# The drifting human genome

**Jianzhi Zhang***

*Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109*

Around the time of the completion of the draft human genome sequence, biologists heatedly debated the number of genes contained in the human genome. In 2003, GeneSweep, an informal gene-count betting pool that began at the 2000 Cold Spring Harbor Laboratory Genome Meeting, announced Lee Rowen of the Institute of Systems Biology in Seattle to be the winner. His bet (25,946) was the lowest and was closest to what many computer programs predicted from the then draft human genome sequence (1). But, in this issue of PNAS, Nozawa, Kawahara, and Nei (2) tell us that asking about the exact number of human genes is meaningless, because different human individuals have different numbers of genes, and it is not uncommon for one person to have 100 more gene copies than another person. They argue that this large variation reflects genomic drift, random changes of gene copy number in evolution.

Genetic variation among humans takes many forms. In the past two decades, most studies have focused on single-nucleotide polymorphisms. For example, the International HapMap Consortium recently published HapMap II, which characterizes >3 million single-nucleotide polymorphisms genotyped in 270 individuals from four human populations of African, Asian, and European origins (3). In the last few years, multiple groups reported a high prevalence of copy number variation (CNV) of DNA segments ranging in size from $10^3$ to $10^6$ nucleotides in humans (4, 5), chimpanzees (6), and mice (7). CNV is generated by segmental duplication and deletion mutations that are apparently quite frequent. Thanks to the rapid advance of genomic technologies, genome-wide human CNV data recently became available, but studies of CNV have been largely descriptive. The work of Nozawa *et al.* (2) puts CNV into the framework of population and evolutionary genetics and is thus an exciting advance in our understanding of CNV in the human genome.

Nozawa *et al.* (2) analyzed a large dataset of human CNV that was based on a survey of the 270 HapMap individuals (4). The dataset included 1,447 CNV regions across the human genome (4), containing 3,144 annotated genes (2). Nozawa *et al.* found that on average, individuals differ from a reference individual at 277 CNV loci. Furthermore, they found a normal distribution of gene copy number among human individuals, with a standard deviation of 54. Between any two individuals, the average gene copy number difference is as high as 61.5. When Africans, Asians, and Europeans were analyzed separately, Africans showed greater variation than Asians and Europeans, consistent with the out-of-Africa model of modern human origins.

Their subsequent analysis focused on three families of chemosensory genes that encode odorant receptors (ORs), type 1 vomeronasal (pheromone) receptors (V1Rs), and bitter taste receptors (T2Rs), respectively, because chemosensory genes have particularly high levels of CNV (2, 4, 8). It is important to note that chemosensation is apparently important and highly refined in mammals, as >5% of all genes in a mammalian genome are devoted to the detection of odorants, pheromones, and tastants. The most interesting analysis Nozawa *et al.* (2) conducted is a comparison of CNV between functional and nonfunctional chemosensory genes. They found that the proportion of genes that have CNV is not significantly different between functional and nonfunctional OR genes. Furthermore, the distribution of OR gene copy number among human individuals is normal for both functional genes and pseudogenes, with almost identical standard deviations. Because pseudogenes are not subject to natural selection, the simplest explanation of this observation is that the CNV of functional OR genes is also neutral. Similar results were found for T2Rs, but the correspondence between functional genes and pseudogenes is less impressive because the T2R family is much smaller than the OR family. For V1Rs, the authors rightly treated all intact and disrupted genes as nonfunctional gene relics because the vomeronasal signal transduction pathway was destroyed in all hominoids (i.e., humans and apes) and Old World monkeys because of the inactivation of *TRPC2*, a crucial channel protein gene, in the common ancestor of hominoids and Old World monkeys (9).

**It is not uncommon for one person to have 100 more gene copies than another person.**

Nozawa *et al.* (2) further extended their analysis from within populations to between populations and between species. This type of analysis is in principle similar to the Hudson–Kreitman–Aguadé (HKA) test of neutrality of single nucleotide changes (10). The HKA test compares the ratios of the intraspecific polymorphism level and the interspecific divergence level between two loci, which are expected to be identical under neutrality. When one of the loci is a functional gene and the other is a pseudogene, rejection of the null hypothesis by the HKA test suggests nonneutral evolution of the functional gene, caused by positive or negative selection. Nozawa *et al.* measured the intraspecific polymorphism of OR gene copy number by its standard deviation among human individuals and measured the interspecific divergence of OR gene copy number by the absolute difference in OR gene number between the reference genome sequences of humans and chimpanzees. Similar measures were used for OR pseudogenes. However, they did not formally test the neutral hypothesis, presumably because it is not possible to compare polymorphism and divergence statistically when one is measured by standard deviation and the other is measured by difference. Inspired by their attempt, I designed a neutrality test in which polymorphism is measured by the number of OR loci that show CNV and divergence is measured by the difference in OR gene number between the reference genome sequences of humans and chimpanzees. Although some information about the intraspecific variation is lost in this treatment as the quantitative level of CNV at each locus is not considered, I can now compare polymorphism and divergence directly. As shown in Table 1, the null hypothesis that OR gene number variation is neutral cannot be rejected ($\chi^2 = 2.06$, $P = 0.15$), although the functional gene copy number

Author contributions: J.Z. wrote the paper.

The author declares no conflict of interest.

See companion article on page 20421.

*E-mail: jianzhi@umich.edu.

**Table 1. Testing the hypothesis of neutral variation of the number of odorant receptor genes**

| Genes | No. of genes with CNV in humans | Gene no. difference between humans and chimpanzees |
|---|---|---|
| Functional | 116 | 16 |
| Pseudogenes | 143 | 11 |

Data are from ref. 2. $\chi^2 = 2.06$, $P = 0.15$.

difference between species appears to be larger than the neutral expectation.

The findings of Nozawa *et al.* (2) and the above neutrality test suggest that both intraspecific variation and interspecific divergence of gene copy number in human OR and other chemosensory receptor families are simply due to chance. For two reasons, this result is not unexpected. First, for intraspecific CNV, the proteins encoded by various copies of the same gene are likely to be extremely similar in sequence and function because of very recent divergences of the different copies. Because dosage is unlikely to be important to chemosensory receptors, the actual copy number of each sensory gene should not affect the fitness of an individual as long as at least one copy is present in the individual. Second, olfaction is believed to play a less important role in humans than in other primates, as reflected by a decreased number of functional OR genes in the human genome compared with those in many other primates (11). As mentioned, V1Rs are not used in humans because of the loss of the vomeronasal signal transduction pathway. Humans also rely less on bitter taste than other apes do, because (*i*) bitter taste is primarily used to detect toxic food, but humans eat less plant tissues and more animal tissues and, thus, encounter fewer toxins than other apes do, and (*ii*) humans cook food, which is an effective method of detoxification. In fact, a previous study (12) showed that purifying selection on human T2R bitter receptor genes appears to have been completely relaxed. Thus, even if the various copies of a given chemosensory gene are slightly different in function, the fitness consequence in humans is expectedly to be extremely small.

Would the result be different had Nozawa *et al.* (2) studied other mammals such as mice? The answer is probably yes. Many studies showed that the numbers of T2R, OR, V1R, and V2R (type 2 vomeronasal receptor) genes vary dramatically across mammals and that this variation is among the largest of all mammalian gene families examined (13–19). For example, of those mammals that have an intact vomeronasal system, platypus has 270 functional V1R genes, whereas the dog has only 8 (15). Furthermore, the number of functional V1R genes correlates well with the morphological complexity of the vomeronasal organ (14, 15). It is thus likely that the interspecific differences of chemosensory receptor gene numbers are not entirely neutral. However, it has also been suggested that in mice and rats, V1R gene duplication is mediated by L1 repetitive elements as these elements densely populate the genomic regions harboring V1R genes (20). Thus, variation in sensory gene number could arise from differential activities of repetitive elements in different species. Taken together, it is likely that a fraction of the interspecific divergence in sensory gene number reflects adaptation, whereas the remaining fraction is neutral. Although it is difficult to assess the size of each fraction at this time, the neutral fraction is likely much higher in humans than in some other mammals that rely heavily on chemosensation.

Beyond the focus on CNV at chemosensory loci, the high level of CNV for many genes across the human genome demonstrates the neutrality or near neutrality of CNV at many loci, and genomic drift apparently occurs. However, both disease-causing CNV (5) and beneficial CNV (21) have been reported, so this type of genetic variation is not entirely neutral in humans. Under the background of neutral CNV, identification of such nonneutral human CNV will enhance our understanding of the genetic basis of disease as well as adaptation. Although genome-wide identification of CNV is rather crude at present, because the breakpoints of each copy and the copy number difference between two individuals at each locus are not precisely determined, the implications of CNV polymorphism and divergence can be reconfirmed and refined when better data become available.

Although it has been some time since the completion of the draft human genome sequence, we continue to discover and explore new types of genomic variation. Fifteen years ago, when I first heard about the Human Genome Project as a senior undergraduate student, I asked my professor "Whose genome is going to be sequenced?" He replied that it does not matter because all human genomes are essentially the same. Now I can tell him, "No, I may have a hundred more genes than you have!"


1. Pennisi E (2003) *Science* 300:1484.
2. Nozawa M, Kawahara Y, Nei M (2007) *Proc Natl Acad Sci USA* 104:20421–20426.
3. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, *et al.* (2007) *Nature* 449:851–861.
4. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, *et al.* (2006) *Nature* 444:444–454.
5. Cooper GM, Nickerson DA, Eichler EE (2007) *Nat Genet* 39:S22–S29.
6. Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, *et al.* (2006) *Proc Natl Acad Sci USA* 103:8006–8011.
7. Li J, Jiang T, Mao JH, Balmain A, Peterson L, Harris C, Rao PH, Havlak P, Gibbs R, Cai WW (2004) *Nat Genet* 36:952–954.
8. Nguyen DQ, Webber C, Ponting CP (2006) *PLoS Genet* 2:e20.
9. Zhang J, Webb DM (2003) *Proc Natl Acad Sci USA* 100:8337–8341.
10. Hudson RR, Kreitman M, Aguadé, M (1987) *Genetics* 116:153–159.
11. Gilad Y, Przeworski M, Lancet D (2004) *PLoS Biol* 2:E5.
12. Wang X, Thomas SD, Zhang J (2004) *Hum Mol Genet* 13:2671–2678.
13. Young JM, Kambere M, Trask BJ, Lane RP (2005) *Genome Res* 15:231–240.
14. Grus WE, Shi P, Zhang YP, Zhang J (2005) *Proc Natl Acad Sci USA* 102:5767–5772.
15. Grus WE, Shi P, Zhang J (2007) *Mol Biol Evol* 24:2153–2157.
16. Shi P, Zhang J (2007) *Genome Res* 17:166–174.
17. Shi P, Zhang J (2006) *Mol Biol Evol* 23:292–300.
18. Niimura Y, Nei M (2006) *J Hum Genet* 51:505–517.
19. Young JM, Trask BJ (2007) *Trends Genet* 23:212–215.
20. Lane RP, Cutforth T, Axel R, Hood L, Trask BJ (2002) *Proc Natl Acad Sci USA* 99:291–296.
21. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, *et al.* (2007) *Nat Genet* 39:1256–1260.