computational biologists and suggests ways to offset them.

Tunicate biologists may sometimes find the proposed names too long, and too constraining. Human gene names and symbols for some favorite developmental regulators may also not speak to them (e.g., Human *RBPJ1* is usually called *Suppressor of Hairless*, which is the name of its ortholog in *Drosophila*). Also, tunicate developmental biologists may prefer naming some genes on the basis of their ascidian loss-of-function phenotype (e.g., *Macho-1*, *Notrlc*) than by reference to Human genes (*Zic-r.a*; *Hand-related*). In such cases, the name that makes most sense could be used in a publication, provided the primary human-derived gene name or symbol is also mentioned in the main text, for instance the first time the gene is discussed. As primary names of genetic elements may sometimes change, traceability requires the inclusion in the materials and methods of the unique identifier of each gene transcript or protein studied, and the precise coordinates of any new *cis*-regulatory element/construct described in the work in a well identified genome assembly. We suggest that this is done in a specific section called "Unique identifiers and coordinates of listed genetic elements." This section will considerably facilitate the work of biocurators.

Although we strongly reduced the number of punctuation marks in names, avoided characters that can be interpreted as delimiters in formatted text files (space, comma, tab) and imposed a fixed number of digits to numerical identifiers, computational biologists may be confronted to some residual issues when using certain strategies to parse the names of genetic elements. The presence of characters ("/",">","[","|") that serve as operators in some programming languages will need to be taken into account in parsing scripts.

LITERATURE CITED

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T. 2003. A uniform system for microRNA annotation. RNA 9: 277–279.

Christiaen L, Stolfi A, Davidson B, Levine M. 2009. Spatio-temporal intersection of Lhx3 and Tbx6 defines the cardiac field through synergistic activation of Mesp. Dev Biol 328:552–560.

Corbo JC, Levine M, Zeller RW. 1997. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Development 124:589–602.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KEM, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang H, Awazu S, Azumi K, Boore J, Branno M, Chin-bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee B, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS. 2002. The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins. Science 298:2157–2167.

Denoeud F, Henriet S, Mungpakdee S, Aury J-M, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Cañestro C, Bouquet JM, Danks G, Poulain J, Campsteijn C, Adamski M, Cross I, Yadetie F, Muffato M, Louis A, Butcher S, Tsagkogeorga G, Konrad A, Singh S, Jensen MF, Cong EH, Eikeseth-Otteraa H, Noel B, Anthouard V, Porcel BM, Kachouri-Lafond R, Nishino A, Ugolini M, Chourrout P, Nishida H, Aasland R, Huzurbazar S, Westhof E, Delsuc F, Lehrach H, Reinhardt R, Weissenbach J, Roy SW, Artiguenave F, Postlethwait JH, Manak JR, Thompson EM, Jaillon O, Du Pasquier L, Boudinot P, Liberles DA, Volff JN, Philippe H, Lenhard B, Roest Crollius H, Wincker P, Chourrout D. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 330:1381–1385.

Di Gregorio A, Corbo JC, Levine M. 2001. The regulation of forkhead/HNF-3 β expression in the Ciona embryo. Dev Biol 229(1):31–43.

Eilbeck K, Lewis S, Mungall C, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The sequence ontology: A tool for the unification of genome annotations. Genome Biol 6:R44.

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. 2013. Genenames.org: The HGNC resources in 2013. Nucleic Acids Res 41:D545–D552.

Hozumi A, Kawai N, Yoshida R, Ogura Y, Ohta N, Satake H, Satoh N, Sasakura Y. 2010. Efficient transposition of a single Minos transposon copy in the genome of the ascidian Ciona intestinalis with a transgenic line expressing transposase in eggs. Dev Dyn 239:1076–1088.

Imai KS, Daido Y, Kusakabe TG, Satou Y. 2012. Cis-acting transcriptional repression establishes a sharp boundary in chordate embryos. Science 337:964–967.

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. 2009. Infrastructure for the life sciences: Design and implementation of the UniProt website. BMC Bioinf 10: 136.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411–412; author reply 414.

Kobayashi M, Takatori N, Nakajima Y, Kumano G, Nishida H, Saiga H. 2010. Spatial and temporal expression of two transcriptional isoforms of Lhx3, a LIM class homeobox gene, during embryogenesis of two phylogenetically remote ascidians, Halocynthia roretzi and Ciona intestinalis. Gene Express Patterns 10(2):98–104.

Kozomara A, Griffiths-Jones S. 2014. miRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68–D73.

Lemaire P. 2011. Evolutionary crossroads in developmental biology: The tunicates. Development 138: 2143–2152.

Lemaire P, Smith WC, Nishida H. 2008. Ascidians and the plasticity of the chordate developmental program. Curr Biol 18:R620–R631.

Nakazawa K, Yamazawa T, Moriyama Y, Ogura Y, Kawai N, Sasakura Y, Saiga H. 2013. Formation of the digestive tract in Ciona intestinalis includes two distinct morphogenic processes between its anterior and posterior parts. Dev Dyn 242:1172–1183.

Sasakura Y, Nakashima K, Awazu S, Matsuoka T, Nakayama A, Azuma J, Satoh N. 2005. Transposon-mediated insertional mutagenesis revealed the functions of animal cellulose synthase in the ascidian Ciona intestinalis. Proc Natl Acad Sci U S A 102: 15134–15139.

Sasakura Y, Kanda M, Ikeda T, Horie T, Kawai N, Ogura Y, Yoshida R, Hozumi A, Satoh N, Fujiwara S. 2012. Retinoic acid-driven Hox1 is required in the epidermis for forming the otic/atrial placodes during ascidian metamorphosis. Development 139:2156–2160.

Satou Y, Mineta K, Ogasawara M, Sasakura Y, Shoguchi E, Ueno K, Yamada L, Matsumoto J, Wasserscheid J, Dewar K, Wiley GB, Macmil SL, Roe BA, Zeller RW, Hastings KEM, Lemaire P, Lindquist E, Endo T, Hotta K, Inaba K. 2008. Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: New insight into intron and operon populations. Genome Biol 9:R152.

Small KS, Brudno M, Hill MM, Sidow A. 2007. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. Genome Biol 8:R41.

Stolfi A, Lowe EK, Racioppi C, Ristoratore F, Brown CT, Swalla BS, Christiaen L. 2014. Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. eLife 10.7554/eLife.03728.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, Ishizuka KJ, Gissi C, Griggio F, Ben-Shlomo R, Corey DM, Penland L, White RA, Weissman IL, Quake SR. 2013. The genome sequence of the colonial chordate, Botryllus schlosseri. eLife 2: e00569.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982.

Wright MW, Bruford EA. 2011. Naming "junk": Human non-protein coding RNA (ncRNA) gene nomenclature. Hum Genom 5:90–98.

Xie J, Zhang M, Zhou T, Hua X, Tang L, Wu W. 2007. Sno/scaRNAbase: A curated database for small nucleolar RNAs and cajal body-specific RNAs. Nucleic Acids Res 35:D183–D187.

Sustained Heterozygosity Across a Self-Incompatibility Locus in an Inbred Ascidian

Yutaka Satou,*[,1] Kazuko Hirayama,[1] Kaoru Mita,[2] Manabu Fujie,[3] Shota Chiba,‡[,1] Reiko Yoshida,[1] Toshinori Endo,[4] Yasunori Sasakura,[2] Kazuo Inaba,[2] and Nori Satoh[1,3]

[1]Department of Zoology, Graduate School of Science, Kyoto University, Sakyo, Kyoto, Japan

[2]Shimoda Marine Research Center, University of Tsukuba, Shimoda, Shizuoka, Japan

[3]Marine Genomics Unit and DNA Sequencing Section, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan

[4]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

‡Present address: Department of Immunology and Microbiology, Meiji University of Integrative Medicine, Nantan-city, Kyoto, Japan

***Corresponding author:** E-mail: yutaka@ascidian.zool.kyoto-u.ac.jp.

Associate editor: Naruya Saitou

Abstract

Because self-incompatibility loci are maintained heterozygous and recombination within self-incompatibility loci would be disadvantageous, self-incompatibility loci are thought to contribute to structural and functional differentiation of chromosomes. Although the hermaphrodite chordate, *Ciona intestinalis*, has two self-incompatibility genes, this incompatibility system is incomplete and self-fertilization occurs under laboratory conditions. Here, we established an inbred strain of *C. intestinalis* by repeated self-fertilization. Decoding genome sequences of sibling animals of this strain identified a 2.4-Mbheterozygous region on chromosome 7. A self-incompatibility gene, *Themis-B*, was encoded within this region. This observation implied that this self-incompatibility locus and the linkage disequilibrium of its flanking region contribute to the formation of the 2.4-Mb heterozygous region, probably through recombination suppression. We showed that different individuals in natural populations had different numbers and different combinations of *Themis-B* variants, and that the rate of self-fertilization varied among these animals. Our result explains why self-fertilization occurs under laboratory conditions. It also supports the concept that the *Themis-B* locus is preferentially retained heterozygous in the inbred line and contributes to the formation of the 2.4-Mb heterozygous region. High structural variations might suppress recombination, and this long heterozygous region might represent a preliminary stage of structural differentiation of chromosomes.

***Key words:* *Ciona intestinalis*, self-incompatibility, inbred line, chromosome evolution.**

Introduction

Balancing selection maintains diversity of self-incompatibility loci (Wright 1939), and self-incompatibility loci are maintained heterozygous, because they prevent self-fertilization. Hence, self-incompatibility loci could cause linkage disequilibrium of flanking regions, and recombination suppression (Uyenoyama 1997), and might eventually cause structural differentiation of chromosomes, as a sex determining gene does in sex chromosomes (Eichler and Sankoff 2003; Fraser and Heitman 2004; Charlesworth et al. 2005; Uyenoyama 2005; Bachtrog 2013). Although this prediction has partially been tested in plants (Kamau and Charlesworth 2005; Llaurens et al. 2009), it could also be tested by inbreeding more directly. First, genomic regions near self-incompatibility loci will be retained heterozygous by linkage disequilibrium. Second, if recombination is suppressed, large regions flanking to self-incompatibility loci will be retained heterozygous.

The genome of *Ciona intestinalis*, which is a hermaphrodite chordate belonging to a sister group of vertebrates (Delsuc et al. 2006), provides a unique opportunity to address this problem. It encodes two self-incompatibility loci, *Themis-A* and *Themis-B* (Harada et al. 2008). These loci are encoded on different chromosomes and are expected to be preferentially maintained heterozygous, because fertilization rarely occurs between sperm and eggs bearing self-incompatibility proteins that are encoded by the same alleles in both loci. Therefore, theoretically, either of the self-incompatibility loci is always heterozygous, and outcrossing is the natural reproductive mode of *C. intestinalis* (Morgan 1944). However, the self-incompatibility system of this animal is however incomplete, and self-fertilization occurs in laboratory conditions.

Each of these incompatibility loci encodes two genes (Harada et al. 2008; Saito et al. 2012); one is expressed in sperm and the other is expressed in eggs. The *Themis-B* locus encodes the genes, *s-Themis-B* and *v-Themis-B*. *s-Themis-B* is expressed from a haploid genome of sperm, and *v-Themis-B* is expressed from a diploid genome of oocytes, and their protein products interact with each other to reject fertilization. These two genes contain a hypervariable region, which enables a specific interaction between s-Themis-B and v-Themis-B proteins encoded by the same allele. The *Themis-A* locus has a structure similar to the

Article

Table 1. Statistics of Sequencing Data and Mapping Data.

	Specimen A	Specimen B
Number of Illumina sequence tags	158,996,466	162,332,248
Total length of Illumina sequence tags	15,693,529,667	16,034,418,242
Number of 454 sequence tags	953,649	860,633
Total length of 454 sequence tags	130,066,159	102,565,614
Total nucleotide length	15,823,595,826	16,136,983,856
Total number of nucleotides mapped onto the reference genome sequence (112,162,187 bases)	11,393,757,651	11,895,122,867
Mean sequence depth	101.5x	106.1x
Total number of nucleotides mapped onto the repeat-masked genome sequence (97,042,271 bases)	10,032,892,695	10,576,941,089
Nucleotide positions where genotypes are called	89,770,299	89,531,196
Heterozygous positions	38,072	39,177
Homozygous positions	89,732,207	82,492,000
Homozygous positions that are different from the reference sequence	991,115	992,810
Ambiguous positions	20	19

Themis-B locus. An incompatible reaction occurs, only if both *Themis-A* and *Themis-B* alleles are matched between sperm and eggs.

To maintain the incompatibility system of *Themis-A* and *Themis-B*, recombination needs to be suppressed within these loci, because such recombination would destroy their function by rendering gametes carrying different alleles incompatible. However, there are no indication that structural differentiation occurs around the *Themis-A* and *Themis-B* loci, because the heterozygosity of regions close to these two loci is not prominently high compared with other regions (Satou et al. 2012). Hence, if these two self-incompatibility loci do evoke structural differentiation of chromosomes, the process remains at a preliminary stage.

In this study, we found that a long heterozygous region in the genome of an inbred line of *C. intestinalis* contained the self-incompatibility *Themis-B* locus. We examined a possibility that this locus contributes to the formation of this long heterozygous region.

Results

Sequencing of Genomes of Two Sibling Animals of an Inbred Line

We established an inbred strain of the tunicate, *C. intestinalis*, from an individual that was caught in Onagawa-Bay, Miyagi Prefecture, Japan. Taking advantage of the incomplete self-incompatibility of this animal (Harada et al. 2008), we repeated self-fertilization. Using sperm, we analyzed genomes of two mature siblings (specimens A and B) obtained after 11 self-fertilization (F11 generation). Using illumina and 454 sequencers, we obtained approximately 1.6×10^8 sequence tags from each of these two sibling genomes (table 1). Over 1×10^{10} bases were mapped onto the nonrepeated region of the reference genome (Dehal et al. 2002; Satou et al. 2008), which corresponded to over 100× sequence coverage. Then, we called genotypes of the 89,770,299 and 89,531,196 nucleotide positions, which correspond to approximately 80% of the reference sequence (table 1).

We found that 0.04% of nucleotide positions were heterozygous in each sibling genome (table 1). Genotypes could not be determined for 20 and 19 positions in the genomes of specimens A and B, respectively, which could be due to misalignments of sequence tags to the reference genome. The observed frequencies of heterozygous sites of specimens A and B were much lower than heterozygosity rates in natural populations (1.1–1.2%) (Dehal et al. 2002; Satou et al. 2012), but higher than the rate expected after 11 generations of self-fertilization assuming neutrality ($1.1–1.2\% \times (1/2)^{11} = 0.00054–0.00059\%$).

Among homozygous sites, 1.1% of genomic positions of the inbred animals were different from the reference (table 1). Thus, each of the four haplotypes of the two siblings, we analyzed differed from the reference by approximately 1.12% (=1.1% + 0.04%/2). This estimate showed great agreement with the mean nucleotide diversity of *C. intestinalis* (Dehal et al. 2002; Satou et al. 2012), confirming the accuracy of genotype calling.

There Are Few Neutral Heterozygous Sites That Would Become Homozygous by Further Inbreeding

To test whether any neutral heterozygous sites or regions were remaining in the genome of the inbred strain, we compared genome sequences between the two sibling specimens. If neutral heterozygous sites or regions remained in the genome of the inbred strain, a quarter of these sites would become homozygous in both specimens, another quarter would remain heterozygous in both specimens, and the rest would become heterozygous in one specimen and homozygous in the other (Hom/Het sites or regions). At 89,068,420 positions where genotypes were determined commonly in both siblings, 2,962 positions were homozygous in specimen A and heterozygous in specimen B, and another 2,820 positions were heterozygous in specimen A and homozygous in specimen B (5,782 Hom/Het sites in total; table 2).

Deep inspection of raw data revealed that at 5,675 of these 5,782 Hom/Het sites, two heterozygous bases were found in aligned tags in both specimens, although in one animal,

Table 2. Comparisons of Heterozygous and Homozygous Sites between the Two Sibling Specimens.

	Number of Sites
Genomic positions whose genotypes were called in both siblings	89,068,420
Homozygous sites that are identical between the siblings	89,028,708
Homozygous sites that are not identical between the siblings	1
Heterozygous sites that are identical between the siblings	33,903
Heterozygous sites that are not identical between the siblings	0
Sites that are homozygous in the specimen A and heterozygous in the specimen B	2,962 (68[a])
Sites that are homozygous in the specimen B and heterozygous in the specimen A	2,820 (39[a])
Ambiguous sites	26

[a]Candidates for bona fide Hom/Het sites (see the main text and Materials and Methods).

frequencies of the secondarily common bases were below the threshold (see Materials and Methods). Thus, these 5,675 (2,894 and 2,781 in specimens A and B, respectively) sites could have been miscalled in either specimen (not bona fide Hom/Het sites). Even if all of these sites were indeed heterozygous, the above estimated frequencies of heterozygous sites are not significantly increased.

The remaining 107 (=5,782–5,675) Hom/Het sites were considered candidates for bona fide Hom/Het sites. Because of genetic linkage, most Hom/Het sites are expected to make clusters on the genome. Of the 107 Hom/Het sites, 56 heterozygous sites were indeed found in a 5,224-bp region of chromosome 5 of specimen B (fig. 1A–C), and therefore these sites likely represent bona fide Hom/Het sites, or multiple copies of this region in one haplotype of the genome of specimen B. In the case of multiple copies, the above 56 sites might not be actual Hom/Het sites, but the corresponding region should be considered as a Hom/Het "region." Indeed, the mean sequence coverage of this region in specimen B (160×) was 1.6 times as deep as that in specimen A (101×). An additional 31 potential Hom/Het sites were found in seven small genomic regions (length = 2, 7, 11, 29, 461, 1,662, and 2,612 bp), although the remaining 20 (=107-56-31) sites were dispersed over the genome. Statistically, half of the neutral heterozygous sites/regions are expected to become homozygous in one sibling and heterozygous in the other. Hence, even if the above 107 sites or 87 (=107-20) sites in eight regions actually represented neutral Hom/Het sites, most heterozygous regions in the parental F10 animal could hardly become homozygous by further inbreeding.

As described earlier, a quarter of neutral heterozygous regions in the parental genome are expected to become homozygous in both siblings, and half of them are expected to be different from each other. There was only one such site at nucleotide position 5,250 of scaffold KHL81 ("G" in specimen A and "C" in specimen B). Direct sequencing of the amplicon of this region of the ancestral F9 genome revealed a homozygous C, but showed no indication of G (fig. 2). Therefore, the G nucleotide most likely occurred by mutation between the F9 and F11 generations. This observation again supports the concept that there are few neutral heterozygous sites that could become homozygous by further inbreeding, and that most heterozygous sites found in the inbred strain will be maintained by further inbreeding.

A 2.4-Mb Region Containing a Self-Incompatibility Locus Is Retained Heterozygous

Although the above analysis showed that most neutral sites became homozygous in the F11 genome, sequence coverage and distribution of heterozygous sites of specimens A and B in nonoverlapping 100-kb chromosomal windows showed that an approximately 2.4-Mb region on chromosome 7 is highly heterozygous (fig. 3A and supplementary fig. S1A, Supplementary Material online). This region contained over 40% of the heterozygous sites we identified, and encoded more than 300 genes. Because the result in the preceding section predicts that this region will not easily become homozygous by further inbreeding, and because the recombination rate in *C. intestinalis* is estimated to be 25–49 kb/cM (Kano et al. 2006), this heterozygous region likely lacks recombination.

We confirmed the high heterogeneity of this 2.4-Mb region on chromosome 7 by sequencing an amplicon of a 783-bp long genomic region of the F4–F9 genomes. These genomes all contained 19 heterozygous sites (supplementary table S1, Supplementary Material online).

The detailed view of the 2.4-Mb region of specimen A in nonoverlapping 10-kb windows showed that there were two peaks of heterozygosity at the 1,840,001–1,850,000-th and 1,920,001–1,930,000-th positions (fig. 3B). This and a still more detailed view of the regions containing these two peaks in nonoverlapping 1-kb windows (fig. 3C) showed that the peak regions encoded a self-incompatibility locus, *Themis-B* (in the reference sequence, there are three *Themis*-B genes). Figure 3C also shows that the observed frequency of heterozygous sites of regions around the *Themis-B* locus, particularly the intervening region, is higher than the mean heterozygosity rate in natural populations (1.1–1.2%; Dehal et al. 2002; Satou et al. 2012) and in more distant regions.

A hypervariable region is contained in the N-terminal end of s-Themis-B (Harada et al. 2008). Although no sequence tags were mapped onto the hypervariable regions of the second and third copies of *Themis-B* genes (middle and right genes in fig. 3C), a considerable number of sequence tags were

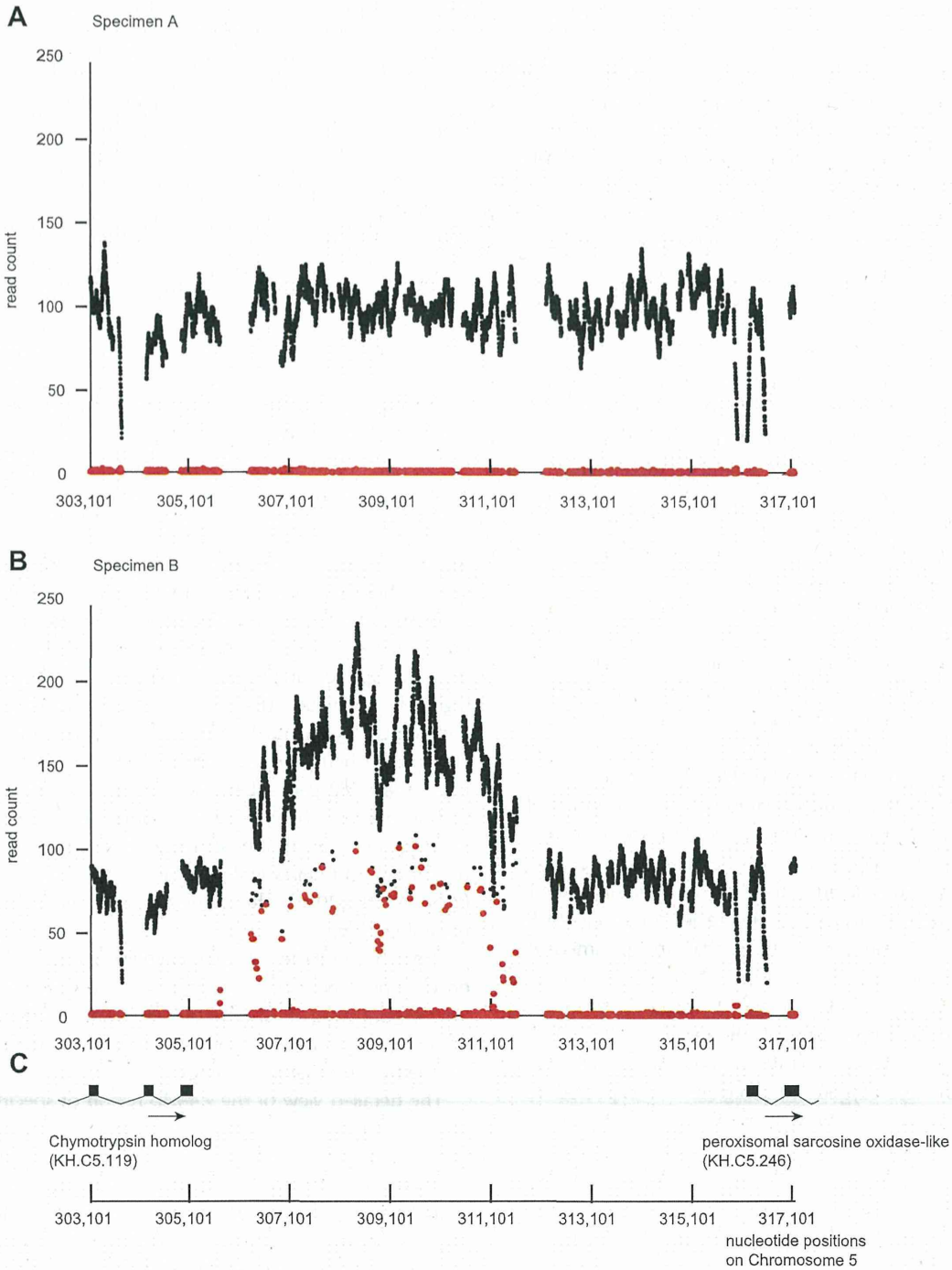

Fig. 1. The number of sequence tags aligned to the region from the 303,101-th to 317,100-th positions on chromosome 5. At each position, the number of the nucleotide that was most frequently found is shown by black dots, and the number of secondarily most frequently identified nucleotide is shown by red dots. Regions where no sequence tags are aligned have no dots. Two siblings are shown in (*A*) and (*B*), respectively. Two genes encoded in this region are indicated in (*C*). Exons are shown by black boxes and introns are shown by thin lines.

mapped onto the hypervariable region of the first copy (left gene in fig. 3C). This *Themis-B* allele appears to be encoded in either haplotype of the inbred strain, because no heterozygous nucleotides were found and the sequence coverage was almost half the average (fig. 3C). Careful manual inspection of Roche 454 tags succeeded in identifying four different alleles of *Themis-B*, which we will further confirm in the following section. In contrast, a substantial number of sequence tags were mapped onto the constant regions of the three *Themis-B* genes. Therefore, the inbred strain likely had multiple copies of *Themis-B*, and sequence tags for their hypervariable regions were not mapped because of their variability.

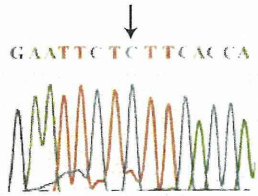

FIG. 2. A chromatogram showing homozygosity of the 5,250-th nucleotide position of scaffold KHL81 of the F9 animal. Although the deep-sequencing data indicated that this position is homozygous G in specimen A and homozygous C in specimen B, direct sequencing of a genomic fragments amplified from the ancestral F9 genome revealed that this position is actually C (an arrow).

In contrast, the genomic region containing the other self-incompatibility locus, *Themis-A*, did not show high heterogeneity (supplementary fig. S1B, Supplementary Material online). In addition, deep inspection of Roche 454 tags succeeded in identifying one allele but not multiple alleles. To confirm this observation, we amplified a genomic fragment containing the hypervariable region of *Themis-A* using a primer for a constant region of the *s-Themis-A* gene with a primer for the upstream region of *Themis-A* from the F9 genome. We obtained only one allele, which was the same as that identified by deep inspection of Roche 454 tags (supplementary fig. S2, Supplementary Material online). Although we could not determine whether the *Themis-A* locus in the F0 genome was homozygous, the above results strongly indicate that the *Themis-A* locus is homozygous in the inbred strain.

Structural Variations of the Self-Incompatibility Locus

The above observation raised the possibility that the *Themis-B* locus, which is thought to be maintained heterozygous, is responsible for the long heterozygous region. Although a previous study speculated that multiple copies of *Themis-B* are artifacts caused by misassembly of the genome (Harada et al. 2008), the above observation implied that multiple copies were indeed encoded in each haplotype of the inbred line.

To confirm that the inbred line has multiple copies of the *Themis-B* locus, we amplified a genomic fragment containing the hypervariable regions between alleles of *s-Themis-B* and *v-Themis-B* using a primer for a constant region of the *s-Themis-B* gene with a primer for the v-*Themis-B* gene. We obtained four different genomic fragments from the F9 genome and determined the nucleotide sequences of them individually (alleles *A*, *B*, *C*, and *D* in supplementary figs. S3–S5, Supplementary Material online). Allele *B* was identical to the allele encoded in the first copy of the reference genome sequence. Thus, the inbred animals had multiple copies of the *Themis-B* locus per haplotype.

To further confirm this result, we determined the copy number of *s-Themis-B* in the genome of the F9 animal using quantitative polymerase chain reaction (qPCR). Under the assumption that two copies of a control *Macho-1* gene are encoded per diploid, the qPCR assay indicated that six copies of *s-Themis-B*, and two copies of another control gene, *FoxA-a*, are encoded in the F9 diploid genome (fig. 4A). Thus, our assay found six copies of four different *Themis-B* alleles in the F9 genome.

Next, we examined the genomes of seven wild animals (wt1–wt7) and found that these animals carried two to ten copies of the *s-Themis-B* gene (fig. 4A). By genotyping of the *Themis-B* locus in the wt1 and wt2 genomes, which have five and seven copies of *Themis-B*, respectively, we identified three (alleles *D–F*) and two (*A* and C) alleles of *Themis-B* (fig. 4B and supplementary figs. S3–S5, Supplementary Material online). Therefore, multiple copies of multiple *Themis-B* variants are likely encoded in these genomes, and are not specific to the inbred strain.

The self-Fertilization Rate in Association with Structural Variations of the Self-Incompatibility Locus

Multiple copies of *Themis-B* imply that the self-incompatibility system is not as simple as previously proposed. Indeed, 79% of eggs obtained from the F9 animal (884 of 1,117) were self-fertilized. We reasoned that structural variations of this locus should be related to the self-fertilization rate, if the *Themis-B* locus is responsible for formation of the 2.4-Mb heterozygous region. For this purpose, we determined the rate of self-fertilization in association with copy numbers and nucleotide sequences of alleles of *Themis-B* in nine different wild type animals (wt8–wt16) (fig. 4B and C and supplementary figs. S3–S5, Supplementary Material online). We identified only one allele (allele *A*) in wt16, the eggs of which were almost completely self-fertilized. This observation indicates that the pair of s-Themis-B and v-Themis-B encoded in this allele does not effectively block self-fertilization. On the other hand, s-Themis-B and v-Themis-B proteins encoded in allele *H*, which was the only allele found in wt10, apparently reject self-fertilization efficiently (fig. 4B and C). This allele *H* is most similar to allele F; s-Themis-B and v-Themis-B proteins encoded in these two alleles were 90% and 81% identical, respectively (supplementary figs. S4 and S5, Supplementary Material online). Two individuals with allele F (wt8 and wt9) showed the lowest self-fertilization rates (fig. 4B and C). Thus, different variants likely have different reactivity, and self-fertility and genotypes of this locus appear correlated.

The high self-fertilization rate of wt15 indicates that alleles *D*, *I*, and *J* do not react very efficiently (fig. 4B and C); if either of these three alleles reacted efficiently, the self-incompatibility reaction would take place (Saito et al. 2012), and no self-fertilization would occur. Wt11 had only alleles *I* and *J*, and showed a lower self-fertilization rate than wt15 (fig. 4B and C). Wt11 had a smaller number of *Themis-B* copies and a smaller number of variants, which might increase the likelihood that a specific type of s-Themis-B expressed in a sperm cell would encounter its counterpart (v-Themis-B) on the egg's vitelline coat. This hypothesis is also supported by the following observation. The self-fertilization rate of wt12, which had three variants including alleles *I* and *J*, was higher than that of wt11, but lower than that of wt15 (fig. 4B and C), whereas the copy number was larger than that of wt11 and smaller than that of wt15. Although we cannot rule out the possibility that some alleles of the *Themis-A* self-incompatibility locus did also not

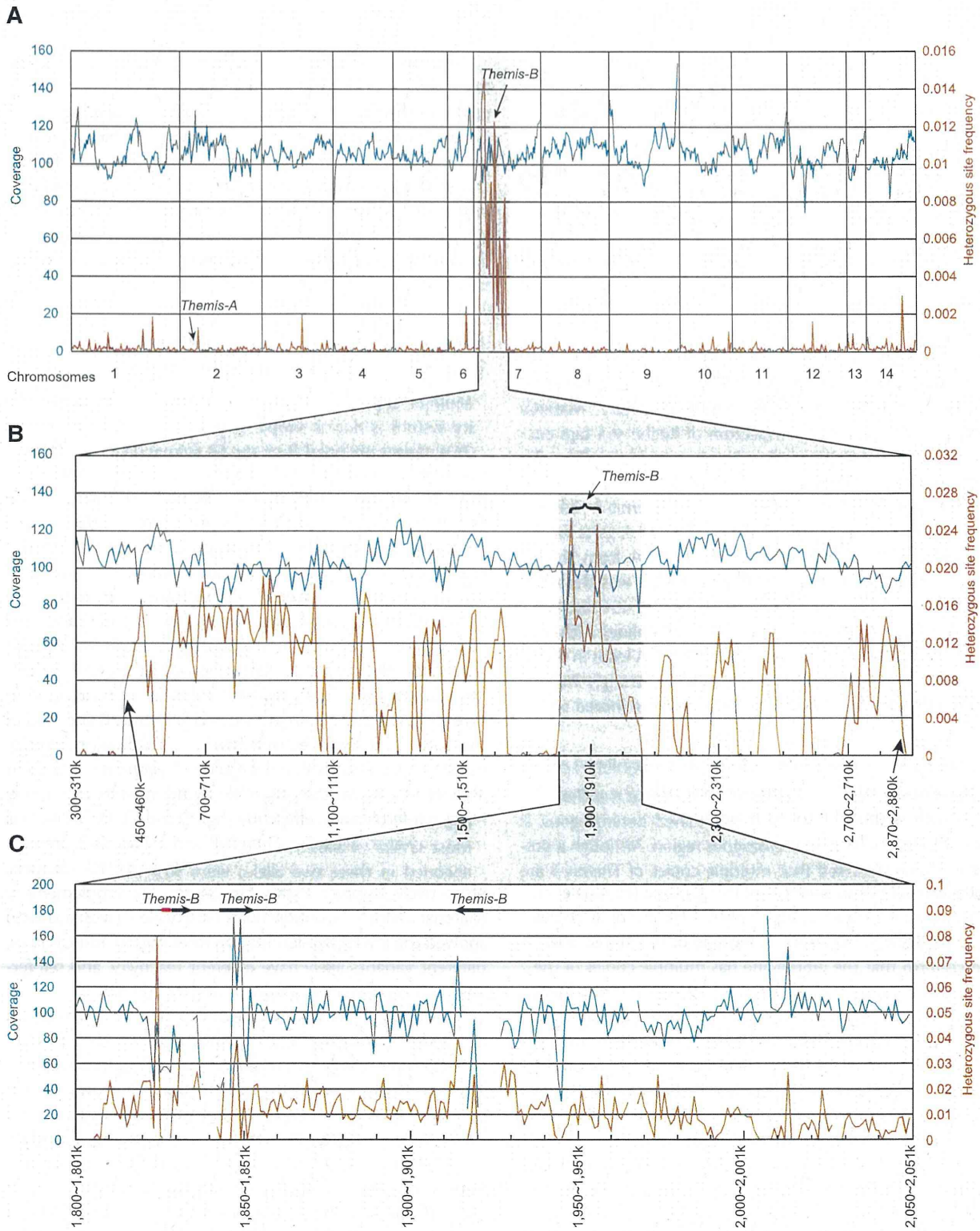

FIG. 3. A chromosomal distribution for heterozygosity in specimen A. (*A*–*C*) Sequence coverage (blue lines) and frequencies of heterozygous sites (red lines) (*A*) in nonoverlapping 100-kb chromosomal windows across the 14 chromosomes, which comprise 68% of the current assembly, (*B*) in nonoverlapping 10-kb windows across the highly heterozygous region on chromosome 7, and (*C*) in nonoverlapping 1-kb windows across the region encoding the *Themis-B* locus. Positions of *Themis-A* and *Themis-B* are shown by arrows in (*A*) and (*B*). In (*C*), the constant regions of three copies of *Themis-B* are indicated by with black arrows. Arrows shows directions of s-*Themis-B* on the genome. The hypervariable region of the first (left) copy of *Themis-B* is shown with a red line.

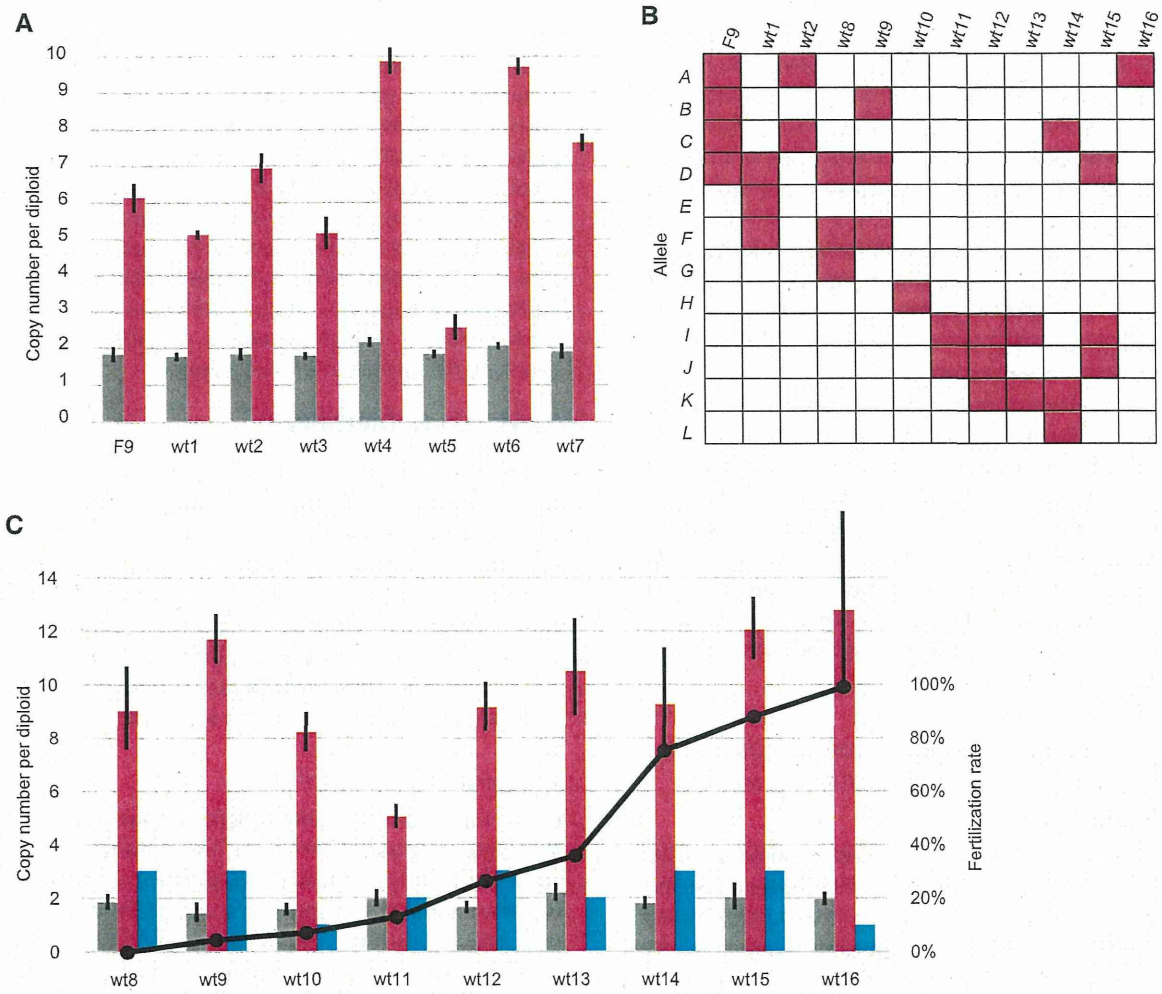

Fig. 4. Allele variants, copy number variation, and variant number variation of the *Themis-B* locus may affect the rate of self-fertilization. (*A*) Copy numbers of *FoxA-a* (gray bars) and *s-Themis-B* (red bars) in eight animals including the F9 inbred animal. Error bars indicate standard errors among triplicates. (*B*) Alleles found in the present study are indicated by red boxes. (*C*) In addition to the copy numbers of *FoxA-a* (gray bars) and *s-Themis-B* (red bars) in nine animals, numbers of *Themis-B* variants identified by PCR cloning (blue bars) and the rate of self-fertilization (a black line) are shown. Error bars on gray and red bars indicate standard errors among triplicates.

react efficiently in wt8–wt16, the above observations strongly indicate that the self-fertilization efficiency is affected by different reactivity among variants, copy number, and allele number of *Themis-B*.

Discussion

Establishment of Inbred Lines

Because of the simplicity and compactness of the genome, as well as simple development, *Ciona* is widely used for experimental biology including developmental and evolutionary studies. Nevertheless, no inbred *Ciona* strain has yet been made available, and animals from natural populations are typically used for experiments. Because this animal has a large gene pool (Satou et al. 2012), genetic variations might often have affected reproduction of results obtained from different animals. Deep genomic sequencing of the inbred strain established here has shown that this strain possesses only 0.04% of heterozygous nucleotides, which is much smaller than the nucleotide diversity rate in natural populations (1.1–1.2%). In addition, decoding of the genomes of two siblings predicted that there are few neutral sites that could potentially become homozygous by further inbreeding. Thus, this strain can now be used as a new biological resource.

The Self-Incompatibility Locus *Themis-B*

We identified 12 *Themis-B* alleles in the genomes of the F9 and eleven wild animals, which were derived from animals caught in Onagawa Bay. Eight of these alleles were shared among multiple individuals. In addition, the *B* allele is found in the reference genome, which is derived from a specimen caught on West Coast of the United States (Dehal et al. 2002). These observations imply that this animal has a limited number of variants of the self-incompatibility gene in natural populations.

The number of variants appears much smaller than the theoretical expectation based on population sizes (Wright

1939; Kimura and Crow 1964); 60–90 variants are expected for even a small population of 10,000 individuals. However, our results strongly indicated that different reactivity among variants, copy number variation, and variant number variation all influence self-fertilization efficiency, and a combination of these factors probably increases the functional complexity of this self-incompatibility locus. At the same time, this system makes self-incompatibility between eggs and sperm of *Ciona* incomplete, such that outcrossing occurs more frequently under natural conditions, where self-sperm and nonself-sperm are available, and self-fertilization occurs in laboratory conditions.

A half of offspring of our inbred line are expected to be homozygous for *Themis-B*. The observation that this locus is maintained heterozygous indicates that the fertility of animals homozygous for *Themis-B* is lower than the fertility of heterozygous animals, and animals homozygous for *Themis-B* rarely generate offspring. In this sense, this incomplete self-incompatibility locus has strong heterozygous advantage in the inbred strain, and likely causes inbreeding depression, which is widely observed in offspring of related individuals (Charlesworth and Willis 2009).

The Large Heterozygous Region Might Represent a Preliminary Stage of Structural Differentiation of Chromosomes

The recombination rate in *C. intestinalis* is estimated to be 25–49 kb/cM (Kano et al. 2006), and three *Themis-B* genes in the reference genome are encoded within approximately 100 kb. Therefore, the 2.4-Mb heterozygous region in inbred animals strongly suggests that recombination rarely occurs in this region, and it seems likely that this recombination suppression is related to the *Themis-B* locus. First, *Themis-B* is retained heterozygous. Previous studies have shown that large genomic regions segregate together with self-incompatibility or mating type loci in other organisms (Boyes et al. 1997; Lahn and Page 1999; Ferris et al. 2002; Lengeler et al. 2002; Liu et al. 2004). Second, any recombination within the hypervariable region of *Themis-B* across alleles would probably impair their function, because v-Themis and s-Themis encoded in each allele specifically react with each other and do not react with those encoded in other alleles. Third, the high degree of nucleotide variability and copy number variations should physically suppress recombination at this locus. Fourth, it is possible that different haplotypes have different arrangements of multiple *Themis-B* copies, including inversions. Such variation may be directly related to recombination suppression in the large regions flanking the *Themis-B* locus, although this hypothesis remains to be tested. Finally, the observation that the frequency of heterozygous sites in the genomic region between the second and third *Themis-B* genes was higher may suggest that this region has started to diverge.

Intervening homozygous regions within the 2.4-Mb heterozygous region (fig. 3*B*), however, indicate that recombination is not completely suppressed. There may be several loci with heterozygous advantage in the 2.4-Mb heterozygous region, although we could not find candidate genes with heterozygous advantage other than *Themis-B*. Linked deleterious mutations might also confer heterozygous advantage. Because the observed frequency of heterozygous sites of regions close to the *Themis-B* locus is higher than those of more distant regions within the 2.4-Mb heterozygous region, genes or genomic regions with heterozygous advantage in distant regions and modifiers of recombination rates would have been acquired under the linkage disequilibrium caused by the *Themis-B* locus. Thus, the 2.4-Mb heterozygous region of chromosome 7 might represent a preliminary stage of structural differentiation of chromosomes.

Materials and Methods

Inbreeding and Sequence Data

A single individual originating from a natural population in Onagawa, Miyagi Prefecture, Japan was chosen as the F0 animal. Although *C. intestinalis* is a hermaphroditic and outcrossing is its normal reproductive mode, self-fertilization occurs under laboratory conditions. Each generation, offspring were obtained by self-crossing. DNA extracted from sperm cells of two F11 animals were individually sequenced with illumina GA2 and Roche 454 sequencers.

Mapping and Genotype Calling

We first trimmed low quality regions (quality values < 25) from sequencing tags with a "TrimmingReads.pl" script (Patel and Jain 2012). Obtained tags were mapped onto the reference genome sequence using ssaha2 (Ning et al. 2001) with default parameters for illumina tags and Roche 454 tags. We used nucleotide positions in nonrepeat regions that were covered by 20 or more tags because of the error-prone nature of sequence tags. Because nucleotide positions that are too deeply covered might represent repeat sequences, we also excluded nucleotide positions that were covered by more than 203 and 212 tags, which were twice as large as the averaged sequence depths of specimens A and B, respectively (table 1).

Before genotype calling, we screened out repeats in the reference genome sequence (KH version) (Satou et al. 2008). We first took 100-bp sliding windows with 50-bp steps and aligned them onto the reference genome using blat (Kent 2002) with the "fastMap" option. Sequence fragments that were more than 50% identical with multiple regions were regarded as repeats. As a result, 15,119,916 nt were found in repeats and were excluded from the subsequent analyses.

Using the mapped sequence tags, genotypes were called using previously published criteria (Harismendy et al. 2009; Nielsen et al. 2011). We called positions with common allele read frequencies over 80% as homozygous, and positions with secondarily common allele read frequencies over 20% as heterozygous, unless ternary common allele read frequencies were over 20%.

PCR Genotyping

We amplified a genomic fragment shown in supplementary table S1, Supplementary Material online, from the F4 to F9