genes

MDPI

Article

Short Insertion and Deletion Discoveries via Whole-Genome Sequencing of 101 Thoroughbred Racehorses

Teruaki Tozaki [1,*,†], Aoi Ohnuma [1,†], Mio Kikuchi [1], Taichiro Ishige [1], Hironaga Kakoi [1], Kei-ichi Hirota [1], Yuji Takahashi [2] and Shun-ichi Nagata [1]

[1] Genetic Analysis Department, Laboratory of Racing Chemistry, 1731-2 Tsurutamachi, Utsunomiya 320-0851, Tochigi, Japan
[2] Equine Research Institute, Japan Racing Association, 1400-4 Shiba, Shimotsuke 329-0412, Tochigi, Japan
* Correspondence: ttozaki@lrc.or.jp
† These authors contributed equally to this work.

Abstract: Thoroughbreds are some of the most famous racehorses worldwide and are currently animals of high economic value. To understand genomic variability in Thoroughbreds, we identified genome-wide insertions and deletions (INDELs) and obtained their allele frequencies in this study. INDELs were obtained from whole-genome sequencing data of 101 Thoroughbred racehorses by mapping sequence reads to the horse reference genome. By integrating individual data, 1,453,349 and 113,047 INDELs were identified in the autosomal (1–31) and X chromosomes, respectively, while 18 INDELs were identified on the mitochondrial genome, totaling 1,566,414 INDELs. Of those, 779,457 loci (49.8%) were novel INDELs, while 786,957 loci (50.2%) were already registered in Ensembl. The sizes of diallelic INDELs ranged from −286 to +476, and the majority, 717,736 (52.14%) and 220,672 (16.03%), were 1-bp and 2-bp variants, respectively. Numerous INDELs were found to have lower frequencies of alternative (Alt) alleles. Many rare variants with low Alt allele frequencies (<0.5%) were also detected. In addition, 5955 loci were genotyped as having a minor allele frequency of 0.5 and being heterogeneous genotypes in all the horses. While short-read sequencing and its mapping to reference genome is a simple way of detecting variants, fake variants may be detected. Therefore, our data help to identify true variants in Thoroughbred horses. The INDEL database we constructed will provide useful information for genetic studies and industrial applications in Thoroughbred horses, including a gene-editing test for gene-doping control and a parentage test using INDELs for horse registration and identification.

Keywords: gene doping; horseracing; INDEL; parentage test; SNV

check for updates

Citation: Tozaki, T.; Ohnuma, A.; Kikuchi, M.; Ishige, T.; Kakoi, H.; Hirota, K.-i.; Takahashi, Y.; Nagata, S.-i. Short Insertion and Deletion Discoveries via Whole-Genome Sequencing of 101 Thoroughbred Racehorses. *Genes* **2023**, *14*, 638. https://doi.org/10.3390/genes14030638

Academic Editor: Carrie Finno

Received: 2 February 2023
Revised: 28 February 2023
Accepted: 2 March 2023
Published: 3 March 2023

Copyright: © 2023 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Thoroughbreds are some of the most famous racehorses worldwide. They had a founding population (of Arabian stallions and British mares) around the 18th century, and have been bred as a closed group for approximately 400 years [1]. In the current racing industry, over 80,000 Thoroughbred racehorses are born worldwide every year [2], and they are currently animals of high economic value.

The horse genome was sequenced and assembled in 2007 as EquCab2.0, in which 2.33-Gb draft sequences were published [3]. The latest version of the horse genome, EquCab3.0, was assembled as a total read length of 2,506,949,475 bp (1–31 and X: 2,409,143,234 bp, unplaced: 97,806,241 bp, Assembly: GCA_002863925.1), in 2019 [4]. By mapping sequence reads obtained from massive parallel sequencing to the reference genome sequences, genome-wide variants have been easily identified. Currently, whole-genome sequencing (WGS) of 88 horses (25 breeds) and 534 horses (46 breeds) has identified approximately 23.6 million and 29.0 million single nucleotide variants (SNVs), respectively [5,6]. In a Thoroughbred population of 101 unrelated horses, 12 million SNVs with their allelic frequencies were identified using WGS [7].

Insertions and deletions (INDELs) are variants consisting of different allele sizes (one or more nucleotides) in the genome. When INDELs occur in amino acid-coding regions, they are generally known to result in a loss of function. Although the existence of INDELs in coding regions is deleterious, their roles and functions in the population are not fully elucidated.

INDELs are highly abundant in human and animal genomes, and approximately 2.4 million and 2.1 million INDELs were identified from 88 and 534 horses, respectively [5,6]. In the Ensembl, 3,461,675 INDELs have been identified in the current horse assembly (https://ftp.ensembl.org/pub/release-108/variation/gvf/equus_caballus/, accessed on 2 February 2023). However, although several Thoroughbred horses were used for identification of genome-wide variants in the previous studies, the number, frequency, type, and size of INDELs in Thoroughbred horses have not been well elucidated.

Gene doping is a practice in horseracing that has been prohibited to maintain integrity [8]. One style of gene doping is to create genetically engineered animals; this has been carried out in many species, including horses [9–11]. The International Stud Book Committee (ISBC, https://www.internationalstudbook.com/, accessed on 2 February 2023) and the International Federation of Horseracing Authorities (IFHA, https://www.ifhaonline.org/, accessed on 2 February 2023) has prohibited the use of genetically engineered horses. In addition, horses born from genetically engineered animals are recognised as engineered horses. Editing (insertion and deletion) of coding genes causes loss of function. For instance, knocking out the *myostatin* gene, which is known as a negative regulator of muscle growth, may affect body composition and racing performance [12,13].

Recently, a gene-editing test was developed to detect artificially edited sequences using the clustered, regularly interspaced, short palindromic repeats/CRISPR-associated proteins (CRISPR/Cas) [14]. This test defined homologous insertions or deletions of novel variants as a type of artificial modification. Therefore, it is also necessary to identify INDELs in current Thoroughbred populations.

Microsatellites have been recently used for parentage testing in horse registration. While about 12–20 markers have been used for construction of a panel, because of their multiple alleles, they have the disadvantage of high mutation rates due to slippage errors. Therefore, construction of a new panel using SNVs or INDELs that have low mutation rates is expected. However, it is difficult to find many polymorphic SNPs and INDELs.

In our previous study [7], we presented only genome-wide SNV detection and their frequency data. This study focused on the detection of genome-wide INDELs and their frequency data. This study aimed to identify INDELs in a Thoroughbred population from 101 WGS data and to construct an INDEL database to provide a reference for genetic studies and industrial applications in horses, including pedigree registration and gene-doping control.

2. Materials and Methods

2.1. Whole-Genome Sequencing Data from 101 Thoroughbred Horses

FASTQ (DDBJ: SAMD00573909 to SAMD00574009) of WGS data (150-bp pair-end reads) from 101 Thoroughbred horses (58 males and 43 females) were used in this study (Table S1). The 101 horses were registered as Thoroughbred horses in Japan; some horses were born in other countries and were imported to Japan [7].

2.2. INDEL Calling and Filtering

Short INDELs were identified using the RESEQ pipeline (Amelieff Co., Minato, Tokyo, Japan). The pipeline was constructed using QCleaner (Amelieff Co.), Burrows–Wheeler Aligner (BWA, version 0.7.17) (https://bio-bwa.sourceforge.net/, accessed on 2 February 2023), Picard (version 2.13.2) (https://sourceforge.net/apps/mediawiki/picard/, accessed on 2 February 2023), GATK HaplotypeCaller (version 4.0.8.1) (https://software.broadinstitute.org/gatk/best-practices/, accessed on 2 February 2023), and SnpEff (version v4_0) (http://pcingola.github.io/SnpEff/download/, accessed on 2 February 2023).

Quality control using QCleaner eliminates sequence reads for the following criteria: reads with a low-quality base (<20 Phred score), a quality value < 20 in 80% of their nucleotides, sequences of over five unknown nucleotides, only <32 bp length sequences, and those that are not mate pairs.

In brief, after qualification, reads obtained from WGS were aligned to the horse reference genome assembly EquCab3.0 (GenBank: GCA_002863925.1) using the BWA with default parameters to obtain a BAM file. Duplicate reads were removed using the Picard tool. GATK HaplotypeCaller detected insertions and deletions, and then filtered using the VariantFiltration program based on the following criteria: cluster WindowSize: 10; MQ0 $\geq$ 4 and ((MQ0/(1.0 $\times$ DP)) > 0.1), DP < 10, QUAL < 30.0, QUAL $\geq$ 30 and QUAL < 50, QD < 1.5, HRun > 5, SB > $-$0.1. Detected INDELs were annotated using SnpEff. Finally, all the annotated information was provided in variant call format (VCF) files. An Integrative Genomics Viewer (Broad Institute) was used for visualising mapping data using BAM files and variant data using VCF files.

2.3. Statistical Analyses of Identified INDELs

The chromosome, position, reference (Ref) allele, alternative (Alt) allele, gene name, HGVSp, annotation, and annotation impact were collected from the VCF files of 101 horses and then integrated by Vcf2sql (Amelieff Co.). The allele frequency of diallelic INDELs on autosomal (1 to 31) chromosomes was calculated from the integrated data by Vcf2sql. Statistical software R (https://www.r-project.org/, accessed on 2 February 2023) and its package (tidyverse package, version 1.3.2) were used for statistical analyses of identified INDELs.

INDEL density was calculated as the number of detected variants in each chromosome multiplied by a scale factor of 1000 to calculate the number of INDELs per 1000 bp.

Identified INDELs were compared with those registered in Ensembl Release 108 (https://ftp.ensembl.org/pub/release-108/variation/gvf/equus_caballus/, accessed on 2 February 2023) using statistical software R.

SNV data identified in the 101 horses were obtained from our previous study [7] and the following site (https://doi.org/10.17605/OSF.IO/PVNCY, accessed on 2 February 2023). To map genome-wide INDELs and SNVs on the horse ideogram, R Ideogram (https://cran.r-project.org/web/packages/RIdeogram/vignettes/RIdeogram.html, accessed on 2 February 2023) was used.

3. Results

3.1. Numbers of Detected INDELs

The mapped region ($\times$10) and coverage of sequence reads were represented as the averages of 2,438,085,403 bp (2,384,904,267–2,557,836,663) and 36.8 coverages (29.5–54.2), respectively (Table S1). These data were considered sufficient for INDEL identification from the Thoroughbred genome because coverage of almost all horses was over 30, which is suitable coverage for identifying genome-wide variants [15]. The number of detected and filtered INDELs was identified with averages of 629,028 (532,978–673,567) and 585,680 (467,271–644,315), respectively (Table S1).

By integrating filtered INDELs, 1,453,349 and 113,047 loci were identified in the autosomal (1–31) and X chromosomes, respectively (Table 1 and Table S2). The majority of INDELs only had two alleles as 'diallelic' INDELs, while the others had three or more alleles as 'multiallelic' INDELs, which were tandem repeat sequences similar to microsatellites. Multiallelic INDELs were identified as 178,641 (12.29%) and 11,081 (9.80%) loci on the autosomal and X chromosomes, respectively (Table 1 and Table S2). Furthermore, 18 INDELs were detected by mapping to the mitochondrial (MT) genome (16,660 bp) (Tables 1 and S2). In total, 1,566,414 INDELs were identified through the genome of the 101 Thoroughbred horses.

Table 1. The number of INDELs detected in 101 Thoroughbred racehorses.

INDEL Category	Chromosomes 1 to 31	Chromosome X	Mitochondria
All INDELs	1,453,349	113,047	18
Diallelic INDELs	1,274,708	101,966	16
Multiallelic INDELs	178,641	11,081	2

Of the 1,566,414 INDELs, 786,957 loci (50.2%) were already registered in the Ensembl (Release 108), while 779,457 loci (49.8%) were novel INDELs (Table S3). Since 3,461,675 INDELs are registered in the Ensembl, the INDELs identified in this study occupied 22.7%.

As described below (see Section 3.4), INDELs with a minor allele frequency (MAF) of 0.5 and heterozygous genotype in all horses were counted independently, because it was unclear whether they were true variants. Among diallelic loci, excluding a MAF of 0.5 and heterozygous genotype in all horses, 676,249 (53.3%) were novel INDELs and 592,504 (46.7%) were registered INDELs (Table S3). Among diallelic loci with MAF of 0.5 and heterozygous genotype in all horses, 3199 (53.7%) were novel INDELs and 2756 (46.3%) were registered INDELs (Table S3). Among multiallelic loci, 44,299 (23.3%) were novel INDELs and 145,425 (76.7%) were registered INDELs (Table S3).

As 1,566,396 INDELs were detected in the Thoroughbred population through chromosomes 1–31 and X, one INDEL was detected every 1538 bp (=2,409,143,234 bp as the total base pairs of all chromosomes/1,566,396 INDELs) on average.

Figure 1A shows the number of INDELs detected on each chromosome. Chromosomes 12 (on average 1.26 in 1000-bp as INDEL density) and 20 (1.10) showed a frequency of 1 or more in 1000 bp (Table S2). Except for these chromosomes, there were <1.0 in 1000 bp (0.54 in chromosome 14 to 0.88 in chromosome X, Table S2).

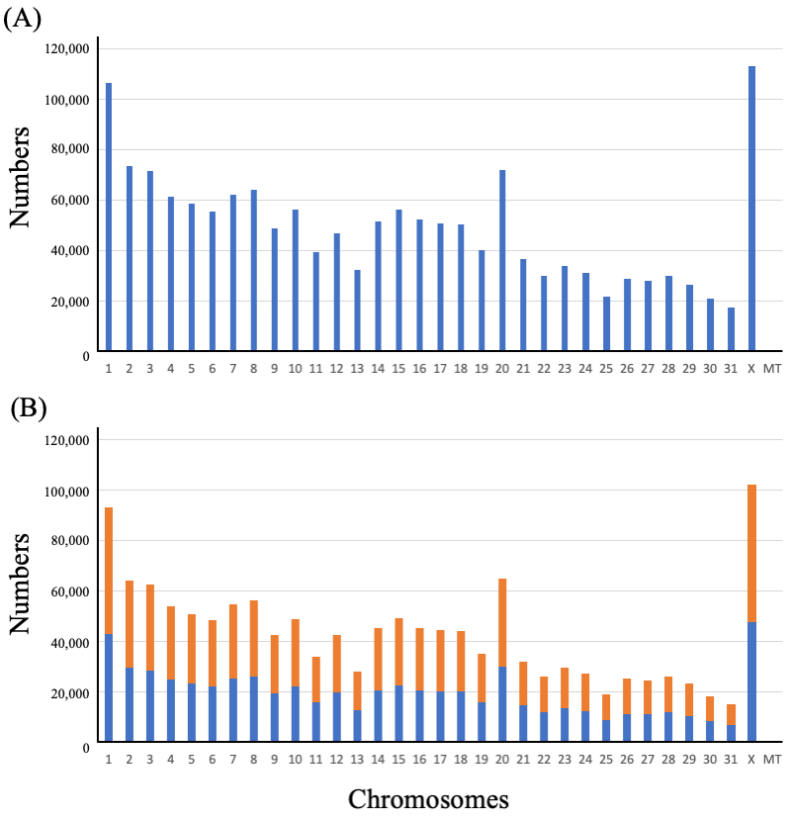


Figure 1. Distribution of insertions and deletions (INDELs) detected in 101 Thoroughbred horses. (**A**) Distribution of all INDELs identified on chromosomes 1–31, X, and MT; and (**B**) Distribution of insertions and deletions of diallelic INDELs on chromosomes 1–31, X, and MT (blue: insertions; orange: deletions).

3.2. Sizes of Detected INDELs

In 1,376,674 diallelic INDELs in autosomal and X chromosomes, 742,723 (54.0%) and 633,951 (46.0%) loci were detected as deletions and insertions, respectively (Figure 1B). The sizes of the detected INDELs (calculated as Alt allele length−Ref allele length) were in the range of −286–+476, and 1-bp, 2-bp, 3-bp, and 4-bp INDELs accounted for the majority, with 717,736 (52.14%), 220,672 (16.03%), 101,756 (7.39%), and 83,780 (6.09%), respectively (Figure 2).

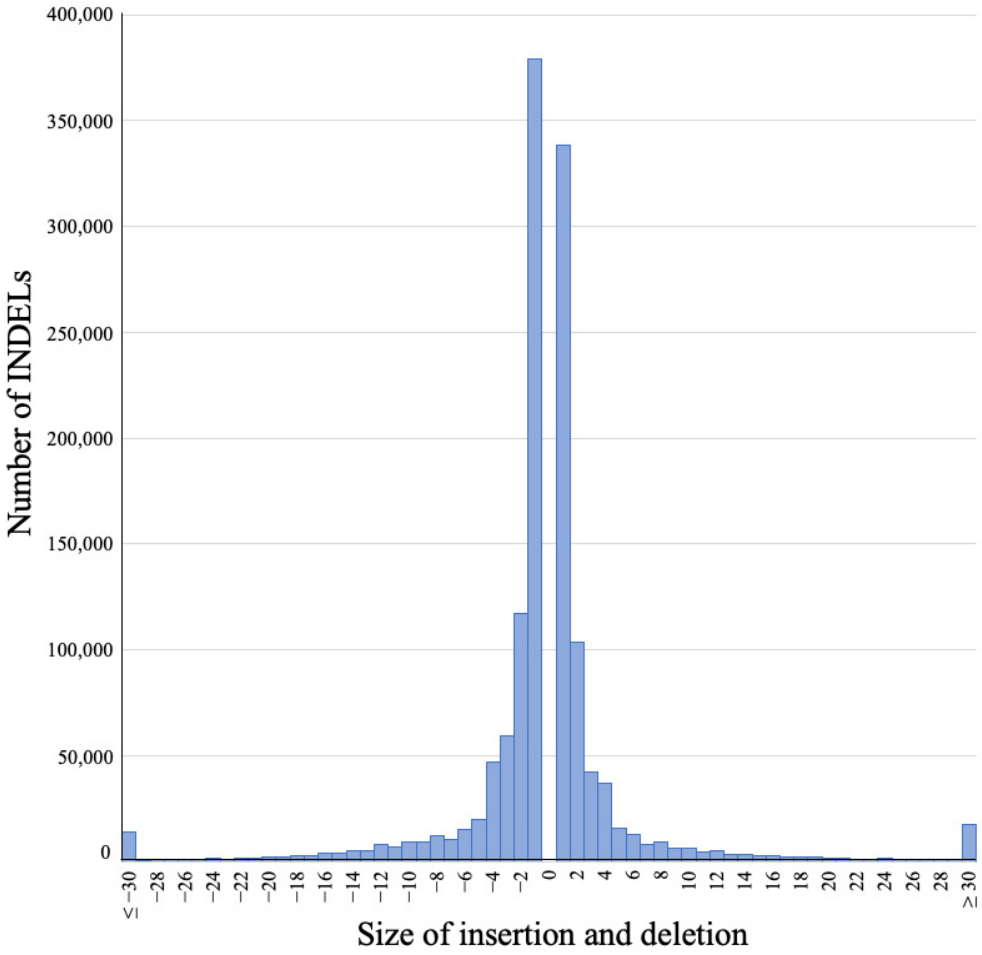


Figure 2. Distribution of insertion and deletion (INDEL) size detected in 101 Thoroughbred horses. The size of INDELs was calculated as Alt allele length–Ref allele length.

3.3. Allelic Frequency Distribution of INDELs Detected

The allelic frequency distributions of 1,274,708 diallelic INDELs identified on autosomes were investigated in 101 Thoroughbred racehorses. The horizontal axis of the figure indicates the number (frequency) of Ref alleles in the population (202 alleles in total), meaning that the right edge shows a MAF of 0.00495 (<0.5%) for Alt allele, the centre shows MAF of 0.5, and the left edge shows a MAF of 0 (Alt/Alt genotypes in all horses). Many INDELs were detected as having a smaller frequency of Alt alleles. In particular, many rare variants with a low Alt allele frequency (<1%) in the population were detected (Figure 3).

In the 1,274,708 diallelic INDELs, 141,686 loci had a MAF of 0.00495 for Alt alleles (see the right edge of Figure 3), whereas 881 loci showed Alt allele homozygotes for all individuals in the population (see the left edge of Figure 3). In addition, a blip was observed at 101 Ref alleles/101 Alt alleles (see the centre of Figure 3), which corresponds to a MAF of 0.5. Interestingly, although 9228 INDELs on autosomal chromosomes (1–31) had a MAF of 0.5, 5955 (64.5%) were of the heterozygous genotype in all horses.

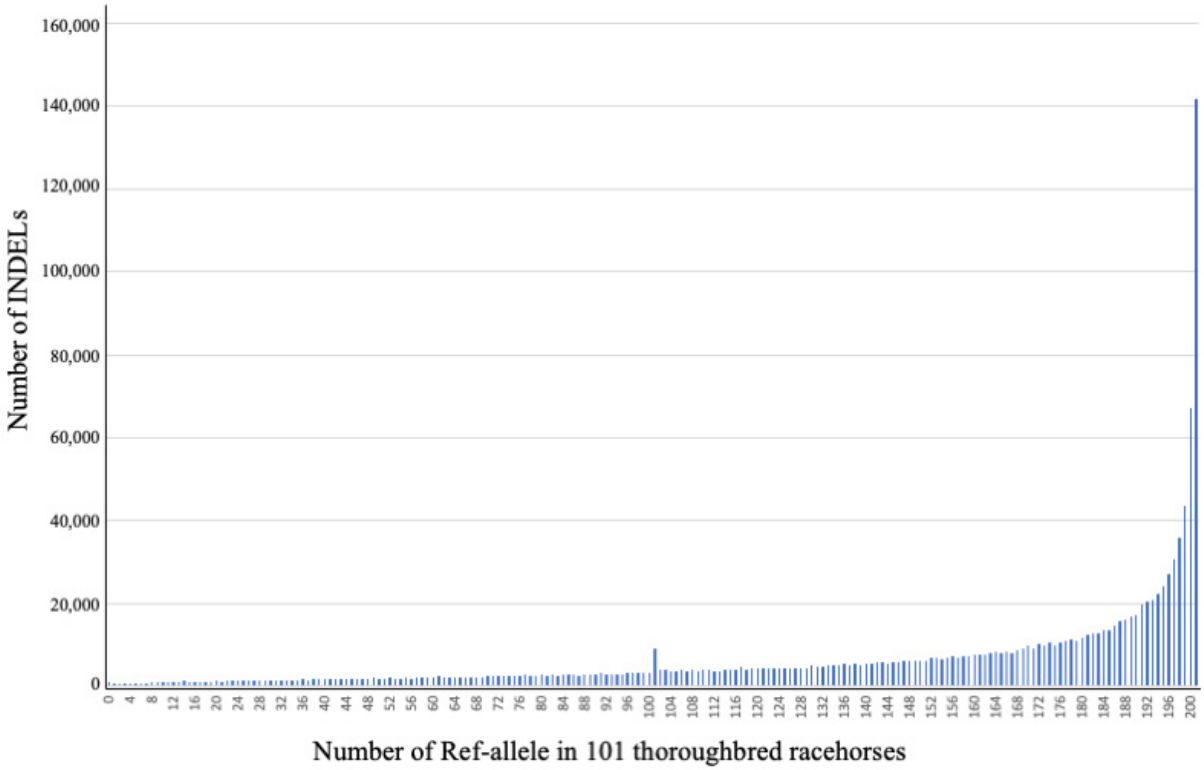


Figure 3. Allelic distribution of 1,274,708 diallelic insertions and deletions (INDELs) identified on autosomes (chromosomes 1–31) in 101 Thoroughbred racehorses. The horizontal axis indicates the number (frequency) of Ref alleles in the population (202 alleles in total), meaning that the right edge shows a MAF of 0.00495 for Alt allele, the centre shows a MAF of 0.5, and the left edge shows a MAF of 0 (Alt/Alt genotypes in all horses).

In the diallelic INDELs identified on chromosomes 1 to 31, excluding variants with a MAF of 0.5 and all heterozygotes in all horses, novel INDELs that were not registered in the Ensembl (Release 108) were mainly distributed in fewer Alt allele counts in 101 Thoroughbred racehorses (Table S4); this is similar to the frequency distribution in Figure 3, meaning that many novel INDELs were detected as rare variants. Interestingly, novel 317 loci were identified as all Alt allele homozygotes in the 101 horses, meaning that all the individuals analysed did not have reference variants for the 317 loci (Table S4).

3.4. Detection of Genomic Regions Having MAF of 0.5 and All Heterozygous

The genomic regions of the 5955 INDELs with a MAF of 0.5 and heterozygous genotype in all horses were visualized on horse chromosomes (Figure 4, blue). In our previous study [7], 58,582 loci on chromosomes 1–31 were identified as SNVs with a MAF of 0.5 and heterozygous genotypes in all horses. These SNVs were also visualized on the horse chromosomes (Figure 4, pink). These genomic regions were densely distributed with SNVs and INDELs, and the regions were similar.

Notably, 1137 of these INDELs were in the pericentromeric region (position: 1–2,261,991) of ECA29. A similar trend was observed in SNVs with a MAF of 0.5 and heterozygous genotype in all horses (chromosome 29: 1–2,293,071). In addition, the other INDELs and SNVs with a MAF of 0.5 and heterozygous genotypes in all horses were widely distributed over short lengths (Figure 4). INDELs with a MAF of 0.5 and heterozygous genotype in all horses presented more densely in chromosome 12 compared to other chromosomes. These regions showed high coverage of mapped reads upon observation with the Integrative Genomics Viewer. Complementing the distribution patterns of SNVs and INDELs would enable the elucidation of more detailed genomic structures.

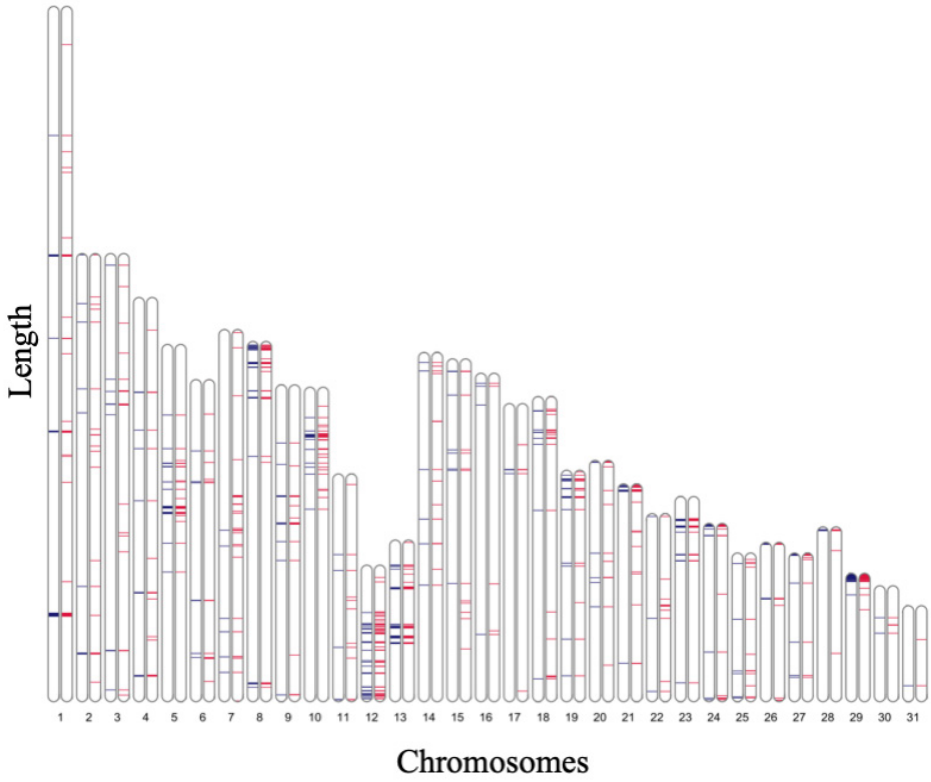


Figure 4. Locations of insertions and deletions (INDELs) and single-nucleotide variants (SNVs) with a MAF of 0.5 and heterozygous genotypes in all horses. Blue: INDELs detected in this study; pink: SNVs detected in our previous study. Many regions are matched between the loci of SNVs and INDELs. The pericentromeric region (position: 1–2,261,991, approximately 2.26 Mb) of ECA29 was densely distributed.

3.5. Characterisation of INDELs Detected

The number of INDELs detected in the functional region of the genome was also investigated (Table 2). Although most INDELs were in the intergenic and intron regions (Table S5), 12,432 loci were found in exons (including non-coding RNA). Of these, 4529, 1871, 5709, and 323 were frameshifts, non-frameshifts, long non-coding RNAs and pseudogenes, respectively. As known non-frameshift INDELs in horses, an INDEL of the agouti signalling protein (*ASIP*) gene was detected with a reference allele frequency of 0.248, while a short interspersed nuclear element (SINE) insertion was not detected in the myostatin (*MSTN*) gene region, based on our detection criteria.

Table 2. Characterisation of insertions and deletions detected in 101 Thoroughbred racehorses.

Region	Chromosomes 1 to 31	Chromosome X	Mitochondria
Upstream	101,434	5633	16
5′-UTR	2071	85	0
Exon	12,432	870	7
Intron	508,310	33,247	0
3′-UTR	3258	227	0
Downstream	96,718	5989	16
Intergenic	754,952	67,934	3

UTR: untranslated region.

Notably, 29 INDELs located in exons of 23 genes all had alternative alleles, causing frameshifts in the 101 horses (Tables S6 and S7). The distribution density (0.640%) of frameshift INDELs was higher than that of non-frameshift INDELs (0.160%) in the 0% Ref allele frequency (left side of Figure 5). While 5 genes were not annotated (EN-

SECAG00000020750, ENSECAG00000022530, ENSECAG00000026851, ENSECAG00000030750, and ENSECAG00000033068), 18 genes were annotated as follows; ankyrin repeat domain 9 (*ANKRD9*), Rho GTPase-activating protein 45 (*ARHGAP45*), cache domain containing 1 (*CACHD1*), cardiotrophin 1 (*CTF1*), EMAP-like 4 (*EML4*), fibrosin-like 1 (*FBRSL1*), HDGF-like 2 (*HDGFL2*), junctophilin 2 (*JPH2*), mitochondrial ribosomal protein L15 (*MRPL15*), nuclear receptor corepressor 2 (*NCOR2*), SH3 and multiple ankyrin repeat domains 3 (*SHANK3*), serine/threonine kinase 11 (*STK11*), trinucleotide repeat-containing 18 (*TNRC18*), UPF1 RNA helicase and ATPase (*UPF1*), zinc finger protein 212 (*ZNF212*), zinc finger protein 282 (*ZNF282*), zinc finger protein 516 (*ZNF516*), and zinc finger protein 853 (*ZNF853*).

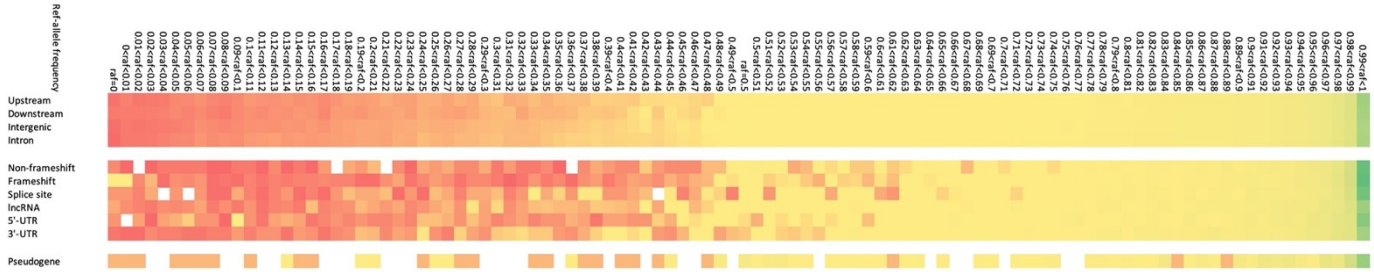


Figure 5. Distribution density of insertions and deletions (INDELs) in genomic functional regions. The distribution density (0.640%: yellow) of frameshift INDELs was higher than that of non-frameshift INDELs (0.160%: red) in the 0% Ref allele frequency (left side of Figure 5). Vertical line: upstream, downstream, intergenic, intron, non-frameshift, frameshift, splice site, long non-coding RNA, 5′-UTR, and 3′-UTR. Horizontal line: reference allele frequency, left side: low frequency, right side: high frequency. Red, yellow, and green show low, middle, and high densities, respectively.

4. Discussion

In this study, 1,566,414 loci were identified as INDELs on autosomal and X chromosomes and mitochondrial genomes. It was confirmed that 779,457 loci (49.8%) were novel INDELs by comparing them with the Ensembl (Release 108). Our current and previous studies [7,16] identified 12 million SNVs, 1.56 million short INDELs, and 62 processed pseudogenes in 101 Thoroughbred horses. Although Thoroughbreds were bred as a closed population from a small founder population, it was demonstrated that current Thoroughbred populations have diverse genome structures.

A recent study estimated a rate of 2.94 INDELs (1–20 bp) and 0.16 structural variants (>20 bp) per generation based on the WGS of 250 human families [17]. In this study, 96.36% of diallelic INDELs were in the size range of 1–20 bp. Therefore, although many rare INDELs were detected in the present study, a greater number of de novo INDELs may be identified by analysing Thoroughbreds other than the 101 horses analysed in this study. However, for de novo INDELs to remain in Thoroughbred populations, they must occur in stallions or broodmares. As few horses are used as breeding stallions or broodmares, many de novo INDELs in other horses will be lost without inheritance. In this study, many INDELs were detected as rare variants (e.g., INDELs with 1 Alt allele count in the 101 horses). These INDELs are variants that may not have remained in the population.

Interestingly, 5955 loci were genotyped with a MAF of 0.5 and heterogeneous genotypes in all the horses. Such genotype distribution is unlikely in a randomly collected population. Because these variants existed in continuous positions with high mapping coverage, fake variants may be detected by mapping of similar sequence reads in multiple regions. While short-read sequencing and its mapping to the reference genome is an easy way to detect variants, fake variants may be detected. Therefore, our data help to identify true variants in Thoroughbreds. As 2756 loci were registered in Ensembl (Release 108), further validation, such as designing their primers and probes, may need to use them as genetic markers.

High INDEL density was observed on chromosomes 12 and 20. Previous studies observed a similar trend on the same chromosomes for SNVs and nonsynonymous substitutions [7,16]. Although the exact reason for this is unknown, one reason is thought to be the existence of metabolic and sensory perception regions on chromosome 12, and immune response and antigen-processing regions on chromosome 20 [18–20]. These genes have several pseudogenes that lose their function because of substitutions, insertions, and deletions in the gene regions. In addition, INDELs were densely populated on chromosome 12, perhaps because of the effect of duplicated regions.

INDELs in coding regions may be frameshift or non-frameshift variants. In this study, the same variant (C/CAGCAGAAAAGA) as a non-frameshift INDEL in the *ASIP* gene was observed, and its genotypes were associated with coat colour [21,22]. In addition, SINE insertion in the *MSTN* gene is associated with optimum race distance and muscle mass in Thoroughbred racehorses [23,24]. However, the SINE insertion was not detected in our analysis pipeline (see Materials and Methods), while reads mapped to the promoter region in *MSTN* had partial sequences soft-clipped by confirmation using the Integrative Genomics Viewer. Although INDELs over 400 bp were detected in this study; the SINEs in the *MSTN* region were not recognised as INDELs by our detection method, because similar sequences were widely distributed in the horse genome.

Frameshift INDELs are generally deleterious and contribute to disease susceptibility, as evidenced by human genomics research that has identified many INDELs related to genetic diseases including cancer [25]. Notably, 29 INDELs detected in exons of 23 genes had Alt alleles causing frameshifts in the 101 horses, indicating that all genotypes were frameshifted in the population. Interestingly, four of them were family genes of zinc finger protein, ZNF212, ZNF282, ZNF516, and ZNF853. Zinc finger proteins are generally involved in gene regulation and development thorough binding to DNA or RNA. Although no functional commonality was observed in the other detected genes, some of them seem to involve in intracellular signalling, such as FBRSL1, NCOR2, and UPF1. Although these genes were important for cellular functions, they may be complemented by different signalling pathways even if the genes did not work well by frameshifting. It is expected that genes containing INDELs will acquire special functions by frameshifting, and that the genome structure of Twilight (the reference genome) will be unique. However, mistakes in genome assembly, gene annotation, and INDEL calling cannot be ruled out. Further detailed studies are needed because INDELs identified in exons may be related to positive selection, genetic diseases, and/or diverse functions in Thoroughbreds.

In human forensic sciences, INDELs have been noted as markers for individual identification testing [26,27] because they are diallelic, small in size, and widely distributed throughout the genome. Furthermore, INDELs have the advantage of lower mutation rates than SNPs and short tandem repeats (STRs) [28]. In the present study, we propose 97,443 INDELs (chromosome: autosomes, allele number: diallelic, size: −2 to −4 and +2 to +4, allele frequency: 0.25 to 0.75, excluding INDELs with all heterozygous genotypes) as marker candidates to construct a panel of horse parentage testing. Registration of Thoroughbred racehorses requires parentage testing using STRs recommended by the International Society for Animal Genetics [29,30]. While parentage testing using SNPs is being developed [31,32], the INDEL panel may also serve as a complementary panel.

Gene doping in the horseracing industry is defined as administration of exogenous genes or therapeutic oligonucleotides to postnatal animals, and the creation of genetically modified animals. These have been prohibited by the IFHA and the ISBC. While the former can be target-specifically detected by quantitative PCR using a hydrolysis probe or sequencing using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry [33,34], the latter has been extremely difficult to detect. One of the reasons for this is that we do not fully understand the genomic diversity of Thoroughbred populations. Therefore, the results of our study we consider to contribute to gene-doping control. Recently, we developed a gene-editing test to detect these racehorses [14]. This test uses the following criterion to identify artificial editing: the presence of homologues of Alt-type

INDELs that were not shown in current Thoroughbred populations. Therefore, as this study analysed and validated the types, sizes, locations, and frequencies of INDELs in the current Thoroughbred population, the results will contribute to the gene-editing test. Interestingly, this study identified many rare INDELs which are of low frequency in the Thoroughbred population. While rare INDELs may be naturally occurring as de novo mutations, the presence of rare INDELs may complicate positive or negative determination in the gene-editing test. If de novo mutations occur in stallions, the mutations are inherited by many offspring. Therefore, the variant database constructed in this study should be regularly updated for accurate gene-editing testing. However, our data may include up to several thousand false-positive variants calls because of current mapping and variant-calling algorithms. Therefore, developing improved variant-calling algorithms is a future research priority and is required for industrial applications such as precise gene-doping control.

The INDEL database we have constructed will provide useful information for genetic studies and industrial applications in Thoroughbred horses, including a gene-editing test for gene-doping control and a parentage test using INDELs for horse registration and identification.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14030638/s1, Table S1: Summary of mapping and INDEL calling; Table S2: Number of INDELs detected at individual chromosomes in 101 Thoroughbred racehorses; Table S3: The number of INDELs identified in this study; Table S4: Distribution of novel diallelic INDELs excluding MAF of 0.5 and heterozygous genotype in all horses; Table S5: Characterisation of INDELs detected at individual chromosomes in 101 Thoroughbred racehorses; Table S6: INDEL distribution based on allele frequency; Table S7: Genes that are homozygous for alternative alleles and are the cause of frameshifts in all 101 horses.

Author Contributions: T.T. and A.O. conceived of and designed the experiments; T.T., A.O., M.K., T.I., H.K., K.-i.H., Y.T. and S.-i.N. performed the experiments and data analyses; T.T. and A.O. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Japan Racing Association (2020–2022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study can be accessed through the Open Science Framework (https://OSF.IO/QN45H/, accessed on 2 February 2023). FASTQ data used in this study are available from the DDBJ (BioProject: PRJDB15140, BioSample: SAMD00573909 to SAMD00574009, DRA015615).

Acknowledgments: We thank Noriko Tanaka for her assistance with this study.

Conflicts of Interest: There are no competing interests, including patents, products in development, or marketed products, to declare concerning this study.

References

1. Bower, M.A.; Campana, M.G.; Whitten, M.; Edwards, C.J.; Jones, H.; Barrett, E.; Cassidy, R.; Nisbet, R.E.; Hill, E.W.; Howe, C.J.; et al. The cosmopolitan maternal heritage of the Thoroughbred racehorse breed shows a significant contribution from British and Irish native mares. *Biol. Lett.* **2011**, *7*, 316–320. [CrossRef] [PubMed]
2. International Stud Book Committee (Resouces/Thoroughbred Horse Statistics). Available online: https://www.internationalstudbook.com/resources/ (accessed on 2 February 2023).
3. Wade, C.M.; Giulotto, E.; Sigurdsson, S.; Zoli, M.; Gnerre, S.; Imsland, F.; Lear, T.L.; Adelson, D.L.; Bailey, E.; Bellone, R.R.; et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **2009**, *326*, 865–867. [CrossRef] [PubMed]
4. Kalbfleisch, T.S.; Rice, E.S.; DePriest, M.S., Jr.; Walenz, B.P.; Hestand, M.S.; Vermeesch, J.R.; O'Connell, B.L.; Fiddes, I.T.; Vershinina, A.O.; Saremi, N.F.; et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. *Commun. Biol.* **2018**, *1*, 197. [CrossRef] [PubMed]
5. Jagannathan, V.; Gerber, V.; Rieder, S.; Tetens, J.; Thaller, G.; Drögemüller, C.; Leeb, T. Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Anim. Genet.* **2019**, *50*, 74–77. [CrossRef]

6. Durward-Akhurst, S.A.; Schaefer, R.J.; Grantham, B.; Carey, W.K.; Mickelson, J.R.; McCue, M.E. Genetic variation and the distribution of variant types in the horse. *Front. Genet.* **2022**, *12*, 758366. [CrossRef]
7. Tozaki, T.; Ohnuma, A.; Kikuchi, M.; Ishige, T.; Kakoi, H.; Hirota, K.; Kusano, K.; Nagata, S. Rare and common variant discovery by whole-genome sequencing of 101 Thoroughbred racehorses. *Sci. Rep.* **2021**, *11*, 16057. [CrossRef]
8. Tozaki, T.; Hamilton, N.A. Control of gene doping in human and horse sports. *Gene Ther.* **2022**, *29*, 107–112. [CrossRef]
9. Moro, L.N.; Viale, D.L.; Bastón, J.I.; Arnold, V.; Suvá, M.; Wiedenmann, E.; Olguín, M.; Miriuka, S.; Vichera, G. Generation of myostatin edited horse embryos using CRISPR/Cas9 technology and somatic cell nuclear transfer. *Sci. Rep.* **2020**, *10*, 15587. [CrossRef]
10. Kim, D.E.; Lee, J.H.; Ji, K.B.; Park, K.S.; Kil, T.Y.; Koo, O.; Kim, M.K. Generation of genome-edited dogs by somatic cell nuclear transfer. *BMC Biotechnol.* **2022**, *22*, 19. [CrossRef]
11. Sheets, T.P.; Park, C.H.; Park, K.E.; Powell, A.; Donovan, D.M.; Telugu, B.P. Somatic Cell Nuclear Transfer Followed by CRIPSR/Cas9 Microinjection Results in Highly Efficient Genome Editing in Cloned Pigs. *Int. J. Mol. Sci.* **2016**, *17*, 2031. [CrossRef]
12. Hill, E.W.; Gu, J.; Eivers, S.S.; Fonseca, R.G.; McGivney, B.A.; Govindarajan, P.; Orr, N.; Katz, L.M.; MacHugh, D.E. A sequence polymorphism in MSTN predicts sprinting ability and racing stamina in thoroughbred horses. *PLoS ONE* **2010**, *5*, e8645. [CrossRef]
13. Tozaki, T.; Sato, F.; Hill, E.W.; Miyake, T.; Endo, Y.; Kakoi, H.; Gawahara, H.; Hirota, K.; Nakano, Y.; Nambo, Y.; et al. Sequence variants at the myostatin gene locus influence the body composition of Thoroughbred horses. *J. Vet. Med. Sci.* **2011**, *73*, 1617–1624. [CrossRef]
14. Tozaki, T.; Ohnuma, A.; Nakamura, K.; Hano, K.; Takasu, M.; Takahashi, Y.; Tamura, N.; Sato, F.; Shimizu, K.; Kikuchi, M.; et al. Detection of indiscriminate genetic manipulation in Thoroughbred racehorses by targeted resequencing for gene-doping control. *Genes* **2022**, *13*, 1589. [CrossRef]
15. Bentley, D.R.; Balasubramanian, S.; Swerdlow, H.P.; Smith, G.P.; Milton, J.; Brown, C.G.; Hall, K.P.; Evers, D.J.; Barnes, C.L.; Bignell, H.R.; et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53–59. [CrossRef] [PubMed]
16. Tozaki, T.; Ohnuma, A.; Kikuchi, M.; Ishige, T.; Kakoi, H.; Hirota, K.; Kusano, K.; Nagata, S. Identification of processed pseudogenes in the genome of Thoroughbred horses: Possibility of gene-doping detection considering the presence of pseudogenes. *Anim. Genet.* **2022**, *53*, 183–192. [CrossRef] [PubMed]
17. Kloosterman, W.P.; Francioli, L.C.; Hormozdiari, F.; Marschall, T.; Hehir-Kwa, J.Y.; Abdellaoui, A.; Lameijer, E.W.; Moed, M.H.; Koval, V.; Renkens, I.; et al. Characteristics of de novo structural changes in the human genome. *Genome Res.* **2015**, *25*, 792–801. [CrossRef] [PubMed]
18. Al Abri, M.A.; Holl, H.M.; Kalla, S.E. Sutter NB, Brooks SA. Whole genome detection of sequence and structural polymorphism in six diverse horses. *PLoS ONE* **2020**, *15*, e0230899. [CrossRef]
19. Miller, D.; Tallmadge, R.L.; Binns, M.; Zhu, B.; Mohamoud, Y.A.; Ahmed, A.; Brooks, S.A.; Antczak, D.F. Polymorphism at expressed DQ and DR loci in five common equine MHC haplotypes. *Immunogenetics* **2017**, *69*, 145–156. [CrossRef]
20. Tallmadge, R.L.; Lear, T.L.; Antczak, D.F. Genomic characterization of MHC class I genes of the horse. *Immunogenetics* **2005**, *57*, 763–774. [CrossRef]
21. Rieder, S.; Taourit, S.; Mariat, D.; Langlois, B.; Guérin, G. Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mamm. Genome* **2001**, *12*, 450–455. [CrossRef]
22. Kakoi, H.; Tozaki, T.; Nagata, S.; Gawahara, H.; Kijima-Suda, I. Development of a method for simultaneously genotyping multiple horse coat colour loci and genetic investigation of basic colour variation in Thoroughbred and Misaki horses in Japan. *J. Anim. Breed. Genet.* **2009**, *126*, 425–431. [CrossRef] [PubMed]
23. Rooney, M.F.; Hill, E.W.; Kelly, V.P.; Porter, R.K. The "speed gene" effect of myostatin arises in Thoroughbred horses due to a promoter proximal SINE insertion. *PLoS ONE* **2018**, *13*, e0205664. [CrossRef]
24. Petersen, J.L.; Valberg, S.J.; Mickelson, J.R.; McCue, M.E. Haplotype diversity in the equine myostatin gene with focus on variants associated with race distance propensity and muscle fiber type proportions. *Anim. Genet.* **2014**, *45*, 827–835. [CrossRef] [PubMed]
25. Chen, J.; Guo, J.T. Structural and functional analysis of somatic coding and UTR indels in breast and lung cancer genomes. *Sci. Rep.* **2021**, *11*, 21178. [CrossRef] [PubMed]
26. Song, F.; Lang, M.; Li, L.; Luo, H.; Hou, Y. Forensic features and genetic background exploration of a new 47-autosomal InDel panel in five representative Han populations residing in Northern China. *Mol. Genet. Genom. Med.* **2020**, *8*, e1224. [CrossRef] [PubMed]
27. Chen, X.; Nie, S.; Hu, L.; Fang, Y.; Cui, W.; Xu, H.; Zhao, C.; Zhu, B.F. Forensic efficacy evaluation and genetic structure exploration of the Yunnan Miao group by a multiplex InDel panel. *Electrophoresis* **2022**, *43*, 1765–1773. [CrossRef] [PubMed]
28. Huang, Y.; Liu, C.; Xiao, C.; Chen, X.; Han, X.; Yi, S.; Huang, D. Mutation analysis of 28 autosomal short tandem repeats in the Chinese Han population. *Mol. Biol. Rep.* **2021**, *48*, 5363–5369. [CrossRef] [PubMed]
29. Kakoi, H.; Nagata, S.; Kurosawa, M. DNA Typing with 17 microsatellites for parentage verification of racehorses in Japan. *Anim. Sci. J.* **2001**, *72*, 453–460. [CrossRef]

30. Tozaki, T.; Kakoi, H.; Mashima, S.; Hirota, K.; Hasegawa, T.; Ishida, N.; Miura, N.; Choi-Miura, N.H.; Tomita, M. Population study and validation of paternity testing for Thoroughbred horses by 15 microsatellite loci. *J. Vet. Med. Sci.* **2001**, *63*, 1191–1197. [CrossRef] [PubMed]
31. Hirota, K.; Kakoi, H.; Gawahara, H.; Hasegawa, T.; Tozaki, T. Construction and validation of parentage testing for thoroughbred horses by 53 single nucleotide polymorphisms. *J. Vet. Med. Sci.* **2010**, *72*, 719–726. [CrossRef]
32. Holl, H.M.; Vanhnasy, J.; Everts, R.E.; Hoefs-Martin, K.; Cook, D.; Brooks, S.A.; Carpenter, M.L.; Bustamante, C.D.; Lafayette, C. Single nucleotide polymorphisms for DNA typing in the domestic horse. *Anim. Genet.* **2017**, *48*, 669–676. [CrossRef] [PubMed]
33. Tozaki, T.; Ohnuma, A.; Kikuchi, M.; Ishige, T.; Kakoi, H.; Hirota, K.; Kusano, K.; Nagata, S. Microfluidic quantitative PCR detection of 12 transgenes from horse plasma for gene doping control. *Genes* **2020**, *11*, 457. [CrossRef] [PubMed]
34. Tozaki, T.; Kwak, H.G.; Nakamura, K.; Takasu, M.; Ishii, H.; Ohnuma, A.; Kikuchi, M.; Ishige, T.; Kakoi, H.; Hirota, K.; et al. Sequence determination of phosphorothioated oligonucleotides using MALDI-TOF mass spectrometry for controlling gene doping in equestrian sports. *Drug Test. Anal.* **2022**, *14*, 175–180. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.