genes

MDPI

Article

Evolutionary History of the Risk of SNPs for Diffuse-Type Gastric Cancer in the Japanese Population

Risa L. Iwasaki [1], Koji Ishiya [2], Hideaki Kanzawa-Kiriyama [3], Yosuke Kawai [4], Jun Gojobori [1] and Yoko Satta [1,*]

1 Department of Evolutionary Studies of Biosystems, SOKENDAI (The Graduate University for Advanced Studies), Kanagawa 240-0193, Japan; iwasaki_risa@soken.ac.jp (R.L.I.); gojobori_jun@soken.ac.jp (J.G.)
2 Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Sapporo 062-8517, Japan; koji.ishiya@aist.go.jp
3 Department of Anthropology, National Museum of Nature and Science, Ibaraki 305-0005, Japan; hkanzawa@kahaku.go.jp
4 Genome Medical Science Project, National Center for Global Health and Medicine, Tokyo 162-8655, Japan; ykawai@ri.ncgm.go.jp
* Correspondence: satta@soken.ac.jp; Tel.: +81-46-858-1574

Received: 11 June 2020; Accepted: 8 July 2020; Published: 10 July 2020

Abstract: A genome wide association study reported that the T allele of rs2294008 in a cancer-related gene, *PSCA*, is a risk allele for diffuse-type gastric cancer. This allele has the highest frequency (0.63) in Japanese in Tokyo (JPT) among 26 populations in the 1000 Genomes Project database. $F_{ST} \approx 0.26$ at this single nucleotide polymorphism is one of the highest between JPT and the genetically close Han Chinese in Beijing (CHB). To understand the evolutionary history of the alleles in *PSCA*, we addressed: (i) whether the C non-risk allele at rs2294008 is under positive selection, and (ii) why the mainland Japanese population has a higher T allele frequency than other populations. We found that haplotypes harboring the C allele are composed of two subhaplotypes. We detected that positive selection on both subhaplotypes has occurred in the East Asian lineage. However, the selection on one of the subhaplotypes in JPT seems to have been relaxed or ceased after divergence from the continental population; this may have caused the elevation of T allele frequency. Based on simulations under the dual structure model (a specific demography for the Japanese) and phylogenetic analysis with ancient DNA, the T allele at rs2294008 might have had high frequency in the Jomon people (one of the ancestral populations of the modern Japanese); this may explain the high T allele frequency in the extant Japanese.

Keywords: genetic differentiation; PSCA; diffuse-type gastric cancer; rs2294008; Japanese; dual-structure model; Jomon people; Yayoi people; soft sweep; hardening selective sweep; selection target change

1. Introduction

Gastric cancer (GC) has a high incidence rate in East Asia and is the third leading cause of cancer death in the world [1]. GC is classified into two types, the diffuse type and the intestinal type [2]. The diffuse type of GC (DGC) is reported as geographically uniformly distributed [3] and the incident rate of DGC has been increasing compared with the intestinal type of GC [4,5]. A previous genome-wide association study identified *PSCA* (prostate stem cell antigen) on chromosome 8 as a susceptible gene for DGC in the Japanese [6]. The study analyzed 925 DGC cases with 1396 controls and reported a strong association with the T allele at rs2294008 in *PSCA*. This association has also been

observed in other populations across the world including Koreans [7,8], Uzbekistanis [9], and even Caucasians [10–12].

rs2294008 is located in the upstream region of *PSCA*, and a T to C transition at this single nucleotide polymorphism (SNP) causes a missense mutation (ATG to ACG) at the presumed translation-initiating codon in the first exon, resulting in a nine-amino-acid truncation of its signal peptide. Tanikawa et al. [13] showed that the T/C variation caused differences in the subcellular localization and stability of PSCA proteins. The longer PSCA containing the risk (T) allele is presumed to be involved in cell proliferation as a signal protein, whereas short PSCA including the non-risk (C) allele downregulates PSCA expression with its protein degradation and subsequent activation of immune responses by antigen presentation [13]. They suggested that individuals with the T allele might have higher risk of GC, including DGC, as a result of accelerated cell proliferation due to suppressed PSCA protein degradation. Additionally, RT-PCR and immunohistochemical analysis revealed that the PSCA expression level is reduced in DGC cells [6]. Luciferase reporter assays also revealed that the T allele of rs2294008 reduced the transcriptional activity of the *PSCA* gene [6]. These findings support the conclusion that the T allele at the rs2294008 SNP can modulate PSCA transcriptional activity and protein functionality.

The frequency of this risk allele of DGC is the highest (0.630) in the Japanese in Tokyo (JPT) among the 26 worldwide populations in the 1000 Genomes Project (1KGP) [14]. Notably, other populations in East Asia showed lower frequencies of the T allele than that in JPT; e.g., 0.248 in Han Chinese in Beijing (CHB) [14], 0.260 in Taiwanese [15], and 0.497 in Koreans [16]. However, it is not clear why the frequency of this risk T allele has been maintained at high frequency only in JPT. In the present study, we explore the evolutionary forces (genetic drift or natural selection) that may be responsible for this high frequency in JPT through: (i) the examination of the signature of natural selection operating on either the T or C alleles at the rs2294008 SNP, (ii) exploring the evolutionary trajectory of the risk allele through computer simulations, and (iii) a phylogenic analysis of ancient genome sequence data from one of the ancestral populations of the extant Japanese population, the Jomon people.

2. Materials and Methods

2.1. Strategy of Analysis

First, using the F_{ST} value between JPT and CHB, we evaluated the extent of T allele frequency difference at rs2294008 compared with genome-wide SNPs. Then, we examined whether the large allele frequency difference could be explained by positive selection on either allele using multiple neutrality tests. We detected a signature of positive selection on the C allele in CHB and then addressed why the non-positively selected T allele could have been maintained at a high frequency in JPT. We tested whether demographic events of JPT could attain the high F_{ST} value at rs2294008 using allele frequency simulation. Moreover, to confirm this demographic effect, using ancient genome sequences of the Jomon people, we investigated their phylogenetic relationship with the extant JPT/CHB.

2.2. Human SNP Data

We retrieved variant call format (VCF) data from two local populations (JPT, CHB) and four metapopulations (East Asian excluding JPT and CHB (non JPT/CHB EAS); European (EUR); South Asian (SAS); African (AFR)) from 1KGP Phase 3 [14] and used the GRCh37 genomic positions (or coordinates) for each SNP. The ancestral/derived alleles are defined in 1KGP [14].

2.3. F_{ST}

We estimated differentiation level by Hudson's F_{ST} [17,18] and examined the relative extent of differentiation at rs2294008 among genome-wide SNPs between JPT and other East Asian populations such as CHB (Table S1). For this calculation, we removed indels, coordinated/overlapped SNPs, and multi-allelic SNPs. The *MHC* region (chr6: 25,726,291-33,368,333) was also removed because it contained many genes

governed by different evolutionary mechanisms such as balancing selection [19]. We drew a Manhattan plot based on the F_{ST} values across the entire genome using the program qqman [20].

2.4. Linkage Disequilibrium Analysis

Linkage disequilibrium (LD) between rs2294008 and other SNPs in flanking regions was examined by D′ [21]. In this calculation, SNPs with minor allele frequency (MAF) < 0.05 were removed and the LD block was determined with the software Haploview (v 4.2) [22] using the criterion of D′ > 0.98 [23].

2.5. Neutrality Tests

2.5.1. Haplotype-Based Tests (EHH, nS_L, and H12)

We performed three haplotype-based tests, extended haplotype homozygosity (EHH), number of segregating sites by length (nS_L), and haplotype homozygosity (H12), for detecting recent or ongoing positive selection on rs2294008 [24–26]. We calculated the EHH for a 400 kb region (chr8: 143,563,622-143,963,622) surrounding rs2294008. Both alleles at rs2294008 (T/C) were used as a single core site for the EHH test. SNPs with MAF < 0.05 were removed in the test. The nS_L (implemented in selscan [27] (v 1.1.0)) was measured as haplotype length linked with a derived C allele based on the number of segregating sites in a haplotype, and not based on a recombination map. We used 303,438 and 377,335 bi-allelic SNPs with MAF > 0.01 on chromosome 8 in JPT and CHB, respectively. We set three maximum window sizes for this analysis: 100 SNPs (default), 1500 SNPs, and no size limitation. H12 detects hard and soft sweeps with homozygosity based on the frequency of the two most common haplotypes [26]. We scanned chromosome 8 using sliding windows of 125 SNPs (for JPT) or 124 SNPs (for CHB), which were equivalent to the number of SNPs in the 21.9 kb LD block (chr8: 143,752,235–143,774,193) in each population, and then calculated H12.

2.5.2. Site Frequency Spectrum-Based Tests (Tajima's D and Fay and Wu's H)

To investigate the signatures of natural selection across the 21.9 kb LD block, we also performed two site frequency spectrum-based tests, Tajima's *D* [28] and normalized Fay and Wu's *H* [29,30], and calculated p-values using DnaSP (v 6.0) [31]. All gaps and sites for which the ancestral alleles are unknown were excluded from the analysis.

2.5.3. Nucleotide Diversity (π)

To evaluate the reduction of genetic diversity due to putative selective sweep, we examined three regions and, for each region, calculated the nucleotide diversity [32] π_C [31] of C haplotypes or π_T of T haplotypes in JPT and CHB separately using DnaSP. The three regions were: (i) the 21.9 kb LD block (chr8: 143,752,235–143,774,193), and its (ii) upstream (chr8: 143,652,235–143,752,234) and (iii) downstream 100 kb flanking regions (chr8: 143,774,194–143,874,193). The number of C haplotype sequences in JPT and CHB was 77 and 155, respectively, and that of T haplotypes was 131 and 49, respectively. In order to avoid the overestimation of a π value caused by the inclusion of a recombinant between C and T haplotypes, we removed two possible recombined haplotypes in CHB. We could not detect any recombined haplotypes in JPT. We then used z tests to compare the π_T or π_C values of the 21.9 kb LD region between JPT and CHB as well as the π_T or π_C values of the LD region with that of flanking 100 kb regions within each population. For the application of false discovery rate (FDR), we corrected the *p*-values of each summary statistic into *q*-values with the Benjamini–Hochberg procedure [33]. Furthermore, we compared the nucleotide diversity of two subhaplotypes, C-A and A-G (see below for the definition of subhaplotypes).

2.5.4. Two-Dimensional Site Frequency Spectrum (2D SFS)

We evaluated intra allelic variability (IAV) within a derived allele (D) group using 2D SFS ($\Phi_{i,j}$ [34]) and its summary statistics, F_c, G_{c0}, and L_{c0} [34,35]. The latter three summary statistics were designed

to detect incomplete selective sweep signals when compared to the null distribution of each summary statistic. Each null distribution was generated under a neutral model using ms [36] with at least 1000 replications under the same demographic parameters used previously [35,37]. For the application of FDR, we applied the Benjamini–Hochberg procedure to the 2D SFS statistics [33]. We judged cases as presenting ongoing positive selection if at least two summary statistics were statistically significant (q-value < 0.05). We tested for selective sweep with rs2294008 in a core region of chr8: 143,755,915–143,770,914 (15 kb) or chr8: 143,755,876–143,771,875 (16 kb) in JPT and CHB, respectively. Each core region was determined by $r^2 > 0.75$ with rs2294008, following prior criteria [34]. In addition, we evaluated the IAV of two subhaplotypes of C within a population. A subhaplotype was defined as the combination of alleles at two sites (rs2976391 (C/A) and rs2978983 (A/G)) linked with C at rs2249008, and named C-A and A-G subhaplotypes, respectively. We evaluated the lower limit of the start of positive selection for the two subhaplotypes, using a mutation rate of 0.5×10^{-9}/site /year [34].

2.5.5. Application of Population Branch Statistics (PBS)

We examined the signal of local adaptation acting on the C allele in CHB and JPT using PBS analysis, which is one of the F_{ST}-based summary statistics. We selected sites for calculating F_{ST} following previously described criteria [38]. Then, using EUR as an outgroup population, we calculated Hudson's F_{ST} per SNP site pairwise among JPT, CHB, and EUR. We calculated PBS values for genome-wide SNPs of JPT and CHB following previous methods [38,39]. Then, we compared the ranking of PBS of rs2294008 in JPT and that in CHB.

2.6. Forward Simulation Using Japanese Demographic Model

We performed forward simulation of allele frequency trajectory at the SNP (T/C) under the Wright–Fisher model [40] of haploidy. This simulation analysis has two conditions under a demographic model (see below, Figure S1): (i) both T and C alleles are neutral, and (ii) the C allele is positively selected only in simulated CHB lineages.

2.6.1. Demographic Parameters in Simulations

The demographic "dual structure model" [41] proposes that the extant genetic and phenotypic diversity in the Japanese populations were caused by the past admixture of two genetically different populations, the Jomon and immigrant Yayoi farmers. The Jomon people are indigenous and inhabited the Japanese archipelago from at least 16,000 years ago (ya) [42], whereas the immigrant Yayoi farmers originated from the Asian continent [43] and migrated to the Japanese archipelago 2500 ya [42] and then admixed with the Jomon people. This model has been supported by previous studies based on ancient or extant genomic data from the Japanese [44,45].

From a common ancestor, the simulated Jomon people (JMN) and simulated ancestral population of the Asian continent (A_CNT) diverged and were isolated for *t1* generations. After this separation, these two populations admixed with a rate r of the JMN component. After *t2* generations of admixture, we calculated Hudson's F_{ST} [17] between simulated JPT, who are descendants of the admixed population, and simulated CHB, who are those of A_CNT. No migration between simulated JPT and simulated CHB was assumed throughout the simulation based on previous studies [43–46].

In the above simulation, we used four variable parameters and three constants. The first parameter was *t1*, and we used five *t1* values (875, 1125, 1375, 1875, and 2375 generations), corresponding to 17,500 to 47,500 years [42,44]. Generation time is assumed to be 20~30 years [47], and we used a shorter generation time to be conservative regarding allele frequency changes in the simulations. The second parameter was r for the admixture proportion of JMN; we used three values: 0.4, 0.2, and 0.1 [43–46]. The third and fourth parameters were the population size for JMN (N_{JMN}) and A_CNT (N_{A_CNT}), respectively. For N_{JMN}, we used 500, 1000, 2000, 4000, 8000, 10,000, 12,500, and 15,000, whereas for N_{A_CNT}, we used 1000, 2000, 4000, 8000, 16,000, 25,000, and 30,000. We added the condition that N_{JMN} was always equal to or smaller than N_{A_CNT}. In contrast, three constants were used: *t2* = 125

generations [42], N_{simJPT} = 12,824 and N_{simCHB} = 29,204 [44]. The number of combinations of these parameters was 570 and each combination was simulated 10,000 times for each initial frequency of T allele (f_i) in the ancestral population, ranging from 0.1 to 0.9 with increments of 0.1. For simulations incorporating selection, we used one additional parameter, the selection coefficient (s). We used $2 \times N_{simCHB} \times s$ = 1, 10, 50, or 100. Selection occurred on the CHB lineage for 126 generations (the time at and after admixture between A_CNT and JMN).

2.6.2. Investigating Possible Causes of Large F_{ST}

For each parameter combination, we drew a histogram for 90,000 simulated F_{ST} values under neutrality. These 90,000 F_{ST} values did not include cases where alleles T or C were fixed in both JMN and A_CNT before admixture or fixed in both the simulated JPT and CHB after admixture. Negative F_{ST} was regarded as zero. The simulated F_{ST} histogram was compared with the empirical F_{ST} data. To do so, we needed a proportion (p_i) of nine f_i in the common ancestor of JMN and A_CNT. Each F_{ST} was weighted by the estimated p_i of the empirical derived allele frequency. p_i was calculated from 570,129 SNPs on chromosome 8 in JPT as p_i = 0.5589, 0.0984, 0.0734, 0.0623, 0.0494, 0.0436, 0.0360, 0.0342, and 0.0438 (i = 1~9).

We then examined whether high T allele frequency and high F_{ST} of a particular SNP are simulated under a neutral state. Among 570 combinations, 410 showed simulated F_{ST} distributions that were not significantly different from the empirical one based on the two-sample Kolmogorov–Smirnov test (p-value $\geq$ 0.05) implemented in the Python package scipy (v. 0.19.1) [48]. Then, of these 410 combinations, we further chose combinations in which there was a SNP with a higher F_{ST} and higher T allele frequency than the observed values. We also checked whether these combinations fulfilled the condition that T allele frequency must be higher in JPT than CHB but that neither alleles are fixed in either population. Additionally, we examined cases of positive selection acting on the C allele in the simulated CHB lineage. We reused 410 combinations that could reproduce similar F_{ST} distributions under neutrality and further simulated a total of 1640 combinations (four selection coefficients on the C allele, each for 410 combinations). Then, we counted the number of cases which attained high F_{ST} (>0.2547) and high T allele frequency (>0.62) in simulations under positive selection or neutrality.

2.7. Analysis using Ancient DNA Sequences from the Jomon People

In order to examine whether the extant T haplotype in JPT was derived from those in the Jomon, we used three reported ancient individual genomes of the Jomon people, the Ikawazu (IK002) [49] and two Funadomari (FUN23, FUN5) [46]. We then constructed a median-joining haplotype network together with the extant JPT and CHB sequences using network (v. 5.0.1.1) [50,51]. A chimpanzee orthologous sequence (Pan tro 3.0:8:145391694:145412716), obtained using the extant human sequence of 21.9 kb LD region as a query by nucleotide blast, was used as an outgroup. We chose reliable sites from the Ikawazu and Funadomari Jomon sequences by the following method. Unreliable sites in IK002 were filtered out with the HaplotypeCaller algorithm implemented in GATK (v. 3.8) with gVCF mode (sites with mapping quality < 20 or base quality < 10). Further, we removed sites which were likely to be derived from sequence error and checked the IK002 coverage of each site. We then produced a consensus individual sequence when the coverage of the genome sequence was low [49,52]. We chose sites with depth > 2 and genotype quality > 1. At a bi-allelic site, the allele which had deeper coverage than the other was chosen. The genomes of both samples of Funadomari were phased according to the previous method [46]. Unreliable sites were filtered out with the HaplotypeCaller algorithm implemented in GATK with gVCF mode and were filtered with the variant quality score recalibration (VQSR) approach. All sites examined in this study were restricted to sites listed in 1KGP-Phase 3 SNPs [6]. Furthermore, we chose sites with genotype quality > 30 and depth > 30 for high coverage sample FUN23 and genotype quality > 20 for the low coverage sample FUN5. FUN23 was treated as a diploid sample but FUN5 was treated as a haploid sequence in the same way as IK002 due to low

coverage. Singleton sites which were unique in the three ancient samples (IK002, FUN23, and FUN5) were determined under the infinite site assumption [46] and were removed from sequences.

3. Results

3.1. The T Allele at rs2294008 is Highly Differentiated between JPT and CHB

JPT has high frequency (0.63) of the risk T allele at rs2294008 for DGC (Table S1). We compared the extent of genetic differentiation, F_{ST}, of rs2294008 to all other genomic SNPs between JPT and CHB (total of 14,653,076 SNPs), and found that the F_{ST} of rs2294008 is significantly high at the genome level (F_{ST}= 0.2547, Figure 1, Figure 2, and Table S2).

The rs2294008 SNP was located in a 21.9 kb LD block (chr8: 143,752,235-143,774,193, D' > 0.98) (Figure 3) that was comprised of 125 SNPs in JPT. Most of the top 50 SNPs (49/50) of genome-wide F_{ST} values are located in this LD block (Table S2). We compared the F_{ST} for this block with other blocks of the same size (21.9 kb) on chromosome 8. The block containing rs2294008 showed the highest F_{ST} value among 6,510 blocks ($F_{ST} = 9.8 \times 10^{-2}$, Figure S2). Therefore, we concluded that rs2294008 and its tightly linked SNPs were significantly differentiated between JPT and CHB.

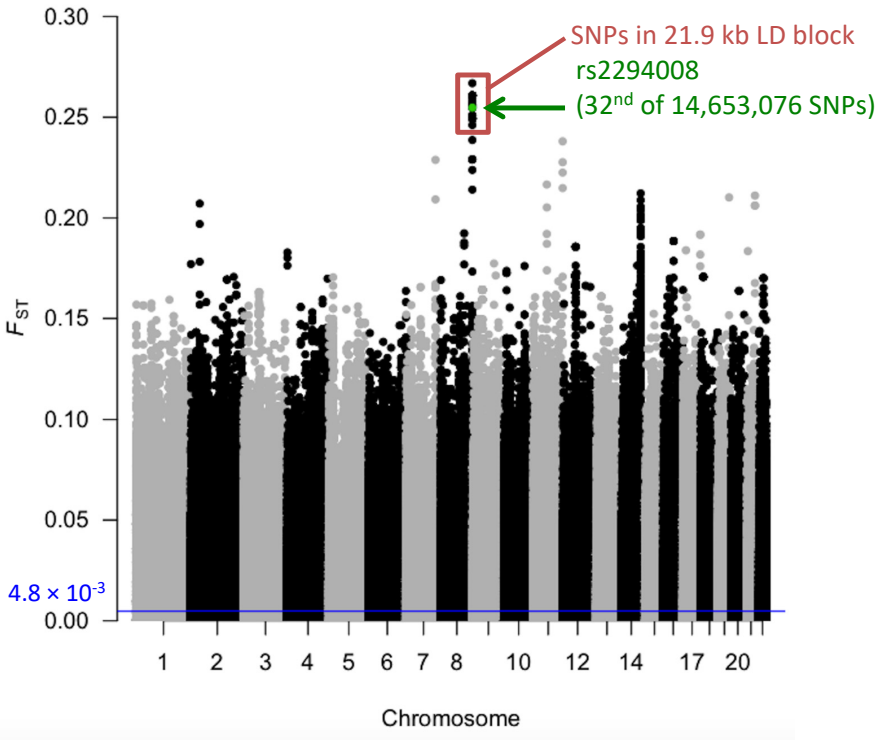

Figure 1. Manhattan plot of F_{ST} values between JPT and CHB. Each dot represents an F_{ST} value of each autosomal SNP. Green dot represents the F_{ST} value at rs2294008 and red box indicates SNPs in the LD block (21.9 kb) containing rs2294008. The average F_{ST} (4.8×10^{-3}) is represented by the blue horizontal line.

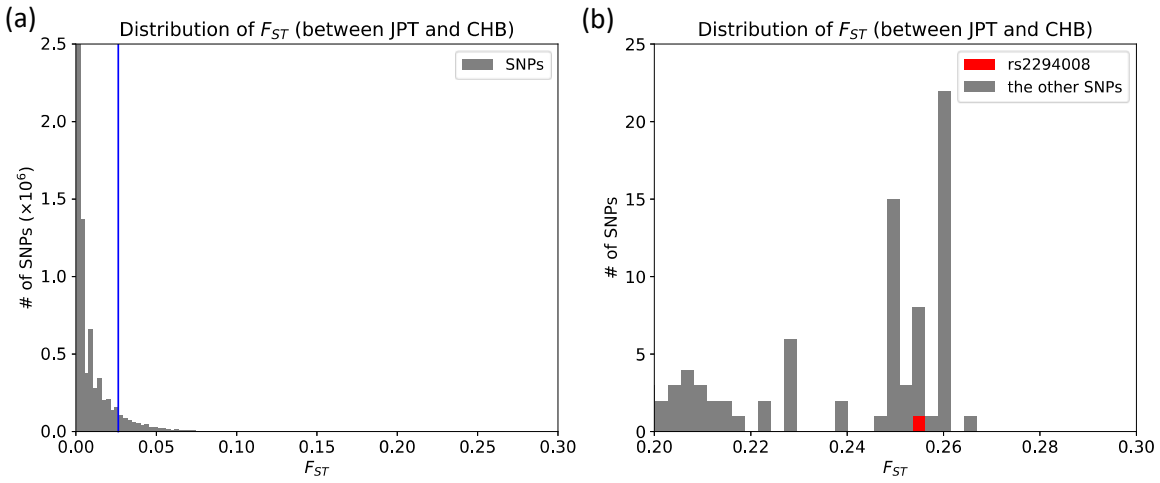

Figure 2. Empirical F_{ST} distribution. (**a**) F_{ST} distribution of the entire genome between JPT and CHB. The blue line represents the fifth percentile of distribution ($F_{ST} \geq 0.0264$); (**b**) Local zoom ($0.20 < F_{ST} < 0.30$) of (**a**). rs2294008 is marked by a red rectangle among 14,653,076 SNPs in the entire genome.

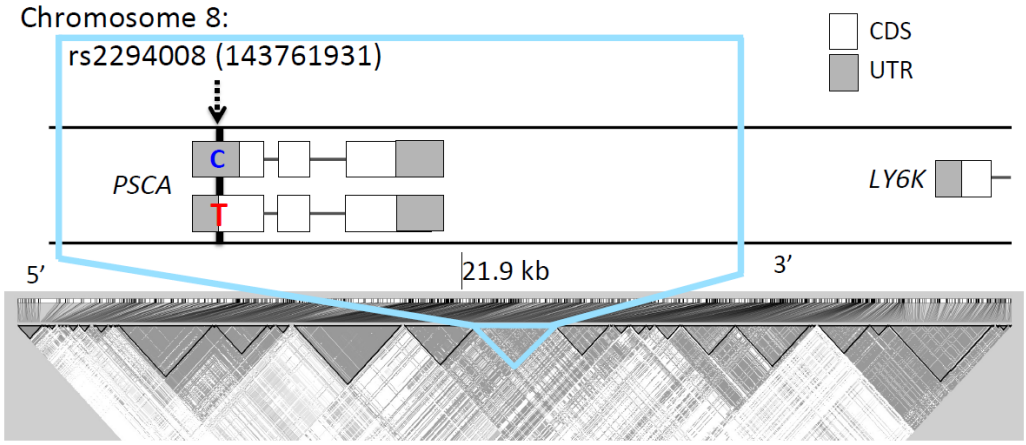

Figure 3. Genomic context (upper) and linkage disequilibrium (lower) of rs2294008. The gray and white squares represent coding sequence (CDS) and untranslated region (UTR) of a gene, respectively. rs2294008 (T/C) is located in a PSCA initiation codon (chr8: 143,761,931). Sequences with the C allele are truncated by nine amino acids compared with sequences with the T allele [13].

3.2. Exploring the Signal of Natural Selection Acting on rs2294008

We found that rs2294008 was one of the most differentiated SNPs between JPT and CHB. This highly differentiated SNP in JPT may be explained by natural selection on the C allele in CHB and/or natural selection on the T allele in JPT. To examine the signature of natural selection, two different types of summary statistics were applied: (i) haplotype-based statistics (EHH, nS_L, and H12) and (ii) site frequency spectrum-based statistics (Tajima's *D*, normalized Fay and Wu's *H*, and PBS). EHH showed no clear signal of selective sweep acting on the T allele at rs2294008 in JPT and CHB (Figure 4a). Moreover, no signal of selective sweep on the C allele was detected in CHB or in JPT (Figure 4a). Similarly, none of the five SNPs with high F_{ST} values (0.2608~0.2669) in the same LD region showed clear signals in either population (Figure S3). Additionally, neither nS_L nor H12 showed any significant signals of hard/soft sweeps on rs2294008 in both JPT and CHB (Figure 4b,c). nS_L of rs2294008 showed an insignificant value, 0.4063 ($p > 0.05$) (Figure 4b). Six of the ten SNPs with high F_{ST} in the 21.9 kb LD block showed a marginally negative nS_L value in CHB ($nS_L = -1.978 \sim -2.106$; $0.01 < p < 0.05$)

(Figure 4b), but the signal for a haplotype containing these SNPs was not clear based on H12 sliding window analysis (H12 = 0.118 ($p > 0.05$) and 0.126 ($p > 0.05$) for JPT and CHB, respectively).

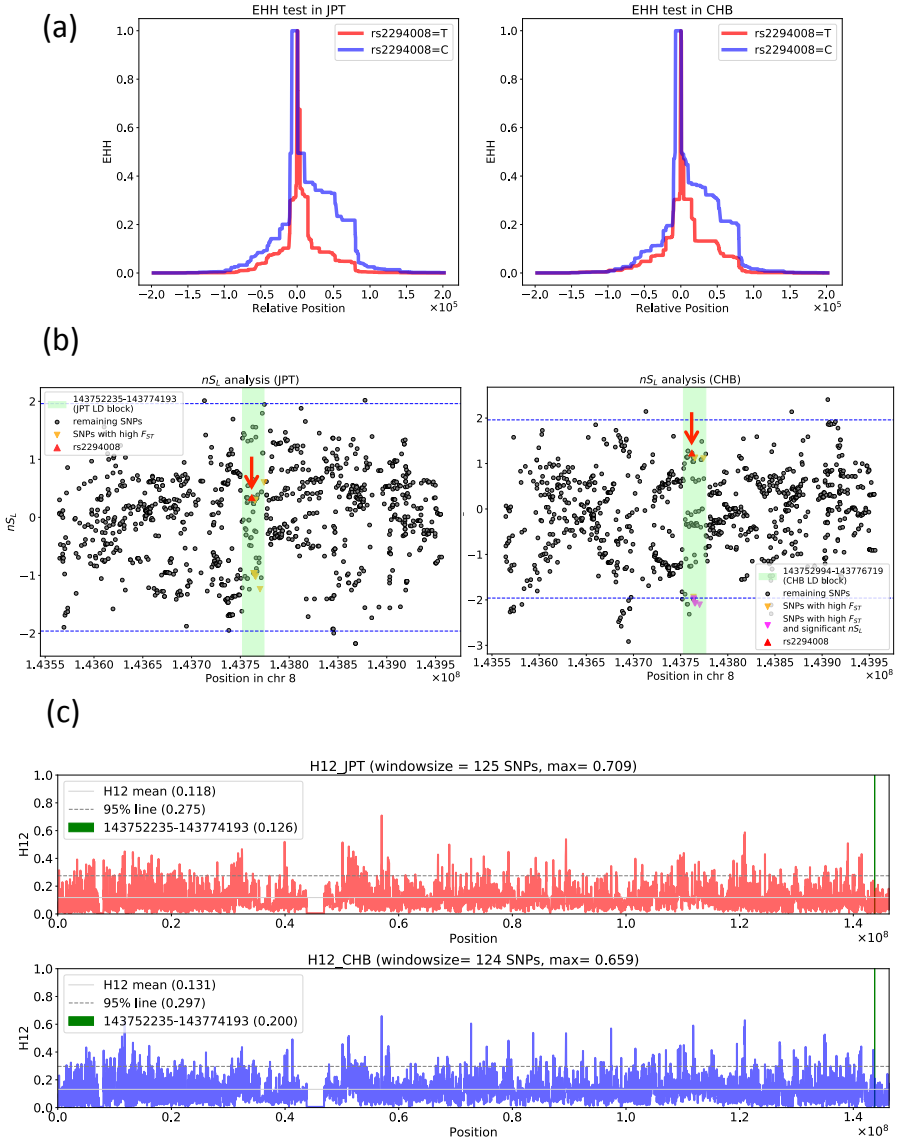

Figure 4. Neutrality tests (Extended haplotype homozygosity (EHH), number of segregating sites by length (nS_L) and haplotype homozygosity (H12)). (**a**) EHH of the region surrounding rs2294008 in JPT (left) and CHB (right). (**b**) nS_L of the region surrounding rs2294008 in JPT (left) and CHB (right). There was no limitation on the maximum size of the window. Each nS_L value at rs2294008 in CHB and JPT is indicated by a red triangle, while the top ten highly differentiated SNPs in CHB are indicated by inverted magenta triangles and the corresponding values in JPT are indicated by orange inverted triangles. The light green belt indicates the LD block region containing rs2294008 (JPT: 143,752,235-143,774,193, CHB: 143,752,994-143,776,719). Blue dashed lines represent 95% confidence intervals. (**c**) Sliding window analysis of H12 surrounding rs2294008 in JPT (upper) and CHB (lower). Light gray solid lines represent the H12 mean values within each population. The gray dashed lines represent the lower 95% line of H12 within each population. The light green line represents the JPT LD block region (143,752,235-143,774,193) used in this analysis. Values in parentheses in labels represent each H12 value.

In the 21.9 kb LD block region including rs2294008, Tajima's *D* showed no clear signature of natural selection (Tajima's $D = 2.031$ ($0.10 > p > 0.05$) and 0.745 ($p > 0.10$) for JPT and CHB, respectively). In contrast, normalized Fay and Wu's *H* detected a weak signal of an excess of high-frequency-derived alleles in CHB (normalized Fay and Wu's $H = -0.319$ ($p > 0.10$) and -2.174 ($0.01 < p < 0.05$) for JPT and CHB, respectively) due to selection or demographic effects, e.g., recent bottleneck or metapopulation structure [53,54]. PBS analysis did not detect any local adaptation signal in JPT or CHB (rs2294008: PBS value = 0.0592 and 0.0685, ranked 23,975th and 11,542th of a total 9,051,837 SNPs, respectively (Figure S4)). These observations suggest that there is no clear signature of positive selection on rs2294008 in either population.

3.3. Two-Dimensional Site Frequency Spectrum (2D SFS)

The newly developed 2D SFS statistics F_c, L_{c0}, and G_{c0} [34,35] were applied to explore a signature of selective sweep on the region containing rs2294008. These summary statistics detected low IAV among sites linked with a putative targeting derived allele (C allele at rs2294008) due to incomplete selective sweep.

3.3.1. Detection of Positive Selection in CHB, but not in JPT

Signatures of selective sweep with the C allele at rs2294008 in both JPT and CHB were examined using 2D SFS statistics. We detected a significant signature in CHB using all three summary statistics (Table 1). Statistics for the C allele were significantly lower than those for neutrality and the T allele. This reveals a significant reduction of IAV for the C allele in CHB and suggests that this allele is very likely to have been positively selected. Conversely, JPT did not show any signature of positive selection acting on either allele (Table 1).

Table 1. Summary statistics of two-dimensional site frequency spectrum (2D SFS).

Population	**JPT**		**CHB**	
Core Region	143755915-143770914		143755876-143771875	
n [#]	208 (C = 77, T = 131)		206 (C = 155, T = 51)	
S [†]	91		88	
Tested Allele	C	T	C	T
Allele Frequency	0.370	0.630	0.752	0.248
F_c [§]	0.167 (0.834)	0.833 (>0.999)	0.352×10^{-1} (0.223×10^{-2}) **	0.869 (>0.999)
G_{c0}	9.60 (0.693)	31.84 (0.975)	1.84 (0.167×10^{-2}) **	25.13 (>0.999)
L_{c0}	0.178×10^{-1} (0.708)	0.259 (>0.999)	0.565×10^{-2} (>0.167×10^{-2}) **	0.130 (>0.999)
G^*_{c0}	22.50	46.23	5.00	30.69
$\gamma^*(10)$	0.500	0.700	0.000	0.962
i_{max}	40	130	8	50
i^*_{max}	0	29	75	21

[§] q-value in parentheses represents under null model. ** represents q-value < 0.01. [#] n represents the number of samples. [†] S represents the number of segregating sites in the tested population.

3.3.2. Selection Mode in CHB and JPT

We examined whether the positive selection signal on the C allele was caused by classic hard sweep or soft sweep. In this study, we followed the definition of soft sweep whereby more than one

distinct selection-targeting haplotype was present in the D group [34]. The summary statistics (G^*_{c0} and $\gamma^*(10)$, i_{max}) for CHB fulfilled the criteria for hard sweep [34]. However, a large value of i^*_{max} of CHB (i^*_{max} = 75) indicated that the D group was classified into several large groups of haplotypes (subhaplotypes). In fact, the two subhaplotypes, classified by two sites (rs2976391 and rs2978983), could be the target of positive selection. We found that 75 chromosomes had derived alleles and the remaining 80 chromosomes had ancestral alleles in the D group at the two sites; we named the former as the A-G subhaplotype and the latter as the C-A subhaplotype based on the allele combination at rs2976391 (C/A) and rs2978983 (A/G).

We examined whether these subhaplotypes in CHB have a signal of positive selection using 2D SFS. Both subhaplotypes had a significant signal of positive selection identified by three summary statistics (Table S3). Then, we compared the π values between the two subhaplotypes in CHB and confirmed the effect of selective sweep. The two subhaplotypes in CHB showed quite similar genetic diversity ($\pi_{A\text{-}G} = 0.4 \times 10^{-4}$, $\pi_{C\text{-}A} = 0.4 \times 10^{-4}$, $\pi_{A\text{-}G\ vs\ C\text{-}A} = 2.4 \times 10^{-4}$). These observations were in good agreement with the results of 2D SFS.

These two subhaplotypes were also observed in JPT and we examined whether they were under positive selection in JPT. We detected a positive selection signature on the C-A subhaplotype, but not the A-G subhaplotype (Table S3). This suggests that within the alleles containing C at this SNP, the C-A subhaplotype under positive selection (hard sweep) may have been masked by the non-selected A-G subhaplotype, resulting in the failure to detect a signal of positive selection when the subhaplotypes were grouped. Taken together, our results support that JPT and CHB were under different selection modes.

3.3.3. History of Natural Selection in JPT and CHB

The upper limit of time of positive selection on the C allele in CHB is equivalent to the divergence time of the C-A and A-G subhaplotypes. We estimated that positive selection on the C allele began at most 240,000 ya, estimated from the divergence between these two subhaplotypes. The lower limit of positive selection beginning on each subhaplotype in CHB was estimated with the time to most recent common ancestor (TMRCA) within each subhaplotype [34]. We estimated that selection occurred circa 30,000 ± 12,000 ya (C-A subhaplotype) or circa 27,000 ± 14,000 ya (A-G subhaplotype) in CHB. Selection on the C-A subhaplotype in JPT occurred circa 11,000 ± 8000 ya. In CHB, this ongoing selection on the two subhaplotypes in the extant population led to the high allele frequency of the rs2294008 C allele. In contrast, selection on the A-G subhaplotype in JPT had relaxed at some point. The selection coefficient of the C-A subhaplotype in JPT was not likely to be strong enough to allow the detection of a signature of classic selective sweep on all C alleles. Consequently, the C allele frequency is maintained at an intermediate range in JPT, which results in a large C allele frequency difference between the two populations.

3.4. Examination of Whether Genetic Drift Can Explain the High F_{ST} Using Forward Simulation

The rs2294008 T allele is the major allele in JPT, even though one of the subhaplotypes with the C allele is likely to be under ongoing positive selection. Considering the possibility that Japanese population-specific demography caused this high allele frequency in JPT, we simulated the T allele frequency trajectory by performing forward simulation under neutrality or selection only in CHB. We constructed a demographic model for JPT and CHB following the "dual structure model" proposed by Hanihara [41] (Figure S1) and examined 570 parameter combinations of four parameters (*t1*, N_{JMN}, N_{A_CNT}, and *r*). Under the neutral state in CHB and JPT, we found three cases from three combinations that satisfy the high F_{ST} and high T allele frequency in simulated JPT (Figure S5). All three cases showed a similar pattern of allele frequency trajectory; the T allele frequencies at the last generation of JMN ranged from 0.94 to 1.0, and dropped by approximately half after admixture. These conditions lead to the actual allele frequency in JPT (approximately 0.6). The length of *t2* (2500 years, 125 generations) was too short to change allele frequency dramatically from those after admixture in the simulated JPT

lineage. In other words, a large difference of allele frequencies between the two extant populations and high T allele frequency in the extant JPT would require high T allele frequency from the JMN lineage. We found that the highest F_{ST} values could be attained if the T allele was nearly fixed in one of the ancestral populations or nearly lost in the other population (Table S4). Additionally, we performed simulations under four selection coefficients on the C allele in the CHB lineage. Based on this, we expected to get a fourfold higher number of combinations/cases relative to that under the neutral state of three cases from three combinations; however, we found an exceedingly high number of cases (35 cases from 24 combinations) (Figure S5 and S6). We also found a high T allele frequency in JMN and high F_{ST} between simulated JPT and CHB (Figure 5, Figure S6, and Table S4). Taken together, the T allele frequency in the Jomon must have been higher than that in the continental immigrant population and so the extant JPT inherited the T allele from the Jomon people. This led to the empirically high F_{ST} of rs2294008 between JPT and CHB, even though the T allele is a risk allele of severe disease.

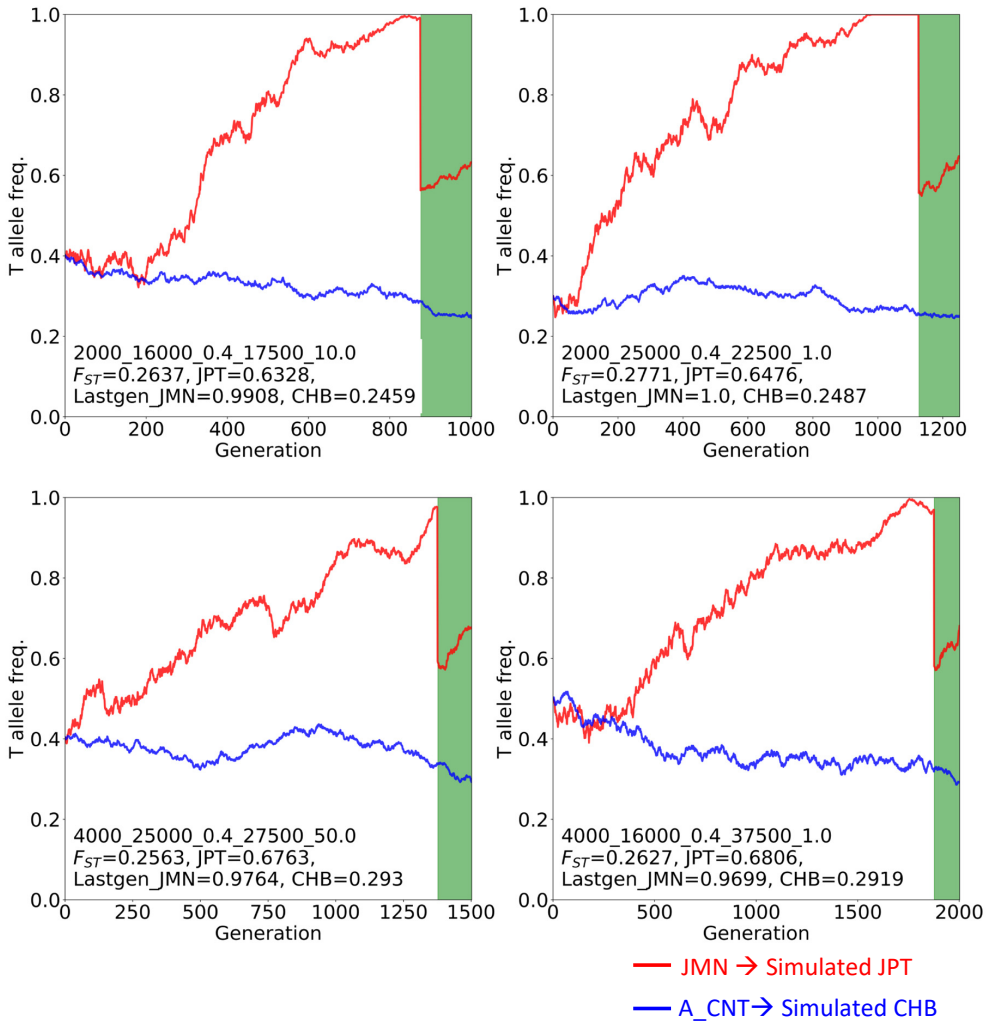

Figure 5. Four T allele frequency trajectories (out of 35) cases which satisfy the high F_{ST} and high T allele frequency in simulated JPT under positive selection on the C allele in the simulated CHB lineage. Red lines represent T allele frequencies of the simulated Jomon population (JMN) and simulated JPT. Blue lines represent those of the simulated ancestral population of the Asian continent (A_CNT) and simulated CHB. Green belts represent post-admixture of JMN and A_CNT. Subtitles display information on the parameters, F_{ST} value, and the T allele frequency of the simulated JPT, of the last generation in the JMN and of the simulated CHB. All 35 trajectories under positive selection and three trajectories under neutrality are shown in Figures S5 and S6, respectively.

3.5. Phylogenic Position of Jomon Haplotypes in the Network of Extant JPT and CHB

We classified the haplotypes of extant JPT and CHB into 93 haplotypes (Table S5); 41 were JPT-specific, whereas 37 were CHB-specific and 15 were shared between JPT and CHB. Of the 41 haplotypes specifically found in JPT, 87.8% (36 haplotypes) were harboring the T allele at rs2294008 (i.e., T haplotypes). Conversely, 27.0% (10 haplotypes) of the 37 CHB-specific haplotypes were T haplotypes. This indicates that the number of T haplotypes specific to JPT was larger than that in CHB, suggesting that T haplotypes in JPT were maintained for a longer time than in CHB.

Based on the simulation results, we hypothesized that the extant Japanese T allele was mainly inherited from the Jomon people. To evaluate this hypothesis, we compared sequences from the extant JPT to those from ancient genomes of Jomon samples, Ikawazu and Funadomari [46,49]. We found that all ancient genomes of Jomon samples had the T allele at rs2294008 as we expected. Then, we constructed a median-joining network and examined the relationship between extant JPT haplotypes and the Jomon haplotypes (Figure 6). The network formed two clusters, T haplotypes and C haplotypes, and all Jomon samples were clustered with the T haplotypes in JPT. Two Jomon samples, FUN23 (Funadomari) and IK002 (Ikawazu), were closely related to extant Japanese-specific haplotypes including haplotype 61 (H_9) for FUN23 and haplotype 30, 31, and 87 (H_13) for IK002. The haplotype of FUN5 was most closely related to that of FUN23 as well as four major haplotype groups in JPT and CHB (haplotypes 19, 42, 43, and 84 (H_4)). Sequences from the Funadomari Jomon were classified into different clusters in the network. FUN23 and FUN5 were reported to not share the same maternal lineage [46], suggesting that these two Jomon individuals likely had a low degree of kinship and that haplotypes had diverged within the population. Although IK002 and FUN5 have low coverage and the phylogenetic position of the haplotypes are uncertain, our results also support that some T haplotypes in the extant JPT are derived from Jomon haplotypes.

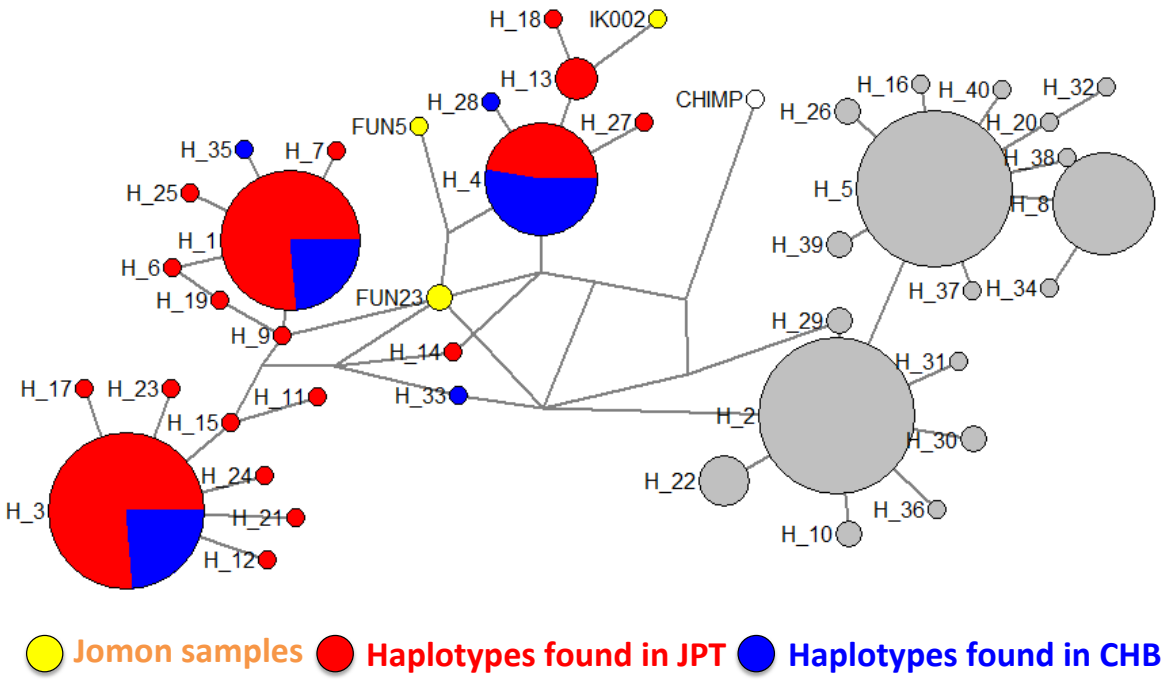

Figure 6. Network of haplotypes from three Jomon samples, extant JPT, extant CHB and the chimpanzee. Haplotypes which have C alleles are represented by gray circles and the other haplotypes are colored (Jomon samples (IK002, FUN5, and FUN23): yellow; extant JPT: red; extant CHB: blue; chimpanzee (CHIMP): white). The size of the circle represents the number of haplotypes. The branch length is not proportional to the number of substitutions. The labels starting with "H" represent the haplotype number defined in our network analysis. The corresponding haplotype numbers defined in extant samples are presented in Table S5.

4. Discussion

A significant association between the T allele at rs2294008 and diffuse-type gastric cancer (DGC) was reported in the Japanese [6]. The risk allele (T) frequency in JPT was found to be higher than that in other East Asian populations in 1KGP (Table S1). Although this SNP showed high genetic differentiation (F_{ST}) between JPT and CHB, most neutrality tests applied did not find any obvious signature of positive selection acting on this SNP or the region containing this SNP, except for 2D SFS (F_c, G_{c0}, and L_{c0}). The 2D SFS tests are specific in detecting incomplete selective sweep based on IAV linked with a putative target site [34]. This method is robust to recombination rate variation (i.e., the existence of recombination hotspots), unlike other haplotype-based selection detection tests. For example, in our study, we examined ten SNPs with high F_{ST} values using the nS_L method (Figure 4b). Only six of the ten SNPs showed a signal of positive selection, although it was marginal. In fact, the distribution of r^2 with rs2294008 showed an abrupt reduction of values outside the 15 kb region (where 2D SFS tests were applied) in JPT (Figure S7). The presence of a hotspot near the putative target (rs2294008 in the present study) would have a high chance of dissipating linkage disequilibrium unless positive selection occurred very recently. Since nS_L has been previously used to detect natural selection in the form of relatively recent soft sweep [25,35,55], both the presence of a hotspot and older soft sweep would lead to the failure to detect natural selection by nS_L. This indicates that 2D SFS is more robust in detecting older soft sweeps than nS_L due to its robustness to recombination rate fluctuation.

When we examined the level of π in JPT and CHB (Table S6), we found that the π_C values of the 21.9 kb region (chr8: 143,752,235–143,774,193) between CHB and JPT were not significantly different from each other (1.6×10^{-4} and 1.5×10^{-4}, respectively ($p = 0.3967$)). The π_C values of the 21.9 kb region in both populations were significantly lower than those of the adjacent 100 kb regions (all $q < 0.01$). In contrast, the π_T values of the 21.9 kb region showed marginal (for CHB) or significant (for JPT) reduction in the upstream flanking region ($0.01 < q < 0.05$ and $q < 0.01$, respectively), but they did not show significant reduction in downstream flanking regions in both populations (all $q > 0.05$). This suggests that the reduction of genetic diversity in the 21.9 kb LD block of the C allele was not caused by mutation rate fluctuation, but rather by a common evolutionary force that occurred in both populations. In other words, the reduction of π_C may have been caused by a common selection event shared between JPT and CHB. Even though only one subhaplotype was selected in JPT, π_C could reflect the operation of natural selection and consequently display a small π_C value. Taken together, the comparison of π_C supported the results from 2D SFS that positive selection operated on one subhaplotype in JPT, and both subhaplotypes in CHB.

The application of 2D SFS clearly indicated a signature of soft sweep in CHB but not in JPT (Table 1). Although the allele configuration between JPT and CHB is very similar, this discrepancy is attributed to a difference in selection mode and selection targets between the two populations. Both the C-A and A-G subhaplotypes were targets of ongoing positive selection in CHB at least 30,000 ya and 27,000 ya, respectively, which may have led to high C allele frequency in CHB. This time frame is consistent with the divergence between the Jomon people and the continental Asian population (15,000~38,000 ya [44–46]). Conversely, in JPT, only the C-A subhaplotype is a target of ongoing positive selection (Table S3) and selection in JPT begun almost simultaneously with that in CHB. These observations suggest that both subhaplotypes were targets of positive selection in the common ancestor of JPT and CHB. However, positive selection on the A-G subhaplotype in JPT was relaxed/ceased at some time point in the Jomon lineage and the selection mode observed in JPT was the hardening of soft selective sweep with the C-A subhaplotype. This suggests that positive selection on the C-A subhaplotype occurred in both JPT and CHB, and was not a result of local adaptation. PBS analysis specifies that unique local adaptation in a specific population is based on the long branch of F_{ST} [38]. If ongoing positive selection had operated in two populations, a long branch would not be limited to a specific population and would lead to less power of detection. Therefore, this caused the failure to detect positive selection in CHB/JPT using PBS analysis. rs2976391 (C/A), one of the two SNPs which defined the two subhaplotypes, is located in the intron of the *PSCA* gene and also overlapped with the

JRK gene [56] (Figure S8). The A allele at rs2976391 is reported to be involved in promotor activity and transcription activity [56] (Tables S7 and S8). Although rs2978983 (A/G) currently has no reported biological function, the C-A subhaplotype may be associated with gene expression function and cause functional variation within a population. In the future, it would be interesting to examine the alteration of *PSCA* gene expression regarding the C haplotype including both subhaplotypes.

We detected C-A and A-G subhaplotypes that are maintained in both JPT and CHB. These two subhaplotypes have also been maintained in four metapopulations (non JPT/CHB EAS, SAS, EUR, and AFR). These haplotypes existing in the extant AFR further supported our estimated divergence time between the two subhaplotypes of ~240,000 ya. We detected the positive selection of subhaplotypes in these populations (Table S3). The C-A subhaplotype showed a signal of ongoing positive selection in the JPT, CHB, non JPT/CHB EAS, and SAS populations. In contrast, the A-G subhaplotype showed signals of ongoing positive selection in the CHB and AFR populations. Neither subhaplotype in EUR showed any signals. Unlike the case of CHB, non JPT/CHB EAS had a weaker signal on the A-G subhaplotype. The A-G subhaplotype is composed of derived alleles (A at rs2978391 and G at rs2978983), suggesting that it is younger than the C-A subhaplotype. Although the C-A subhaplotype is an ancestral subhaplotype, the extent of IAV was lower than the A-G subhaplotype in non-AFR populations (Table S3). This suggests that the signal of positive selection is stronger on the C-A subhaplotype than the A-G subhaplotype. Our study of the C haplotype in JPT identified two main findings. First, the C-A subhaplotype shared ancestral alleles with the T haplotype and it was difficult to detect unique derived SNPs (which are possible targets for the haplotype) in the C-A subhaplotype using a 2D SFS approach, suggesting that the combination of the C-A subhaplotype with the C allele at rs2249008 is a potential target of positive selection and would have a biologically important function, when compared to the C allele alone. Second, the A-G subhaplotype frequency is very similar to that of the C-A subhaplotype and this masks the signal of unique derived alleles in the C-A subhaplotype. Interestingly, the A-G subhaplotype showed a stronger signal than the C-A subhaplotype only in AFR. These findings suggest that the target of positive selection may have changed or reactivated during human population history. The status of positive selection may not have been necessarily inherited from an ancestor to the offspring populations (Figure S9). The lower limit of the beginning of positive selection was estimated as circa 30,000 to 11,000 ya (C-A subhaplotype) or circa 35,000 to 21,000 ya (A-G subhaplotype) among several metapopulations under positive selection. It is notable that the lower limit of the beginning of positive selection was similar among global populations, even though these populations have been isolated from each other for a long period and continued to differentiate. This further supports the idea that the positive selection target changed or was reactivated.

In the forward simulation performed, we examined whether the F_{ST} value in each simulation was equal to or greater than that of rs2294408 (0.2547). The allele frequency trajectory revealed that high T allele frequency in the simulated JPT and large F_{ST} value required high T allele frequency in the ancestral population before admixture both under the neutral state and under positive selection on CHB lineages. This high T allele frequency implied that the Jomon people were likely to have had a high frequency of the T allele. In fact, in the extant Ryukyuan and Ainu people, who have a higher Jomon genetic component than the mainland Japanese [45], the frequency of the T allele on rs2294008, or its tightly linked adjacent allele, was higher than that in JPT (Ryukyuan: 0.701 (rs2294008) [57], Ainu: 0.975 (rs2976396, tightly linked with the rs2294008 T allele) [58]). To examine the relationship between the extant JPT and Jomon haplotypes, we checked the phylogenetic position of the Jomon haplotypes Ikawazu (IK002) and Funadomari (FUN5 and FUN23). Two of these Jomon samples (IK002 and FUN23) are closely related to the extant JPT-specific T haplotypes. FUN5 is also closely related to either T haplotypes that are extant JPT-specific or those widely shared between JPT and CHB. These observations support the possibility that the Jomon people had high T allele frequency and that the T allele in some extant Japanese haplotypes is derived from the Jomon haplotype.

Based on our analyses, we reconstructed the trajectory of the C and T haplotypes at rs2294008 along the Japanese population history (Figure S9). Positive selection on both C-A and A-G subhaplotypes

began in the common ancestor of East Asians (including the ancestor of the Jomon). The ancestor of the Jomon diverged from ancestral populations in the East Asian continent from 15,000 to 38,000 ya [44–46], and migrated to the Japanese archipelago. Positive selection on the A-G subhaplotype then ceased or relaxed within the Jomon at some point, and this elevated the haplotype frequency containing T at rs2294008 within the Jomon. In contrast, the C allele frequency in the population was elevated in the ancestor of East Asian populations due to positive selection on the C haplotype. The ancestor of immigrant Yayoi farmers (who were genetically close to the extant Koreans) diverged from an ancestor of extant East Asian populations between 3000 and 3600 ya [59]. Immigrant Yayoi farmers migrated to the Japanese archipelago between 2500 and 3000 ya [42,60] and admixed with the Jomon who had high T allele frequency. Thus, although positive selection on both C-A and A-G subhaplotypes occurred in the ancestor of populations in the East Asian lineage, positive selection acted on only the C-A subhaplotype in the extant Japanese, and subsequently a high T allele frequency was observed in the Japanese. In future work, we will examine this scenario in more detail including Korean samples genetically close to the Japanese.

Although we cannot conclude whether selection on the A-G subhaplotype in the Japanese was relaxed or completely stopped, our study clearly showed that selection processes acting on rs2294008 differ between genetically close populations, such as JPT and CHB. This difference may be extended to other East Asian populations, SAS, EUR, and AFR. The relaxation or cessation of positive selection on the A-G subhaplotype may be a unique trait of the Japanese among the studied East Asian populations.

The associations between biological functions and rs2294008 and its adjacent region are complex, and interpreting the impact on the fitness of the T and C allele is difficult. For example, the T allele is reported as the risk allele of DGC, but the C allele (non-risk allele against DGC) is also reported as a risk allele for duodenal ulcers in Japanese [13] and Caucasian populations [12] based on genome-wide association studies. Thus, the difference between the haplotypes in biological function as well as selection modes/coefficients may have resulted in the functional variation of this region, which includes rs2294008, among genetically close East Asian populations.

The above discussion is summarized as follows:

(i) Selection operated on the C allele (the non-risk allele) in the common ancestor of the Han Chinese and the Jomon people. The mode of positive selection in the Japanese is complex; selection on the A-G subhaplotype ceased or relaxed at some point along the Japanese lineage, but ongoing selection occurred on the C-A suphaplotype. Relaxation or cessation of positive selection on the A-G subhaplotype may have led to low frequency of the C allele in the extant JPT.

(ii) The ancestral population (the Jomon people) had a high T allele frequency, which led to a high T allele frequency in the extant Japanese, even though the Jomon people experienced admixture with immigrant Yayoi farmers. These factors result in the large T/C allele frequency difference between JPT and CHB.

This study is the first to report the complex positive selection modes acting at rs2294008 among human populations, even between JPT and their genetically closest population. Our findings support that positive selection targets can change during human history over relatively short periods.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/7/775/s1. Figure S1: Demographic model of the Jomon (JMN), an ancestral population inhabiting the Asian continent (A_CNT), simulated Japanese in Tokyo (simulated JPT) and simulated Han Chinese in Beijing (simulated CHB); Figure S2: F_{ST} distribution per 21.9 kb windows; Figure S3: SNPs with high F_{ST} as core for EHH; Figure S4: Distribution of PBS values and position of rs2294008 for JPT (**a**) and CHB (**b**) based on 9,051,837 genome-wide PBS values; Figure S5: Trajectories of T allele frequency at SNPs with high T allele frequency and high F_{ST} under neutral state in the CHB lineage; Figure S6: Trajectories of T allele frequency at SNPs with high T allele frequency and high F_{ST} under positive selection on the C allele in the JPT lineage; Figure S7: The decay of r^2 with rs2294008; Figure S8: Genomic context of rs2976391; Figure S9: Positive selection status among global populations; Table S1: T allele frequency and F_{ST} values at rs2294008 in 1KGP populations; Table S2: List of F_{ST} of top 50 SNPs; Table S3: 2D SFS analysis of two subhaplotypes in global populations; Table S4: Differentiation level classified by allele frequencies of the last generation of JMN and A_CNT; Table S5: List of extant JPT and CHB haplotypes in 21.9 kb

LD block; Table S6: Nucleotide diversity at LD block and flanking 100 kb regions; Table S7: Genes and transcript consequences of rs2976391; Table S8: Motif feature consequences of rs2976391.

Author Contributions: Conceptualization, R.L.I., J.G., and Y.S.; methodology, R.L.I., J.G., and Y.S.; software, R.L.I.; validation, R.L.I., J.G., and Y.S.; formal analysis, R.L.I. and Y.S.; investigation, R.L.I.; resources, K.I., H.K.-K., Y.K., J.G., and Y.S.; data curation, R.L.I.; writing—original draft preparation, R.L.I.; writing—review and editing, R.L.I., K.I., H.K.-K., Y.K., J.G., and Y.S.; visualization, R.L.I.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The research assistant fee was funded by SOKENDAI. This study was partially supported by SOKENDAI.

Acknowledgments: The authors would like to thank Tatsuya Ota, Hitomi Hongo and Naoyuki Takahata for expert advice and discussion in this project, as well as all colleagues in our laboratory. The authors would like to thank the Asian DNA Repository Consortium (for providing allele frequency data for the Ainu population), Takehiro Sato and Ryosuke Kimura (for providing allele frequency data for the Ryukyuan population), Wen-Ya Ko (for providing allele frequency data for the Taiwanese), Ituro Inoue (for providing allele frequency data for the Korean population through Hyoung Doo Shin), Hiroki Ohta (for providing the Ikawazu Jomon sequence), and Quintin Lau (for English editing). This work was supported in part by The Graduate University for Advanced Studies, SOKENDAI.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef]
2. Laurén, P. The two histological main types of gastric carcinoma: Diffuse and so-called intestinal-type carcinoma. *Acta Pathol. Microbiol. Scand.* **1965**, *64*, 31–49. [CrossRef]
3. Crew, K.D.; Neugut, A.I. Epidemiology of gastric cancer. *World J. Gastroenterol.* **2006**, *12*, 354–362. [CrossRef]
4. Henson, D.E.; Dittus, C.; Younes, M.; Nguyen, H.; Albores-Saavedra, J. Differential trends in the intestinal and diffuse types of gastric carcinoma in the United States, 1973–2000: Increase in the signet ring cell type. *Arch. Pathol. Lab. Med.* **2004**, *128*, 765–770.
5. Miyahara, R.; Niwa, Y.; Matsuura, T.; Maeda, O.; Ando, T.; Ohmiya, N.; Itoh, A.; Hirooka, Y.; Goto, H. Prevalence and prognosis of gastric cancer detected by screening in a large Japanese population: Data from a single institute over 30 years. *J. Gastroenterol. Hepatol.* **2007**, *22*, 1435–1442. [CrossRef] [PubMed]
6. The Study Group of Millennium Genome Project for Cancer; Sakamoto, H.; Yoshimura, K.; Saeki, N.; Katai, H.; Shimoda, T.; Matsuno, Y.; Saito, D.; Sugimura, H.; Tanioka, F.; et al. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat. Genet.* **2008**, *40*, 730–740. [CrossRef]
7. Song, H.-R.; Kim, H.N.; Piao, J.-M.; Kweon, S.-S.; Choi, J.-S.; Bae, W.-K.; Chung, I.J.; Park, Y.-K.; Kim, S.-H.; Choi, Y.-D.; et al. Association of a common genetic variant in prostate stem-cell antigen with gastric cancer susceptibility in a Korean population. *Mol. Carcino.* **2011**, *50*, 871–875. [CrossRef]
8. Park, B.; Yang, S.; Lee, J.; Woo, H.D.; Choi, I.J.; Kim, Y.W.; Ryu, K.W.; Kim, Y.-I.; Kim, J. Genome-Wide Association of Genetic Variation in the PSCA Gene with Gastric Cancer Susceptibility in a Korean Population. *Cancer Res. Treat.* **2019**, *51*, 748–757. [CrossRef]
9. Turdikulova, S.; Dalimova, D.; Abdurakhimov, A.; Adilov, B.; Navruzov, S.; Yusupbekov, A.; Djuraev, M.; Abdujapparov, S.; Egamberdiev, D.; Mukhamedov, R. Association of rs2294008 and rs9297976 Polymorphisms in PSCA Gene with Gastric Cancer Susceptibility in Uzbekistan. *Cent. Asian J. Glob. Health* **2016**, *5*, 227. [CrossRef]
10. Lochhead, P.; Frank, B.; Hold, G.L.; Rabkin, C.S.; Ng, M.T.H.; Vaughan, T.L.; Risch, H.A.; Gammon, M.D.; Lissowska, J.; Weck, M.N.; et al. Genetic Variation in the Prostate Stem Cell Antigen Gene and Upper Gastrointestinal Cancer in White Individuals. *Gastroenterology* **2011**, *140*, 435–441. [CrossRef]
11. Sala, N.; Muñoz, X.; Travier, N.; Agudo, A.; Duell, E.J.; Moreno, V.; Overvad, K.; Tjønneland, A.; Boutron-Ruault, M.C.; Clavel-Chapelon, F.; et al. Prostate stem-cell antigen gene is associated with diffuse and intestinal gastric cancer in Caucasians: Results from the EPIC-EURGAST study. *Int. J. Cancer* **2012**, *130*, 2417–2427. [CrossRef]

12. García-González, M.A.; Bujanda, L.; Quintero, E.; Santolaria, S.; Benito, R.; Strunk, M.; Sopena, F.; Thomson, C.; Perez-Aisa, A.; Nicolás-Pérez, D.; et al. Association ofPSCArs2294008 gene variants with poor prognosis and increased susceptibility to gastric cancer and decreased risk of duodenal ulcer disease. *Int. J. Cancer* **2015**, *137*, 1362–1373. [CrossRef]
13. Tanikawa, C.; Urabe, Y.; Matsuo, K.; Kubo, M.; Takahashi, A.; Ito, H.; Tajima, K.; Kamatani, N.; Nakamura, Y.; Matsuda, K. A genome-wide association study identifies two susceptibility loci for duodenal ulcer in the Japanese population. *Nat. Genet.* **2012**, *44*, 430–434. [CrossRef] [PubMed]
14. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]
15. Chen, C.-H.; Yang, J.-H.; Chiang, C.W.; Hsiung, C.-N.; Wu, P.-E.; Chang, L.-C.; Chu, H.-W.; Chang, J.; Song, I.-W.; Yang, S.-L.; et al. Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum. Mol. Genet.* **2016**, *25*, 5321–5331. [CrossRef]
16. Bae, J.S.; Cheong, H.S.; Kim, J.-O.; Lee, S.O.; Kim, E.-M.; Lee, H.W.; Kim, S.; Kim, J.-W.; Cui, T.; Inoue, I.; et al. Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. *Biochem. Biophys. Res. Commun.* **2008**, *373*, 593–596. [CrossRef]
17. Hudson, R.R.; Slatkin, M.; Maddison, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **1992**, *132*, 583–589. [PubMed]
18. Bhatia, G.; Patterson, N.; Sankararaman, S.; Escott-Price, V. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **2013**, *23*, 1514–1521. [CrossRef]
19. Takahata, N.; Satta, Y.; Klein, Y. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics.* **1992**, *130*, 925–938.
20. Turner, S.D. Qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **2018**, *3*, 1–2. [CrossRef]
21. Hartl, D.L.; Clark, A.G. *Principles of Population Genetics*, 4th ed.; Sinauer Associates: Sunderland, MA, USA, 2007; Chapter 2; pp. 45–88. ISBN 978-0-87893-308-2.
22. Barrett, J.C.; Maller, J.; Daly, M.J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinfomatics* **2005**, *21*, 263–265. [CrossRef] [PubMed]
23. Gabriel, S.B.; Schaffner, S.F.; Nguyen, H.; Moore, J.M.; Roy, J.; Blumenstiel, B.; Higgins, J.; DeFelice, M.; Lochner, A.; Faggart, M.; et al. The Structure of Haplotype Blocks in the Human Genome. *Science* **2002**, *296*, 2225–2229. [CrossRef] [PubMed]
24. Sabeti, P.C.; Reich, D.E.; Higgins, J.M.; Levine, H.Z.P.; Richter, D.J.; Schaffner, S.F.; Gabriel, S.B.; Platko, J.V.; Patterson, N.J.; McDonald, G.J.; et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **2002**, *419*, 832–837. [CrossRef] [PubMed]
25. Ferrer-Admetlla, A.; Liang, M.; Korneliussen, T.S.; Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Boil. Evol.* **2014**, *31*, 1275–1291. [CrossRef]
26. Garud, N.R.; Messer, P.W.; Buzbas, E.O.; Petrov, D.A. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet.* **2015**, *11*, e1005004. [CrossRef] [PubMed]
27. Szpiech, Z.A.; Hernandez, R.D. Selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Mol. Boil. Evol.* **2014**, *31*, 2824–2827. [CrossRef]
28. Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **1989**, *123*, 585–595.
29. Fay, J.; Wu, C.-I. Hitchhiking Under Positive Darwinian Selection. *Genetics* **2000**, *155*, 1405–1413.
30. Zeng, K.; Fu, Y.-X.; Shi, S.; Wu, C.-I. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics* **2006**, *174*, 1431–1439. [CrossRef]
31. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Boil. Evol.* **2017**, *34*, 3299–3302. [CrossRef]
32. Nei, M.; Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **1979**, *76*, 5269–5273. [CrossRef] [PubMed]
33. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. [CrossRef]

34. Satta, Y.; Zheng, W.; Nishiyama, K.V.; Iwasaki, R.L.; Hayakawa, T.; Fujito, N.T.; Takahata, N. Two-dimensional site frequency spectrum for detecting, classifying and dating incomplete selective sweeps. *Genes Genet. Syst.* **2019**, *94*, 283–300. [CrossRef] [PubMed]
35. Fujito, N.T.; Satta, Y.; Hayakawa, T.; Takahata, N. A new inference method for detecting an ongoing selective sweep. *Genes Genet. Syst.* **2018**, *93*, 149–161. [CrossRef] [PubMed]
36. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **2002**, *18*, 337–338. [CrossRef]
37. Schaffner, S.F.; Foo, C.; Gabriel, S.; Reich, D.; Daly, M.J.; Altshuler, D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **2005**, *15*, 1576–1583. [CrossRef]
38. Yi, X.; Liang, Y.; Huerta-Sanchez, E.; Jin, X.; Cuo, Z.X.P.; Pool, J.; Xu, X.; Jiang, H.; Vinckenbosch, N.; Korneliussen, T.; et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **2010**, *329*, 75–78. [CrossRef]
39. Cavalli-Sforza, L.L. Human diversity. In Proceedings of the 12th International Congress on Genetics, Tokyo, Japan, 19–28 August 1968; pp. 405–416.
40. Ewens, W.J. *Mathematical Population Genetics 1. Biomathematics*; Springer: New York, NY, USA, 1979; Volume 9.
41. Hanihara, K. Dual Structure Model for the Population History of the Japanese. *Japan Rev.* **1991**, *2*, 1–33. [CrossRef]
42. Habu, J. *Ancient Jomon of Japan*; Cambridge University Press: Cambridge, UK, 2004; ISBN 0521776708.
43. Jinam, T.A.; Kanzawa-Kiriyama, H.; Inoue, I.; Tokunaga, K.; Omoto, K.; Saitou, N. Unique characteristics of the Ainu population in Northern Japan. *J. Hum. Genet.* **2015**, *60*, 565–571. [CrossRef]
44. Nakagome, S.; Sato, T.; Ishida, H.; Hanihara, T.; Yamaguchi, T.; Kimura, R.; Mano, S.; Oota, H. The Asian DNA Repository Consortium. Model-Based Verification of Hypotheses on the Origin of Modern Japanese Revisited by Bayesian Inference Based on Genome-Wide SNP Data. *Mol. Boil. Evol.* **2015**, *32*, 1533–1543. [CrossRef]
45. Kanzawa-Kiriyama, H.; Kryukov, K.; Jinam, T.A.; Hosomichi, K.; Saso, A.; Suwa, G.; Ueda, S.; Yoneda, M.; Tajima, A.; Shinoda, K.-I.; et al. A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. *J. Hum. Genet.* **2016**, *62*, 213–221. [CrossRef]
46. Kanzawa-Kiriyama, H.; Jinam, T.A.; Kawai, Y.; Sato, T.; Hosomichi, K.; Tajima, A.; Adachi, N.; Matsumura, H.; Kryukov, K.; Saitou, N.; et al. Late Jomon male and female genome sequences from the Funadomari site in Hokkaido, Japan. *Anthr. Sci.* **2019**, *127*, 83–108. [CrossRef]
47. Fenner, J.N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthr.* **2005**, *128*, 415–423. [CrossRef]
48. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Method* **2020**, *17*, 261–272. [CrossRef]
49. McColl, H.; Racimo, F.; Vinner, L.; Demeter, F.; Gakuhari, T.; Moreno-Mayar, J.V.; Van Driem, G.; Wilken, U.G.; Seguin-Orlando, A.; Castro, C.D.L.F.; et al. The prehistoric peopling of Southeast Asia. *Science* **2018**, *361*, 88–92. [CrossRef]
50. Available online: https://www.fluxus-engineering.com/index.htm (accessed on 1 July 2020).
51. Bandelt, H.-J.; Forster, P.; Röhl, A. Median-Joining Networks for Inferring Intraspecific Phylogenies. *Mol. Biol. Evol.* **1999**, *16*, 37–48. [CrossRef]
52. Sikora, M.; Seguin-Orlando, A.; Sousa, V.C.; Albrechtsen, A.; Korneliussen, T.; Ko, A.; Rasmussen, S.; Dupanloup, I.; Nigst, P.; Bosch, M.D.; et al. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* **2017**, *358*, 659–662. [CrossRef]
53. Wakeley, J.; Alicar, N. Gene genealogies in a metapopulation. *Genetics* **2001**, *159*, 893–905.
54. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **2002**, *160*, 1179–1189.
55. Nakayama, K.; Ohashi, J.; Watanabe, K.; Munkhtulga, L.; Iwamoto, S. Evidence for Very Recent Positive Selection in Mongolians. *Mol. Boil. Evol.* **2017**, *34*, 1936–1946. [CrossRef]
56. Cunningham, F.; Achuthan, P.; Akanni, W.; Allen, J.; Amode, M.R.; Armean, I.M.; Bennett, R.; Bhai, J.; Billis, K.; Boddu, S.; et al. Ensembl 2019. *Nucleic Acids Res.* **2018**, *47*, D745–D751. [CrossRef]

57. Sato, T.; Nakagome, S.; Watanabe, C.; Yamaguchi, K.; Kawaguchi, A.; Koganebuchi, K.; Haneji, K.; Yamaguchi, T.; Hanihara, T.; Yamamoto, K.; et al. Genome-Wide SNP Analysis Reveals Population Structure and Demographic History of the Ryukyu Islanders in the Southern Part of the Japanese Archipelago. *Mol. Boil. Evol.* **2014**, *31*, 2929–2940. [CrossRef] [PubMed]
58. Japanese Archipelago Human Population Genetics Consortium; Jinam, T.; Nishida, N.; Hirai, M.; Kawamura, S.; Oota, H.; Umetsu, K.; Kimura, R.; Ohashi, J.; Tajima, A.; et al. The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J. Hum. Genet.* **2012**, *57*, 787–795. [CrossRef] [PubMed]
59. Wang, Y.; Lu, D.; Chung, Y.-J.; Xu, S. Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* **2018**, *155*, 19. [CrossRef] [PubMed]
60. Fujio, S. *History of Yayoi Period*; Kodansha: Tokyo, Japan, 2015; ISBN 9784062883306. (In Japanese)

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).