consistent with balancing selection to maintain allelic differences in intact OR genes. These results demonstrate that the targeted capture by probes with diversity-oriented design is far more effective than a whole-genome approach to retrieve OR genes and achieve high-depth sequencing and thus to reveal polymorphisms for the OR multigene family. The composition of OR genes in the human reference genome hg38 does not necessarily represent those in many humans, implying higher perceptual variation than previously thought. The current study inspires further investigation with a similar approach at a global scale.

Key words: olfactory receptors, Japanese population, targeted capture, pseudogene, copy number variation

Introduction

Humans are generally considered to be visually oriented animals and the importance of their olfactory sense is not widely appreciated. However, recent studies have challenged this view, revealing significant roles of olfaction in daily life and communication (Hoover, 2010; Bushdid et al.,

2014; McGann, 2017; Roberts et al., 2020). Olfactory receptor (OR) genes comprise the largest gene family in vertebrates. They belong to the G-protein-coupled receptor (GPCR) supergene family and have a seven-transmembrane structure (Buck and Axel, 1991; Zhang and Firestein, 2002). OR genes are single-coding-exon genes lacking introns with a typical size of ~930 bp. Data on the diversity of OR genes in human populations is accumulating, with records of prevalent single-nucleotide polymorphisms (SNP), short insertions and deletions (indels), intact (functional)/disrupted (pseudogene) polymorphisms, and copy number variations (CNVs), all of which have been associated with olfactory perceptual variations (Menashe et al., 2002; Gilad and Lancet, 2003; Keller et al., 2007; Menashe et al., 2007; Hasin et al., 2008; Keller et al., 2012; Jaeger et al., 2013;

* Correspondence to: Shoji Kawamura, Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Bioscience BLDG Room 502, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan.
E-mail: kawamura@edu.k.u-tokyo.ac.jp

Published online 20 January 2022
in J-STAGE (www.jstage.jst.go.jp) DOI: 10.1537/ase.211024

McRae et al., 2013; Wooding, 2013; Hoover et al., 2015; Secundo et al., 2015; Majid and Kruspe, 2018; Olofsson and Wilson, 2018; Trimmer et al., 2019; Concas et al., 2021).

Humans have ~400 intact and ~440 disrupted OR genes in the genome (Glusman et al., 2001; Hasin et al., 2008; Niimura et al., 2018). Because duplicated genes are similar in DNA sequence, and creating assemblies is inherently problematic, many multigene families assemble poorly with low coverage in whole-genome sequence (WGS) databases using short-read massive parallel ('next-generation') sequencing (NGS) (Sims et al., 2014; Yohe et al., 2020). Although the long-read sequencing would be an option for sequencing a multigene family in a genomic region (Larsen et al., 2014; Eaton et al., 2021), combined use of short-read NGS is still desired to increase sequencing accuracy. However, its application to ecological, evolutionary, or population studies is often limited due to the difficulty of collecting a sufficient amount of high-molecular-weight genomic DNA samples. Since OR genes represent ~1% of genomic coding sequences and only ~0.0125% of the whole genome in mammals (Lane et al., 2001; Mainland et al., 2013), application of the WGS approach to sequence OR genes would also be labor- and cost-ineffective.

Targeted capture is a strategy to enrich subsets of the genome by hybridizing nucleotide probes designed from reference sequences with fragmented genomic DNA of study subjects. The DNA, now enriched in the regions of interest, is sent for subsequent sequencing. This strategy has emerged as a powerful alternative to WGS (Jones and Good, 2016). The growing use of targeted capture with short-read NGS demonstrates its potential to address a range of research questions (Schweizer et al., 2016). Targeted capture has been successfully applied to sequence OR genes with high coverage and to identify variation with high sensitivity and specificity (Mainland et al., 2013; Yohe et al., 2020).

Building on this previous research, we focus on the Japanese population as a case study of OR gene family variation. The Japanese population is a hybrid between the Yayoi people, farmers who migrated from the East Eurasian Continent ~3000 years ago, and the Jomon people, indigenous hunter-gatherers who were likely to be the descendants of late-Paleolithic people in the Japanese archipelago (Hanihara, 1991; Adachi et al., 2021; Jinam et al., 2021; Koganebuchi and Oota, 2021; Osada and Kawai, 2021). Ancient DNA analyses inferred that the Jomon people diverged from the current East Asians earlier than 26000 years ago (McColl et al., 2018; Gakuhari et al., 2020). A recent genome analysis showed that local populations in the Japanese archipelago exhibited varying degrees of Jomon ancestry in their genome with a geographical gradient (Watanabe et al., 2021). For example, people in the Ryukyu Islands, located at the southwestern tip of the Japanese archipelago, have inherited more components of Jomon genomes than people in the central part of mainland Japan (Matsukusa et al., 2010; Koganebuchi et al., 2012; Sato et al., 2014). However, it remains to be elucidated how much genetic variation of the OR gene family is in the Japanese population and whether variation is associated with the genomic ancestry.

We employed targeted capture using probes designed mainly from the human reference genome hg38 and con-

ducted short-read NGS. The hg38 sequence represents a haploid genome sequence at each nucleotide site from a source individual. In addition to the annotated OR genes in hg38, we also included non-annotated OR sequences and nearly intact OR gene sequences in hg38 and the OR gene sequences absent in hg38 but present in the chimpanzee reference genome Pantro3.0 for the probe design to enable a comprehensive survey that would capture OR genes possibly missing in hg38 due to genetic polymorphism present in the human population. In order to proxy neutral variation, we additionally captured single-copy and non-protein-coding genome regions as 'neutral' control references in the same genomic DNA samples to evaluate genetic diversity of the OR gene family and the neutrality of the variation in the Japanese population.

In the present study, we test whether targeted capture by probes with a diversity-oriented design is more effective than a whole-genome approach to retrieve OR genes and achieve high-depth sequencing. We then examine whether there is any difference of the OR gene composition between this study and the reference human genome data and how novel the identified variation in our study is. We then test whether the application of the same methods to neutral references is effective to reveal natural selection on the OR gene family. We finally evaluate whether OR variation is associated with difference of genomic ancestry in Japan. Regarding the current study as a test case, we assess prospects of this approach applied to a global scale.

Materials and Methods

Genomic DNA Samples

We examined genomic DNA from 69 anonymized, unrelated healthy Japanese men. The DNA samples from 54 individuals out of the 69 were purchased from the Japanese B Cell DNA Bank (JBCDB), as representing mainland Japanese samples. The JBCDB is managed by the Health Science Research Resources Bank (HSRRB)/Japanese Collection of Research Bioresources (JCRB) Cell Bank in National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN). The cell sources were originated from Epstein-Barr virus-transformed B-lymphoblast cell lines which were established by the Japan Biological Informatics Consortium (JBIC) in collaboration with the Tokai University School of Medicine and the Institute for Genome Research of the University of Tokushima, and by the Pharma SNP consortium (PSC) in collaboration with the Institute of Rheumatology of the Tokyo Women's Medical University (Kamatani et al., 2004; Ozeki et al., 2006; Sakamoto et al., 2007; Takata et al., 2008; Nishimoto et al., 2010). The DNA samples from the remaining 15 individuals were from the Ryukyu Islands. These samples were previously reported (Matsukusa et al., 2010) and were from blood samples collected in Miyako and Ishigaki Islands (five individuals each) and from saliva samples collected in main-land Okinawa (five individuals). This research using these sets of samples was approved by the ethical committee at the Graduate School of Frontier Sciences of the University of Tokyo, under approval numbers 20-354 and 20-355.

OR sequence source information for designing probes

The sequence sources of the OR genes for designing probes in this study are listed in Supplementary Table 1. In the table, the gene name is given on the basis of its chromosome number, the OR gene cluster number on the chromosome, and the gene order in the cluster (Niimura and Nei, 2003; Niimura et al., 2018). For example, *HsOR1.4.2* represents the human (*Homo sapiens*) OR gene located on chromosome 1, in OR cluster 4 of the chromosome, and the second gene in the cluster. Pseudogenes are indicated by suffix 'P' or 'P0' after the gene name. The gene names provided by the Human Olfactory Receptor Data Explorer (HORDE) (<https://genome.weizmann.ac.il/horde/>, last accessed 21 October 2021) are also given for clarity. The gene-residing chromosome or scaffold, gene position and probe position coordinates in databases, and transcriptional direction relative to the coordinates are given. Categories are also given to each gene as 'F' for intact (putatively functional) genes, 'PF' for possibly functional genes, and 'P0' and 'P1' for 'nearly intact' pseudogenes as explained below. For each OR gene, the 100-bp flanking sequences to the coding region were included in the probe position coordinates. Repeat sequences in the flanking regions were deleted, if any, and the deleted lengths are also indicated.

The nucleotide probes for OR genes were designed for the following sequence categories:

(1) Three hundred and ninety-eight intact (putatively functional) OR genes in the human reference genome hg38 (abbreviated as hg38-intact in this study) (gene numbers 1–398 in Supplementary Table 1). These are labelled as 'F' in the Category column of Supplementary Table 1. Three hundred and ninety-seven genes are autosomal and one gene is X-non-pseudoautosomal.

(2) Four sequences that are only in 'alt' (hg38-alt) (gene numbers 399–402 in Supplementary Table 1). These are labelled as 'F (alt)' in the Category column of Supplementary Table 1. All four are autosomal. The hg38 contains haplotype information as 'alt' that is similar to OR genes but not included as OR genes. Most of the sequences in the 'alt' are identical to or only slightly different from the OR genes in hg38 (thus are considered as alleles of the hg38-intact genes). But these four sequences are only in 'alt' and the amino acid identity of them are 89–95% to the most similar standard ones. The sequences which correspond to these four genes have been identified in HORDE (Supplementary Table 1) with 99.4–100% identity. The possibly orthologous sequences also exist in the chimpanzee (*Pan troglodytes*) with 97.3–98.7% identity (GenBank accession numbers XM_527318.5, XM_521959.3, XM_009440401.2, XM_001158962.3 for the gene numbers 399–402, respectively) and bonobo (*Pan paniscus*) with 97.4–98.9% identity (XM_003829755.3, XM_003806071.2, XM_003806072.3, XM_003806807.2).

(3) Ninety-nine 'nearly intact' sequences (hg38-pseudo) (gene numbers 403–501 in Supplementary Table 1). All 99 genes are autosomal. The criteria for an intact gene in designing probes was that (i) its open reading frame (ORF), starting with a methionine codon and ending with a stop codon, shows a significant similarity to known OR genes and (ii) there are only two or fewer missing amino acids in the

highly conserved amino acid sites (Niimura, 2013; Niimura et al., 2018) in the alignment of Class 1 or Class 2 OR family. By extending the criteria, 'nearly intact' sequences were subcategorized as follows:

(3-1) Four sequences of 'possibly intact' OR genes (Category 'PF' in Supplementary Table 1). These sequences are not interrupted by stop codons or frameshifts, but there are 3–5 missing amino acids in the highly conserved sites.

(3-2) Ninety-one sequences of Category 'P0' in Supplementary Table 1. These sequences are interrupted by one stop codon or one frameshift, and there are two or fewer missing amino acids in the highly conserved sites (Niimura et al., 2018).

(3-3) Four sequences of Category 'P1' in Supplementary Table 1. These sequences are interrupted by one stop codon or one frameshift, and there are 3–5 missing amino acids in the highly conserved sites.

(4) Fifty-three intact OR genes in the chimpanzee genome database Pantro3.0 with no orthologous genes in hg38 which meet (1)–(3) (gene numbers 502–554 in Supplementary Table 1). These are labelled as 'F' in the Category column of Supplementary Table 1.

Neutral reference sequence source information for designing probes

Eighty-five single-copy non-protein-coding sequences (82 autosomal and 3 X-non-pseudoautosomal) were selected from the human reference genome hg19 assembly using the Neutral Region Explorer (Arbiza et al., 2012) as a control reference to evaluate copy number variation and selective neutrality (Supplementary Table 2). We selected regions without repeat sequences and with a length longer than 1.0 kb, distance to the nearest gene longer than 0.2 centimorgan (cM) in autosomes or 0.1 cM in the X chromosome, and a minimum recombination rate of 0.9 cM/Mb under the Neutral Region Explorer. At the stages of variant calling and copy number calling, these sequence positions in hg19 were converted to those in hg38 using the UCSC LiftOver (Kent et al., 2002). Because the genomic region of *hg19_NR50* was missing in hg38, the hg19 sequence was directly used for the procedures.

Probe synthesis, targeted capture, and NGS

Synthesis of the in-solution biotinylated RNA baits (myBaits®) was outsourced to Biodiscovery, LLC (Ann Arbor, MI) as the targeted capture probes (Gnirke et al., 2009). Each probe was 120 nucleotides (nt) in length and was overlapped by 60 nt with adjacent probes (i.e. 2 × tiling). Construction of the DNA sequencing library, targeted capture, and NGS were conducted by the Centre for Health Genomics and Informatics and UCDNA Service, University of Calgary, Canada. The DNA sequencing library was constructed by fragmenting genomic DNA into ~500 bp and by attaching adaptor oligonucleotides to the fragmented DNA. Genomic DNA samples with distinct adaptors were pooled together in groups of six. The library was heat-denatured in the presence of adapter-specific blocking oligonucleotides. The library and blockers were lowered to the hybridization temperature, allowing blockers to hybridize to the library adaptors. Biotinylated RNA baits were introduced and al-

lowed to hybridize to targets for several hours. Bait–target hybrids were pulled out of the solution with streptavidin-coated magnetic beads. Beads were stringently washed several times to remove non-hybridized and non-specifically-hybridized molecules. The captured DNA library was released from the beads and amplified by polymerase chain reaction (PCR) using complementary adaptors. The massive parallel sequencer used in this study was the Illumina Next-Seq with 300 cycles and 150-bp paired-end sequencing.

Processing of NGS read data

The procedure is outlined in Figure 1.

(1) The PRINSEQ tool was used for quality control of raw NGS reads (Schmieder and Edwards, 2011). For the FASTQ-formatted reads with the base quality score emitted by the Illumina sequencer at each nucleotide position, any nucleotide position with quality score below 20 at either edge of a read was trimmed away. Any read with a mean quality score less than 20 and any read shorter than 20 nucleotides were also filtered out.

(2) The BWA MEM tool was used to map the filtered reads to the human reference genome assembly hg38 and *hg19_NR50* as well as the 53 OR orthologous gene sequences from the chimpanzee reference genome assembly Pantro3.0. The tool is specialized for mapping highly similar reads on a reference genome (Zheng-Bradley et al., 2017), which is the case with OR genes because of the large amount of highly

similar paralogous genes.

(3) After the mapping, by applying the MarkDuplicatesSpark of the Genome Analysis Toolkit (GATK) (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013), we identified read pairs that are likely to have originated from duplicates of the same original DNA fragments through artefactual processes such as PCR amplification. Only a single read pair within each set of duplicates was used for the subsequent analyses.

(4) We then masked known variant sites reported in the HapMap 3.3 (International HapMap Consortium, 2005, 2007), the 1000 Genomes Omni 2.5, the high-confidence SNP sites in the 1000 Genomes Phase 1 (1000 Genomes Project Consortium, 2012, 2015) and the dbsnp 138 dataset to avoid removing these sites from our variant dataset by error in the subsequent base quality score recalibration and variant discovery steps.

(5) We conducted the Base Quality Score Recalibration (BQSR) using the BaseRecalibrator, a machine-learning algorithm of GATK, to detect patterns of systematic errors in the base quality scores and correct the errors possibly originating from biochemical processes during library preparation and sequencing, from manufacturing defects in the chips, or from instrumentation defects in the sequencer.

(6) For SNP and indel discovery, short variant discovery pipeline (4.1.2.0) of GATK was used. The HaplotypeCaller was used to identify haplotypes of SNPs and indels relative

Figure 1. Flowchart of the NGS read analyses.

to the reference sequence based on the physical linkage of variants in NGS reads per sample. Then, variant information from all samples was joined to make a variant database using the GenomicsDBImport of GATK. We then used the GenotypeGVCFs of GATK to extract multi-sample variant calls in the Variant Call Format (VCF) (Osada and Kawai, 2021).

(7) We then conducted the Variant Quality Score Recalibration (VQSR) using VariantRecalibrator and ApplyVQSR, machine-learning algorithms of GATK which incorporate known variant site information. As the source information on true variants, we used the HapMap 3.3 (International HapMap Consortium, 2005, 2007), the 1000 Genomes Omni 2.5, the high-confidence SNP sites in the 1000 Genomes Phase 1 (1000 Genomes Project Consortium, 2012, 2015), and dbSNP 138. Any variants within the first percentile of the variant quality score log odds in the variant recalibration were filtered out.

(8) We further refined variant calls using the CalculateGenotypePosteriors of GATK which evaluates genotype likelihood as the genotype quality score for each variant site in each study individual. Any variant sites with a genotype quality score <20 were filtered out by the VariantFiltration of GATK. We used the VariantAnnotator of GATK to annotate possible de novo mutations. After completing the previous steps, the resultant VCF dataset can be regarded with high confidence and was used in downstream analyses.

(9) Intact OR gene sequences are ~930 bp long (encoding ~310 amino acids) (Niimura, 2013). EMBOSS getorf was used to find ORFs (Rice et al., 2000). We regarded any gene with maximum ORF <750 bp (encoding <250 amino acids long) as a disrupted gene unlikely to form a functional structure, following Niimura (2013). We also examined if an OR sequence is capable of forming the 7-transmembrane (7-TM) structure, a hallmark of GPCRs by using TMHMM 2.0 (Krogh et al., 2001). We regarded any gene with an ORF of 750 bp or longer but without proper 7-TM structure as a disrupted gene. Thus, only OR genes with an ORF of at least 750 bp and capable of forming a 7-TM structure were considered as intact genes. For intact OR genes, non-synonymous variants were identified using SnpEff (Cingolani et al., 2012).

(10) The number of reads at one nucleotide site was averaged among all nucleotide sites (sequencing depth) in an OR gene or a neutral reference region per individual using the Samtools bedcov (Li et al., 2009). The sequence depths were further averaged among carrier individuals for every gene/region or among genes/regions for every carrier.

(11) The germline copy number variant discovery pipeline (4.1.9.0) of GATK was used to identify CNV of the hg38-OR genes and of hg38-liftover neutral references among individuals using the Bayesian model (Van der Auwera et al., 2013). In the first step, we instructed the PreprocessIntervals of GATK which part of the hg38 data was our study targets. We used the CollectReadCounts of GATK to count the depth of each target region for each individual for CNV analysis. In the next step, we used the AnnotateIntervals and the FilterIntervals of GATK to remove genes/regions judged as outliers in terms of high GC content because higher GC content can facilitate targeted capture and cause higher depth regardless of its copy number.

(12) We used the DetermineGermlineContigPloidy of GATK to preset the ploidy level for each chromosome based on chromosomal locality and depth information of our study gene/region using the default prior ploidy probabilities provided by GATK for autosomes and X chromosomes. We used the GermlineCNVCaller of GATK to estimate copy number state (ploidy) of each OR gene and neutral reference in terms of $0n$, $1n$, $2n$, $3n$, and so on, for every individual based on its sequencing depth relative to the mean depth among all OR genes and neutral references from the same individual. PostprocessGermlineCNVCalls was then used to extract the copy number state of each sequence in VCF file format. The GATK pipeline treats the mean depth as the single-copy diploid standard because majority of the genes and genome regions are autosomal. When the depth of a gene/region of interest is evaluated not to significantly differ from the mean depth, its copy number is estimated to be $2n$.

Mapping and assemblage for non-hg38 reference sequences

For the 53 OR genes absent in hg38 but present in chimpanzee reference genome Pantro3.0 and for the neutral reference *hg19_NR50*, another set of mapping of all PRINSEQ-controlled reads was conducted to a custom reference set containing hg38 sequences, the 53 Pantro3.0 sequences, and the *hg19_NR50* sequence. After step (1) in the 'Processing of NGS read data' section, reads mapped to the 53 OR gene sequences at step (2) were extracted using a BED file (Quinlan and Hall, 2010). Extracted reads were assembled using AbySS (Simpson et al., 2009; Jackman et al., 2017) at a custom k -mer value of 55 which was considered optimal after testing several values. After assemblage, a custom BLAST was performed between 53 Pantro 3.0 OR gene sequences and their assembled sequences to confirm their orthology (Johnson et al., 2008; NCBI Resource Coordinators, 2016). Replacing the Pantro 3.0 sequences with the assembled human sequences as references, we repeated step (2) and went through steps (3)–(10), except for steps (4), (5), and (7). Steps (4), (5), and (7) are not supported for sequences that do not exist in hg38. Regarding *hg19_NR50* sequence, steps (1)–(10) except for (4), (5), (7), and (9), were conducted. The GATK pipeline for the CNV analysis (steps (11) and (12)) was not applied to the non-hg38 sequences for simplicity.

Statistical analyses

All statistical analyses were performed using GraphPad QuickCalcs, Microsoft Excel for Microsoft 365 MSO, or IBM SPSS Statistics 22.0.

Tajima's D analysis

Tajima's D evaluates the difference in nucleotide diversity and the number of SNPs adjusted by the number of sequences for each gene or genome region (Tajima, 1989). To evaluate if a balancing selection is operating on OR genes to maintain allelic diversity in a population, we examined whether Tajima's D values of OR genes are distributed towards larger values than those of neutral references. We calculated Tajima's D for each OR gene and neutral reference sequence using a custom shell script with GATK and

VCF-kit utilities (Van der Auwera et al., 2013; Cook and Andersen, 2017). We confirmed from the quantile–quantile (QQ) plot that the distribution of Tajima's D values did not greatly deviate from a normal distribution. Then, we performed a parametric independent-sample t -test to compare the mean Tajima's D values between OR genes and neutral references. We also performed a parametric one-sample t -test to compare the mean Tajima's D values with 0.

Results

OR gene retrieval by targeted capture

The 398 hg38-intact OR genes, the four hg38-alt OR genes, the 99 hg38-pseudo OR genes, and the 53 Pantro3.0 OR genes were examined in 69 individuals. All of the 501 hg38 OR genes were retrieved from all or most of the 69 individuals. On the other hand, among the 53 Pantro3.0 OR

genes, 52 genes were not detected in any of the 69 individuals. One gene (*Chimpanzee_CM000325.3_5526517_5527518+*) was detected in one individual in an intact form and was not in any of the other individuals. Thus, a total of 502 OR genes were retrieved. Eighty-four out of the 85 neutral references were retrieved from all the 69 individuals. For one region (*hg19_NR49*), NGS reads were only patchily mapped to the reference in all individuals. Thus, the *hg19_NR49* was excluded from subsequent analyses.

Sequencing depth

The numbers of NGS reads at one nucleotide site averaged among all nucleotide sites (sequencing depth) in every OR gene or every neutral reference region per individual are listed in Supplementary Table 3 and plotted in Figure 2A. Data are presented for the 500 autosomal hg38 OR genes (mean, 414.5), the 81 autosomal neutral references (mean,

Figure 2. Per-site sequencing depths. (A) Box plots of sequencing depths for all data throughout genes/regions and throughout carrier individuals in each gene/region category. The data from the targeted capture are categorized into the autosomal OR genes (OR-A), autosomal neutral references (NR-A), an X-chromosomal OR gene (OR-X), X-chromosomal neutral references (NR-X), and a Pantro 3.0 OR gene detected in this study (OR-P). (B) Comparison of mean depths among targeted captures OR-A, NR-A, OR-X, NR-X, and OR-P, the whole-genome sequencing in the 1000 Genomes project (1K WGS), and the whole-exome sequencing in the 1000 Genomes project (1K WES). (C) Box plots of sequencing depths averaged throughout carrier individuals for each gene/region in each gene/region category. (D) Box plots of sequencing depths averaged throughout genes/regions in each gene/region category for each carrier individual. In all box plots, the per-site depth smaller than 1.0 was regarded as absence of the gene/region and are not included. The first quartile and the third quartile values are indicated nearby the lower and upper edges of the box, respectively, the median value nearby the horizontal line in the box, and the mean nearby the symbol 'X.' The upper whisker value represents the largest within 1.5 times the interquartile range above the third quartile, and the lower whisker value represents the smallest within 1.5 times the interquartile range below the first quartile. Outliers are indicated by dots. The median was included in the calculation to determine quartile values. The number of data points are indicated with n.

283.7), the one X-chromosomal hg38 OR gene (mean, 231.6) and the three X-chromosomal neutral references (mean, 120.4) of the 69 individuals, as well as the Pantro3.0 OR gene (*Chimpanzee_CM000325.3_5526517_5527518+*) found in one individual (84.6). A per-site depth <1.0 was regarded as absence of the gene/region and these reads were not included in the table (shown as blanks) or in the plot.

The mean read depths of the autosomal OR genes and the autosomal neutral references were significantly larger than those in the 1000 Genomes Project WGS (1K WGS) (7.4) and in the whole exome sequencing data (1K WES) (65.7) in which autosomal genome regions are a major component (1000 Genomes Project Consortium, 2015) (Figure 2B). Depths of the X-chromosomal OR gene and the X-chromosomal neutral references were roughly half of the autosomal counterparts (Figure 2B). Since samples were all from males, this is consistent with expectation. The low read depth of the one Pantro3.0 OR gene was comparable to those of X-chromosomal regions. The larger depth of OR genes than neutral references is consistent with frequent observation of CNV in OR genes, shown in a later section.

For each of the 500 autosomal hg38 OR genes, the 81 autosomal neutral references, the one X-chromosomal hg38 OR gene, and the three X-chromosomal neutral references, sequencing depths were averaged among the gene/region carriers and are plotted in Figure 2C, together with the sequencing depth of the *Chimpanzee_CM000325.3_5526517_5527518+* found in one carrier. Likewise, for each carrier, the sequencing depths were averaged among the 500 autosomal hg38 OR genes, among the 81 autosomal neutral references, and among the three X-chromosomal neutral references, and are plotted in Figure 2D together with the sequencing depths of the one X-chromosomal hg38 OR gene in its carriers and that of *Chimpanzee_CM000325.3_5526517_5527518+* found in the one carrier. In all plots (Figure 2A, C, D), a

broad range of depths was evident. However, depths in the interquartile range and the averaged depths in the whisker range were larger or comparable to the mean depth of 1K WES (65.7) as shown in Figure 2B.

Variant sites

Table 1 shows the number of variant sites (SNPs, insertions, and deletions relative to the references) detected in the study individuals with breakdown by JBCDB and Ryukyu samples. In the 398 hg38-intact OR genes, there were 1566 variant sites among all individuals (1518 SNPs, 16 insertions, and 32 deletions) of which 80 (77 SNPs, 1 insertion, and 2 deletions) were novel, i.e. not previously reported in HapMap, 1000 Genomes, or dbSNP projects. In the four hg38-alt OR genes, there were 18 variant sites among all individuals (17 SNPs and 1 deletion), all of which were novel. In the 99 hg38-pseudo OR genes, there were 385 variant sites among all individuals (364 SNPs, 9 insertions, and 12 deletions) of which 45 (40 SNPs, 1 insertion, and 4 deletions) were novel. In the one Pantro3.0 OR gene (*Chimpanzee_CM000325.3_5526517_5527518+*) retrieved from one individual, there were no heterozygous sites. In the 84 neutral references, there were 1362 variant sites among all individuals (1265 SNPs, 29 insertions, and 68 deletions) of which 193 (162 SNPs, 8 insertions, and 23 deletions) were novel.

Among all individuals, there were 1142 amino-acid-altering (non-synonymous: NS) SNPs in the 398 hg38-intact OR genes, of which 60 were novel (Table 1). There were 14 NS SNPs in the four hg38-alt OR genes, all of which were novel.

Among all individuals, the mean numbers of SNPs per 1000 bp (SNP density) were 4.16 in the 398 hg38-intact OR genes, 4.76 in the four hg38-alt OR genes, 4.10 in the 99 hg38-pseudo OR genes, and 5.36 in the 84 neutral referenc-

Table 1. The number of variant sites and SNP density observed in Japanese study samples

	398 hg38-intact OR genes	4 hg38-alt OR genes	99 hg38-pseudo OR genes	84 neutral references
SNP (all)	1518 (77)	17 (17)	364 (40)	1265 (162)
JBCDB	1381 (58)	17 (17)	332 (31)	1183 (134)
Ryukyu	1081 (20)	6 (6)	256 (9)	839 (34)
NS SNP (all)	1142 (60)	14 (14)	NA	NA
JBCDB	994 (13)	14 (14)	NA	NA
Ryukyu	754 (4)	0	NA	NA
Insertions (all)	16 (1)	0	9 (1)	29 (8)
JBCDB	14 (1)	0	8 (1)	25 (6)
Ryukyu	12 (0)	0	6 (0)	20 (3)
Deletions (all)	32 (2)	1 (1)	12 (4)	68 (23)
JBCDB	30 (0)	1 (1)	11 (3)	61 (18)
Ryukyu	20 (2)	1 (1)	8 (1)	35 (4)
Total (all)	1566 (80)	18 (18)	385 (45)	1362 (193)
JBCDB	1425 (59)	18 (18)	351 (35)	1243 (158)
Ryukyu	1113 (22)	7 (7)	270 (10)	911 (41)
SNP density (all)	4.16	4.76	4.10	5.36
JBCDB	3.78	4.76	3.73	4.74
Ryukyu	2.96	1.85	2.87	3.47

The number of novel variants is indicated in parentheses. NS, non-synonymous. SNP includes non-synonymous SNPs.

es. These results are also shown in Table 1. The density was highest in neutral references, which is consistent with expectations given the difference in functional constraint.

The number of variant sites is dependent on the number of samples. Accordingly, values for JBCDB (54 samples) were no larger than those for the Ryukyu samples (15 samples).

Disruption of OR genes

Among the 398 hg38-intact OR genes, 38 genes were polymorphic with disrupted and intact alleles ('segregating pseudogenes') in the 69 individuals (Supplementary Table 1). The other 360 genes in this gene group were intact in all individuals. The four hg38-alt OR genes were intact in all 69 individuals. Among the 99 hg38-pseudo OR genes, three genes (*HsOR7.6.17P0*, *HsOR14.1.9P0*, and *HsOR14.1.28P0*) from category P1 (a 'nearly intact' pseudogene category) and one gene (*HsOR19.3.8P*) from category PF ('possibly functional') were segregating pseudogenes among the 69

individuals (Supplementary Table 1), while one gene (*HsOR1.5.11P0*) from category P0 (another 'nearly intact' pseudogene category) was intact in all individuals. The other 94 genes in this gene group were disrupted in all individuals.

Frequencies of disrupted alleles of the 42 OR segregating pseudogenes are shown in Supplementary Table 4 for all samples, JBCDB samples, and Ryukyu samples. Figure 3 presents a graphical representation of all samples. While frequencies of disrupted alleles among all samples were relatively high (26–77%) in the four genes from the hg38-pseudo category (*HsOR7.6.17P0*, *HsOR14.1.9P0*, *HsOR14.1.28P0*, and *HsOR19.3.8P*), those of many hg38-intact OR genes were also high (over 10% in 11 genes and maximally 67%) (Figure 3). These frequencies were overall similar between JBCDB and Ryukyu samples (Supplementary Table 4), although a larger sample size would be required for more detailed comparison between the populations. It was noted that the OR segregating pseudogenes

Figure 3. Frequencies of disrupted alleles of the OR segregating pseudogenes in the 69 study individuals. Genes are arrayed in the order of the gene name representing the chromosome number, the OR cluster number in the chromosome, and the gene order in the cluster (see Materials and methods). The OR genes in the hg38-pseudo category are marked with a star. Chromosome numbers are shown together with the OR cluster number beside the gene names when two or more genes were found to be segregating pseudogenes in one cluster.

were often found in the same cluster of OR genes and thus were closely located in the genome (Figure 3).

Copy number variation

Three OR genes (*HsOR8.1.1*, *Human_chr6_GL000252v2_alt_682794_683756+* and *Human_chr11_JH159137v1_alt_18234_19151+*) and one neutral reference (*hg19_NR77*) were excluded from CNV evaluation due to their high GC content (see Materials and methods). The copy number of the one X-chromosomal OR gene and the three X-chromosomal neutral references were estimated to be $1n$, as expected for males in all individuals but one. The one exception was estimated to be $2n$ on the one X-chromosomal OR gene and the three X-chromosomal neutral references. Regarding the 80 autosomal neutral references, all individuals were estimated to be $2n$ as expected for the single-copy standard.

Regarding the autosomal OR genes, the copy numbers varied among all individuals in 62 genes (Supplementary Table 5). The genotype $0n$, the gene-deletion homozygote, was observed in 23 OR genes. The genotype $1n$, the gene-deletion/normal heterozygote, was observed in 16 OR genes. The genotype $3n$ was assumed to be a gene-duplication/normal heterozygote (and not the more complicated gene-triplication/gene-deletion combination) and was observed in 33 OR genes. The genotype $4n$ was likewise assumed to be a gene-duplication homozygote (and not more the complex gene-triplication/normal heterozygote or gene-quadruplicate/gene-deletion heterozygote) and was observed in five OR genes. The genotype $5n$ was assumed to be the gene-duplication/gene-triplication heterozygote (and not the more complex gene-quadruplicate/normal heterozygotes or pentaplication/deletion heterozygote) and was observed in seven OR genes. The genotype $2n$ was simply assumed to be a normal homozygote (and not the more complex gene-duplication/gene-deletion heterozygote).

The CNV allele frequencies of autosomal OR genes are shown in Supplementary Table 5 for all samples, JBCDB samples, and Ryukyu samples. Figure 4 is the graphical representation of all samples. Among all samples, the gene-deletion allele was observed in 32 OR genes, the gene-duplication allele was observed in 37 OR genes, and the gene-triplication allele was observed in seven OR genes. It was noted that the seven OR genes with a triplication allele also carried deletion and duplication alleles (Figure 4). These frequencies were overall similar between JBCDB and Ryukyu samples (Supplementary Table 5) although a larger sample size would be required for more detailed comparison between the populations. It was also noted that OR genes with CNV were often found in the same cluster of OR genes and were thus closely located in the genome (Figure 4).

Tajima's D

We examined Tajima's D for the pool of all samples ('all'), and for the JBCDB samples and Ryukyu samples separately. For each sample group, Tajima's D was examined for (1) 320 intact autosomal OR genes with no CNV and no disrupted alleles, (2) 82 disrupted autosomal OR genes with no CNV and no intact alleles, and (3) 81 autosomal neutral references (all are with no CNV). In Supplementary Table 6, these Tajima's D values are arranged in high to

low order. Genes or regions with no SNPs in a sample group were excluded from the analysis of the group and are shown as blanks in the table. The distribution of these Tajima's D values is shown in Figure 5.

In the 'all' sample group (69 individuals) (Figure 5A), among the 320 intact OR genes 268 genes were variable, among the 82 disrupted OR genes 62 genes were variable, and among the 81 neutral reference regions 79 regions were variable. There was no significant difference of Tajima's D distribution detected between the three gene/region categories. However, the mean Tajima's D of the intact OR genes was 0.32 and was significantly higher than 0 ($P < 0.0001$; Bonferroni corrected P -value threshold: 0.01667) while the mean of the disrupted OR genes (0.22) and of the neutral references (0.05) was not significantly different from 0.

In the JBCDB sample group (54 individuals) (Figure 5B), among the 320 intact OR genes 268 genes were variable, among the 82 disrupted OR genes 61 genes were variable, and among the 81 neutral reference regions 79 regions were variable. There was no significant difference of Tajima's D distribution detected between the three gene/region categories. However, the mean Tajima's D of the intact OR genes was 0.33 and was significantly higher than 0 ($P < 0.0001$, Bonferroni corrected P -value threshold: 0.01667) while the mean of the disrupted OR genes (0.25) and of the neutral references (0.05) was not significantly different from 0.

In the Ryukyu sample group (15 individuals) (Figure 5C), among the 320 intact OR genes 180 genes were variable, among the 82 disrupted OR genes 57 genes were variable, and among the 81 neutral reference regions 77 regions were variable. There was no significant difference of Tajima's D distribution detected between the three gene/region categories. However, the mean Tajima's D of the intact OR genes and of the disrupted OR genes was 0.41 and 0.34, respectively, and was significantly higher than 0 ($P < 0.0001$ and $P = 0.0155$, respectively; Bonferroni corrected P -value threshold: 0.01667) while the mean of the neutral references (0.18) was not significantly different from 0.

Among neutral references, Tajima's D value of *hg19_NR12* was the top in both JBCDB and Ryukyu individuals (Supplementary Table 6). Similarly, among disrupted OR genes, *HsOR1.4.6P0*, *HsOR11.3.15P0*, and *HsOR3.3.10P0* were top three in both JBCDB and Ryukyu individuals. Among intact OR genes, *HsOR11.3.61* and *HsOR11.3.84* were the top two in both JBCDB and Ryukyu individuals.

Discussion

In the present study, we applied a targeted capture approach, paired with short-read NGS, to examine genetic variation in 69 anonymized unrelated Japanese individuals for 398 intact OR genes, four intact but unannotated OR genes and 99 'nearly-intact' OR genes in the human reference genome hg38 and 53 intact OR genes in the chimpanzee genome database Pantro3.0 with no orthologous genes in hg38. We showed that a targeted capture by probes with a diversity-oriented design was far more effective than a whole-genome approach to retrieve OR genes and achieve high-depth sequencing and, thus, to reveal intact/disrupted polymorphisms and CNVs for the OR multigene family. We

Figure 4. The deletion, duplication, and triplication allele frequencies of autosomal OR genes in the 69 study individuals. Genes are arrayed in the order of the gene name representing the chromosome, the OR cluster number in the chromosome, and the gene order in the cluster (see Materials and methods). The OR genes in the hg38-pseudo category are marked with a star. Chromosome numbers are shown together with the OR cluster number beside the gene names when two or more genes were found to exhibit CNV in one cluster.

found that the composition of OR genes in the standard human genome data was not necessarily representative of those in humans and that many of the OR gene variants identified in this study were not previously reported, implying the po-

tential for higher perceptual variation in humans than previously believed.

We retrieved one of the Pantro3.0 OR genes (*Chimpanzee_CM000325.3_5526517_5527518+*) from one individu-

Figure 5. Box plots of Tajima's D values for intact OR genes, disrupted OR genes, and neutral references. (A) Plots for all 69 individuals. (B) Plots for 54 JBCDB individuals. (C) Plots for 15 Ryukyu individuals. For all plots, only autosomal genes/regions without CNV are considered. In addition, for intact OR genes only genes without disrupted alleles are considered, and for disrupted OR genes only genes without intact alleles are considered. n represents the number of genes/regions which exhibit SNPs and thus are considered for the Tajima's D analysis. The first quartile and the third quartile values are indicated near the lower and upper edges of the box, respectively, the median value near the horizontal line in the box, and the mean near the symbol 'X'. The upper whisker value represents the largest within 1.5 times the interquartile range above the third quartile, and the lower whisker value represents the smallest within 1.5 times the interquartile range below the first quartile. Outliers are indicated by dots. The median was included in the calculation to determine quartile values.

al as an intact gene. Its relatively low sequencing depth and lack of heterozygous sites is indicative of its heterozygous state with a deletion allele and is consistent with its low incidence. Although rare, its existence demonstrates the importance and effectiveness of population surveys using targeted capture based on a probe set that includes genes from close evolutionary relatives.

We noted that the sequencing depths of the autosomal OR genes and the autosomal neutral references were roughly double of those of the X-chromosomal OR gene and the X-chromosomal neutral references (Figure 2), which is consistent with their ploidy difference in males. Since all samples were from males, this observation supports the supposition that sequencing depth can reflect the copy number difference in our samples. We also noted that the sequencing depths in this study (mean depths, 84.6 (OR-P) ~414.5 (OR-A) (Figure 2B)) were significantly larger than that of the WGS (7.4) from the 1000 Genomes Project. Because many multigene families assemble poorly with low coverage in WGS databases using short-read NGS (Sims et al., 2014; Yohe et al., 2020) and because OR genes represent only ~0.0125% of the whole genome in mammals (Lane et al., 2001; Mainland et al., 2013), the targeted capture approach used in this study should improve the reliability and efficiency of identification of the OR multigene family and its genetic variation. We also identified a wealth of novel variant sites in our dataset including non-synonymous SNPs that were not previously reported in HapMap, 1000 Genomes, or dbSNP projects.

We showed that 38 of 42 segregating pseudogenes were in the hg38-intact OR gene category (Supplementary Table 1).

In HORDE, there are 55 segregating pseudogenes among the 501 hg38 OR genes included in this study (Supplementary Table 1). Forty-three of the 55 segregating pseudogenes are in the hg38-intact category and one is in hg38-alt category (Supplementary Table 4). The frequency of disrupted alleles in many hg38-intact OR genes in this study was high (>10% in 11 genes) (Figure 3). Thus, a majority of segregating pseudogenes are in the hg38-intact OR gene category.

On the other hand, a small fraction (four genes) of the 42 segregating pseudogenes were in the hg38-pseudo OR gene category (*HsOR7.6.17P0*, *HsOR14.1.9P0*, *HsOR14.1.28P0*, and *HsOR19.3.8P*) (Supplementary Table 4). Among most of these genes, intact alleles are not minor (Figure 3). One hg38-pseudo gene, *HsOR1.5.11P0*, was even intact in all individuals. These observations support the view that the OR gene repertoire is highly variable, and the human reference genome database represents only one example of many OR gene repertoires in human populations.

Among the 42 segregating pseudogenes in this study, 30 genes are not reported as segregating pseudogenes in HORDE. On the other hand, among the 55 segregating pseudogenes in HORDE within the 501 hg38 OR genes considered in this study, 43 genes are not segregating pseudogenes in our study: 34 genes are intact in all individuals and 9 genes are disrupted in all individuals. This observation also supports the notion that intact/disrupted composition of human OR genes is highly variable.

Among the 42 segregating pseudogenes in this study, 12 genes are reported as segregating pseudogenes in HORDE as well. Intact/disruption polymorphism of these genes could be worldwide; *HsOR11.3.63* is one such gene. The

frequency of the disrupted allele is high (62.3%) in our samples (Supplementary Table 4). This gene is also reported to be disrupted in Denisovans and Neanderthals (Hughes et al., 2014) and is intact in the chimpanzee genome database PanTro3.0. Thus, the pseudogenization of this gene appears to be ancient, at the time of hominin common ancestor. Worldwide distribution of its intact and disrupted alleles implies a sensory role of the presence/absence variation of this OR as well as of the other 11 genes.

While segregating pseudogenes are distributed throughout chromosomal OR gene clusters, we noted that some of the segregating pseudogenes appeared to be closely located in the same clusters (Figure 3). However, disruptions occur mainly due to small indels and non-sense nucleotide changes, and a possible cause of systematic incidence of these mutations is currently unclear.

We found CNV in 62 autosomal and one X-chromosomal OR gene. These CNVs were created by gene-deletion, gene-duplication, and gene-triplication alleles. In most cases, CNV was due to either gene-deletion allele or gene-duplication allele. However, when a gene-triplication allele was found, the gene was also found to carry gene-deletion and gene-duplication alleles (Figure 4). We also noted that genes with CNV were often found in the same OR gene clusters (Figure 4). CNVs were observed in both hg38-intact and hg38-pseudo OR gene categories. Frequencies of CNVs appear not to be correlated with these categories. These observations could suggest a chromatin-structural cause of CNV. Frequencies of CNVs were variable among OR genes. Thus, CNV could also contribute to olfactory variation.

We conducted Tajima's *D* analysis for autosomal OR genes and autosomal neutral references to avoid complication derived from differences in the population size between the autosomes and the X chromosome. For intact OR genes we further confined our analyses to genes without CNV and without disrupted alleles to avoid complication derived from paralogous nucleotide differences and from differences in the selective constraint between intact and disrupted alleles. For the same reason, for disrupted OR genes, we chose those without CNV and without intact alleles. All autosomal neutral references were without CNV in our study samples. Between JBCDB and Ryukyu sample groups, the same neutral reference (*hg19_NR12*) and the same disrupted OR genes (*HsOR1.4.6P0*, *HsOR11.3.15P0*, and *HsOR3.3.10P0*) were ranked highest in Tajima's *D* value. Because of the expected neutrality of both categories, this coincidence implies that the two sample groups are not genetically distinct enough for us to detect population-specific genetic traits in our study genes in spite of the supposed difference of genomic ancestry between mainland Japan and Ryukyu islands.

The mean Tajima's *D* values of intact OR genes were overall larger than zero in the three sample groups, while those of neutral references were not distinct from zero, although there was no significant difference of Tajima's *D* distribution detected between the intact OR gene category and the neutral reference. This is consistent with prediction that genetic variation of OR genes is maintained by natural selection (balancing selection), although this is not very explicit in the current dataset. Among intact OR genes, *HsOR11.3.61* and *HsOR11.3.84* were top two in both

JBCDB and Ryukyu individuals.

Although our sample size is not large and we could not detect clear genetic differences between the mainland Japanese and Ryukyu samples, with the large number of OR genes and neutral references studied we have gained significant insights into the genetic variability of the OR multigene family in a Japanese population. The targeted capture by probes with a diversity-oriented design is far more effective than a whole-genome approach to retrieve OR genes and achieve high-depth sequencing. Because of the large variation and the novelty of the variation of the OR gene family found in this study, olfactory perceptual variation in humans could also be larger than previously imagined, supporting a notion that everyone experiences their own unique 'flavor world' (McRae et al., 2013). In vitro expression assays (Zhuang et al., 2009; Adipietro et al., 2012) of the OR genes featured above for possible odorant ligands could promote our understanding of olfactory perceptual variation. The novel non-synonymous SNPs should also be an important source for functional tests that seek to explore olfactory perceptual variation within and between populations. The current study demonstrates the promise of using targeted capture with diversity-oriented probes and short-read NGS to reveal the exceptional diversity of olfactory sense among human populations worldwide.

Acknowledgments

This research was supported by a Grant-in-Aid for Scientific Research (A) (18H04005 and 15H02421) of Japan Society for the Promotion of Science (JSPS) to S.K., a Todai Fellowship to M.S.A., the JST-Mirai Program (JPMJMI19D1), Japan, to K.T. and the National Science and Engineering Research Council of Canada and Canada Research Chairs program to A.D.M. We are grateful to Kuniaki Haneji and Takashi Toma for assisting with collection of samples from Ryukyu individuals, and to Mr Muhammad Abdullah Nabeel of the College of Veterinary and Animal Sciences Jhang, Pakistan for a custom program for Tajima's *D* analysis.

References

- Adachi N., Kanzawa-Kiriyama H., Nara T., Kakuda T., Nishida I., and Shinoda K.-I. (2021) Ancient genomes from the initial Jomon period: new insights into the genetic history of the Japanese archipelago. *Anthropological Science*, 129: 13–22.
- Adipietro K.A., Mainland J.D., and Matsunami H. (2012) Functional evolution of Mammalian odorant receptors. *PLoS Genetics*, 8: e1002821.
- Arbiza L., Zhong E., and Keinan A. (2012) NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics*, 13: 301.
- Buck L. and Axel R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65: 175–187.
- Bushdid C., Magnasco M.O., Vossell L.B., and Keller A. (2014) Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343: 1370–1372.
- Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S.J., Lu X., and Ruden D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6: 80–92.

- Concas M.P., Cocca M., Francescato M., Battistuzzi T., Spedicati B., Feresin A., Morgan A., Gasparini P., and Girotto G. (2021) The role of knockout olfactory receptor genes in odor discrimination. *Genes*, 12: 631.
- Cook D.E. and Andersen E.C. (2017) VCF-kit: assorted utilities for the variant call format. *Bioinformatics*, 33: 1581–1582.
- DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43: 491–498.
- Eaton K.M., Bernal M.A., Backenstose N.J.C., Yule D.L., and Krabbenhoft T.J. (2021) Nanopore amplicon sequencing reveals molecular convergence and local adaptation of rhodopsin in Great Lakes salmonids. *Genome Biology and Evolution*, 13: evaa237.
- Gakuhari T., Nakagome S., Rasmussen S., Allentoft M.E., Sato T., Korneliussen T., Chuineagain B.N., Matsumae H., Koganebuchi K., Schmidt R., et al. (2020) Ancient Jomon genome sequence analysis sheds light on migration patterns of early East Asian populations. *Communications Biology*, 3: 437.
- Gilad Y. and Lancet D. (2003) Population differences in the human functional olfactory repertoire. *Molecular Biology and Evolution*, 20: 307–314.
- Glusman G., Yanai I., Rubin I., and Lancet D. (2001) The complete human olfactory subgenome. *Genome Research*, 11: 685–702.
- Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., Fennell T., Giannoukos G., Fisher S., Russ C., et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27: 182–189.
- Hanihara K. (1991) Dual structure model for the population history of the Japanese. *Japan Review*, 2: 1–33.
- Hasin Y., Olender T., Khen M., Gonzaga-Jauregui C., Kim P.M., Urban A.E., Snyder M., Gerstein M.B., Lancet D., and Korbel J.O. (2008) High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genetics*, 4: e1000249.
- Hoover K.C. (2010) Smell with inspiration: the evolutionary significance of olfaction. *American Journal of Physical Anthropology*, 143: 63–74.
- Hoover K.C., Gokcumen O., Qureshy Z., Bruguera E., Savangsuksa A., Cobb M., and Matsunami H. (2015) Global survey of variation in a human olfactory receptor gene reveals signatures of non-neutral evolution. *Chemical Senses*, 40: 481–488.
- Hughes G.M., Teeling E.C., and Higgins D.G. (2014) Loss of olfactory receptor function in hominin evolution. *PLoS One*, 9: e84714.
- International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, 437: 1299–1320.
- International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449: 851–861.
- Jackman S.D., Vandervalk B.P., Mohamadi H., Chu J., Yeo S., Hammond S.A., Jahesh G., Khan H., Coombe L., Warren R.L., et al. (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, 27: 768–777.
- Jaeger S.R., McRae J.F., Bava C.M., Beresford M.K., Hunter D., Jia Y., Chheang S.L., Jin D., Peng M., Gamble J.C., et al. (2013) A Mendelian trait for olfactory sensitivity affects odor experience and food selection. *Current Biology*, 23: 1601–1605.
- Jinam T.A., Kawai Y., and Saitou N. (2021) Modern human DNA analyses with special reference to the inner dual-structure model of Yaponesian. *Anthropological Science*, 129: 3–11.
- Johnson M., Zaretskaya I., Raytselis Y., Merezuk Y., McGinnis S., and Madden T.L. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36 (Web Server issue): W5–9.
- Jones M.R. and Good J.M. (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25: 185–202.
- Kamatani N., Sekine A., Kitamoto T., Iida A., Saito S., Kogame A., Inoue E., Kawamoto M., Harigai M., and Nakamura Y. (2004) Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. *American Journal of Human Genetics*, 75: 190–203.
- Keller A., Zhuang H., Chi Q., Vosshall L.B., and Matsunami H. (2007) Genetic variation in a human odorant receptor alters odour perception. *Nature*, 449: 468–472.
- Keller A., Hempstead M., Gomez I.A., Gilbert A.N., and Vosshall L.B. (2012) An olfactory demography of a diverse metropolitan population. *BMC Neuroscience*, 13: 122.
- Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., and Haussler D. (2002) The human genome browser at UCSC. *Genome Research*, 12: 996–1006.
- Koganebuchi K. and Oota H. (2021) Paleogenomics of human remains in East Asia and Yaponesia focusing on current advances and future directions. *Anthropological Science*, 129: 59–69.
- Koganebuchi K., Katsumura T., Nakagome S., Ishida H., Kawamura S., Oota H., and The Asian Archival DNA Repository Consortium (2012) Autosomal and Y-chromosomal STR markers reveal a close relationship between Hokkaido Ainu and Ryukyu islanders. *Anthropological Science*, 120: 199–208.
- Krogh A., Larsson B., von Heijne G., and Sonnhammer E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305: 567–580.
- Lane R.P., Cutforth T., Young J., Athanasiou M., Friedman C., Rowen L., Evans G., Axel R., Hood L., and Task B.J. (2001) Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proceedings of the National Academy of Sciences of the USA*, 98: 7390–7395.
- Larsen P.A., Heilman A.M., and Yoder A.D. (2014) The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. *BMC Genomics*, 15: 720.
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., and Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25: 2078–2079.
- Mainland J.D., Willer J.R., Matsunami H., and Katsanis N. (2013) Next-generation sequencing of the human olfactory receptors. *Methods in Molecular Biology*, 1003: 133–147.
- Majid A. and Kruspe N. (2018) Hunter-gatherer olfaction is special. *Current Biology*, 28: 409–413.
- Matsukusa H., Oota H., Haneji K., Toma T., Kawamura S., and Ishida H. (2010) A genetic analysis of the Sakishima islanders reveals no relationship with Taiwan aborigines but shared ancestry with Ainu and main-island Japanese. *American Journal of Physical Anthropology*, 142: 211–223.
- McColl H., Racimo F., Vinner L., Demeter F., Gakuhari T., Moreno-Mayar J.V., van Driem G., Gram Wilken U., Seguin-Orlando A., de la Fuente Castro C., et al. (2018) The prehistoric peopling of Southeast Asia. *Science*, 361: 88–92.
- McGann J.P. (2017) Poor human olfaction is a 19th-century myth. *Science*, 356: aam7263.
- McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernysky A., Garimella K., Altshuler D., Gabriel S., Daly M., et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20: 1297–1303.
- McRae J.F., Jaeger S.R., Bava C.M., Beresford M.K., Hunter D., Jia Y., Chheang S.L., Jin D., Peng M., Gamble J.C., et al. (2013) Identification of regions associated with variation in sensitivity to food-related odors in the human genome. *Current Biology*, 23: 1596–1600.
- Menashe I., Man O., Lancet D., and Gilad Y. (2002) Population

- differences in haplotype structure within a human olfactory receptor gene cluster. *Human Molecular Genetics*, 11: 1381–1390.
- Menashe I., Abaffy T., Hasin Y., Goshen S., Yahalom V., Luetje C.W., and Lancet D. (2007) Genetic elucidation of human hyperosmia to isovaleric acid. *PLoS Biology*, 5: e284.
- NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44: D7–D19.
- Niimura Y. (2013) Identification of olfactory receptor genes from mammalian genome sequences. *Methods in Molecular Biology*, 1003: 39–49.
- Niimura Y. and Nei M. (2003) Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences of the USA*, 100: 12235–12240.
- Niimura Y., Matsui A., and Touhara K. (2018) Acceleration of olfactory receptor gene loss in primate evolution: possible link to anatomical change in sensory systems and dietary transition. *Molecular Biology and Evolution*, 35: 1437–1450.
- Nishimoto K., Kochi Y., Ikari K., Yamamoto K., Suzuki A., Shimane K., Nakamura Y., Yano K., Iikuni N., Tsukahara S., et al. (2010) Association study of *TRAF1-C5* polymorphisms with susceptibility to rheumatoid arthritis and systemic lupus erythematosus in Japanese. *Annals of Rheumatic Diseases*, 69: 368–373.
- Olofsson J.K. and Wilson D.A. (2018) Human olfaction: it takes two villages. *Current Biology*, 28: R108–R110.
- Osada N. and Kawai Y. (2021) Exploring models of human migration to the Japanese archipelago using genome-wide genetic data. *Anthropological Science*, 129: 45–58.
- Ozeki T., Furuya Y., Nagano C., Matsui C., Takayanagi R., Yokoyama H., and Yamada Y. (2006) Analysis of linkage between lymphotoxin alpha haplotype and polymorphisms in 5'-flanking region of tumor necrosis factor alpha gene associated with efficacy of infliximab for Crohn's disease patients. *Mutation Research*, 602: 170–174.
- Quinlan A.R. and Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26: 841–842.
- Rice P., Longden I., and Bleasby A. (2000) EMBOS: the European Molecular Biology Open Software Suite. *Trends In Genetics*, 16: 276–277.
- Roberts S.C., Havlicek J., and Schaal B. (2020) Human olfactory communication: current challenges and future prospects. *Philosophical Transactions of the Royal Society of London. Series B: Biological sciences*, B 375: 20190258.
- Sakamoto Y., Inoue H., Keshavarz P., Miyawaki K., Yamaguchi Y., Moritani M., Kunika K., Nakamura N., Yoshikawa T., Yasui N., et al. (2007) SNPs in the *KCNJ11-ABCC8* gene locus are associated with type 2 diabetes and blood pressure levels in the Japanese population. *Journal of Human Genetics*, 52: 781–793.
- Sato T., Nakagome S., Watanabe C., Yamaguchi K., Kawaguchi A., Koganebuchi K., Haneji K., Yamaguchi T., Hanihara T., Yamamoto K., et al. (2014) Genome-wide SNP analysis reveals population structure and demographic history of the Ryukyu islanders in the southern part of the Japanese archipelago. *Molecular Biology and Evolution*, 31: 2929–2940.
- Schmieder R. and Edwards R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27: 863–864.
- Schweizer R.M., Robinson J., Harrigan R., Silva P., Galverni M., Musiani M., Green R.E., Novembre J., and Wayne R.K. (2016) Targeted capture and resequencing of 1040 genes revealed environmentally driven functional variation in grey wolves. *Molecular Ecology*, 25: 357–379.
- Secundo L., Snitz K., Weissler K., Pinchover L., Shoenfeld Y., Loewenthal R., Agmon-Levin N., Frumin I., Bar-Zvi D., Shushan S., et al. (2015) Individual olfactory perception reveals meaningful nonolfactory genetic information. *Proceedings of the National Academy of Sciences of the USA*, 112: 8750–8755.
- Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J., and Birol I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19: 1117–1123.
- Sims D., Sudbery I., Illott N.E., Heger A., and Ponting C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15: 121–132.
- Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123: 585–595.
- Takata Y., Inoue H., Sato A., Tsugawa K., Miyatake K., Hamada D., Shinomiya F., Nakano S., Yasui N., Tanahashi T., et al. (2008) Replication of reported genetic associations of *PADI4*, *FCRL3*, *SLC22A4* and *RUNX1* genes with rheumatoid arthritis: results of an independent Japanese population and evidence from meta-analysis of East Asian studies. *Journal of Human Genetics*, 53: 163–173.
- 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491: 56–65.
- 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, 526: 68–74.
- Trimmer C., Keller A., Murphy N.R., Snyder L.L., Willer J.R., Nagai M.H., Katsanis N., Vossell L.B., Matsunami H., and Mainland J.D. (2019) Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proceedings of the National Academy of Sciences of the USA*, 116: 9475–9480.
- Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., Del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43: 11.10.11–11.10.33.
- Watanabe Y., Isshiki M., and Ohashi J. (2021) Prefecture-level population structure of the Japanese based on SNP genotypes of 11,069 individuals. *Journal of Human Genetics*, 66: 431–437.
- Wooding S. (2013) Olfaction: it makes a world of scents. *Current Biology*, 23: R677–679.
- Yohe L.R., Davies K.T.J., Simmons N.B., Sears K.E., Dumont E.R., Rossiter S.J., and Dávalos L.M. (2020) Evaluating the performance of targeted sequence capture, RNA-Seq, and degenerate-primer PCR cloning for sequencing the largest mammalian multigene family. *Molecular Ecology Resources*, 20: 140–153.
- Zhang X. and Firestein S. (2002) The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience*, 5: 124–133.
- Zheng-Bradley X., Streeter I., Fairley S., Richardson D., Clarke L., Flicek P., and Genomes Project C. (2017) Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, 6: 1–8.
- Zhuang H.Y., Chien M.S., and Matsunami H. (2009) Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates. *Proceedings of the National Academy of Sciences of the USA*, 106: 21247–21251.