

- hypoplasia as fovea plana. *Arch Ophthalmol* 2008; 126:907-13. [PMID: 18625935]
27. Byeon SH, Kang SY. Interpretation of outer retina appearance in high-resolution optical coherence tomography. *Am J Ophthalmol* 2009; 147:185-6. [PMID: 19100358]
28. Lim JI, Tan O, Fawzi AA, Hopkins JJ, Gil-Flamer JH, Huang D. A pilot study of Fourier-domain optical coherence tomography of retinal dystrophy patients. *Am J Ophthalmol* 2008; 146:417-26. [PMID: 18635153]
29. Kondo M, Miyake Y. Assessment of local cone on- and off-pathway function using multifocal ERG technique. *Doc Ophthalmol* 2000; 100:139-54. [PMID: 11142743]
30. Miyake Y. What can we know from focal macular ERG? *Jpn J Clin Ophthalmol*. 2002; 56:680-8.
31. Okuno T, Oku H, Kondo M, Miyake Y, Sugasawa J, Utsumi T, Ikeda T. Abnormalities of visual-evoked potentials and pupillary light reflexes in a family with autosomal dominant occult macular dystrophy. *Clin Experiment Ophthalmol* 2007; 35:781-3. [PMID: 17997791]
32. Hanazono G, Ohde H, Shinoda K, Tsunoda K, Tsubota K, Miyake Y. Pattern-reversal visual-evoked potential in patients with occult macular dystrophy. *Clin Ophthalmol*. 2010; 4:1515-20. [PMID: 21191449]
33. Miyake Y. Occult macular dystrophy. *Electrodiagnosis of retinal diseases*. Tokyo, Japan: springer-Verlag; 2006:153-159
34. Sieving PA, Murayama K, Naarendorp F. Push-pull model of the primate photopic electroretinogram: a role for hyperpolarizing neurons in shaping the b-wave. *Vis Neurosci* 1994; 11:519-32. [PMID: 8038126]
35. Sieving PA. 'Unilateral cone dystrophy': ERG changes implicate abnormal signaling by hyperpolarizing bipolar and/or horizontal cells. *Trans Am Ophthalmol Soc* 1994; 92:459-71. [PMID: 7886877]

Articles are provided courtesy of Emory University and the Zhongshan Ophthalmic Center, Sun Yat-sen University, P.R. China. The print version of this article was created on 20 April 2012. This reflects all typographical corrections and errata to the article through that date. Details of any changes may be found in the online version of the article.

A New Database (GCD) on Genome Composition for Eukaryote and Prokaryote Genome Sequences and Their Initial Analyses

Kirill Kryukov^{1,2}, Kenta Sumiyama¹, Kazuho Ikeo^{2,3}, Takashi Gojobori^{2,3}, and Naruya Saitou^{1,*}

¹Division of Population Genetics, National Institute of Genetics, Mishima, Japan

²Genome Network Project, National Institute of Genetics, Mishima, Japan

³DNA Data Analysis Laboratory, National Institute of Genetics, Mishima, Japan

*Corresponding author: E-mail: saitounr@lab.nig.ac.jp.

Accepted: 8 March 2012

Abstract

Eukaryote genomes contain many noncoding regions, and they are quite complex. To understand these complexities, we constructed a database, Genome Composition Database, for the whole genome composition statistics for 101 eukaryote genome data, as well as more than 1,000 prokaryote genomes. Frequencies of all possible one to ten oligonucleotides were counted for each genome, and these observed values were compared with expected values computed under observed oligonucleotide frequencies of length 1–4. Deviations from expected values were much larger for eukaryotes than prokaryotes, except for fungal genomes. Mammalian genomes showed the largest deviation among animals. The results of comparison are available online at <http://esper.lab.nig.ac.jp/genome-composition-database/>.

Key words: GCD, oligonucleotide frequency, alignment-free sequence comparison.

Introduction

Noncoding regions are the major part of eukaryote genomes, and most of them are believed to evolve neutrally (Kimura 1983). Under this assumption, we expect that the frequency of a particular short oligonucleotide, or DNA word, of 10 bp or shorter should be primarily determined through accumulation of neutral mutations, and the total set of frequencies of all DNA words of certain length should follow some simple statistical rules. Oligonucleotide frequencies of one genome can provide a useful mechanism of genome comparison (Karlín 2005), including phylogeny reconstruction (Takahashi et al. 2009). Most frequently, such comparisons are based on a dinucleotide composition model (Karlín and Mrazek 1997; Gentles and Karlín 2001) or on self-organizing maps (Abe et al. 2003). It may be better to examine longer oligonucleotide compositions. We created a series of statistical models predicting the frequencies of word of up to 4 nt in a genome. We retrieved all available complete eukaryote and prokaryote genomes, constructed such models for them,

and compared the actual word frequencies with those predicted by the models to determine the discrepancy.

Here, we present a database, called Genome Composition Database (GCD), which shows how accurately each genome can be approximated by a model. The GCD also provides the sequences of over- and underrepresented DNA words. The unique point of this database is that it allows to compare compositional complexity of genomes and to analyze over- or underrepresentation of particular oligonucleotides.

Materials and Methods

Available complete genomes were collected from NCBI (<http://www.ncbi.nlm.nih.gov/>; Wheeler et al. 2007), Ensembl (<http://uswest.ensembl.org/>; Flicek et al. 2012), University of California–Santa Cruz (<http://genome.ucsc.edu/>; Fujita et al. 2011), FlyBase (<http://flybase.org/>; McQuilton et al. 2012), and WormBase (<http://www.wormbase.org/>; Harris 2010). Genome sequences of a total of 1,228 species (101 eukaryotes, 1,043 eubacteria, and 84 archaea, as of

June 2010) were used to construct the database. For every genome, we created a series of five composition models: uniform (composition of A, C, G, and T are set to be all 25%), mononucleotide, dinucleotide, trinucleotide, and tetranucleotide. Each composition model is based on the total size and word frequencies of an actual genome.

The uniform composition model has just one parameter—genome size. The mononucleotide model has two parameters—genome size and GC content. We use both DNA strands to perform the word counting, so the number of G bases is always same with number of C, same for A and T, and each DNA word has the same frequency with its reversed complementary counterpart. Among the 16 dinucleotides, there are 12 that differ from their reversed complementary dinucleotide and 4 that are identical to their reversed complementary one (CG, GC, AT, and TA). Therefore, the first group of dinucleotides can be described with six frequencies (12/2) and the second—with four. Subtracting one, and adding the genome size, we obtain ten parameters for the dinucleotide model. In case of trinucleotide frequencies, none of the trinucleotides are identical to their reversed complementary counterpart, so the model has $4^3/2 = 32$ parameters. In tetranucleotide case, there are 16 tetranucleotides that are identical to their reversed complementary counterparts, so the tetranucleotide model has $(4^4 - 16)/2 + 16 = 136$ parameters.

For a genome G of total length M and a DNA word w , a composition model can be used to compute $p(w)$, which is the probability of observing w at any particular position in the genome. For example, the uniform composition model gives

$$p(w) = \frac{1}{4^L}, \quad (1)$$

where L is the length of w . The mononucleotide composition model predicts

$$p(w) = \prod_{i=1}^L \frac{F(w_i) + F[C(w_i)]}{2M}, \quad (2)$$

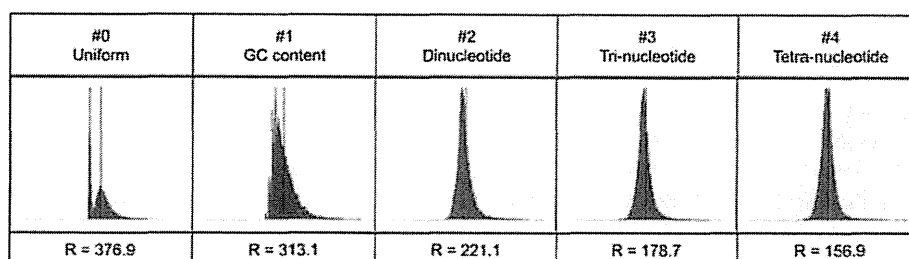


FIG. 1.—Histograms of relative abundances of all oligonucleotides of 8 bp in human genome, according to the five composition models. The R value computed for each model is used as a horizontal scaling factor. The vertical red line corresponds to the expected frequency. The words placed to the left of the line are underrepresented and to the right—overrepresented.

Table 1
 R Value Comparison for Selected Species

	Model				
	Uniform	Mono	Di	Tri	Tetra
<i>Escherichia coli</i> E24377A	9.5	9.4	7.6	5.3	3.2
<i>Saccharomyces cerevisiae</i> (baker's yeast)	18.7	9.0	6.2	5.0	3.4
<i>Arabidopsis thaliana</i> (thale cress)	72.7	33.6	23.7	18.6	13.9
<i>Drosophila melanogaster</i> (fruit fly)	59.7	41.3	29.9	23.1	19.3
<i>Oryzias latipes</i> (medaka)	165.9	115.8	71.2	49.5	37.3
<i>Anolis carolinensis</i> (lizard)	251.1	188.9	130.4	110.0	92.1
<i>Mus musculus</i> (mouse)	343.9	309.0	219.0	145.1	122.8

Note.—This table compares the R values of *E. coli*, yeast, plant, fruit fly, fish, lizard, and mouse, respectively, for each of the five models we used, based on words of 8 bp.

where w_i is the i th nucleotide of w , $F(x)$ is the observed frequency of x in the genome sequence, and $C(x)$ is the complementary sequence to x . Using the same principle, $p(w)$ from dinucleotide, trinucleotide, and tetranucleotide composition models can be computed.

The model expectation of the frequency of word w in both strands of the modeled genome is then given as follows:

$$E(w) = 2Mp(w). \quad (3)$$

Then, we can define the deviation of the observed frequency from the expected frequency:

$$d(w) = F(w) - E(w). \quad (4)$$

Because each of the composition models assumes independence of different genome positions from each other, $E(w)$ follows the binomial distribution, and its variance can be computed as follows:

$$\sigma_{E(w)}^2 = 2Mp(w)[1 - p(w)]. \quad (5)$$

The standard deviation of $E(w)$ is its square root.

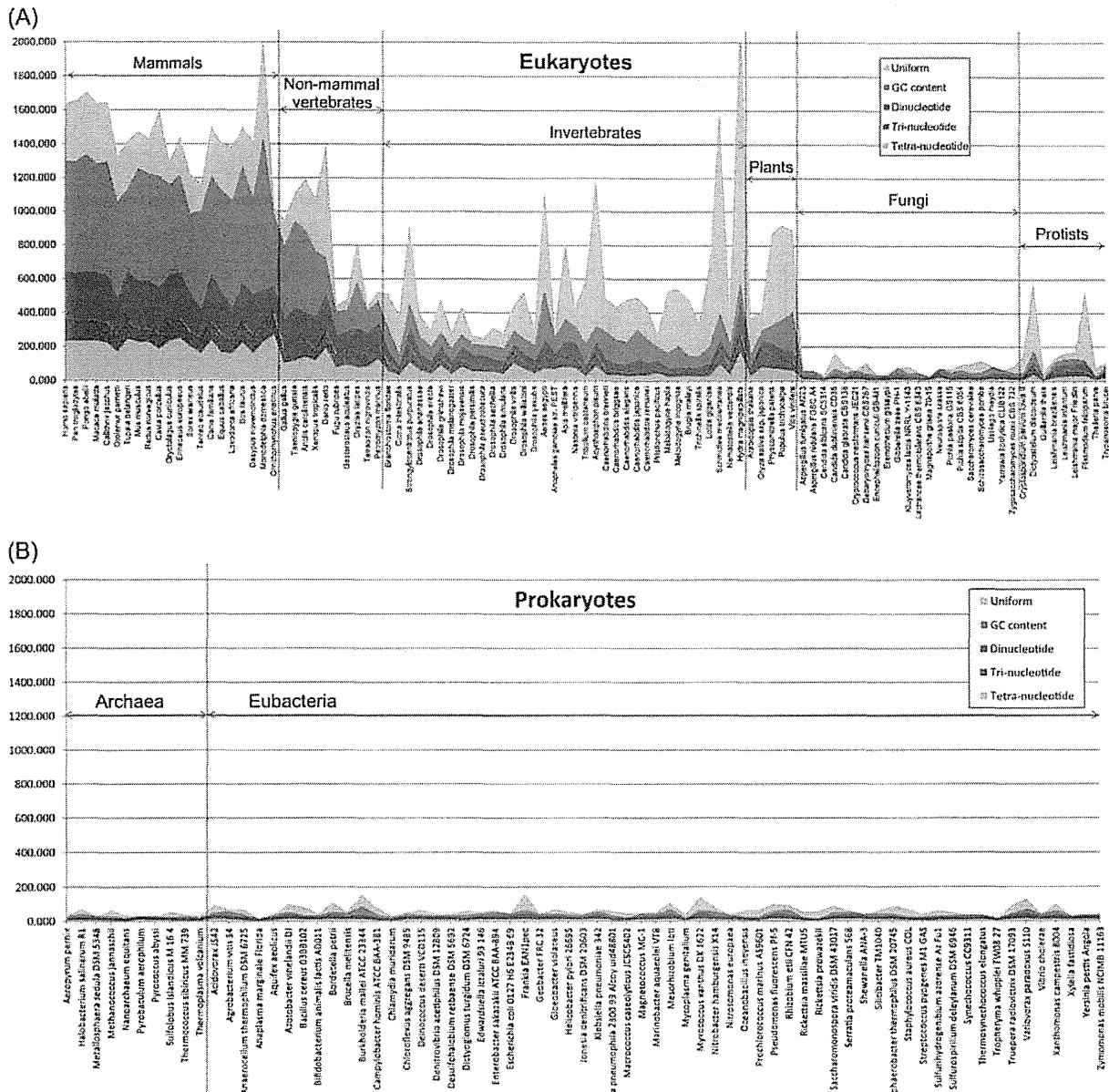


FIG. 2.—Comparison of R values based on oligonucleotides of 5 bp and all five composition models. (A) Eukaryote genomes (all available in public databases by October 2010). (B) Representative prokaryote (both eubacteria and archaea) genomes.

We then can define the relative abundance of w , under this particular model, as follows:

$$r(w) = \frac{d(w)}{\sigma_{E(w)}} \quad (6)$$

This $r(w)$ is 0 for DNA words, occurring in the genome with exactly the same frequency, as predicted by the composition model. $r(w)$ is positive when the actual frequency is

larger than expected by the model. In such cases, we describe that w is overrepresented in the genome, according to this model. When the actual frequency is smaller than expected by the model, $r(w)$ is negative, and w is underrepresented.

Now we can summarize the overall magnitude of over- or underrepresentation of all DNA words of length L in the genome (using a particular composition model of choice) as follows:

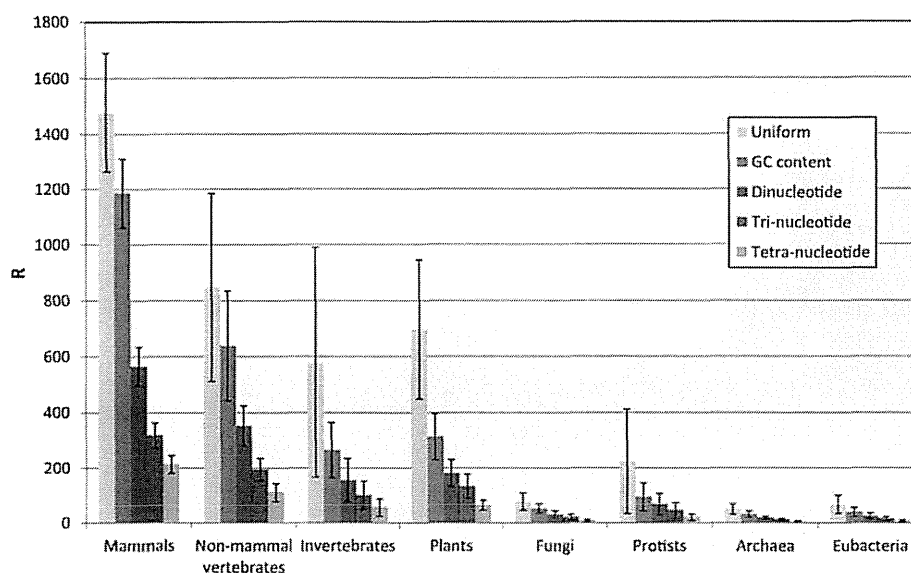


FIG. 3.—Average R values for different groups of organisms, with standard deviations, using five different composition models. Standard deviation for each value is displayed.

$$R = \sigma_{r(w)} = \sqrt{\frac{\sum_{w \in W} [r(w) - \bar{r}]^2}{4^L}}, \quad (7)$$

where W is the set of all DNA words of length L and $\bar{r} = \frac{\sum_{w \in W} r(w)}{4^L}$.

Because R is the standard deviation of a sample of all $r(w)$ for a particular word length L , the unit of R is the same with that of $r(w)$, which is $\sigma_{E(w)}$ (standard deviation of the word frequency, predicted by the model). For each w , $R\sigma_{E(w)}$ gives the relative number of occurrences of w , which would make w averagely rare or abundant.

R is computed for a particular genome, composition model, and L and summarizes the ability of the composition model to predict the frequencies of words of length L in the

genome. Large R implies that many w 's have large absolute values of $r(w)$, which means that their actual frequencies are far from those expected by the model. Thus, a large value of R signifies that the model's ability to describe the actual genome is poor.

A good composition model has small value of R , with R being 0 for the perfect model. An example of such perfect model is the L -bp composition model used to predict the frequencies of words of the same length L bp or shorter. For instance, the dinucleotide composition model has the exact information about dinucleotide frequencies, so it gives perfect predictions for 1-bp or 2-bp word frequencies, resulting in R value of 0.

For the longer words, R is typically much larger than 0 for nonrandom sequences. On the other hand, when a random sequence is modeled using any composition model, the

Table 2
Underrepresented Oligonucleotides of 10 bp, Example from Human Genome

Rank	Oligonucleotide	Actual Observed Frequency	Frequency Predicted by the Model	Deviation from the Expected Frequency, in Model's Standard Deviations
1	tataaaaaa (ttttttata)	45,933	115,110	-203.9
2	aaatttttc (gaaaaaattt)	29,389	89,480	-200.9
3	ttttttggg (cccaaaaaaa)	19,774	72,956	-196.9
4	aaaaatttt	103,832	185,936	-190.4
5	ttttttgga (tcccaaaaaa)	14,119	60,161	-187.7
6	aaaattttc (gaaaaatttt)	33,460	89,480	-187.3
7	aaaaaatat (atatttttt)	80,964	153,706	-185.6
8	aaaaaattc (gaaattttt)	34,571	89,480	-183.6
9	aaaaatttg (caaaattttt)	33,265	87,274	-182.8
10	aaaaatttg (caaaattttt)	33,454	87,274	-182.2

NOTE.—Showing ten most underrepresented oligonucleotides, according to the tetranucleotide composition model. Both the actual and the expected frequency are given for both DNA strands combined, so each word's frequency is identical with that of its reversed complementary counterpart (given in parentheses).

Table 3
Overrepresented Oligonucleotides of 10 bp, Example from Human Genome

Rank	Oligonucleotide	Actual Observed Frequency	Frequency Predicted by the Model	Deviation from the Expected Frequency, in Model's Standard Deviations
1	acacacacac (gtgtgtgtgt)	1,161,477	9,207	12008.1
2	tgtgtgtgtg (cacacacaca)	1,169,668	12,946	10166.1
3	cctgtaatcc (ggattacagg)	835,133	6,999	9898.3
4	ctgtaatccc (gggattacag)	825,499	7,235	9619.4
5	aaaaaaaaaa (ttttttttt)	5,951,413	380,529	9031.2
6	ctgggattac (gtaatcccag)	802,262	7,934	8917.5
7	tglaatccca (tgggattaca)	856,563	11,024	8053.0
8	taatcccagc (gctgggatta)	839,950	10,726	8006.5
9	gattacaggc (gcctgtaatc)	628,774	7,004	7429.1
10	tgcaagtgcg (gctcactgca)	580,240	7,705	6522.3

Note.—Showing ten most overrepresented oligonucleotides, according to the tetranucleotide composition model.

actual variances of the word frequencies are the same with the variances predicted by the model; therefore, R is close to 1 in this case (approaching 1 as the sequence becomes longer).

This is also the case for semirandom sequences, where the deviation from uniform randomness is at most as complex (controlled by at most as many parameters) as the model used to analyze the sequence. For example, a semirandom GC-biased sequence can be accurately modeled by the nucleotide composition model, or any more complex model, but not by the uniform composition model. The R values obtained with the uniform composition model for such sequence are much larger than 1, whereas other models still produce R close to 1. Thus, the R values directly reflect compositional complexity of the sequence.

Figure 1 illustrates this by showing the example histograms of relative abundances for all words of length 8 in the human genome, using five different models. The strange bimodal-looking shape of the uniform model histogram results from the extreme depletion of CpG dinucleotide in mammalian (including human) genomes. Any 8-bp word containing CpG will appear as strongly underrepresented when comparing the actual frequencies with those predicted by the uniform model. So, all such words contribute to the left peak on the histogram, whereas words without CpG form the other peak, in agreement with the model.

We computed R for all five composition models for available complete genomes, both eukaryotes and prokaryotes. Table 1 shows R values for seven representative species. We then extracted unusually rare and unusually abundant words, which we define as those having $|r(w)| > R$. These DNA words, together with the corresponding statistics, are available for viewing and downloading at the GCD online.

Next, we analyzed the spacing patterns of individual DNA words in complete genomes. Looking at all occurrences of a particular DNA word in the genome, we can extract the distances between the genomic locations of every two

neighboring occurrences and use this set of distances as a spacing data set for this particular word. Sample parameters (mean, standard deviation, skewness, and kurtosis) are computed for such data set. What would be the physical meaning of those parameters? The mean distance approximately equals to the genome size divided by total number of occurrences, so it correlates with the reciprocal of the word frequency. Standard deviation shows how evenly is a particular word distributed in the genome. Skewness shows whether extremely unusual spacing values for this word tend to be large or small. Kurtosis shows if the word tends to form clusters and the density of those clusters relative to the distance between them.

Taking a particular parameter for all words of length L , we get a sample of 4^L values. The nature of this sample would characterize the genome as a whole. Furthermore, selecting only subset of DNA words with parameters falling into particular ranges, we can extract interesting DNA words.

In order to verify the models and better understand the parameters, we constructed a range of semirandom sequences using a random sequence generator (Kryukov K, unpublished data). Each semirandom sequence was based on particular real genome used as template (e.g., the human genome): It had the same size with the template genome, and it imitated N -bp composition of the template genome, with N ranging from 1 to 4. Thus, we constructed four semirandom genomes based on a single actual genome sequence. We used genomes of five species as templates: human, *Anolis carolinensis* (lizard), *Xenopus tropicalis* (frog), *Oryzias latipes* (fish), and *Drosophila melanogaster* (fruit fly). The resulting 20 semirandom genomes were added into the GCD.

Results

Figure 2 shows the comparison of R values for 101 eukaryote genomes used in this study, as well as representative prokaryote genomes, computed for 5 bp oligonucleotides.

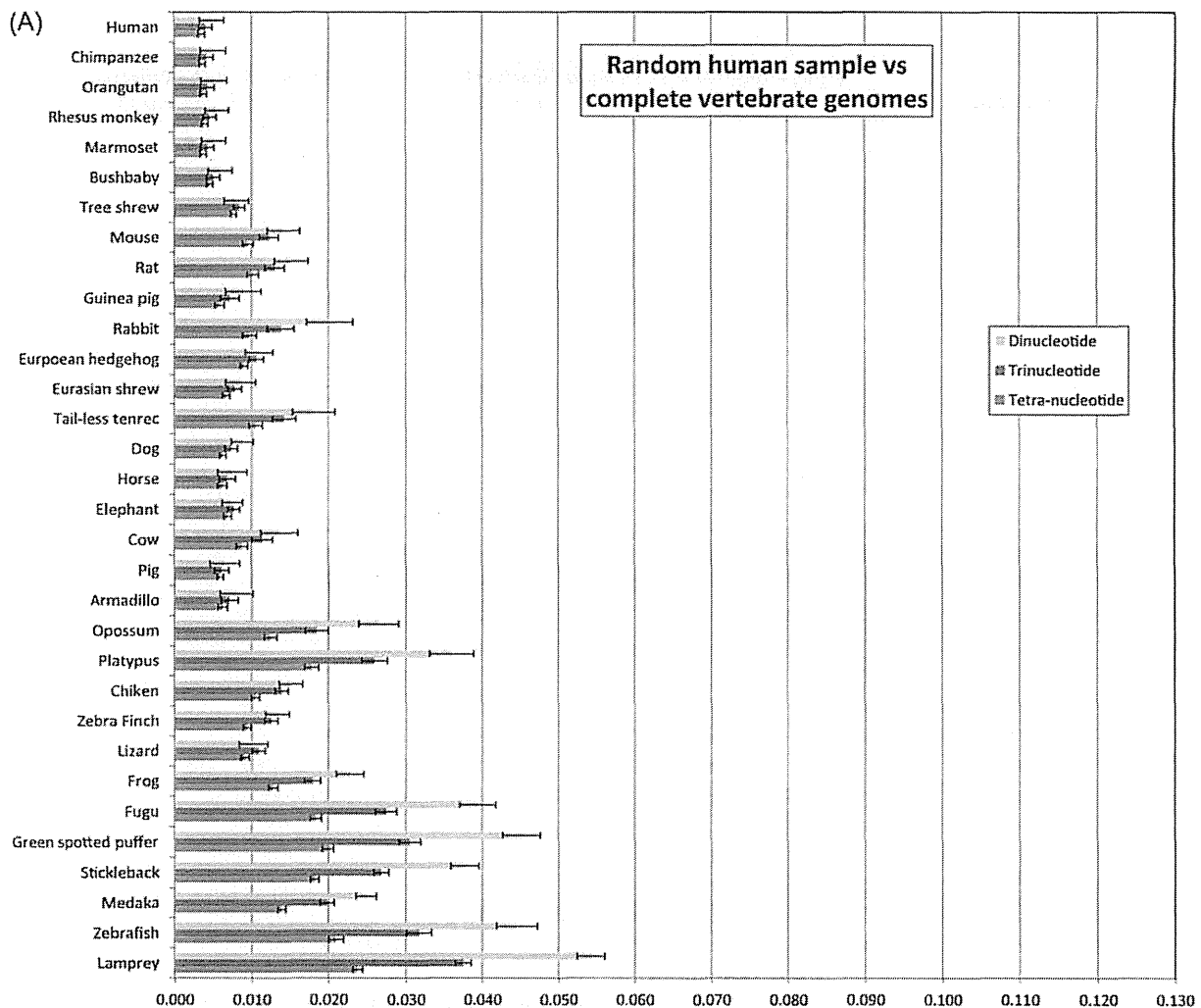


FIG. 4.—Euclidean distances between composition vectors (oligonucleotide frequencies) of sample data sets and complete vertebrate genomes for three composition models (dinucleotide, trinucleotide, and tetranucleotide). (A) When sampled data set is human genome. One thousand samples were used, where each sample consisted of 481 sequences of 262 bp each (for a total size of each sample same with the UCE data set), taken from the random locations in the complete human genome. Also, panel (A) shows the standard deviations of the distances. (B) The composition of the UCE data set is compared with that of complete vertebrate genomes. (C) The composition of human miRNA seed sequences is compared with that of complete vertebrate genomes.

Such R values represent how well different composition models can predict 5-bp composition of the genome. Panel A shows all eukaryote genomes and panel B shows representative prokaryote genomes. Supplementary figure 1 (Supplementary Material online) shows comparison of all prokaryote genomes included in this study. R values of five composition models are displayed as differently colored areas. As can be seen, R varies greatly among species and groups of species. Mammals are compositionally more complex than nonmammal vertebrates, land vertebrates are more complex than fishes, and fishes are more complex than most invertebrates and plants, which are still more complex

than fungi and protists. Compositional genome complexity of prokaryotes, represented by R values, is comparable with that of fungi.

Figure 3 shows the average R values for different groups of organisms, with standard deviation. Under all five composition models, statistically significant difference is observed between the R values of mammals and nonmammal vertebrates (Mann–Whitney $P < 0.001$, see supplementary table 1, Supplementary Material online for test results). Statistically significant difference is also observed between nonmammal vertebrates and invertebrates. Interestingly, R values of invertebrates are close to those of

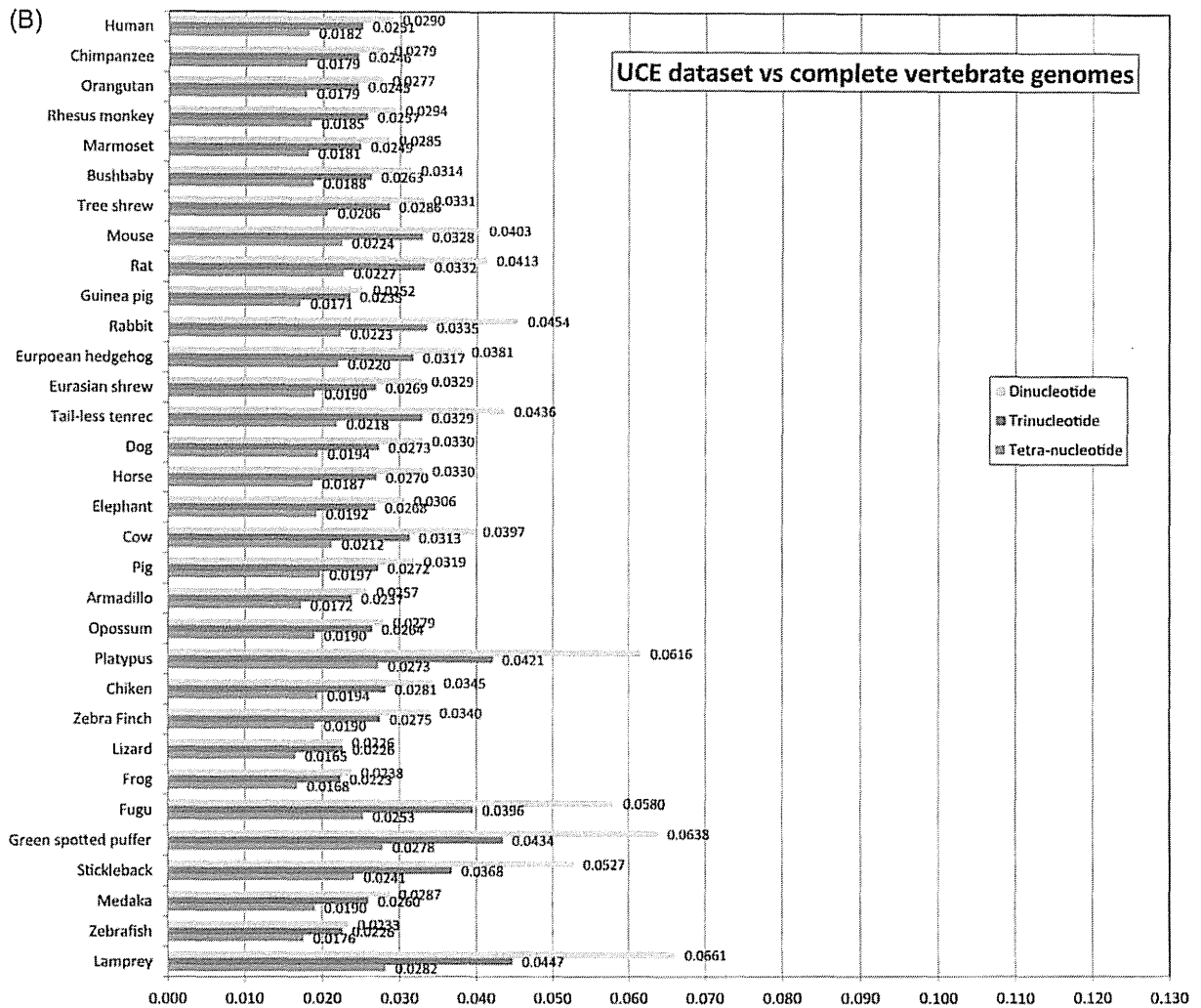


Fig. 4.—Continued

plants and significantly higher than those of fungi, protists, or prokaryotes (archaea and eubacteria). In terms of *R* values, fungi genomes are more similar to those of prokaryotes than those of other eukaryotes.

Significantly, over- and underrepresented DNA words may be biologically important. Tables 2 and 3 show the partial lists of under- and overrepresented words of 10 bp in human genome, using tetranucleotide composition model. The complete lists of under- and overrepresented words, for every of the included genomes, for each of the five composition models, and for DNA words of up to 10 bp for eukaryotes and 8 bp for prokaryotes, are available at the GCD online. Both the actual and the expected frequency are given for both DNA strands combined, so each word's frequency is identical with that of its reversed complementary counterpart (given in parentheses).

Other than the reporting the general compositional complexity, the GCD can be used to compute the distances between the composition vectors of various complete genomes and submitted sequences (similar to the method taken by Takahashi et al. 2009). We used this tool to analyze three classes of human sequences: random sample from the human genome, conserved sequences of unknown function, and conserved functionally important sequences. Although sequences from these three classes are all found in the human genome, they have different nature and evolutionary history, allowing interesting comparison. The UCE data set (human–mouse–rat ultraconserved elements, 481 sequence, 126 kbp in total, Bejerano et al. 2004) was used as the data set of conserved sequences of unknown function. Human microRNA (miRNA) seed sequences (1,100 sequences from

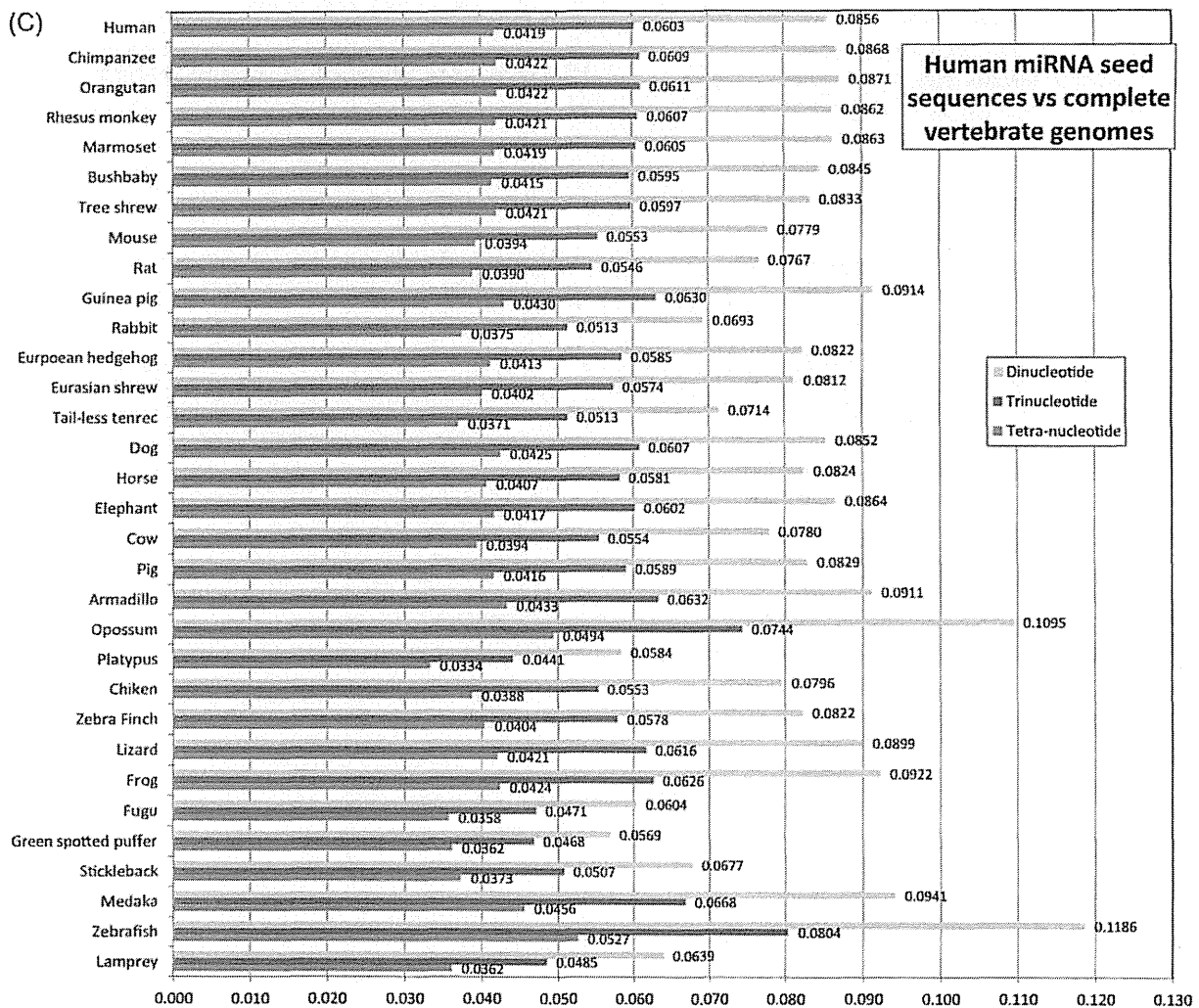


FIG. 4.—Continued

miRBase, 7.7 kbp in total, Kozomara and Griffiths-Jones 2011) were used as functionally important conserved sequences.

Figure 4A shows the average Euclidean distances between the composition vectors obtained from randomly sampled human sequence and composition vectors of complete vertebrate genomes. Each sample was chosen to have the same number of sequences and average sequence length with the UCE data set: 481 sequences, 262 bp each. One thousand such samples were produced. Di-, tri-, and tetranucleotide composition vectors are used for comparison. As expected, primate genomes are the closest to human sample, and more diverged species show progressively larger distances, with some fluctuations.

Figure 4B shows the comparison for human–mouse–rat ultraconserved elements. The compositional distances between the UCE and the complete vertebrate genomes

appear to be relatively uniform among vertebrates and much larger than those for the random human sample. Interestingly, these sequences appear to be compositionally close to lizard, fish, and frog.

Figure 4C shows the compositional distances between human miRNA sequence data set and complete vertebrate genomes. Again the distances are uniformly large. Platypus and the fishes are compositionally the closest to this data set.

To further investigate the differences between these three data sets, we computed the average distances by combining the genomes into four groups (fig. 5). The distances show a steep increase in case of random human sample (fig. 5A), while much more uniformity can be seen for UCE and miRNA seed data sets (fig. 5B and C).

Figure 6 shows the plots for the pairs of spacing parameters, taken for 8 bp oligonucleotides for six species—human,

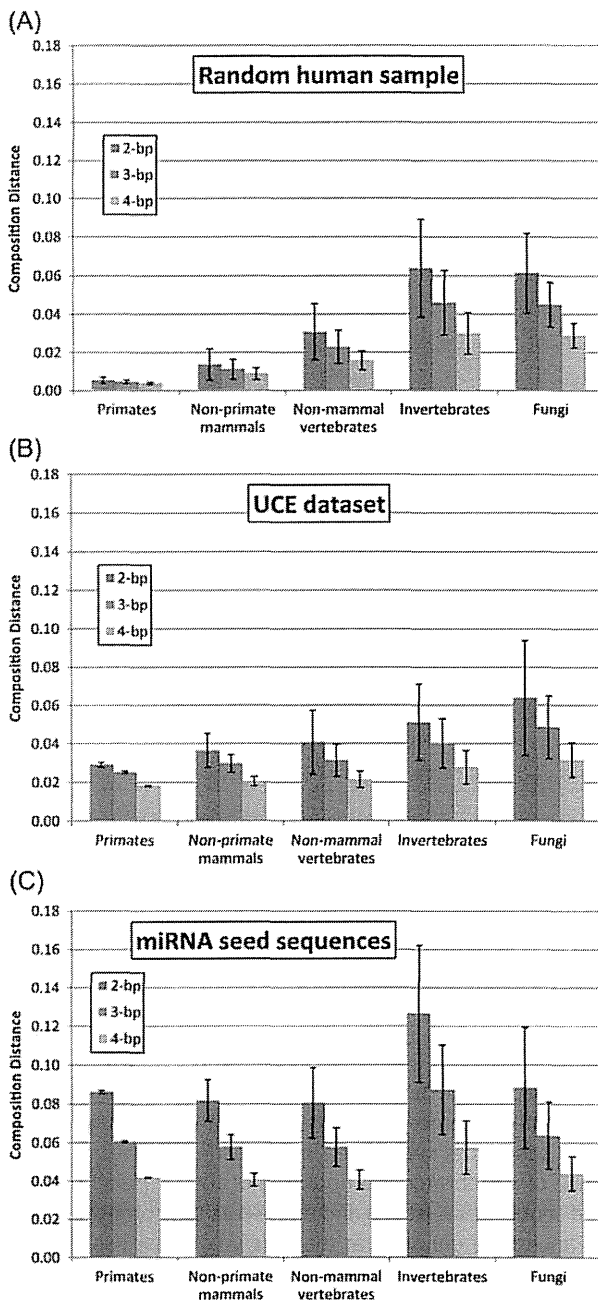


FIG. 5.—Average compositional distance (Euclidean distances between the composition vectors) between sample data sets and complete genomes grouped into four groups. Panels (A, B, and C) correspond to panels (A, B, and C) of figure 4. Standard deviations of the distances are shown for all cases.

lizard, fish, fruit fly, yeast, and *Escherichia coli*. Although the interpretation is difficult, more structure can be seen in the plots of more complex organisms.

Figure 7 shows spacing plots for four random genomes (generated using human genome as a template), the com-

plete actual human genome and the repeat-masked version of the human genome. Repeat-masked is included because complexity is often associated with repetitive sequences. In case of the “Hs Random 1” sequence, discrete elements appear in the figure. Those elements correspond to the groups of DNA words containing different number of GC. With GC contents being the only parameter for constructing the sequence, DNA words with the same number of GC will have exactly same compositional properties, blurred only by randomness of the sequence. In case of “Hs Random 2” similar grouping happens, this time depending on number of CpG each particular word may contain. Going into more complex semirandom sequence, the discreteness becomes less clear, and the plots are getting closer to that for the real human genome. Still significant difference remains between the plots of semirandom and real sequences and very little difference between the plots of repeat-masked and the complete human genome.

Discussion

The GCD provides a convenient measure of relative complexity of various genomes from statistical point of view. A genome is compositionally simple if its composition can be accurately described by a simple model. A set of R values for various word length and models can tell us how complex a particular genome is?

As figure 2 shows, R values become smaller with the increase of model complexity—as expected, a more complex model can describe genome composition more accurately, which results in smaller discrepancy. We observe that, generally speaking, R values are related to the general complexity of the organism. Remarkably, even tetranucleotide compositional models are unable to give good predictions of 5-bp composition in case of complex genomes, particularly for mammals and land vertebrates.

Figure 3 confirms that compositional complexity of a genome is in good correlation with general complexity of the organism. Mammalian genomes are significantly more compositionally complex than genomes of any other organisms. Compositional discrepancy R computed with different composition models seems to be useful as a measure of compositional complexity of the genome.

The extremely rare and extremely abundant sequences, as shown in tables 2 and 3, suggest the possible mechanisms of creating compositional complexity. The most underrepresented 10 bp DNA words (using tetranucleotide composition model) seem to be found on the boundary of mononucleotide repeats, particularly poly-A to poly-T boundary (words 1, 2, 4, 6, 7, 8, 9, 10 in table 2) also poly-A to poly-C (words 3 and 5 in table 2). This means that such boundary is much less common, than suggested by the 4-bp composition.

Among the top overrepresented words, there are poly-A (word 5 in table 3), dinucleotide repeats (words 1 and 2 in

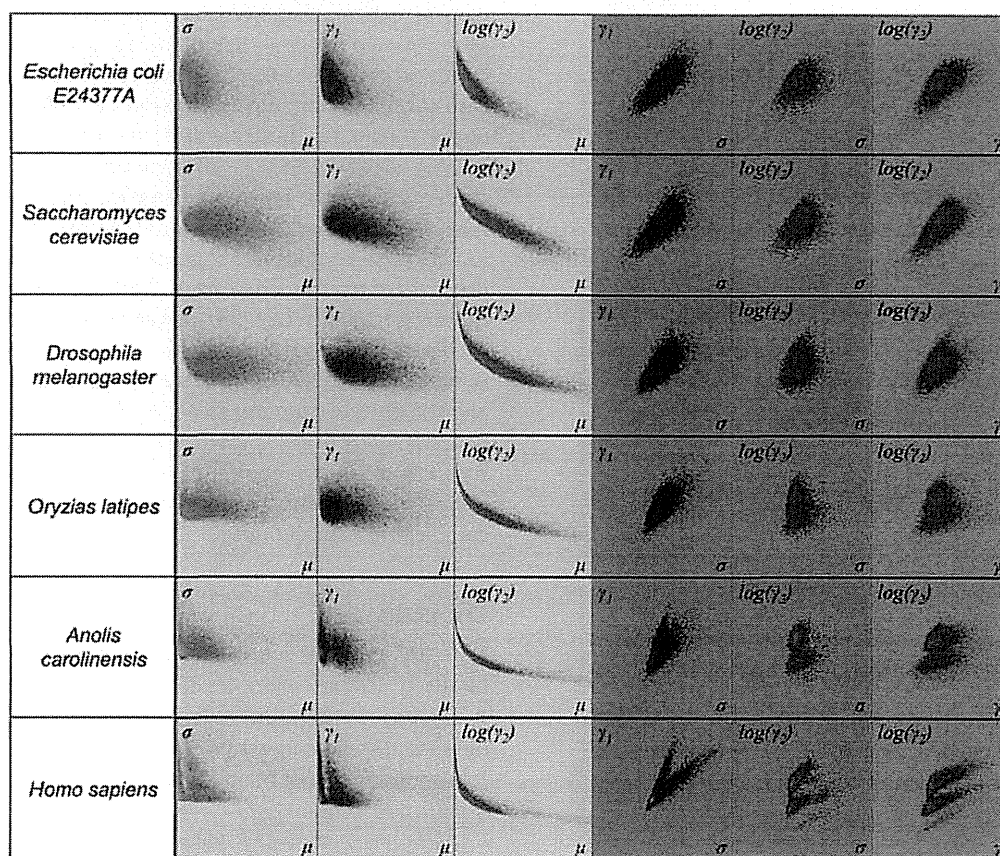


FIG. 6.—Plots of the spacing distribution parameters for six species, based on oligonucleotides of 8 bp. Each row represents one genome. Different columns show plots for different pairs of parameters, from left to right: mean spacing (x axis) versus standard deviation (y axis), mean (x) versus skewness (y), mean (x) versus $\log(\text{kurtosis})$ (y), standard deviation (x) versus skewness (y), standard deviation (x) versus $\log(\text{kurtosis})$ (y), and skewness (x) versus $\log(\text{kurtosis})$ (y). Each dot in the plot represents a particular 8 bp DNA word, so 48 words constitute the data set in each case.

table 3), as well as fragments of sequence “gcctgtaatcc-cagc” (words 3, 4, 6, 7, 8, 9 in table 3), which has about 800,000 occurrences in the human genome compared with the expected number of about 7,000–10,000. This sequence being unusual is already reported by Valle (1993); however, no explanation for the cause was given.

Figure 4 shows the compositional distances between three sequence data sets (human sample, UCE, and miRNA seeds) and vertebrate genomes. Figure 5 summarizes the distances for organism groups, including invertebrates. Although in all three cases, the sequences are contained in the human genome, the compositional distances of those sequences to various genomes show very different pictures. The random sample behaves as expected—the compositional distance is increasing with the increase of divergence from human. However, UCE and miRNA seed data sets show more or less uniform compositional distances from various vertebrate genomes. This suggests that those sequences became conserved before the emergence of mammals. In case

of miRNA seed sequences, the composition distances to all vertebrate genomes are more or less uniform, suggesting those sequences were fixed much earlier than the emergence of vertebrates. Composition of the UCE and miRNA seed sequences is frozen and represents the composition of the ancestral genome, at the time where the fixation occurred. The compositional distance from the current day vertebrates is larger for miRNA seed data set because the miRNA fixation occurred much earlier, so larger compositional distance exists between the ancestral genome and current day genomes. Thus, this allows us to discuss the composition of pre-mammal vertebrate genome (in case of UCE data set) and early animal genome (in case of miRNA seeds).

Oligonucleotide spacing patterns, summarized as sample parameters and displayed as scatterplots (figs. 6 and 7), provide a further interesting view into the compositional complexity. It is apparent that the human genome is very different from the semirandom sequences that imitate only

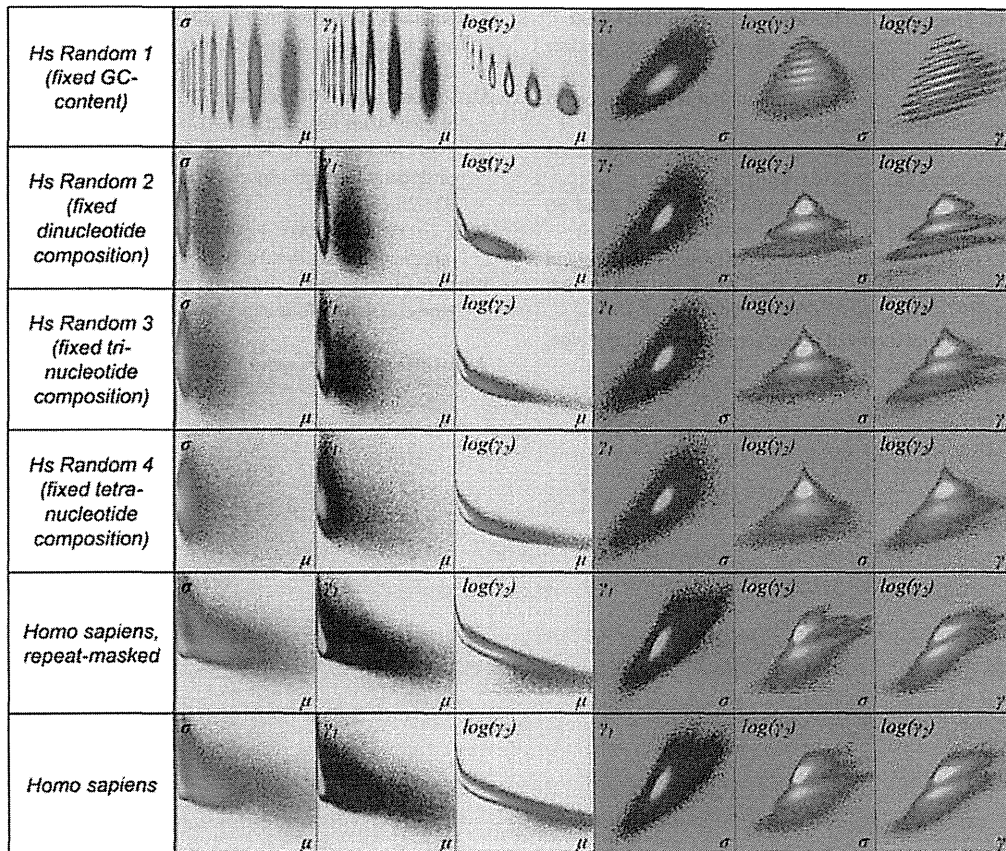


FIG. 7.—Plot of the spacing distribution parameters for semirandom sequences, compared with the real human genome, as well as repeat-masked one. Based on oligonucleotides of 10 bp. Semirandom genomes 1, 2, 3, and 4 are constructed using 1-, 2-, 3-, and 4-bp composition of the actual human genome.

some compositional properties of the actual genome. Often we attribute complexity to the abundant repetitive elements in the vertebrate genome. However, the spacing scatterplots for the repeat-masked human genome looks similar to those of the complete genome and different from those based on the semirandom sequences. It remains to be seen whether the apparent complexity results from the isochore structure of the mammalian genomes (Bernardi et al. 1985), from decaying ancient repeats, or from some other mechanism.

The online GCD provides the means of comparing the compositional complexity of various complete genome and extracting unusual DNA words. The composition parameters computed using five models, as well as histograms, are available. Also spacing patterns, summarized as parameter histograms and 2D scatterplots, are included. In addition that database features a facility for submitting a sequence data set and performing composition analysis and comparison with various complete genomes.

Compositional models that we used in this study only utilize the word frequencies as parameters. The natural next challenge is to design an integrated composition model, which would be based on both frequencies and spacing pat-

terns. Such model would better approximate the genome and thus would allow focusing more closely on the real source of complexity.

Supplementary Material

Supplementary figure 1 and table 1 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, Japan, to N.S. K.K. was additionally supported by the Human Genome Network Project.

Literature Cited

- Abe T, et al. 2003. Informatics for unveiling hidden genome signatures. *Genome Res.* 13:693–702.
- Bejerano G, et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.
- Bernardi G, et al. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.

- Flicek P, et al. 2012. Ensembl 2012. *Nucleic Acids Res.* 40:D84–D90.
- Fujita PA, et al. 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39:D876–D882.
- Gentles AJ, Karlin S. 2001. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* 11:540–546.
- Harris TW. 2010. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* 38:D463–D467.
- Karlin S. 2005. Statistical signals in bioinformatics. *Proc Natl Acad Sci U S A.* 102:13355–13362.
- Karlin S, Mrazek J. 1997. Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci U S A.* 94:10227–10232.
- Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.
- Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39:D152–D157.
- McQuilton P, et al. 2012. FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res.* 40:D706–D714.
- Takahashi M, Kryukov K, Saitou N. 2009. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics* 93:525–533.
- Valle G. 1993. Discover 1: a new program to search for unusually represented DNA motifs. *Nucleic Acids Res.* 21:5152–5156.
- Wheeler DL, et al. 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35:D5–D12.

Associate editor: Eugene Koonin

Case of acute zonal occult outer retinopathy with abnormal pattern visual evoked potentials

Yuzhu Chai¹
Hiroko Yamazaki¹
Kaoru Fujinami²
Kazushige Tsunoda²
Shuichi Yamamoto³

¹Department of Ophthalmology, Kohnodai Hospital, National Center for Global Health and Medicine, Chiba, Japan; ²National Institute of Sensory Organs, Tokyo, Japan; ³Department of Ophthalmology and Visual Science, Chiba University Graduate School of Medicine, Chiba, Japan

Abstract: Electrophysiological and morphological findings were studied in a case of acute zonal occult outer retinopathy (AZOOR) showing abnormal pattern visual evoked potentials (VEPs) at the onset and significant functional recovery in the natural course. A 21-year-old woman presented with acute onset of photopsia and a large scotoma in the right eye of 2 weeks duration. Her visual acuity was 20/20 in both eyes with no ophthalmoscopic and fluorescein angiographic abnormalities. However, a relative afferent pupillary defect and an enlarged blind spot were found in the right eye. The pattern VEPs were severely reduced when the right eye was stimulated. The amplitudes of both rod and cone full-field electroretinographics (ERGs) were reduced in the right eye. The amplitudes of the multifocal ERGs were reduced in the area of the enlarged blind spot. Irregularities in the inner segment/outer segment (IS/OS) line of the photoreceptors were observed over the nasal fovea by optical coherence tomography (OCT). The patient was followed without treatment. The enlarged blind spot disappeared in 3 months after the onset. At 5 months, reappearance of the IS/OS line was detected by OCT. At 6 months, the P₁₀₀ recovered to normal values. At 1 year, the reduced full-field ERGs were almost normal size and the multifocal ERGs in the area corresponding to the enlarged blind spot were also improved. ERG findings are crucial for differentiating AZOOR from retrobulbar neuritis, especially in patients with abnormal pattern VEPs. The pattern VEPs, full-field ERGs, multifocal ERGs, and OCT images can be abnormal in the early phase of AZOOR, but they can all improve during the natural course.

Keywords: AZOOR, pattern VEP, full-field ERG, multifocal ERG, OCT

Introduction

Acute zonal occult outer retinopathy (AZOOR), first reported in 1993 by Gass,¹ is characterized by an acute onset of photopsia, scotoma, minimal funduscopic changes, and electroretinographic (ERG) abnormalities affecting one or both eyes. The presence of abnormal ERGs is important for the diagnosis of AZOOR, and the ERG findings suggest a dysfunction of the photoreceptors.¹⁻⁵ Recent optical coherence tomographic (OCT) studies have shown morphological alterations of the photoreceptors.⁶⁻⁹

Although a viral or autoimmune etiology has been suspected, no cause is readily identifiable in this group of generally healthy patients. Some patients with AZOOR have been misdiagnosed with optic nerve disorders because they had an afferent pupillary defect, a scotoma, and no obvious fundus abnormalities.¹⁻³

It has been reported that the natural course of AZOOR is varied,¹⁻³ although there is still no known treatment. There are few detailed reports about AZOOR patients who showed subjective and objective improvements in their visual function.

Correspondence: Yuzhu Chai
Department of Ophthalmology, Kohnodai
Hospital, National Center for
Global Health and Medicine,
1-7-1 Kohnodai Ichikawa City,
Chiba 2728516, Japan
Tel +81 47 372 3501
Fax +81 47 372 1858
Email chai_yzhu@yahoocorp.jp

We report a case of AZOOR, showing abnormal pattern visual evoked potentials (VEPs) at the onset and significant functional recovery in the natural course.

Case report

A healthy 21-year-old woman reported that she had a sudden onset of photopsia and a large scotoma in the right eye on August 14, 2008. She visited her ophthalmologist on August 15, and the initial examination showed no abnormalities of the fundus in both eyes but a large scotoma was detected by static perimetry in the right eye. She was referred to us for further examination on August 29. She reported that she had no systemic problems and was not taking any medications. Her best-corrected visual acuity was 20/20 bilaterally, and the refractive error was -9.5 diopters in the right eye and -8.5 diopters in the left eye. However, she had a relatively afferent pupillary defect in the right eye. The results of ophthalmoscopy (Figure 1A), fluorescein angiography (Figure 1B), blood screening, and brain magnetic resonance imaging (MRI) were within normal limits. Static perimetry with the Humphrey field analyzer showed an enlarged blind spot in the right eye (Figure 1C). The pattern VEPs elicited

by transient and steady-state stimuli to the right eye were severely reduced (Figure 2). She was suspected of having retrobulbar neuritis at this point of time. VEPs were recorded by Nihon Kohden MEB-2200 Neuropack (Tokyo, Japan) with the active electrode placed at Oz. The reference electrode was located at the right earlobe and the ground electrode at the left earlobe. The visual stimulus was a black and white check board pattern generated on a television monitor. The check size was 20 minutes of arc. The contrast was 80%, and the mean luminance was kept at 50 cd/m^2 . The stimulus field of the pattern was 7×11 degrees. The pattern was reversed at three reversals per second for transient VEP, and 10 reversals per second for steady-state VEP. The electrodes were connected to a preamplifier with a bandpass of 1 to 100 Hz, and for each measurement, 100 responses were averaged. The patient fixed on a point in the center of the pattern monocularly from an observing distance of 170 cm, with an undilated pupil under full refractive correction.

To eliminate the possibility that the patient had AZOOR, full-field ERGs and multifocal ERGs were recorded. The amplitudes of the rod and cone responses in the right eye were reduced to about 50% of those in

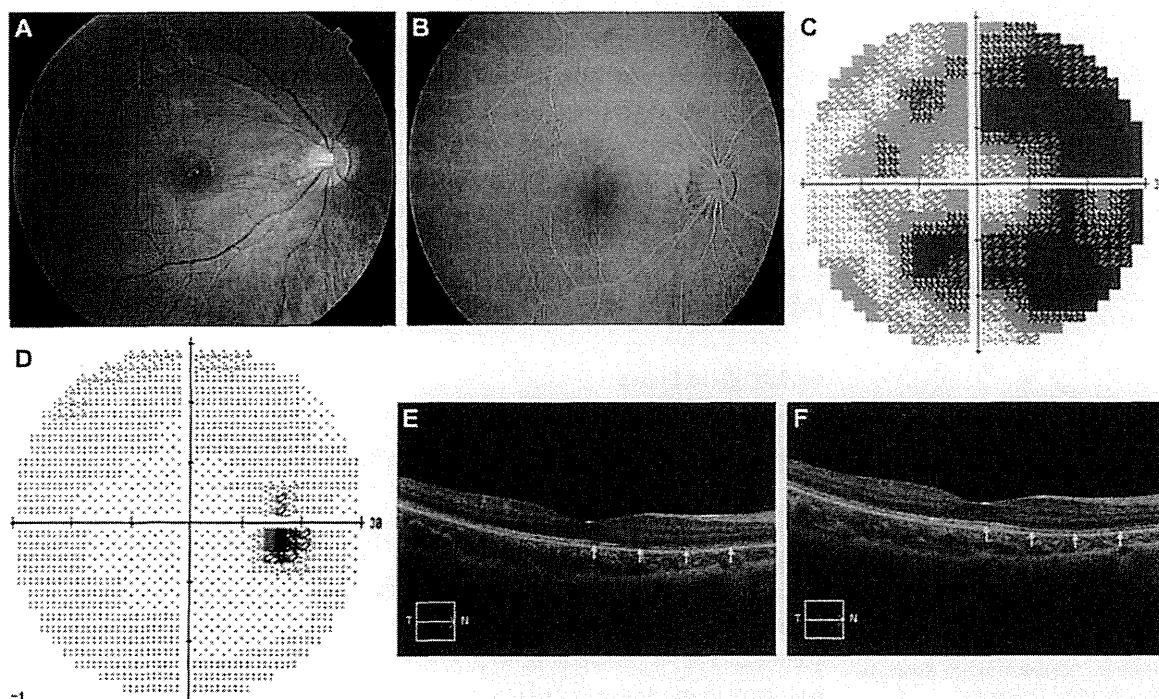


Figure 1 Fundus photograph, fluorescein angiographic image, Humphrey static perimetry, and Optical coherence tomography (OCT) of the right eye. (A) Fundus photograph at the onset showing that the retina was normal. (B) Fluorescein angiography at the onset showing normal vascular pattern and no leakage. (C) Humphrey static perimetry at the onset showing enlarged blind spot (30-2 strategy MD -21.97 dB). (D) Humphrey static perimetry 3 months after the onset showing marginally enlarged blind spot (30-2 strategy MD -1.90 dB). (E) OCT image at the onset showing irregular inner segment/outer segment (IS/OS) line over the nasal fovea. (F) OCT image 5 months after the onset showing reappearance of IS/OS line over the nasal fovea.

Abbreviation: MD, mean defect.

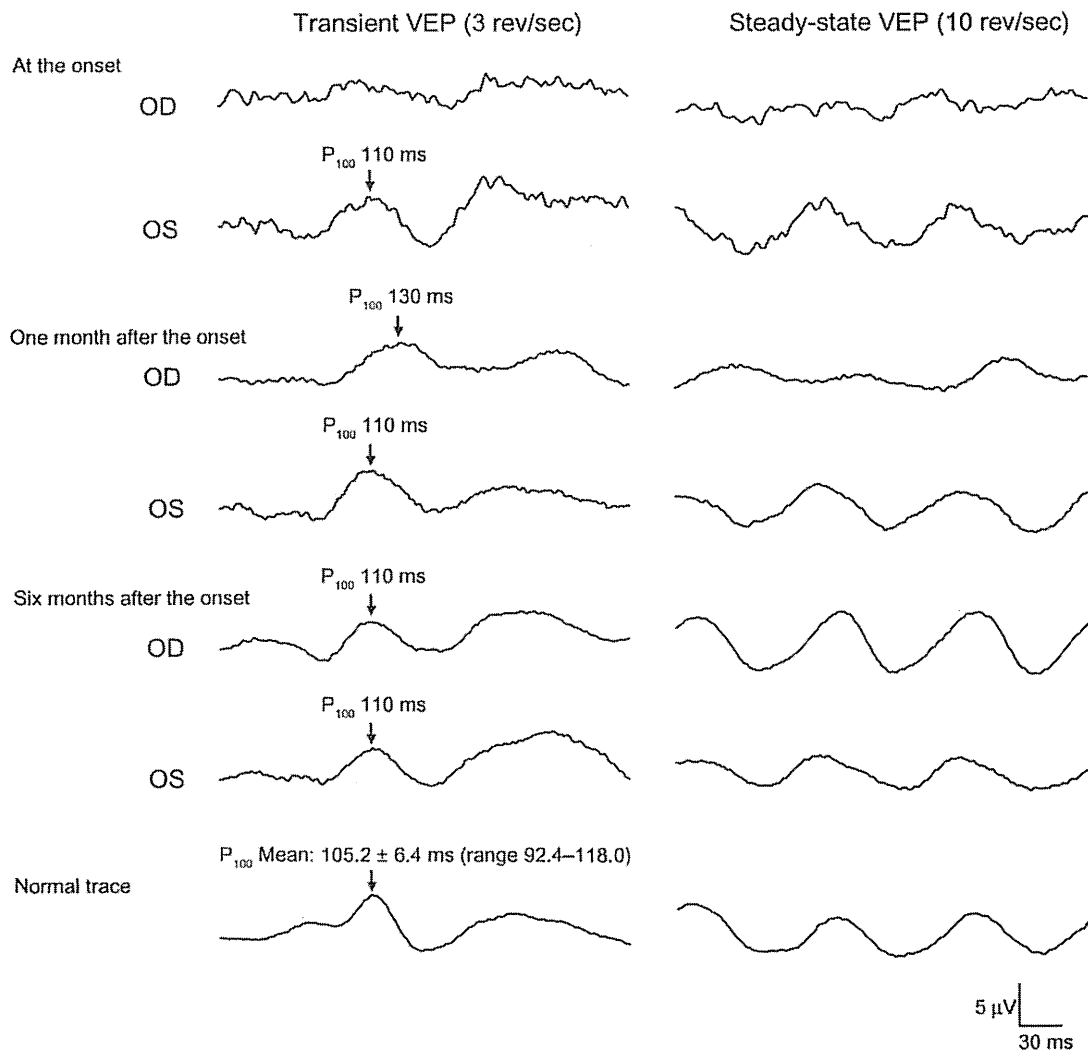


Figure 2 Pattern visual evoked potentials (VEPs). At the onset, pattern VEPs are severely reduced in the right eye. One month after the onset, the P₁₀₀ component of the pattern VEPs has a prolonged latency of 130 milliseconds. Six months after the onset, the P₁₀₀ recovers to 110 milliseconds. **Abbreviations:** OD, right eye; OS, left eye.

the left eye (Figure 3). The multifocal ERGs recorded from the area of the enlarged blind spot were reduced (Figure 4). Irregularities in the inner segment/outer segment (IS/OS) line of the photoreceptors over the nasal fovea were observed by Fourier-domain OCT (HD-OCT; Carl Zeiss Meditec, Oberkochen, Germany) (Figure 1E). Serum anti-recoverin was negative. She was diagnosed with AZOOR from these findings. Full-field ERGs were recorded by Nihon Kohden Neuropack 9400, following dilation of the pupils and 30 minutes of dark adaptation. A contact lens electrode was used. The reference electrode was placed at the forehead and the ground electrode at the earlobe. The flash strength was 0.01 cd s m⁻² for rod response, 80.0 cd s m⁻² for combined rod–cone response and

oscillatory potentials, and 3.0 cd s m⁻² for cone response and flicker. Light adaptation and background luminance was 25 cd/m². Responses were amplified with a bandpass of 0.2 to 1000 Hz. First-order Kernel multifocal ERGs were recorded with the Visual Evoked Response Imaging System (VERIS science 4.1; EDI, San Mateo, CA). A Burian-Allen bipolar contact lens electrode was used. The visual stimuli consisted of 61 and 103 hexagonal elements with an overall subtense of approximately 60°. The luminance of each hexagon was independently modulated between black (3.5 cd/m²) and white (138.0 cd/m²) according to a binary M-sequence at 75 Hz. The surround luminance was 70.8 cd/m².

The patient was followed without any treatment, and more comprehensive examinations were made in her follow-up

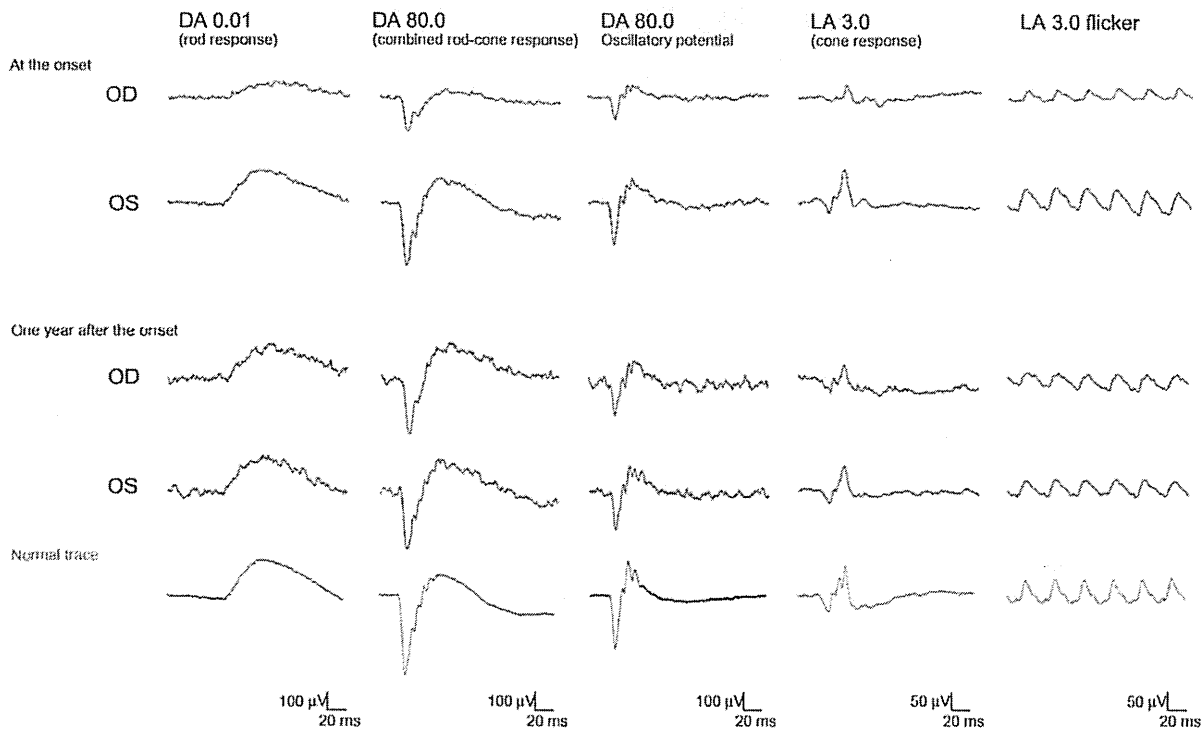


Figure 3 Full-field electroretinographics (ERGs). At the onset, the amplitudes of the rod and cone responses in the right eye are reduced to about 50% of those in the left eye. At the 1 year follow-up examination, the reduced full-field ERGs are improved to be approximately the same amplitudes as those from the left eye. Abbreviations: DA, dark adapted; LA, light adapted; OD, right eye; OS, left eye.

examinations. The enlarged blind spot in static perimetry disappeared 3 months after our initial examination (Figure 1D). The P_{100} component of the pattern VEPs had a prolonged latency of 130 milliseconds 1 month later. At 6 months, the P_{100} recovered to 110 milliseconds (mean: 105.2 ± 6.4 milliseconds; normal range: 92.4–118.0 milliseconds) (Figure 2). At 5 months, reappearance of the IS/OS line was detected by OCT over the nasal fovea (Figure 1F). At the 1 year follow-up examination, her best-corrected visual acuity was 20/20 in both eyes. The reduced full-field ERGs were improved to be approximately the same amplitudes as those from the left eye (Figure 3). A mild improvement of the reduced multifocal ERGs was also observed (Figure 4). The other findings had not worsened.

Discussion

In 1993, Gass introduced AZOOR to describe a previously unrecognized syndrome occurring predominantly in young females. In his original series of 13 patients, affected individuals typically presented with acute onset of photopsia, scotoma, minimal fundoscopic changes, and ERG abnormalities affecting one or both eyes.¹ The presence of abnormal ERGs is essential for the diagnosis of AZOOR. Gass et al

reported that electroretinographic amplitudes were depressed in all 90 affected eyes.² Jacobson et al reported 24 patients with AZOOR showing abnormal ERGs. Interocular asymmetry was a prominent feature.⁴ Francis et al reported that electrophysiology demonstrated a consistent pattern of dysfunction both at the photoreceptor/retinal pigment epithelial complex but also at inner retinal levels in 28 patients with AZOOR.⁵

Our case presented with a sudden onset of photopsia and a large scotoma in the right eye at the onset. The normal fundus, large scotoma, and the afferent pupillary defect made us suspect retrobulbar neuritis at first. But the absence of retrobulbar pain at rest or on eye movement, which is a common symptom in retrobulbar neuritis,¹⁰ and the presence of photopsia, which is a typical symptom of AZOOR,^{1–5} made us suspect AZOOR at the same time. Further examinations were taken to make a definite diagnosis.

Our patient had an extinguished P_{100} component of the pattern VEPs at the onset, although her visual acuity was 20/20. Fluorescein angiography showed normal vascular pattern and no leakage. MRI showed no changes in the optic nerve. Strikingly, in full-field ERGs, the amplitudes of both rod and cone responses were reduced in the right eye. The amplitudes of the multifocal ERGs were reduced in the area

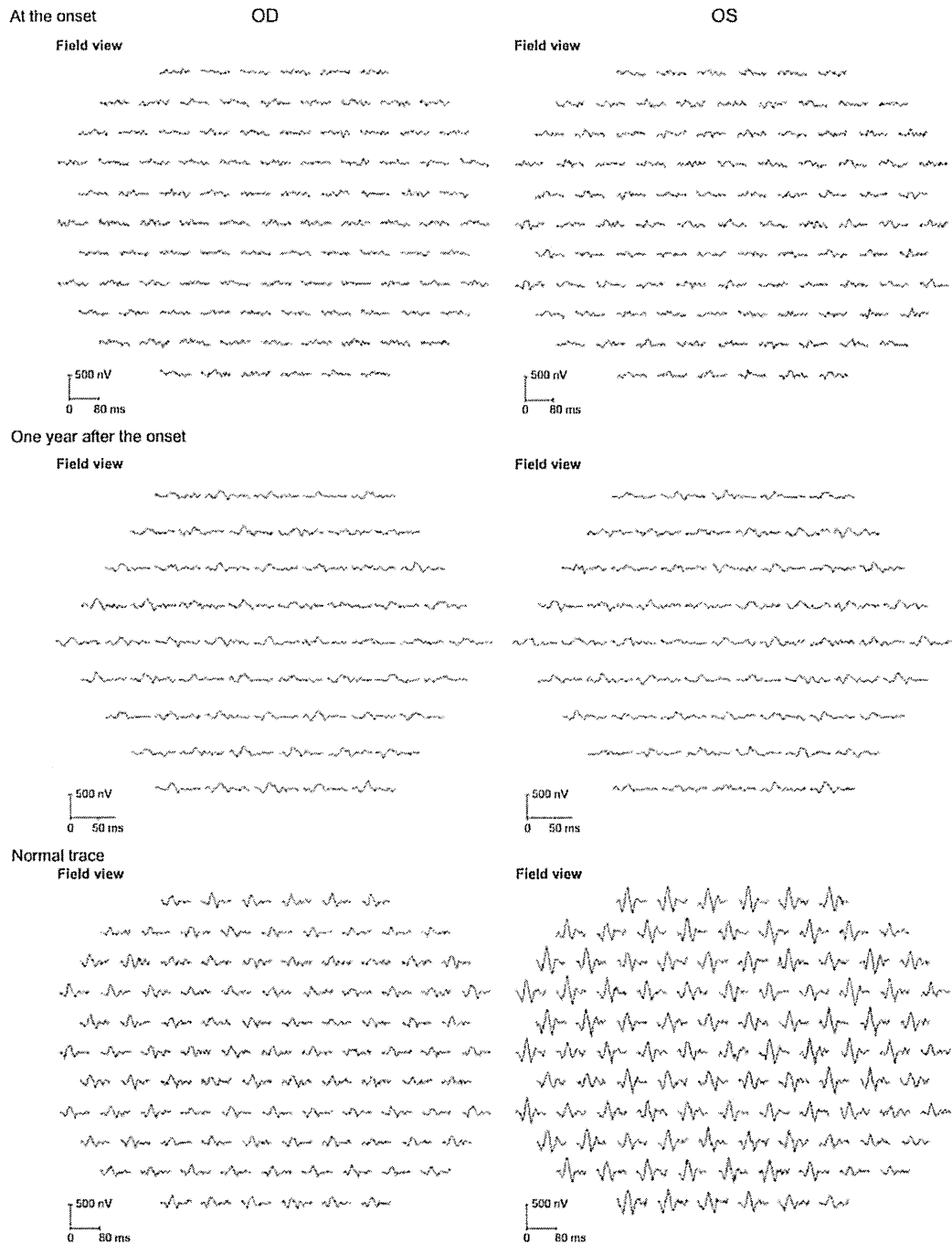


Figure 4 Multifocal electroretinographics (ERGs). At the onset, the multifocal ERGs are reduced in the right eye. At the 1 year follow-up examination, the multifocal ERGs have improved but are still reduced especially from the temporal retina. Shorter duration protocol with 61 hexagonal elements was used for the latest recording, because acceptable responses to 103 hexagonal elements could not be obtained due to fatigue of the patient during the recording.
Abbreviations: OD, right eye; OS, left eye.

of the enlarged blind spot. She was diagnosed with AZOOR because of ERG abnormalities. Serum anti-recoverin was negative, which was helpful to reduce the possibility that she had cancer-associated retinopathy. Irregularities in the

IS/OS line of the photoreceptors were observed by OCT. The findings of OCT helped make the diagnosis.

It is known that a delayed latency of the pattern VEPs is not a specific sign of optic neuropathy because it is also found

in eyes with retinal diseases.^{11–15} However, there were few reports of AZOOR showing abnormal VEPs. Gass reported that 80% of AZOOR patients had normal VEPs, and only one patient with poor visual acuity of 20/70 in the right eye and 20/300 in the left eye had abnormal VEPs.¹ Takai et al reported five AZOOR patients who had no delay in the VEPs.⁷ Patients of AZOOR who had normal pattern VEPs with normal visual acuity have been reported.^{16,17} Although significant abnormalities of the VEPs have rarely been reported in eyes with AZOOR, our case showed that AZOOR can be associated with significantly delayed VEPs. The mechanisms for the altered pattern VEPs without a decrease in the visual acuity was not determined in our case. The delayed VEP might be explained by reduced macular sensitivities. However, the possibility of inner retina or optic nerve involvement cannot be completely excluded.

Gass et al reported that in the presence of normal fundi, the most frequent misdiagnosis was retrobulbar neuritis, and there was a median of 17 months delay in the diagnosis of AZOOR.² In addition to the afferent pupillary defect, abnormal VEPs could mislead the ophthalmologist to a diagnosis of optic nerve disease. This case report serves to alert the ophthalmologist to consider the diagnosis of AZOOR and consider recording ERGs in individuals presenting with unexplained scotoma, particularly where photopsia are a prominent feature.

It has been reported that the natural course of AZOOR is varied.^{2,3} Gass et al reported that visual field loss stabilized within 6 months in 37 patients (72%), progressed stepwise in two patients (4%), and partly improved in 12 patients (24%).² Degenerations of the photoreceptor outer segment have been detected by OCT at the convalescent stage.^{6–9} There are few detailed reports about the clinical course of AZOOR patients who show some recovery. Yasuda et al reported a case with mild improvement of the multifocal ERGs, but morphological changes were not demonstrated in the report.¹⁸ Spaide et al reported restoration of the IS/OS line in the areas of improved visual field, but electrophysiological alterations of these patients were not demonstrated in the report.⁹

Our case had a recovery of retinal function as assessed by not only visual field tests, but also full-field ERGs and multifocal ERGs. Morphological improvements were confirmed by OCT in parallel. Although it is difficult to compare the order or degree of improvement of each parameter, our case demonstrated that subjective and objective functional recovery could occur in the eyes with AZOOR. The irregular IS/OS line at the onset might have reflected incomplete loss

of photoreceptor. In this case, the incomplete impairment of photoreceptor might have been associated with the visual function recovery. Further studies are needed.

Conclusion

We reported a case of AZOOR showing profoundly abnormal pattern VEPs at the onset, and significant functional recovery in the natural course. ERG findings are crucial for differentiating AZOOR from retrobulbar neuritis, especially in patients with abnormal pattern VEPs. The pattern VEPs, full-field ERGs, multifocal ERGs, and OCT images can be abnormal in the early phase of AZOOR, but they can all improve during the natural course. Further studies and long-term follow-up are needed to better understand this disorder, and these findings will hopefully reduce the number of misdiagnoses and unnecessary treatments.

Disclosure

The authors report no conflicts of interest in this work.

References

- Gass JDM. Acute zonal occult outer retinopathy. *J Clin Neuroophthalmol*. 1993;13(2):79–97.
- Gass JD, Agarwal A, Scott IU. Acute zonal occult outer retinopathy: a long-term follow-up study. *Am J Ophthalmol*. 2002;134(3):329–339.
- Monson DM, Smith JR. Acute zonal occult outer retinopathy. *Surv ophthalmol*. 2011;56(1):23–35.
- Jacobson SG, Morales DS, Sun XK, et al. Pattern of retinal dysfunction in acute zonal occult outer retinopathy. *Ophthalmology*. 1995;102(8):1187–1198.
- Francis PJ, Marinescu A, Fitzke FW, Bird AC, Holder GE. Acute zonal occult outer retinopathy: towards a set of diagnostic criteria. *Br J Ophthalmol*. 2005;89(1):70–73.
- Fujiwara T, Imamura Y, Giovinazzo VJ, Spaide RF. Fundus autofluorescence and optical coherence tomographic findings in acute zonal occult outer retinopathy. *Retina*. 2010;30(8):1206–1216.
- Takai Y, Ishiko S, Kagokawa H, Fukui K, Takahashi A, Yoshida A. Morphological study of acute zonal occult outer retinopathy by multiplanar optical coherence tomography. *Acta Ophthalmol*. 2009;87(4):408–418.
- Li D, Kishi S. Loss of photoreceptor outer segment in acute zonal occult outer retinopathy. *Arch Ophthalmol*. 2007;125(9):1194–1200.
- Spaide RF, Koizumi H, Freund KB. Photoreceptor outer segment abnormalities as a cause of blind spot enlargement in acute zonal occult outer retinopathy-complex diseases. *Am J Ophthalmol*. 2008;146(1):111–120.
- Optic Neuritis Study Group. The clinical profile of acute optic neuritis. Experience of the Optic Neuritis Treatment Trial. *Arch Ophthalmol*. 1991;109(12):1673–1678.
- Lennerstrand G. Delayed visual evoked cortical potentials in retinal disease. *Acta Ophthalmol*. 1982;60(4):497–504.
- Bass SJ, Sherman J, Bodis-Wollner I, Nath S. Visual evoked potentials in macular disease. *Invest Ophthalmol Vis Sci*. 1985;26(8):1071–1074.
- Negishi C, Takasoh M, Fujimoto N, Tsuyama Y, Adachi-Usami E. Visual evoked potentials in relation to visual acuity in macular disease. *Acta Ophthalmol Scand*. 2001;79(3):271–276.

14. Shimada Y, Adachi-Usami E, Murayama K. How are macular changes reflected in pattern visually evoked cortical potentials? *Acta Ophthalmol Scand*. 1997;75(3):277–280.
15. Holder GE. Electrophysiological assessment of optic nerve disease. *Eye (Lond)*. 2004;18(11):1133–1143.
16. Lee AG, Prager TC. Acute zonal occult outer retinopathy. *Acta Ophthalmol Scand*. 1996;74(1):93–95.
17. Arai M, Nao-i N, Sawada A, Hayashida T. Multifocal electroretinogram indicates visual field loss in acute zonal occult outer retinopathy. *Am J Ophthalmol*. 1998;126(3):466–469.
18. Yasuda K, Shimura M, Noro M, Nakazawa M, Tamai M. Clinical course of acute retinal zonal occult outer retinopathy in visual field and multifocal electroretinogram. *Br J Ophthalmol*. 1999;83(9):1089–1090.

Clinical Ophthalmology

Publish your work in this journal

Clinical Ophthalmology is an international, peer-reviewed journal covering all subspecialties within ophthalmology. Key topics include: Optometry; Visual science; Pharmacology and drug therapy in eye diseases; Basic Sciences; Primary and Secondary eye care; Patient Safety and Quality of Care Improvements. This journal is indexed on

Submit your manuscript here: <http://www.dovepress.com/clinical-ophthalmology-journal>

Dovepress

PubMed Central and CAS, and is the official journal of The Society of Clinical Ophthalmology (SCO). The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.