was expressed in rod and cone photoreceptors [10]. Because the amino acid sequence of human RP1L1 is only 39% identical to that of the mouse, researchers have suggested that the primate RP1L1 might have different functional roles in the cone photoreceptors of the retina than that of other species [10].

We have identified a new mutation in the *RP1L1* gene in a patient with clinical characteristics of OMD: abnormal focal macular ERGs and blurring of the IS/OS junction and the disappearance of the COST line in SD-OCT images. The fundus examination, fluorescein angiograms, and full-field ERGs were normal in this case. The mutation is an amino acid substitution of cysteine for serine in exon 4 of the *RP1L1* gene that has not been reported in the Single Nucleotide Polymorphism (SNP) database and was also not detected in any of the 294 normal control alleles. The serine at position 1199 is well conserved among the RP1L1 family in other species. Four out of five computational assessment tools (PolyPhen-2, SIFT, PMut, Align GVGD, and MutationTaster) predicted that this mutation is damaging to the protein function. A segregation of the mutation and the disease was found in one affected member and one unaffected member of the same family.

METHODS

The protocol conformed to the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board of the Nippon Medical School and the ethics review committees of the National Hospital Organization Tokyo Medical Center. Written informed consent was obtained from all patients after the nature and possible consequences of the study were explained.

Clinical studies: The ophthalmological examinations included best-corrected visual acuity (BCVA) measurements, refraction, slit-lamp biomicroscopy, ophthalmoscopy, fundus photography, perimetry, SD-OCT, fluorescein angiography (FA), full-field ERGs, focal ERGs, and mfERGs. The visual fields were determined with the Goldman perimetry and the Humphrey Visual Field Analyzer (model 745i; Carl Zeiss Meditec, Inc., Dublin, CA). The Swedish interactive threshold algorithm standard strategy was used with program 30–2 of the Humphrey Visual Field Analyzer. The OCT images were recorded using a SD-OCT (Carl Zeiss Meditec) on this patient and normal controls. Full-field scotopic and photopic ERGs were recorded using an extended testing protocol incorporating the International Society for Clinical Electrophysiology of Vision standards [14]. The full-field ERGs were used to assess retinal function under scotopic and photopic states.

Focal macular electroretinograms: Focal macular ERGs were recorded with a commercial Focal Macular ERG system (ER80; Kowa Company, Tokyo, Japan, and PuREC; Mayo Company, Nagoya, Japan) using a bipolar contact lens electrode (MY type Electrode; Mayo Company). The stimulus and background lights were integrated into an infrared fundus camera [15-17]. The size of the stimulus spot was 15° in diameter and was placed on the macula by observing the infrared image of the retina on a monitor. The white stimulus and background illumination were generated by light-emitting diodes that had maximal spectral emissions at 440 to 460 nm and 550 to 580 nm, respectively. The luminances of the stimuli and background were 115.7 cd/m^2 and 8.0 cd/m^2. The duration of the stimulation was 100 ms. The responses were amplified and filtered with digital band pass filters from 5 to 200 Hz. Three hundred responses were summed with a stimulus frequency of 5 Hz. The a-wave, b-wave, d-wave, and oscillatory potentials (OPs) were evaluated.

Multifocal electroretinograms: The mfERGs were recorded using a commercial mfERG system (LE-4000, Tomey, Nagoya, Japan; LE4100; Mayo Company, Inazawa, Japan). This system uses basically the same technology as the Visual Evoked Response Imaging System [18]. The visual stimuli consisted of 37 hexagonal elements with an overall subtense of approximately 50°. The luminance of each hexagon was independently modulated between black (2.47 cd/m^2) and white (200.4 cd/m^2) according to a binary m-sequence at 75 Hz. The surround luminance was set at 75.4 cd/m^2.

Mutation analysis: Blood samples were collected from the patient, and genomic DNA was isolated from peripheral white blood cells using a blood DNA isolation kit (NucleoSpin Blood XL; Macherey Nagel, Düren, Germany). The DNA was used as the template to amplify the *RP1L1* gene. Coding regions and flanking introns of the *RP1L1* gene were amplified with polymerase chain reaction (PCR) using primers produced by Greiner Bio-One (Tokyo, Japan). Primer sequences are listed in Table 1. The PCR products were purified (ExoSAP-IT; USB Corp., Cleveland, OH) and were used as the template for sequencing. Both strands were sequenced on an automated sequencer (Bio Matrix Research; Chiba, Japan). The identified mutations and coding polymorphisms were assayed in 294 control chromosomes from 147 healthy Japanese individuals with direct sequencing except the length polymorphism region. To sequence the length polymorphism region of the *RP1L1* gene, the amplified PCR products were subcloned into the StrataClone PCR cloning vector (Stratagene; La Jolla, CA). At least five cloned products from this case and 20 control individuals were sequenced on an automated sequencer.

Computational assessment of missense mutation: The effect of a missense mutation on the encoded protein was predicted with the PolyPhen-2, SIFT, PMut, Align GVGD, and MutationTaster online tools [19-24]. PolyPhen-2 is a software tool that predicts the possible impact of amino acid substitutions on the structure and function of human proteins using straightforward physical and evolutionary comparative

Table 1. Sequences of oligonucleotide primers used in this study and PCR product size.

Fragment name	Forward primer (5′-3′)	Reverse primer (5′-3′)	Product size (bp)
RP1L1–2A	GAGACAGGAAATGCCAATCC	CCGCAACTGCTGAGCAGTGG	471
RP1L1–2B	CCTCTGCTCTGATAAGAAGC	TCCATGTGAGTATTTTGACC	373
RP1L1–3	CCTCCAGCTAGTGATAGAGG	GATTGACAGTACTGAGAAGG	498
RP1L1–4A	TTCCTTTATCCTGATGCTGC	CCAAAGACTTCCCTGCATCC	509
RP1L1–4B	TGTGGGAGGGCTACCCTTGG	GCTGACGAGTCCGAAGAAGC	508
RP1L1–4C	CTATGCATAGATGGAGCAGG	GTTACAGAGGAGTCCAGTGG	536
RP1L1–4D	CAATGTCCTCACCCAGCAGC	TCCAACCTGCAGAACCAAGG	494
RP1L1–4E	GACTCCTGCTCAAAATCTGG	GGACACCCTCTCCTGATTGG	784
RP1L1–4F	GGACAGCAGTCCCTGGAAGG	ACTGCACCGCCTCTTCTTGC	937
RP1L1–4G	AAACACAGTGCAAGAAGAGG	AGGCTCAAGCTGGGAGCCACTCTGC	variable
RP1L1–4H	GGGAAAGGCTCCCAGGAAGATGACC	TTCTGCACCTTCTGACTCTGGCTGG	1470
RP1L1–4I	CACAGAGGAACCCACAGAGC	GAGAAGGCCGAGAGGTTTCG	522
RP1L1–4J	CAAGAGAGAGCTCCAGAAGC	TCTGTTGAGTCTCTGGCTCC	547
RP1L1–4K	GACAAAGATCCCAAACTCGG	AGAGTCAGAAGATGTAGAGG	836
RP1L1–4L	TGAAGGGGAGATGCAAGAGG	GAGTGGGCCTGTCCTCAGGGACTGG	821
RP1L1–4M	AGGCTTCTGAAAGCAGCAGC	ACTATGGACATCTCCAGTGG	517

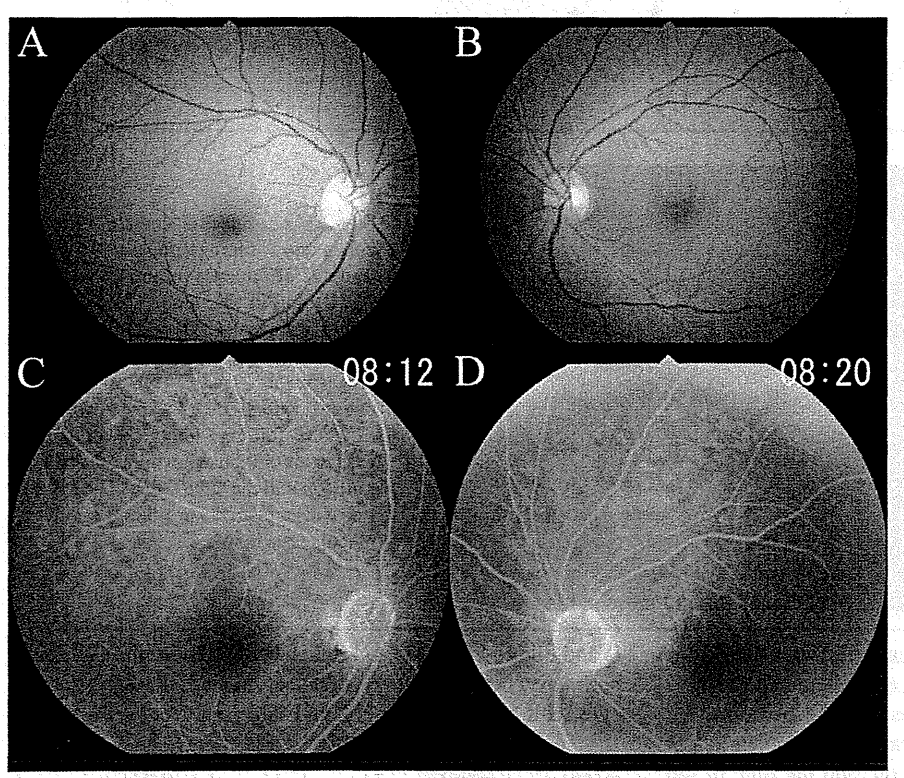

Figure 1. Fundus photographs (**A**, **B**) and fluorescein angiograms (**C**, **D**) of this case showing no abnormal findings.

considerations. SIFT generates multiple alignments of the sequence over different species to look at the conserved sequences of a gene; it assesses the conserved amino acid positions and analyzes the effect of missense changes on the conserved structure of proteins over the course of evolution. The SIFT tool assigns a score to the mutations, and a score of <0.05 is considered potentially damaging. PMut is software aimed at annotating and predicting pathological mutations. Align GVGD combines the biophysical characteristics of amino acids and protein multiple sequence alignments to predict where missense substitutions in genes of interest fall in a spectrum from enriched deleterious to enriched neutral. MutationTaster evaluates the disease-causing potential of sequence alterations.

Statistical analysis: We calculated the 95% confidence intervals (CI) of the results of the focal macular ERGs of normal controls. There were 25 men and 21 women whose age ranged from 23 to 60 years (mean, 38.04±8.33 years) in

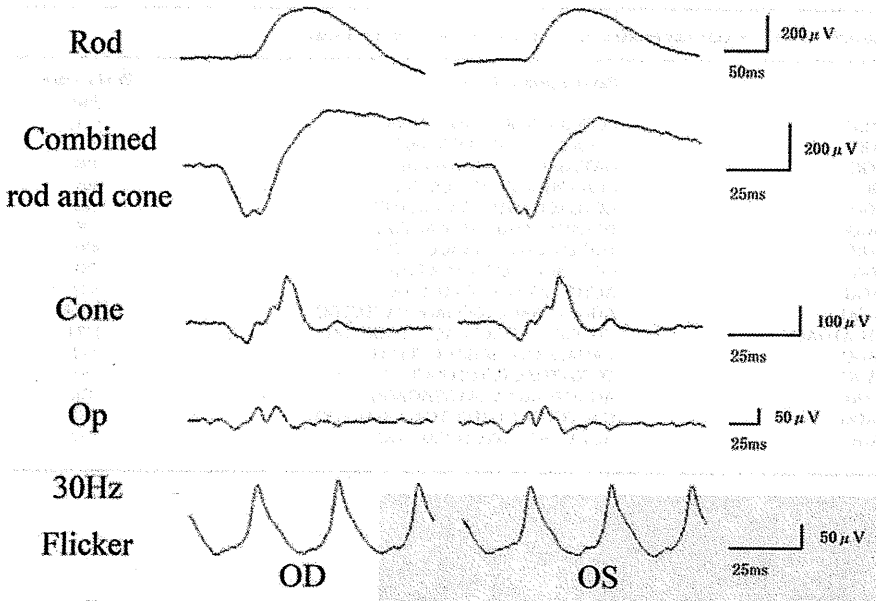

Figure 2. Full-field electroretinograms (ERGs) recorded according to the International Society for Clinical Electrophysiology of Vision (ISCEV) standards protocol in this case. The rod, combined rod-cone, cone, oscillatory potentials, and 30-Hz flicker full-field ERGs are shown. The results of full-field ERGs are within the normal limits in this case.

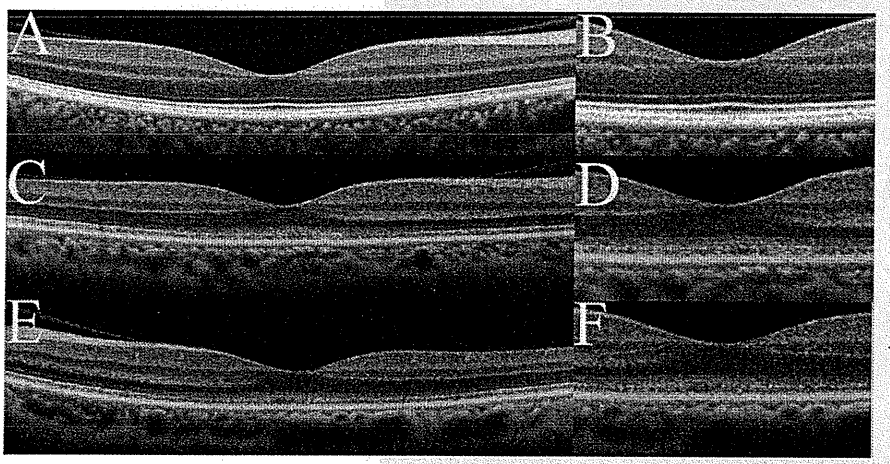

Figure 3. Spectral-domain optical coherence tomography (SD-OCT) findings of the eyes in normal controls (**A**, **B**) and in this case (**C-F**). Images from right eyes (**C**, **D**) and left eyes (**E**, **F**) are shown. Images at lower magnification (**A**, **C**, **E**) and higher magnification (**B**, **D**, **F**) are shown. The SD-OCT findings for the eyes in this case show obvious blurring of the IS/OS junction and the COST line. The COST line disappeared in the peripheral macula area in this case.

this control group. We recorded focal macular ERGs from either of the eyes of normal controls and calculated the 95% CI of the amplitudes of the a-waves and the b-waves, the implicit time of the a-waves and b-waves, the potentials at 70 ms after the stimulus was turned on, and the time of the recovery of the b-wave to the baseline.

RESULTS

Case report: A 52-year-old woman complained of a gradual decrease in vision in both eyes during the past two to three years. Family history revealed no other members with any eye diseases, including her parents who were deceased. Her BCVAs were 20/63 in the right eye and 20/50 in the left eye. The fundus examination, fluorescein angiography, and full-field ERG results were within the normal limits (Figure 1A-D and Figure 2). The visual fields were full with the Goldman perimetry, but a relative central scotoma was detected in both eyes with the Humphrey Visual Field Analyzer.

Spectral domain optical coherence tomography: The SD-OCT images of this case showed a blurred IS/OS junction and COST line at the foveal center (Figure 3D,F). In the peripheral macula area, the COST line was absent, and only the blurred IS/OS junction was visible in this case (Figure 3C,E).

Focal macular electroretinograms and multifocal electroretinograms: A severe reduction in the a-waves of the focal macular ERGs was found in this case (Figure 4). Although the b-waves were large, their shapes were abnormal. The b-waves rose to a peak, and the potential was maintained longer than normal. The plateau region of the b-wave was significantly elevated above the baseline potential (Figure 4,

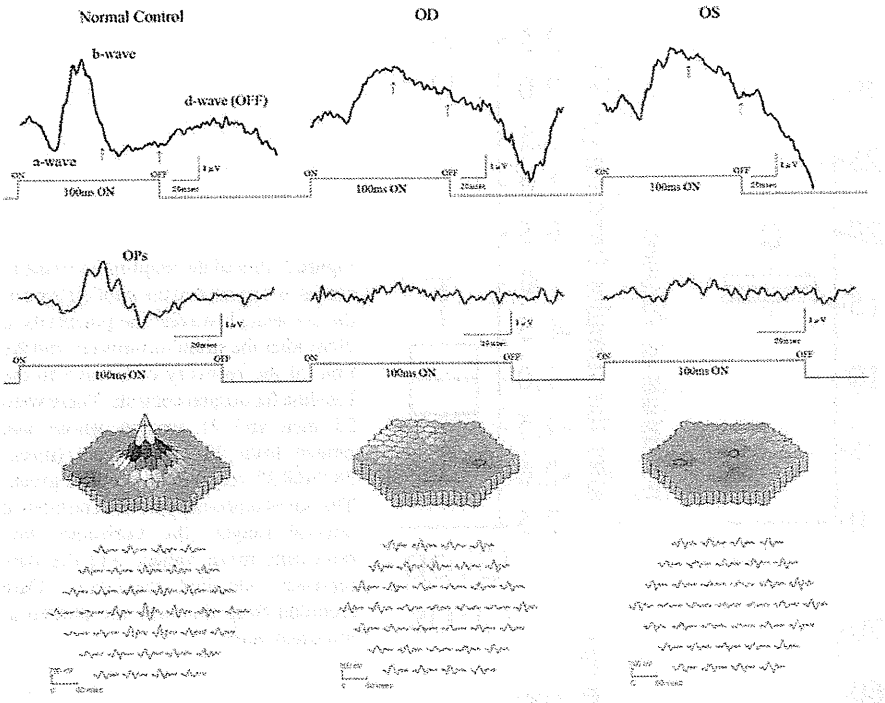

Figure 4. Results of focal macular electroretinograms (ERGs) and multifocal ERGs. Focal macular ERGs and oscillatory potentials recorded from a normal subject and this case are shown (top). The amplitude of the a-wave of this case was severely reduced, and the plateau region was significantly elevated (arrows). The topographic map and the local responses of multifocal ERGs recorded from the normal subject and this case are shown (bottom). The amplitudes in the foveal area were severely reduced in this case.

arrow). To analyze this characteristic, we quantified the potentials at 70 ms after the stimulus was turned on, and the recovery time of the descending slope of b-wave to the baseline from the peak of the b-wave. We calculated the 95% confidence intervals (CI) for the amplitudes of the a-waves and b-waves, the implicit times of the a-waves and b-waves, the potentials at 70 ms after the stimulus turns on, and the time of the recovery of the b-waves to the baseline obtained from the normal controls (Figure 5). Among these six parameters, the amplitudes of the a-waves, the implicit times of the b-waves, the potentials at 70 ms after the stimulus was turned on, and the time of the recovery of the descending slope of the b-wave to the baseline obtained from both eyes of this case were outside the range of the standard deviation and the 95% CI of the normal controls (Figure 5). Especially, the amplitudes of the a-waves, the potentials at 70 ms after the stimulus was turned on, and the time of the recovery of the descending slope of the b-wave to the baseline obtained from this case were severely affected. The amplitudes of the mfERGs in the foveal area were severely reduced in this case (Figure 4).

Molecular genetic findings: Mutation analysis of the *RP1L1* gene in this case showed three missense mutations. There was a c.2578 C>T in exon 4 with a substitution of tryptophan (TGG) for arginine (CGG) at amino acid position 860, a c. 3596 C>G in exon 4 with a substitution of cysteine (TGT) for serine (TCT) at amino acid position 1199, and a c. 4484 C>G in exon 4 with a substitution of arginine (CGC) for proline (CCC) at amino acid position 1495. The amino acid substitution at position 860 and 1495 has already been reported in the SNP database and is found in a high percentage of the normal population. A mutation at amino acid position 1199 has not been reported in the SNP database or in earlier reports (Figure 6A). The serine at position 1199 is well conserved among the RP1L1 family in other species (Figure 6B). This mutation was predicted to be probably damaging with a score of 0.999 by PolyPhen-2. The SIFT tool analysis revealed a score of 0 and predicted that the replaced amino acid is potentially damaging and would not be tolerated. PMut predicted that this mutation is pathological. Align GVGD predicted this mutation as class C65, which means it most likely interferes with the protein function. Out of five computational assessments, only MutationTaster predicted this mutation as a polymorphism. We confirmed that the mutation in this case was segregated with the disease in one affected member and one unaffected member of the family (Figure 6C). The unaffected member of the family in Case 1 underwent clinical examination, including BCVAs, slit-lamp biomicroscopy, fundus ophthalmoscopy, OCT, and focal ERGs. All examination findings were normal. This mutation was not present in 300 control alleles. This mutation p.S1199C has been registered in GenBank with accession number AB684329.

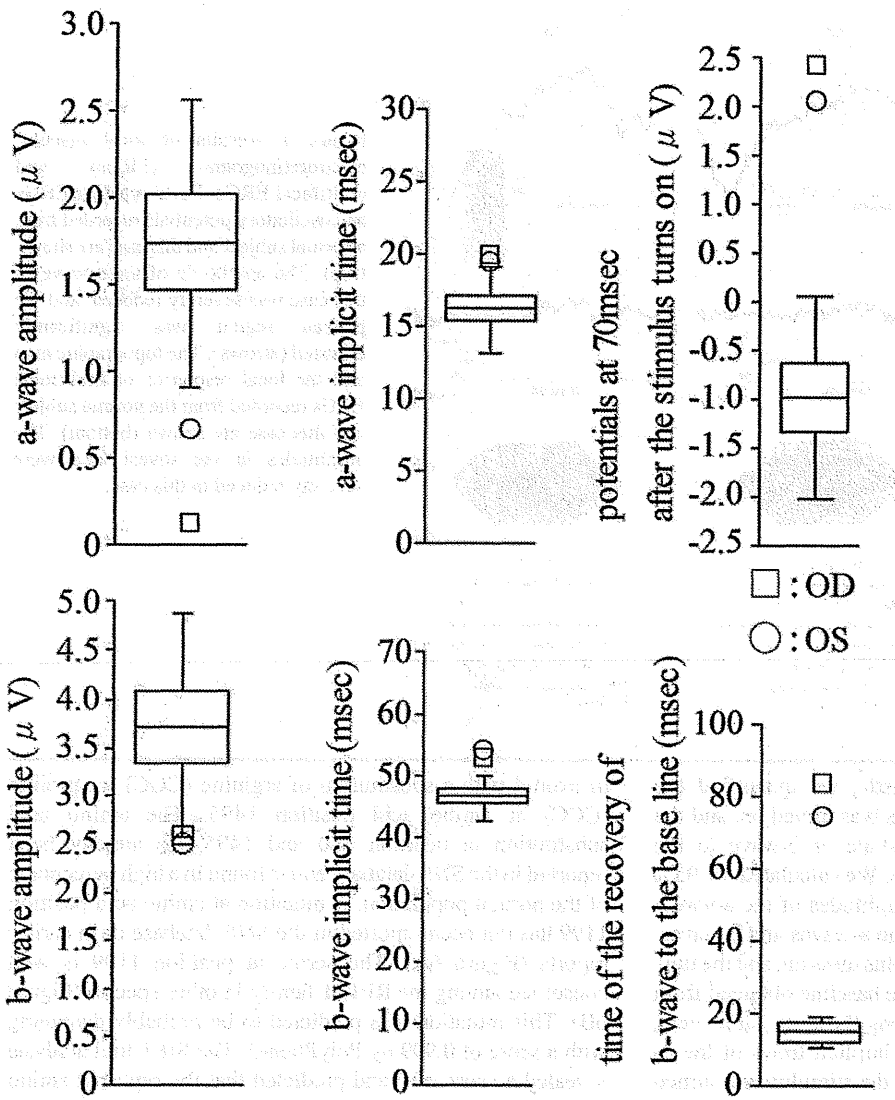

Figure 5. Plot of the amplitudes of the a-waves, b-waves, and the implicit time of the a-waves, b-waves, the potentials at 70ms after the stimulus turns on, and the time of the recovery of b-wave to the baseline for normal controls. There were 25 men and 21 women whose age ranged from 23 to 60 years (mean, 38.04±8.33 years) in this control group. The boxes represent the 95% confidence interval ranges, the horizontal line represents mean values, and the bars represent standard deviation. Data recorded from this case are plotted at indicated mark.

Bowne et al. [11] reported that *RP1L1* mRNA is variable due to the presence of a 48 bp polymorphic coding repeat. They reported that as many as six 48 bp repeats have been observed in normal controls. In this case, one allele contains a 48 bp repeat, and the other allele contains three 48 bp repeats (Figure 6D). There are variations of only two amino acids in the length polymorphism region from this case compared to the reference sequence (NP_849188). One variation with the substitution of E to G in the 14th amino acid of the length polymorphism region was in a previous report [12] (AAN86962, AAN86963, and AAN86964). The other variation with the substitution of G to V in the ninth amino acid of the length polymorphism region was found in more than 10 normal control alleles from a Japanese population. These variations of the length polymorphisms of *RP1L1* with one and three repeats have been registered in GenBank with accession numbers AB684331 and AB684332, respectively.

DISCUSSION

The mutation found in the *RP1L1* gene in this case was a missense mutation with cysteine substituted for serine at amino acid position 1199. This residue is well conserved among the RP1L1 family in other species, suggesting the importance of this amino acid residue for RP1L1 function. Four out of five computational analysis tools predicted this mutation is damaging to the protein function. We did not find this mutation in the sister of the patient with normal vision, although she was the only other family member we were able to test. To decide whether this mutation was pathogenic, we need to examine more family members and a larger number

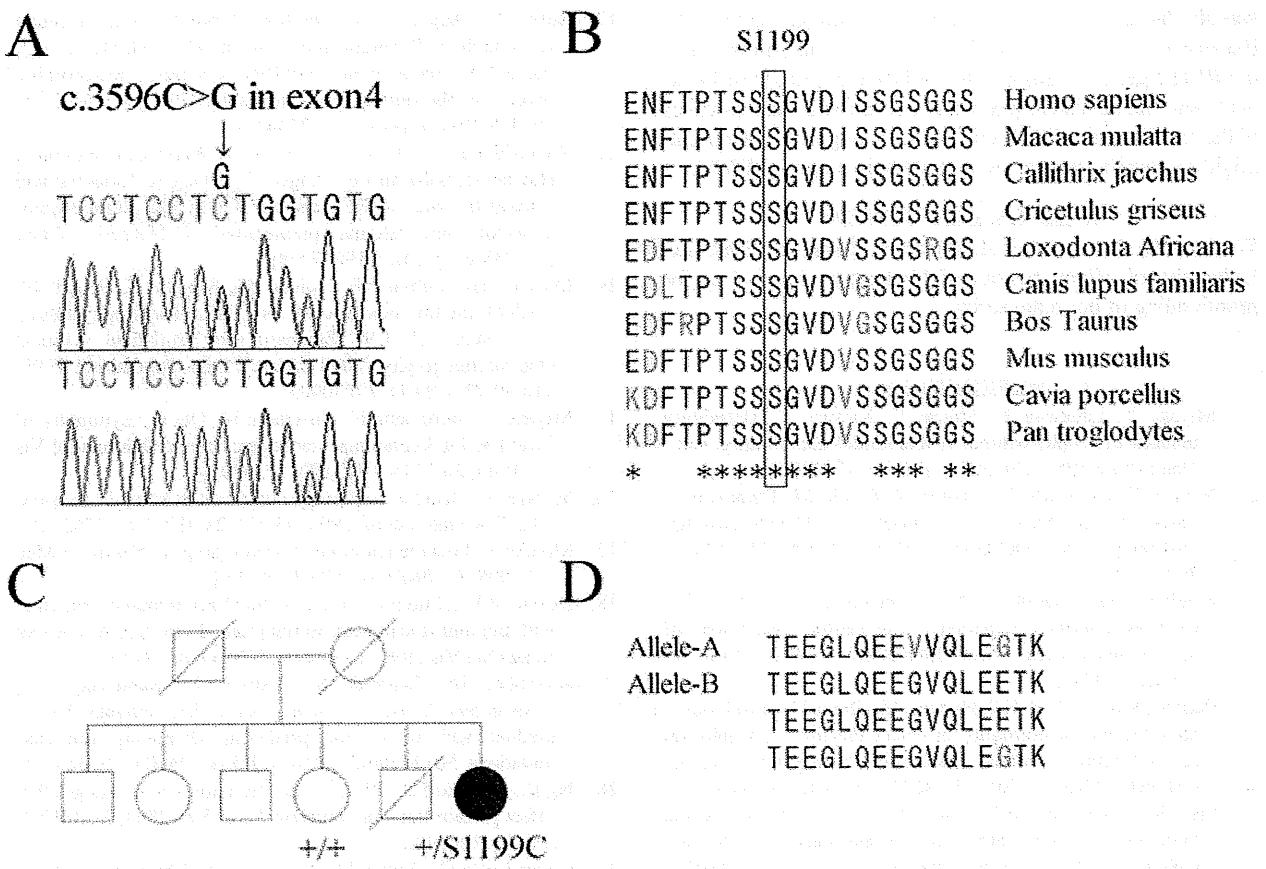

Figure 6. DNA analysis for c.3596C>G mutation and deduced amino acids of length polymorphism region of the RP1-like protein 1 (*RP1L1*) gene and the pedigree of the family with *RP1L1* gene mutation. **A**: Sequence chromatograms for this case (top) and the normal control (bottom) are shown. This case had a c.3596 C>G mutation in exon 4. **B**: Alignment of S1199 in the RP1L1 family proteins. Amino acid-sequence alignments of RP1L1 from 10 species reported in the NCBI database are shown. Amino acid residues of S1199 in humans and conserved residues from other species are boxed. The asterisks indicate completely conserved residues. S1199 is well conserved in all species reported. **C**: We confirmed that the mutation in Case 1 was segregated with the disease in one affected member and one unaffected member of the family. **D**: Deduced amino acids (AA) of repeated regions of the RP1L1 length polymorphism. In this case, one allele contains a 16 AA, and the other allele contains three 16 AA repeats. Variations of amino acids from reference sequence of RP1L1 are shown in red. Those variations are within normal limits.

of normal controls. However, the phenotype of this case was typical of OMD, and thus the mutation in this case was most likely pathogenic.

The photoreceptor IS/OS junction and the COST line can be detected in the SD-OCT images of normal eyes [25-28]. Recently, several degrees of disruption of the IS/OS junction and/or COST line in the SD-OCT images of patients with OMD have been reported [4-8]. In our case, the IS/OS junction and the COST line appeared blurred in the SD-OCT images similar to previous reports.

Researchers have emphasized that the key to differentiating OMD from other diseases, such as optic neuritis or psychological disorders, is the recording of focal macular ERGs from the central retina [1-3]. Focal macular ERGs have a unique waveform when elicited by long-duration stimuli [29]. As shown in this patient, the waveform of focal macular ERGs recorded from patients with OMD with long-duration stimuli had a depolarizing pattern, simulating the ERG waveforms observed after the hyperpolarizing bipolar cell activity is blocked [30-33]. Researchers have demonstrated that by blocking hyperpolarizing bipolar cells with cis-2,3-piperidine dicarboxylic acid or kynurenic acid in monkeys, the a- and d-waves of photopic ERGs become smaller and the plateau between the b- and d-waves remains elevated above the baseline potential [34]. Full-field cone ERG in some human retinal dystrophies show a similar depolarizing pattern [29,35]. Kondo et al. [29] reported similar focal macular ERGs elicited with 100 ms stimuli from a patient with glittering crystalline deposits in the posterior fundus. The waveform of the focal macular ERGs of this case

was similar to those reported for patients with OMD [31-33]. Because this case had a putative disease-causing mutation of the *RP1L1* gene, we suggest the reduced amplitude of the a-wave and the persistent plateau between the b- and d-waves of the focal macular ERGs elicited with long-duration stimuli might be specific markers that could help diagnose OMD.

ACKNOWLEDGMENTS

We thank Dr. Duco Hamasaki (Bascom Palmer Eye Institute, University of Miami School of Medicine, Miami, FL) for proofreading of our manuscript.

REFERENCES

1. Miyake Y, Ichikawa K, Shiose Y, Kawase Y. Hereditary macular dystrophy without visible fundus abnormality. Am J Ophthalmol 1989; 108:292-9. [PMID: 2774037]
2. Miyake Y, Horiguchi M, Tomita N, Kondo M, Tanikawa A, Takahashi H, Suzuki S, Terasaki H. Occult macular dystrophy. Am J Ophthalmol 1996; 122:644-53. [PMID: 8909203]
3. Matthews GP, Sandberg MA, Berson EL. Foveal cone electroretinograms in patients with central visual loss of unexplained etiology. Arch Ophthalmol 1992; 110:1568-70. [PMID: 1444913]
4. Park SJ, Woo SJ, Park KH, Hwang JM, Chung H. Morphologic photoreceptor abnormality in occult macular dystrophy on spectral-domain optical coherence tomography. Invest Ophthalmol Vis Sci 2010; 51:3673-9. [PMID: 20164460]
5. Sisk RA, Berrocal AM, Lam BL. Loss of Foveal Cone Photoreceptor Outer Segments in Occult Macular Dystrophy. Ophthalmic Surg Lasers Imaging 2010; 9:1-3. [PMID: 20337322]
6. Kim YG, Baek SH, Moon SW, Lee HK, Kim US. Analysis of spectral domain optical coherence tomography findings in occult macular dystrophy. Acta Ophthalmol 2011; 89:e52-6. [PMID: 20560888]
7. Fujinami K, Tsunoda K, Hanazono G, Shinoda K, Ohde H, Miyake Y. Fundus autofluorescence in autosomal dominant occult macular dystrophy. Arch Ophthalmol 2011; 129:597-602. [PMID: 21555613]
8. Tsunoda K, Usui T, Hatase T, Yamai S, Fujinami K, Hanazono G, Shinoda K, Ohde H, Akahori M, Iwata T, Myake Y. Clinical characteristics of occult macular dystrophy in family with mutation of RP1L1 gene. Retina. 2012 [PMID: 22466457]In press
9. Lyons JS. Non-familial occult macular dystrophy. Doc Ophthalmol 2005; 111:49-56. [PMID: 16502307]
10. Akahori M, Tsunoda K, Miyake Y, Fukuda Y, Ishiura H, Tsuji S, Usui T, Hatase T, Nakamura M, Ohde H, Itabashi T, Okamoto H, Takada Y, Iwata T. Dominant mutations in RP1L1 are responsible for occult macular dystrophy. Am J Hum Genet 2010, 87:424-9. [PMID: 20826268]
11. Conte I, Lestingi M, den Hollander A, Alfano G, Ziviello C, Pugliese M, Circolo D, Caccioppoli C, Ciccodicola A, Banfi S. Identification and characterization of the retinitis pigmentosa 1-like1 gene (RP1L1): a novel candidate for retinal degenerations. Eur J Hum Genet 2003; 11:155-62. [PMID: 12634863]
12. Bowne SJ, Daiger SP, Malone KA, Heckenlively JR, Kennan A, Humphries P, Hughbanks-Wheaton D, Birch DG, Liu Q, Pierce EA. Characterization of RP1L1, a highly polymorphic paralog of the retinitis pigmentosa 1 (RP1) gene. Mol Vis 2003; 9:129-37. [PMID. 12724644]
13. Yamashita T, Liu J, Gao J, LeNoue S, Wang C, Kaminoh J, Bowne SJ, Sullivan LS, Daiger SP, Zhang K. Essential and synergistic roles of RP1 and RP1L1 in rod photoreceptor axoneme and retinitis pigmentosa. J Neurosci 2009; 29:9748-60. [PMID: 19657028]
14. Marmor MF, Fulton AB, Holder GE, Miyake Y, Brigell M, Bach M, International Society for Clinical Electrophysiology of Vision. ISCEV Standard for full-field clinical electroretinography (2008 update). Doc Ophthalmol 2009; 118:69-77. [PMID: 19030905]
15. Miyake Y, Shiroyama N, Horiguchi M, Ota I. Asymmetry of focal ERG in human macular region. Invest Ophthalmol Vis Sci 1989; 30:1743-9. [PMID: 2759790]
16. Miyake Y. Macular oscillatory potentials in humans: macular OPs. Doc Ophthalmol 1990; 75:111-24. [PMID: 2276312]
17. Miyake Y. Focal macular electroretinography. Nagoya J Med Sci 1998; 61:79-84. [PMID: 9879190]
18. Bearse MA Jr, Sutter EE. Imaging localized retinal dysfunction with the multifocal electroretinogram. J Opt Soc Am A Opt Image Sci Vis 1996; 13.634-40. [PMID: 8627420]
19. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010; 7:248-9. [PMID: 20354512]
20. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003; 31:3812-4. [PMID: 12824425]
21. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 2002; 315:771-86. [PMID: 11812146]
22. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 2006; 43:295-305. [PMID: 16014699]
23. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res 2006; 34:1317-25. [PMID: 16522644]
24. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010; 7.575-6. [PMID: 20676075]
25. Srinivasan VJ, Monson BK, Wojtkowski M, Bilonick RA, Gorczynska I, Chen R, Duker JS, Schuman JS, Fujimoto JG. Characterization of outer retinal morphology with high-speed, ultrahigh-resolution optical coherence tomography. Invest Ophthalmol Vis Sci 2008; 49:1571-9. [PMID: 18385077]
26. Marmor MF, Choi SS, Zawadzki RJ, Werner JS. Visual insignificance of the foveal pit: reassessment of foveal

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. Elctrodiagnosis 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 Sumiyama[1], Kazuho Ikeo[2,3], Takashi Gojobori[2,3], and Naruya Saitou[1,*]

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

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

[3]DNA 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 (Karlin 2005), including phylogeny reconstruction (Takahashi et al. 2009). Most frequently, such comparisons are based on a dinucleotide composition model (Karlin and Mrazek 1997; Gentles and Karlin 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

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

Table 1
R Value Comparison for Selected Species

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

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

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

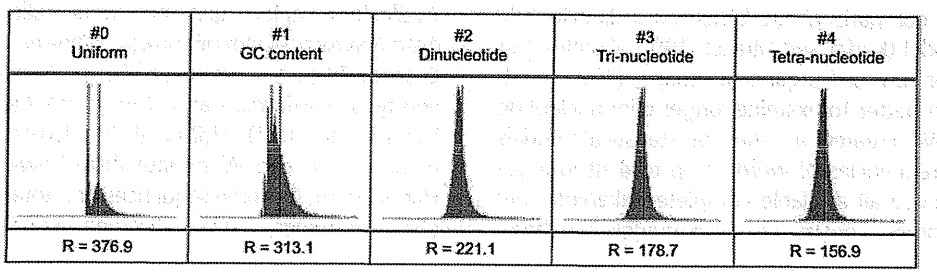

FIG. 1.—Histograms of relative abundances of all oligonucleotides of 8 bp in human genome, according to the five composition models. The *R* value computed for each model is used as a horizontal scaling factor. The vertical red line corresponds to the expected frequency. The words placed to the left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

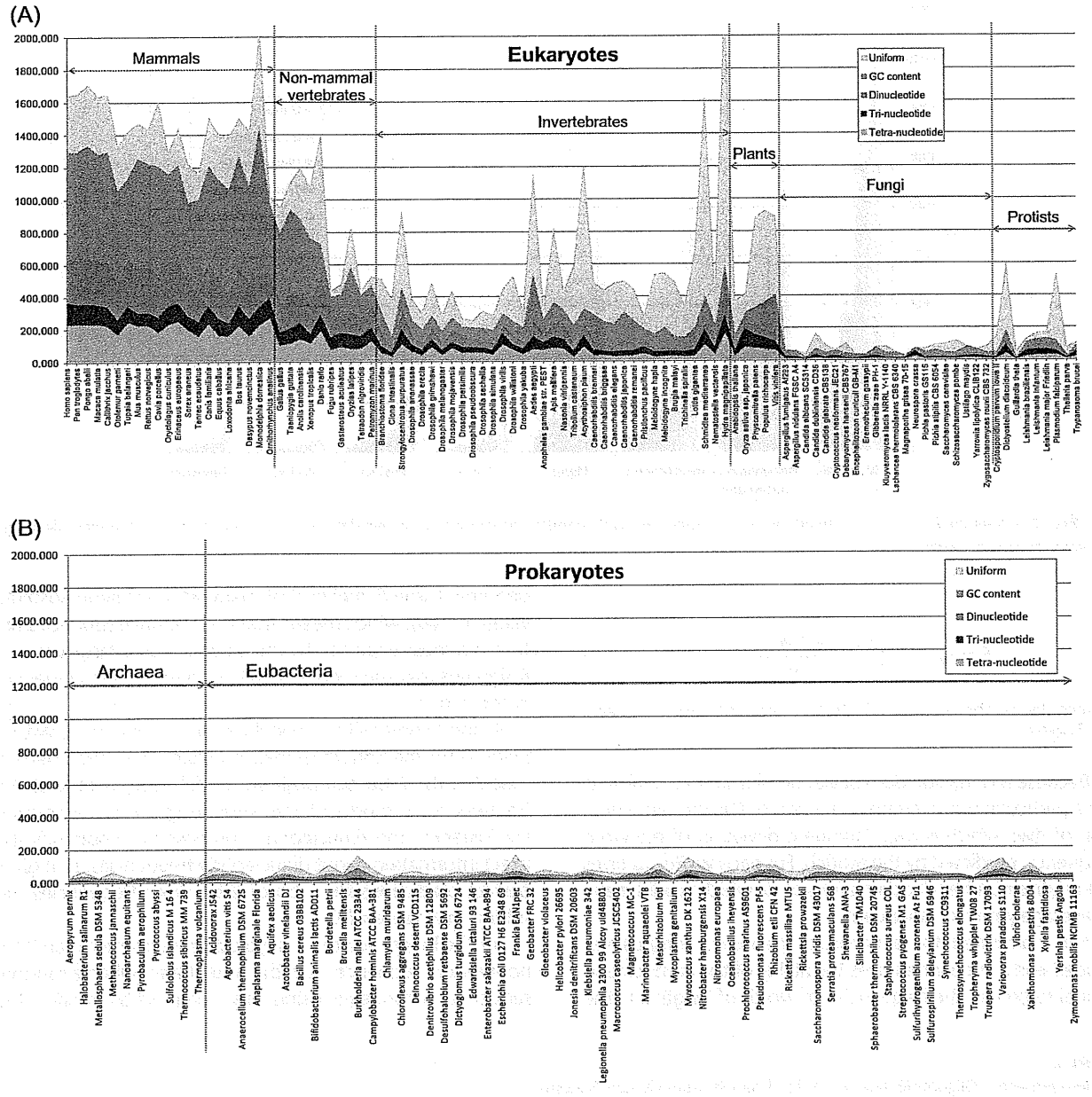

Fig. 2.—Comparison of *R* values based on oligonucleotides of 5 bp and all five composition models. (*A*) Eukaryote genomes (all available in public databases by October 2010). (*B*) Representative prokaryote (both eubacteria and archaea) genomes.

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:

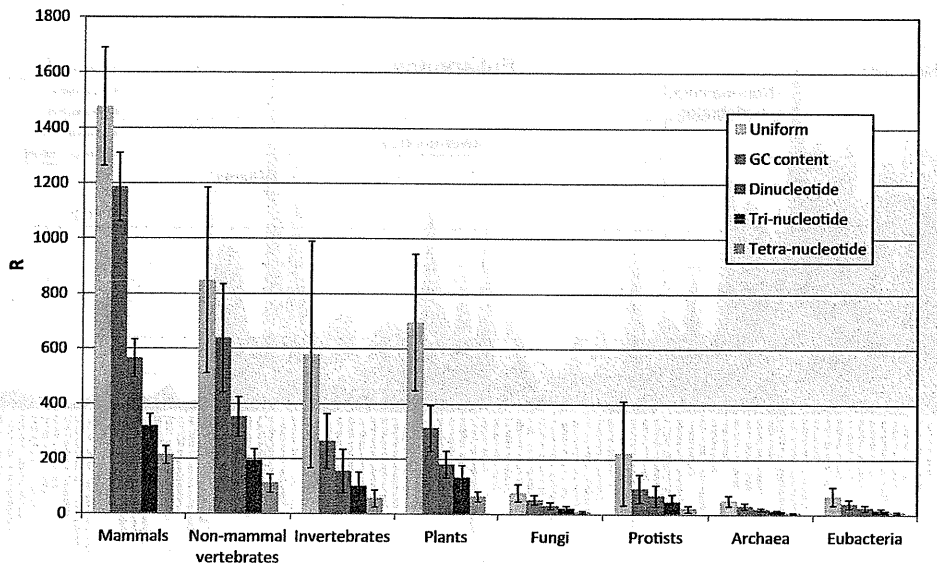

Fig. 3.—Average *R* values for different groups of organisms, with standard deviations, using five different composition models. Standard deviation for each value is displayed.

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

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

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	tataaaaaaa (ttttttata)	45,933	115,110	−203.9
2	aaatttttc (gaaaaaattt)	29,389	89,480	−200.9
3	tttttttggg (cccaaaaaaa)	19,774	72,956	−196.9
4	aaaaattttt	103,832	185,936	−190.4
5	ttttttggga (tcccaaaaaa)	14,119	60,161	−187.7
6	aaaattttc (gaaaaatttt)	33,460	89,480	−187.3
7	aaaaaaatat (atattttttt)	80,964	153,706	−185.6
8	aaaaaatttc (gaaatttttt)	34,571	89,480	−183.6
9	aaaaattttg (caaaattttt)	33,265	87,274	−182.8
10	aaaaaatttg (caaatttttt)	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	tgtaatccca (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	tgcagtgagc (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 or 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.

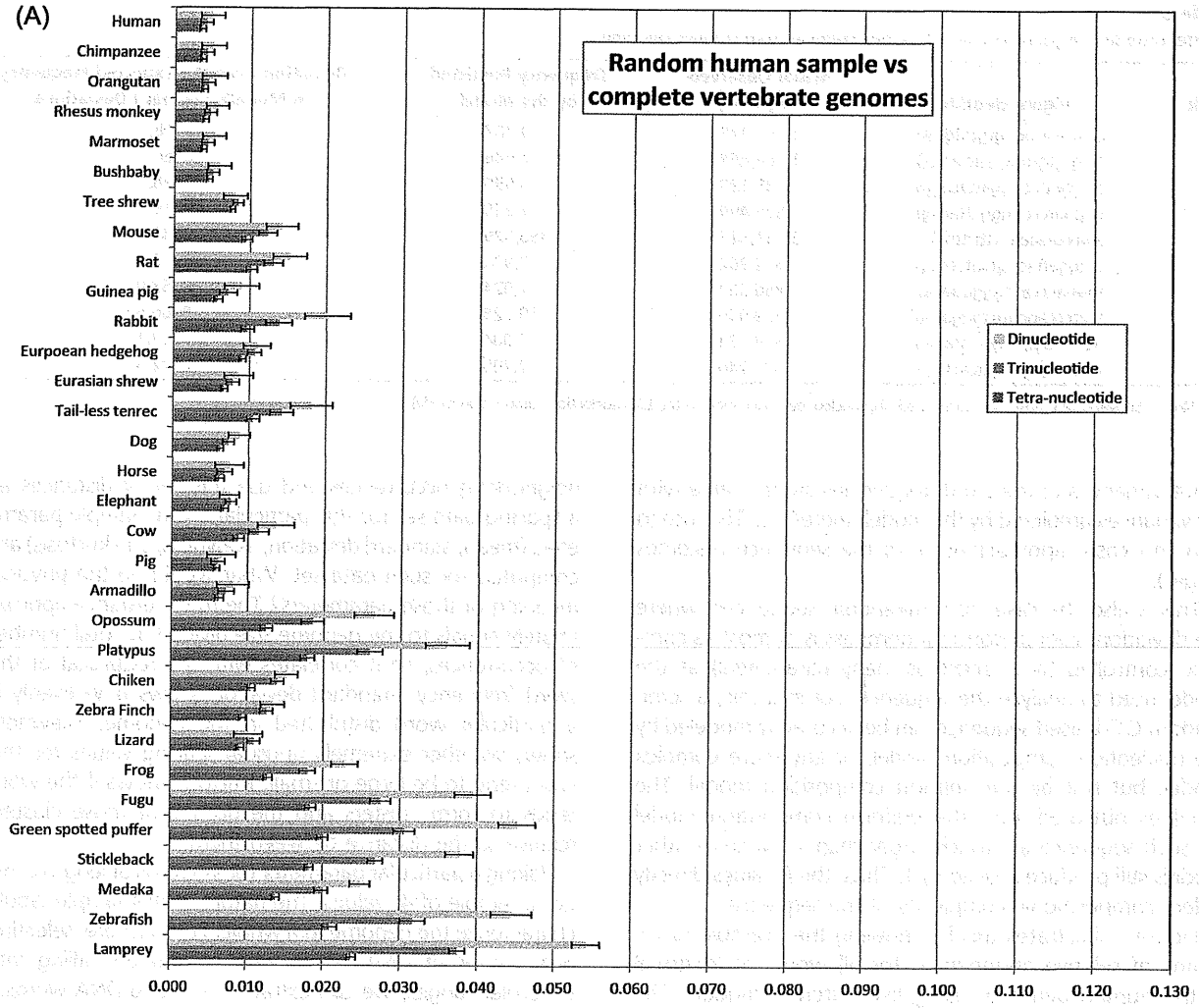

Fig. 4.—Euclidean distances between composition vectors (oligonucleotide frequencies) of sample data sets and complete vertebrate genomes for three composition models (dinucleotide, trinucleotide, and tetranucleotide). (*A*) When sampled data set is human genome. One thousand samples were used, where 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

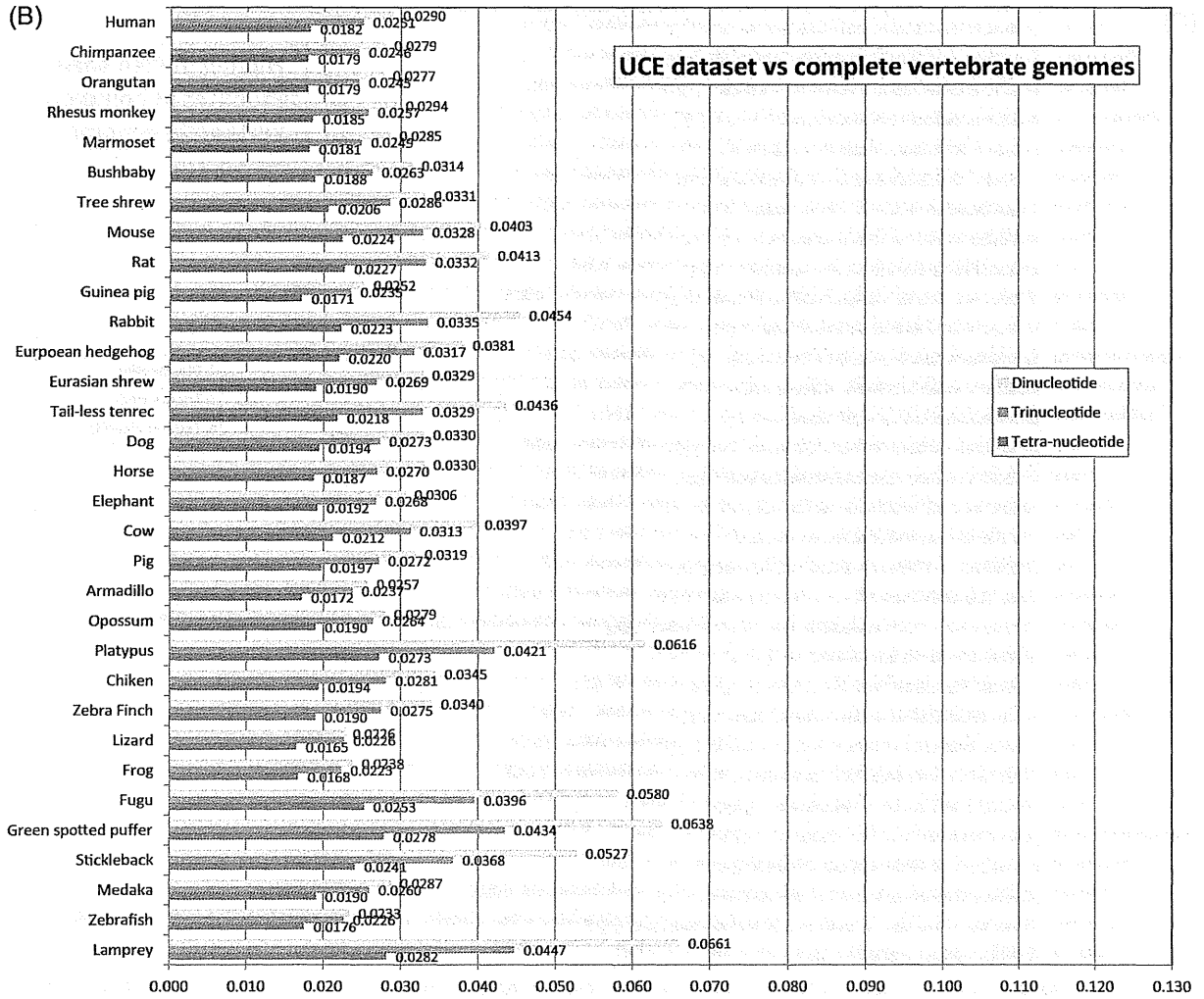

FIG. 4.—Continued

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

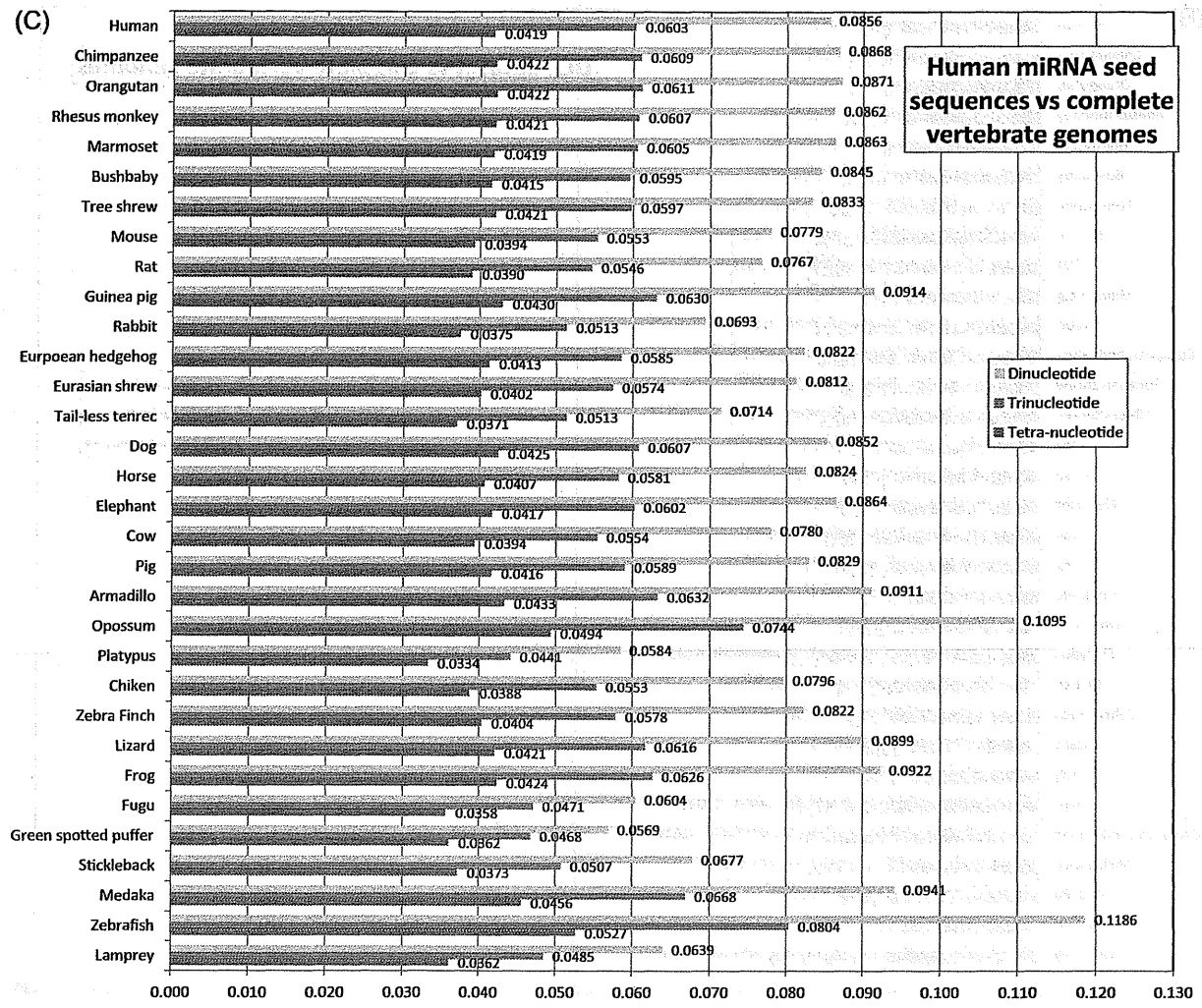

Fig. 4.—Continued

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

Figure 4*A* 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 4*B* 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 4*C* 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. 5*A*), while much more uniformity can be seen for UCE and miRNA seed data sets (fig. 5*B* and *C*).

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

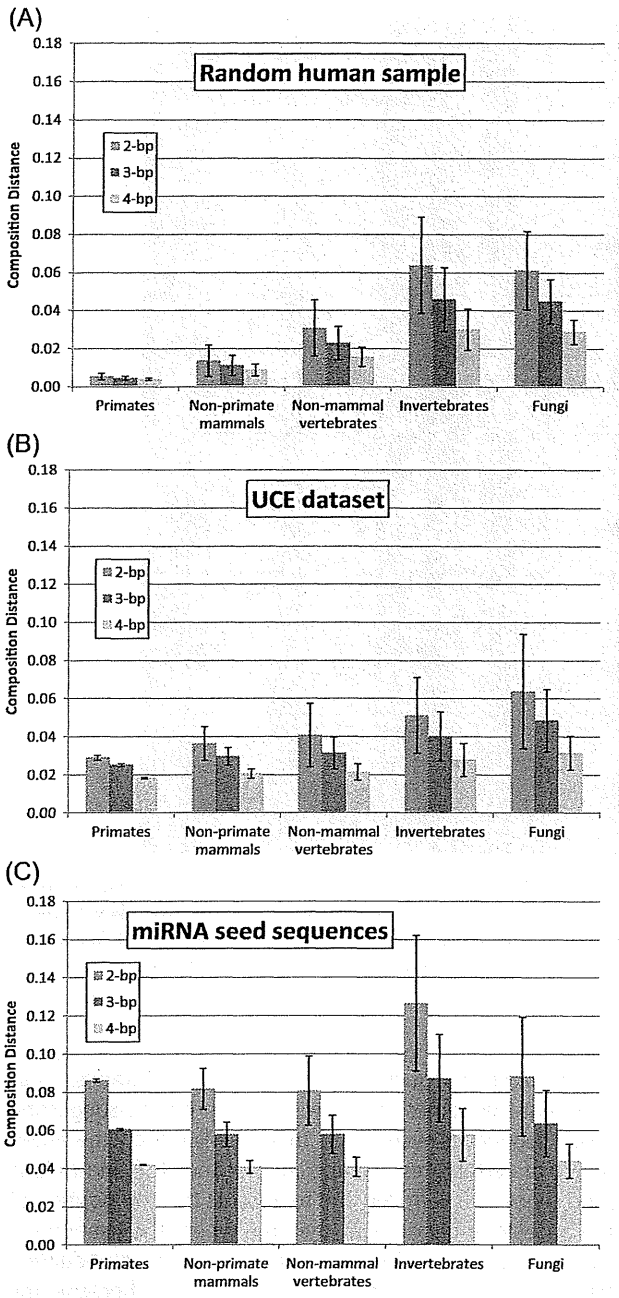

FIG. 5.—Average compositional distance (Euclidean distances between the composition vectors) between sample data sets and complete genomes grouped into four groups. Panels (*A*, *B*, and *C*) correspond to panels (*A*, *B*, and *C*) of figure 4. Standard deviations of the distances are shown for all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

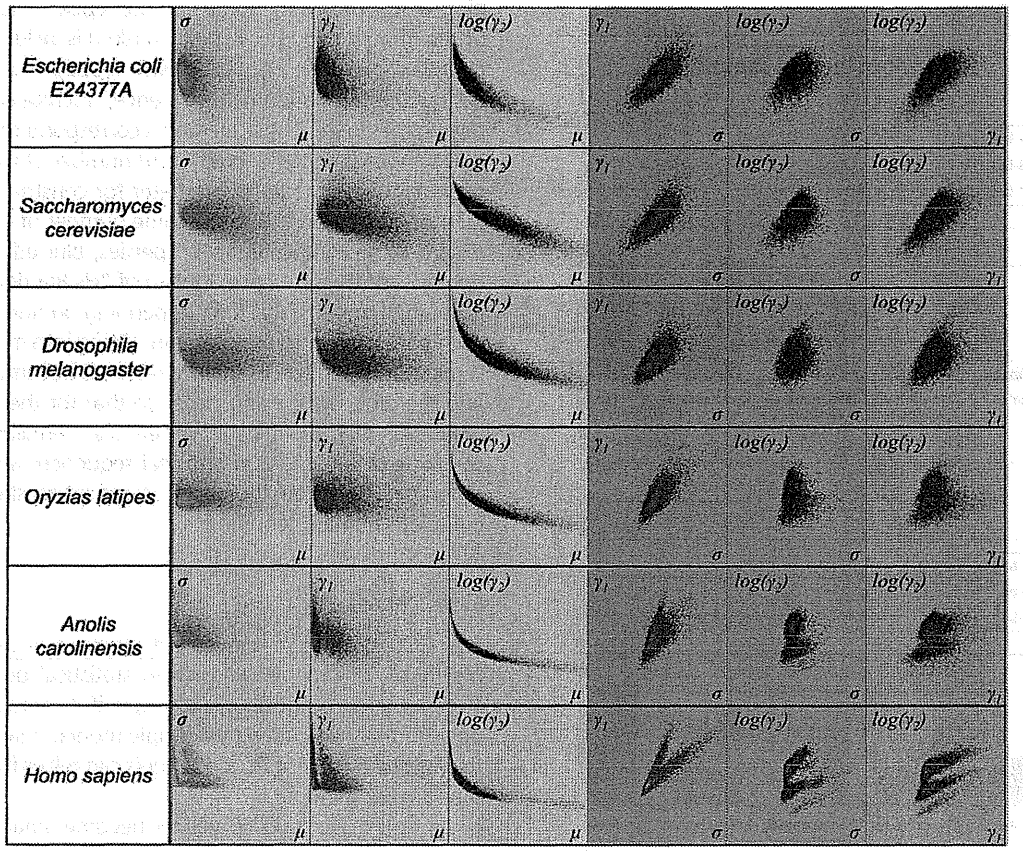

Fig. 6.—Plots of the spacing distribution parameters for six species, based on oligonucleotides of 8 bp. Each row represents one genome. Different columns show plots for different pairs of parameters, from left to right: mean spacing (*x* axis) versus standard deviation (*y* axis), mean (*x*) versus skewness (*y*), mean (*x*) versus log(kurtosis) (*y*), standard deviation (*x*) versus skewness (*y*), standard deviation (*x*) versus log(kurtosis) (*y*), and skewness (*x*) versus log(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 "gcctgtaatcccagc" (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 ancestoral 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 ancestoral genome and current day genomes. Thus, this allows us to discuss the composition of premammal 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

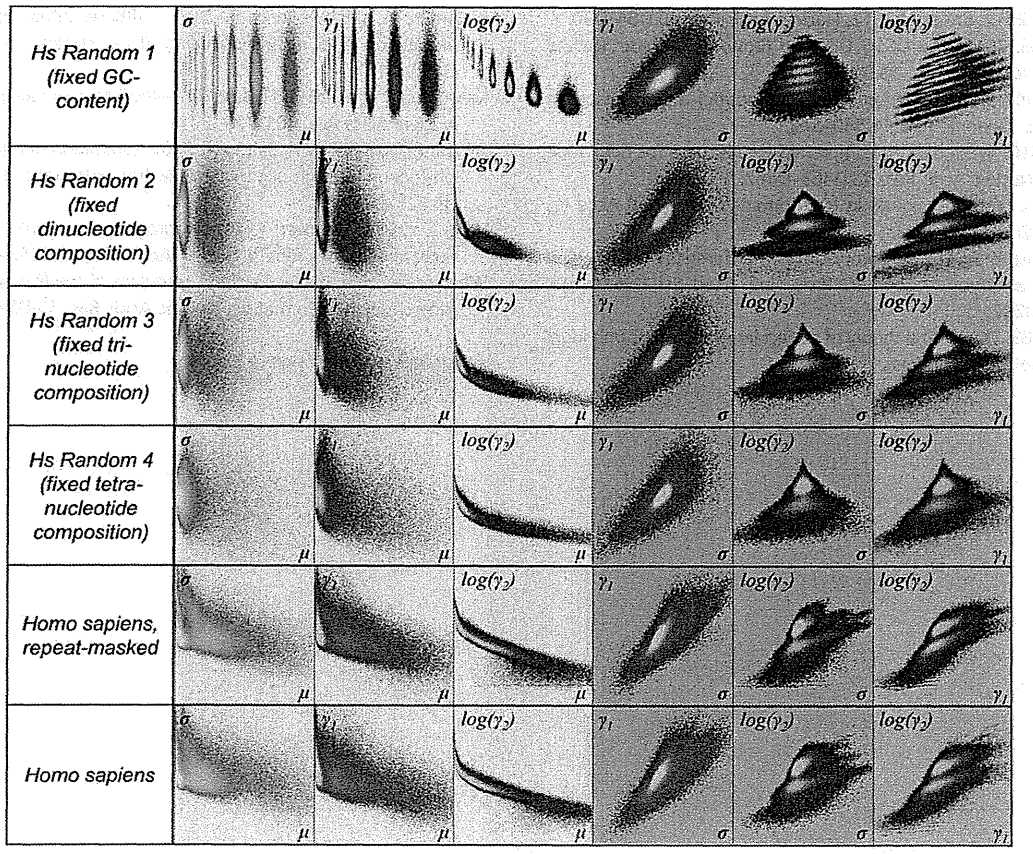

Fig. 7.—Plot of the spacing distribution parameters for semirandom sequences, compared with the real human genome, as well as repeat-masked one. Based on oligonucleotides of 10 bp. Semirandom genomes 1, 2, 3, and 4 are constructed using 1-, 2-, 3-, and 4-bp composition of the actual human genome.

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 patterns. 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