論 文 の 内 容 の 要 旨

論文題目　Genetic variation of olfactory receptor multigene family in humans
（人類集団における嗅覚受容体多重遺伝子族の遺伝的多様性）

氏　名　アクター　モハマッド　シュエブ

Background

Sense of smell (olfaction) is one of five basic senses in humans. Olfaction is initiated by olfactory receptors (ORs). ORs belong to G-protein coupled receptor family. OR family is the largest gene family comprising roughly 840 genes as per human reference genome hg38. Out of these OR genes, only 398 are intact. Rest of OR genes are pseudogenized. OR gene family is the most enriched with pseudogenes among gene families. Many OR genes are reported to be "segregating pseudogenes", i.e., polymorphic in terms of having pseudogenes as alleles in addition to intact alleles. Many OR genes also show copy number variations (CNV). Humans are diverse in living environment, historical subsistence, and culture, which could have influenced the olfactory diversity and adaptative evolution. However, accuracy and reliability of identification of OR gene repertoire, segregation pseudogenes, and CNV in human populations remains elusive. This is largely due to reliance on the publicly available whole genome sequence (WGS) data, which are often incomplete with low sequencing depth, resulting in misjudgments of gene identification and characterization especially for multigene families. In this study, I employed a targeted capture approach, probing OR genes followed by massive-parallel ("Next-generation") sequencing (NGS) to obtain high sequencing depth for human populations with diverse ethnicity and historical subsistence.

Materials and Methods

Study human populations: Study populations included 18 ethnic groups of Asian-, African-, and European-origins comprising 401 individuals. Asian study populations included Japanese and Filipinos. Japanese samples are from mainland Honshu (n=54) possibly representing more "Yayoi"-derived genomic background, as well as Hokkaido (Ainu) (n=29) and Ryukyu islands (n=15) possibly representing more "Jomon"-derived genomic background. Filipino populations included "Negrito" populations, indigenous historical hunter-gatherers, i.e., Aeta (n=56), Agta (n=17), Batak (n=19), and Mamanwa (n=23), and farmers, i.e., Manobo (n=28), Visayan (n=18), and Tagalog (n=22). African study populations included Chagga, Hausa and Biaka and Mbuti Pygmies, 15 individuals each, and European study populations included Dane, Irish, Adygei and Russians, 15 individuals each, representing diverse historical subsistence.

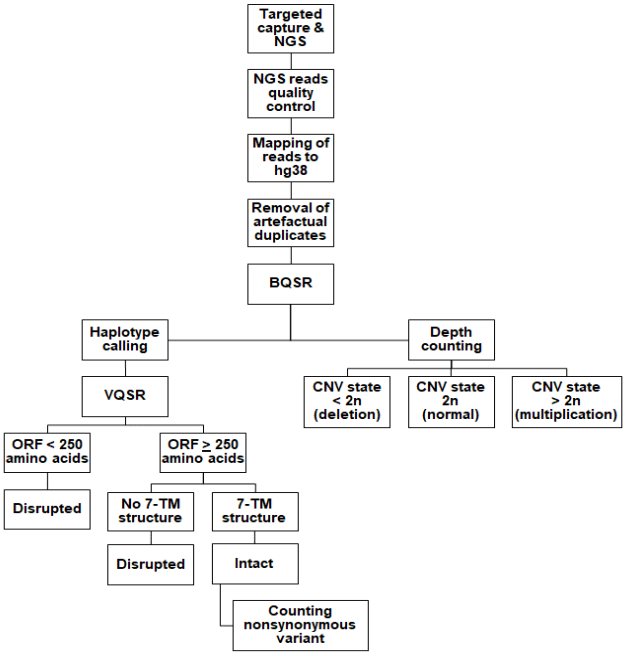

Figure 1: An overview of methods used in this study

Design of probes: I designed probes for the targeted capture mainly from the human reference genome hg38 which represents a haploid genome sequence at each nucleotide site from only one source individual. In addition to the annotated 398 OR genes in hg38, I also included four non-annotated OR sequences ("alt"s) and 99 nearly intact OR pseudogenes as well as the 53 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, I additionally designed probes for 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 human populations.

Targeted capture, NGS & bioinformatics: Synthesis of the in-solution biotinylated RNA baits (myBaits®) was outsourced to Biodiscovery, LLC (Ann Arbor, MI) as the targeted capture probes. Captured DNA fragments were sequenced using Illumina NextSeq for 150-bp paired ends with 300 Cycles. Figure 1 illustrates a flowchart of NGS reads processing. Sequencing reads were quality controlled and mapped to custom reference containing human reference genome hg38 and 53 chimpanzee OR orthologous gene sequences. The Genome Analysis Toolkit (GATK), germline short variant discovery, and germline copy number variant discovery pipelines were applied to call nucleotide and copy number variants. Reads mapped to 53 OR orthologous genes were extracted and assembled using AbySS. A consensus sequence of each OR gene was examined for open reading frame and seven-transmembrane protein structure using EMBOSS getorf and TMHMM 2.0. Principal component and Tajima's *D* analyses were conducted for population differentiation and natural selection evaluations respectively.

Results and Discussion

High depth sequence retrieval: I retrieved all hg38 OR genes and 84 neutral reference sequences from all individuals. Out of 53 chimpanzee OR sequences, one gene was retrieved as an intact OR and two were retrieved as disrupted genes. The mean depths per nucleotide site of OR and neutral reference sequences were 295 and 204, respectively, which were significantly higher than the mean depth of whole genome (7.4) and whole exome (65.7) sequences of the 1000 Genomes project (Figure 2).

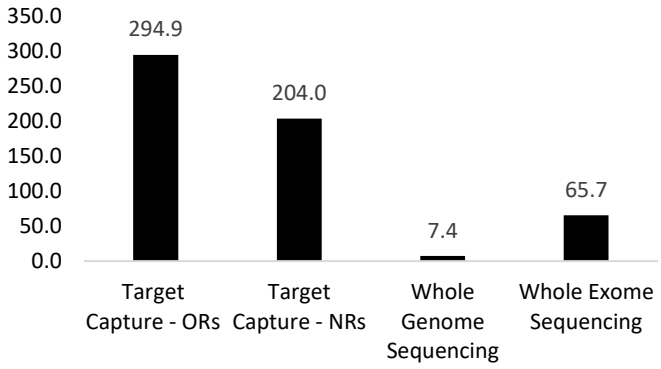

Figure 2: Comparison of mean depth of ORs and NRs with mean depth of WGS and WES

Variant sites: Variant calling of OR and neutral reference sequences revealed SNPs, insertions, and deletions many of which were novel. Frequency of these variants is given in Table 1. SNP density in each OR category and neutral reference was also calculated. Neutral reference showed a higher SNP density followed by OR pseudogenes as expected due to functional constraint.

Table 1: Frequency of variants called among ORs and neutral variants

	SNPs	Insertions	Deletions	Novel variants (%)	SNP Density (per kb)
Intact ORs	4504	50	143	10.2	12.5
Alt ORs	40	0	2	100	11.1
Pseudo ORs	1187	20	45	20.8	13.3
Neutral Reference	3735	70	209	25.1	15.6

Population differentiation: Variants data from ORs and neutral references were further analyzed using principal components analysis (PCA). PCA revealed OR repertoire is clearly differentiated between Asian, African, and European populations while neutral references were not between Asian and European populations as clearly as ORs. This implies adaptive differentiation of OR genes to the local environments over the neutral demographic differentiation.

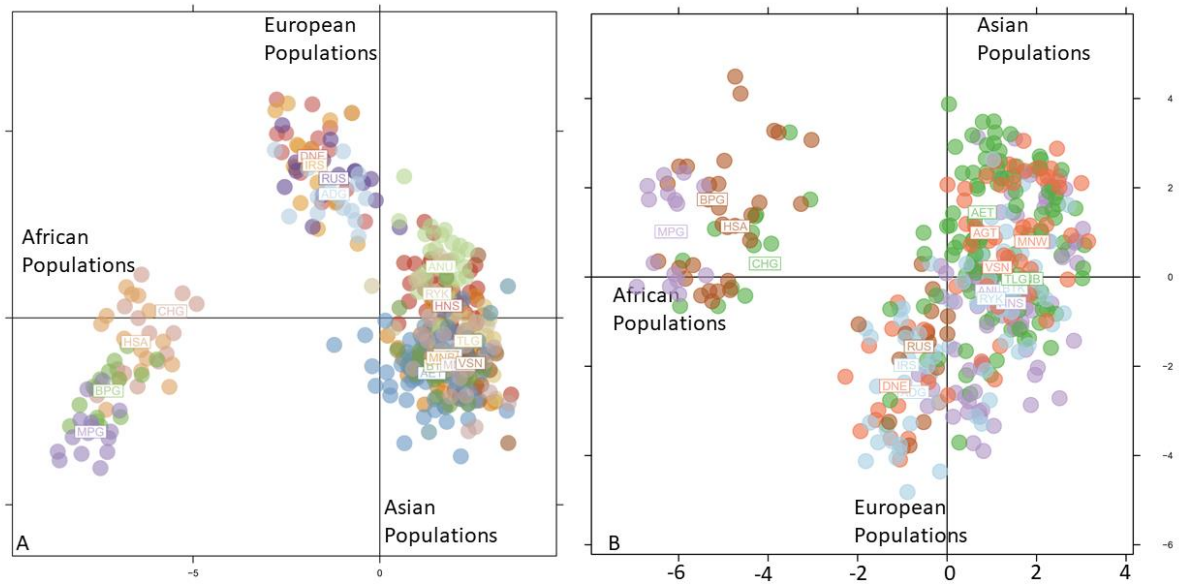

Figure 3: Principal components analysis. A: olfactory receptors; B: neutral reference

Intact/Disrupted Polymorphism: Among three categories of ORs, I looked for intact and disrupted OR genes and found 134 OR genes were in intact/disrupted polymorphism (i.e., segregating pseudogenes). Many of these polymorphisms were population specific. Out of these 134 OR genes, one OR gene belongs to Alt OR gene category and five belong to hg38 pseudogene category. Disrupted allele frequencies of the 134 intact/disrupted polymorphic OR genes are showed in Figure 4 in chromosomal order. Many hg38 intact OR genes also showed high disrupted allele frequencies. I also analyzed mean number of intact ORs in each study individual using one way ANOVA followed by Tukey's post-hoc analysis if there's any difference between populations. The numbers of intact OR genes within African, Asian and European individuals were significantly differentiated in each population.

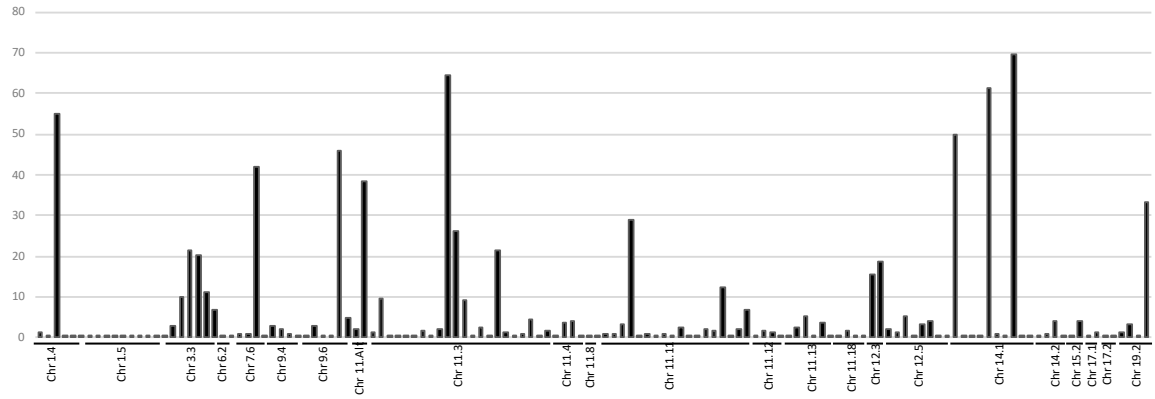

Figure 4: Disrupted allele frequency among ORs showing intact/disrupted polymorphism

Copy number variation: In addition to intact/pseudogene polymorphism, OR genes also showed CNV; results are displayed in chromosomal order in Figure 5. OR genes were found to show 0*n*, 1*n*, 2*n*, 3*n*, 4*n* and 5*n* ploidy genotypes. Based on it, deletion, duplication, and triplication allelic compositions were inferred in each individual. Among studied individuals, 176 OR genes were observed to show CNV. Out of these 176, 140 genes belong to hg38-intact category, 2 belong to Alt OR sequences and 34 belong to 99 nearly intact hg38-pseudogene category. It was noted that OR genes locating in close genomic proximity are more prone to CNV together suggestive of chromatin structural causes for CNV.

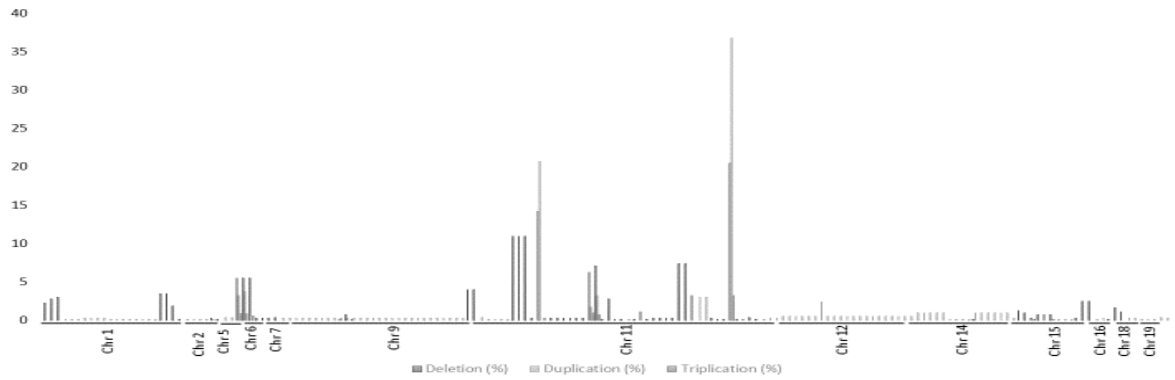

Figure 5: CNV polymorphism presented in chromosomal order

Balancing selection in ORs: Tajima's *D* was calculated for each OR gene and neutral region. Although distribution of Tajima's *D* values was not significantly different between OR genes and neutral references, the mean Tajima's *D* of OR genes was significantly larger than zero while that of neutral reference was not, being consistent with operation of balancing selection maintaining genetic diversity of OR genes over neutral variation in each population.

Effect of subsistence on OR repertoire: There is reported evidence of higher odor name proficiency among historical hunter-gatherer populations of East Asia. However, genetic differentiation of OR genes between "Negrito" and non-Negrito and within each population was as large as that of neutral references. This differentiation is an indication of role of historical subsistence in shaping OR repertoire among populations with different subsistence but living together.

Conclusion:

Targeted capture followed by NGS is significantly better for obtaining high-depth and reliable sequencing data than using publicly available whole-genome sequence data for the study of multigene families like ORs. I showed that OR multigene family is highly variable among different populations and that the public reference genome is not representative of human OR gene repertoire. I found 134 OR genes to show intact/disrupted polymorphism in contrast to human reference genome. In addition to intact/disrupted polymorphism, 164 OR genes comprising both intact and disrupted gene were found to exhibit CNV. Although not very explicitly, Tajima's *D* analysis implied that genetic variation of ORs was maintained by balancing selection. OR repertoire of hunter-gatherer populations was differentiated from non-hunter-gatherers living in similar environments implying role of historical subsistence.

Publications:

Akhtar, M. S., Ashino, R., Oota, H., Ishida, H., Niimura, Y., Touhara, K., Melin, A. D. & Kawamura, S. 2022. Genetic variation of olfactory receptor gene family in a Japanese population. Anthropological Science, 211024.