BMC Genomic Data

RESEARCH

Open Access

Identification of the genetic background of laboratory rats through amplicon-based next-generation sequencing for single-nucleotide polymorphism genotyping

Meng Lu[1], Kai Li[1], Yuxun Zhou[1] and Junhua Xiao[1*]

Abstract

Background Laboratory rats, as model animals, have been extensively used in the fields of life science and medicine. It is crucial to routinely monitor the genetic background of laboratory rats. The conventional approach relies on gel electrophoresis and capillary electrophoresis (CE) technologies. However, the experimental and data analysis procedures for both of these methods are time consuming and costly.

Results We established a single-nucleotide polymorphism (SNP) typing scheme using multiplex polymerase chain reaction (PCR) and next-generation sequencing (NGS) to address the genetic background ambiguity in laboratory rats. This methodology involved three rounds of PCR and two rounds of magnetic bead selection to improve the quality of the sequencing data. We simultaneously analysed 100 laboratory rats (including rats of 5 inbred strains and 2 in-house closed colonies), and the sequencing depth varied from an average of 108.25 to 5189.89, with sample uniformity ranging from 82.5 to 97.5%. A total of 98.9% of the amplicons were successfully genotyped (≥ 30 reads). Genetic background analysis revealed that all 38 experimental rats from the 5 inbred strains were successfully identified (without a heterozygous allele). For the 2 in-house closed colonies, the average heterozygosity (0.162 and 0.169) deviated from the typical range of 0.5–0.7, indicating a departure from the ideal heterozygosity level. Additionally, we employed multiplex PCR-CE to validate the NGS-based method, which yielded consistent results for all the rat strains. These results demonstrated that this approach significantly improves efficiency, saves time, reduces costs and ensures accuracy.

Conclusion By utilizing NGS technology, our developed method leverages SNP genotyping for genetic background identification in laboratory rats, demonstrating advantages in terms of labour efficiency and cost-effectiveness, thereby rendering it well suited for projects involving extensive sample cohorts.

Keywords Multiplex PCR, Next-generation sequencing, Laboratory rat, SNP genotyping, Genetic background

*Correspondence:
Junhua Xiao
xiaojunhua@dhu.edu.cn
[1]College of Biological Science and Medical Engineering, Donghua University, 2999 Renmin North Road, Shanghai 201620, China

Background

The laboratory rat (*Rattus norvegicus*) is an important model organism that has provided new insights into disease mechanisms, in toxicological research and for the development of new compounds [1, 2]. Laboratory rats can be classified into inbred, chromosome substitution, congenic, closed colony and gene-edited strains on the basis of their genetic features [3]. The genetic stability of laboratory rats is an important factor for the accuracy and reliability of experiments. However, owing to the enormous number of strains and differences in feeding conditions, many substrains with differences in genetic information, especially inbred substrains, are produced from the same strains. Therefore, it is necessary to monitor genetic quality regularly during the process of raising these animals.

An ideal genotyping approach should be simple and highly cost-effective and have high throughput [4]. Currently, microsatellites and single-nucleotide polymorphisms (SNPs) are considered the gold standard for genetic quality control in laboratory animals, as they can be used to distinguish different genetic backgrounds [5]. Microsatellites, also known as short tandem repeats (STRs) and simple sequence length polymorphisms (SSLPs), are used in genetic monitoring programs for laboratory rats because of their cost-effectiveness and ease of typing [6, 7]. An STR is an array of repeat motifs with a repetition size of 1 kb or less and a width ranging from 2 to 6 bp [8]. To date, more than 50,000 STR markers have been identified and characterized for amplifying STR sequences from genomic DNA via singleplex PCR or multiplex PCR, after which the PCR products are analysed via agarose or polyacrylamide gel electrophoresis or capillary electrophoresis (CE) [6, 7, 9–11]. However, to distinguish loci labelled with the same fluorescent tag, they must be separated by size. Additionally, during the analysis of STRs using CE, the occurrence of peak shifting poses a challenge, potentially resulting in typing inaccuracies. Moreover, these methods are costly and require complex operations, making them unsuitable for large cohorts.

SNPs are single-nucleotide variations in individual genomes. Almost all SNPs are biallelic, and they can be homozygous (A/A or C/C) or heterozygous (A/C) in an individual. Over 13 million SNPs have been identified in 27 rat strains [12]. Over 66% of the SNPs are present in all rat strains [13, 14]. Therefore, for genetic background identification in laboratory rats, the use of highly abundant, high-density, stably heritable and cost-effective SNPs is the ideal genotyping approach to replace STR genotyping. Currently, all SNP genotyping approaches available have intrinsic advantages and disadvantages. Traditional methods, such as PCR- restriction fragment length polymorphism (PCR-RFLP), PCR-single strand conformation polymorphism (PCR-SSCP), and PCR-ligase detection reaction (PCR-LDR), are rarely used because of their low throughput, long cycle time, high cost, and cumbersome operation procedures [4, 15, 16]. Therefore, it is necessary to develop a cost-effective and simple approach for the genetic monitoring of laboratory rats.

With technological development, next-generation sequencing (NGS) has emerged as a highly useful tool for the simultaneous analysis of multiplex target regions in large samples with high data quality [17]. NGS, owing to its high cost-effectiveness, high sensitivity and flexibility and short turnaround time, can overcome the limitations of traditional SNP analysis approaches, which require analysis of fragment size or temperature and chemical gradients to denature DNA molecules. Hence, an increasing number of researchers are focusing on the application of NGS for genetic quality assurance, such as in wheat, peanut and *Brassica napus* [18–22].

In this study, we developed a genetic monitoring method employing NGS technology to enhance efficiency and reduce costs in identifying the genetic background of laboratory rats through the genotyping of 119 SNP markers. This method was developed using comprehensive SNP information, and the detection range included common laboratory rat strains. In this method, target regions are enriched via a low number of multiplex PCR cycles, and the PCR products are further enriched via universal primers to increase the yield of the targeted products. Unique index primers are then ligated to the ends of PCR products from each sample via an index PCR to distinguish SNPs among various samples during sequencing. We subsequently selected 100 laboratory rats, including rats of 5 inbred strains (including an in-house inbred strain) and 2 in-house closed colonies, to validate our strategy on the Illumina sequencing platform. This approach yields desirable results in terms of genetic background identification for laboratory rat strains. Furthermore, we designed specific primers on the basis of the STR information obtained by genetic quality control [23] and subsequently constructed four sets of multiplex STR–PCR assays. The PCR products were then detected via CE, and the CE results were consistent with the NGS data. This study presents a high-throughput approach for monitoring the genetic background of common laboratory rats via SNP genotyping.

Materials and methods

Sample collection and genomic DNA extraction

Tails of 100 rats of the SHR (4 females and 4 males), GK (4 females and 4 males), F344 (4 females and 4 males), and WKY (3 females and 3 males) strains, an in-house inbred Wistar strain (4 females and 4 males, Wistarjin), an in-house Wistar closed colony (12 females and 10

males, Wistaryuan), and an in-house SD closed colony (40 males) were collected at 8 weeks of age from Slac Laboratory Animal Co., Ltd. (Shanghai, China). DNA extraction was performed using the Eltbio Mag Blood DNA Small Extraction Kit (Eltbio, Shanghai, China) according to the manufacturer's instructions.

Selection of SNP and STR loci

A total of 123 SNP loci for inbred rat strains were selected from 9,665,340 SNVs of 27 rat strains in the Rat Genome Database (http://rgd.mcw.edu/) (reference genome: mRatBN7.2). The selected SNPs met the following requirements: the SNP loci were (i) annotated; (ii) distant from STRs; (iii) excluded low heterozygosity (≤0.4); (iv) existed uniquely in the genome; (v) distant from repetitive structures, CNVs, long homopolymers, rich in polymorphisms, and extreme GC contents; and (vi) had pairwise distances between SNPs≥1 Mb.

A total of 26 autosomal STR loci were selected for analysis based on Genetic quality control [23]. These loci included 5 inbred rat loci (D3Wox9, D11mgh3, D12Mit2, Apoc3, and PA2S) and 24 closed colony rat loci (D1Rat345, D1Mgh14, D2Wox15, D2Mgh26, D3Wox9, D4Arb10, D4mit15, D6Mit1, D7Mgh3, D8Rat14, D9Mit2, D10Wox12, D11Mgh3, D11Wox3, D12Mit2, LCA, ALB, D15Mit3, MBPA, ACRM, TILP, D19Rat58, TNF, and PRPS2).

Primer design

For primer design for the SNPs, the sequences of 123 targeted regions containing SNP loci were downloaded from the Rat Genome Database (RGD, https://rgd.mcw.edu) (reference genome: mRatBN7.2). Primer3 v4.1.0 [24] was used to design specific primer pairs, each of which may be used to amplify a 170–240 nt amplicon that contains one SNP locus. Some options of Primer3 were modified to improve primer specificity. The following parameters were used: the optimum primer length range was set to 21 to 24 nt. The optimum annealing temperature (Tm) range was set to 60 to 64°C, and the maximum Tm difference was 2°C. The GC content range was set to 40–60%. Default settings were used for other parameters. The gene-specific primer regions were linked to universal sequences at the 5' end of the specific primers, corresponding to adapters compatible with the Illumina sequencing system (Supplementary Table S1).

For the PCR design of universal PCR, the universal primers had the same sequences as the specific primers (forward primer: 5'-ACGACGTGTCGAGTTCAGGT-3'; reverse primer: 5'-CAGTGAGTCGCCACAGGTCA-3').

The index primers were designed by Primer3 v4.1.0 [24] to distinguish among different samples and for sequencing. The index primers included P5 or P7 sequences, barcode sequences, sequencing primers, and universal sequences. The P5 or P7 sequences and sequencing primers were suitable for the Illumina sequencing system, and barcode sequences were used to distinguish among samples, universal sequences were used to add the index information to the 5 'end of PCR products. The primer information is shown in Supplementary Table S2.

The primer sets for the STR loci were designed via Primer3 v4.1.0 [24]. The primer design parameters used were as follows: primer length, 21–24 nt; Tm, 60–68°C; maximum Tm difference, 2°C; GC content, 40–60%; and product length, 110–450 nt. Other parameters were set to default values. For probe design, the 5' ends of the forward primers of each set were labelled with the fluorescent dye FAM or HEX for detection. The primer and probe information is shown in Supplementary Table S3.

Library preparation

Low cycle number for multiplex PCR

A low cycle number for multiplex PCR was employed for amplifying the target region, as described previously [25]. A 25 μL PCR mixture comprising 0.02 μM each primer, 5 μL of 5× multiplex PCR mix (Novoprotein, Suzhou, Jiangsu, China), and 50 ng of rat genomic DNA was prepared. In the first round of multiplex PCR, the cycling conditions included a predenaturation step at 94 °C for 5 min, followed by 3 cycles of 98 °C for 15 s, 60 °C for 30 min, and 72 °C for 2 min. The second round of PCR was performed with 4 cycles of 98 °C for 15 s, 60 °C for 2 min, and 72 °C for 2 min.

Two rounds of magnetic bead selection

The magnetic bead selection step was implemented as described previously [25]. A total volume of 48 μL, comprising 12 μL of PCR products, 18 μL of carrier DNA I (Sangon, Shanghai, China), 18 μL of ddH2O, and 31 μL of magnetic beads (Eltbio, Shanghai, China), was added to a tube and mixed by pipetting 30 times. The mixture was incubated at room temperature (RT) for 5 min, after which the reaction vessel was placed on a magnetic stand to isolate the supernatant. This supernatant was then combined with 11 μL of magnetic beads to isolate the targeted products. After the magnetic beads were washed with 85% ethanol for 30 s at RT, the products were eluted with 20 μL of Tris-EDTA (TE) buffer solution. The eluted products were subsequently purified again by mixing 15 μL of the eluate with 12 μL of carrier DNA II (Sangon, Shanghai, China), 21 μL of ddH2O and 44 μL of magnetic beads, and the elution procedure was repeated. Finally, the purified products were eluted in 20 μL of TE and stored at -20 °C until use.

Universal PCR

A 20 μL PCR mixture was prepared by combining 5 μL of 5 × multiplex PCR mix (Novoprotein, Suzhou, Jiangsu, China), 2 μL of universal primer mix (5 μM), 2 μL of the previously selected products, and 11 μL of RNase-free water. The thermocycling conditions were as follows: primary denaturation at 95 °C for 5 min, followed by 20 cycles of 98 °C for 15 s, 60 °C for 60 s, 70 °C for 30 s, and 72 °C for 30 s. The PCR products were analysed via electrophoresis on a 3% agarose gel containing Tris-borate EDTA buffer (w/v).

Index PCR

A 25 μL reaction mixture was prepared by combining 10 μL of PCR buffer (NiuHigh, Suzhou, Jiangsu, China), 2 μL of I5 primer mix (0.5 μM), 2 μL of I7 primer mix (0.5 μM), 2 μL of adapter primer mix (2.5 μM), 2 μL of previous PCR products, 0.25 μL of EzAmp® MPX Taq DNA Enzyme (NiuHigh, Suzhou, Jiangsu, China) and 6.75 μL of RNase-free water. The PCR cycles were as follows: primary denaturation at 94 °C for 15 min; 4 cycles of 98 °C for 15 s, 65 °C for 90 s, 70 °C for 30 s, and 72 °C for 30 s; and 16 cycles of 98 °C for 15 s, 68 °C for 45 s, 70 °C for 30 s, and 72 °C for 30 s.

Uniformity assessment by real-time fluorescence quantitative PCR

A real-time fluorescence quantitative PCR (qPCR) experiment was performed to evaluate library uniformity via SYBR Green-based detection on a SLAN-96 S real-time quantitative PCR detection system (Hongshi, Shanghai, China) with specific amplification primers (Supplementary Table S1). The index PCR products were diluted 60-fold with RNase-free water to be used as templates for qPCR. Each reaction mixture had a final volume of 20 μL, comprising 2.5 μL of template, 4 μL of each primer (5 μM), 3.5 μL of RNase-free water, and 10 μL of NovoStart SYBR qPCR SuperMix Plus (Novoprotein, Suzhou, Jiangzu, China). The amplification profile consisted of an initial predenaturation step at 94 °C for 5 min, followed by 35 cycles of 94 °C for 15 s, 60 °C for 30 s, and 70 °C for 45 s, with fluorescence signal collection at 70 °C. The cycle threshold (Ct) value indicates the cycle number at which the fluorescence exceeds the fixed threshold.

The percentage of amplicons with Ct value differences ≤ 5 serves as an indicator of amplicon uniformity [25]. The mean Ct ($\bar{Ct}$) value was calculated for all amplicons, followed by computation of the deviation of each amplicon (ΔCt) via the formula $\Delta Ct = Ct - \bar{Ct}$. For amplicon uniformity assessment, a ΔCt within the range of -2.5 to +2.5 was deemed favourable; otherwise, it was considered unfavourable.

Sequencing

The quality-controlled libraries were mixed in equal volumes in a tube and purified via the DNA FC Magnetic Beads Kit (Eltbio, Shanghai, China) according to the provided protocol. The size distribution of the purified amplicon libraries was analysed via an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). The quantification of the DNA libraries was conducted with an Agilent DNF-915 Reagent Kit (Agilent, Santa Clara, CA, USA) on an Applied Biosystems 7500 real-time PCR system (Applied Biosystems, Foster City, CA, USA). The libraries were subsequently loaded onto a standard flow-cell on the Illumina NovaSeq 6000 system (2×150 cycles) following standard Illumina protocols.

Data processing and analysis

In the Linux system, the segregation of the sequencing data was carried out according to the index combination information specified in Supplementary Table S4, using the FASTX-toolkit v0.0.14 with a parameter permitting a maximum 1-bp mismatch in the index sequence [26]. We performed quality control of the raw reads via FASTQC [27]. The adapters for the raw sequences were trimmed via cuptadapt (version: 4.4) software, generating clean data for each sample [28]. The clean data were aligned with the reference rat genome (mRatBN7.2) using the Burrows–Wheeler MEM algorithm (BWA-MEM version 0.7.17-r1188) with the default parameters [29]. SAMtools (version: 1.9) software was then employed to convert the SAM file to a mpileup file, enabling the retrieval of depth-of-coverage data for each SNP locus in the reference genome mRatBN7.2 [30]. For SNP calling, the SNPs with a sequencing depth ≤ 30 × were filtered out. All the statistical analyses and data visualization were performed via GraphPad Prism software 5 (GraphPad, La Jolla, CA, USA).

Traditional method

CE is the most common method used to detect STR loci. This method involves analysing different STR loci that are amplified in one or more tubes through PCR. To validate the accuracy of the identification of rat strains via SNP genotyping, the nucleotide sequences of STR loci were determined via CE following genetic quality control [23]. For each PCR, a 20 μL reaction mixture containing 40 ng of genomic DNA, 10 μL of NHID® Direct Multiplex PCR Mix III (Nuhigh, Suzhou, Jiangsu, China), and 2 μL of primer mixture (5 μM) was used. The PCRs were carried out with initial denaturation at 95 °C for 15 min, followed by 35 cycles of denaturation at 95 °C for 30 s, annealing at 60 °C for 1 min and 30 s, extension at 72 °C for 1 min and a final extension at 72 °C for 10 min. The PCR products were then sequenced via the BigDye Terminator v3.1 Cycle Sequencing Kit and an ABI 3730XL DNA analyser

(Thermo Fisher, formerly Savant; MA, USA) following the manufacturer's protocol. GeneMapper (version: 4.0, Thermo Fisher, formerly Savant; MA, USA) was used to analyse the allele sizes.

Genetic diversity analysis

PowerMarker (version 3.25) [31] software was used to calculate the polymorphism content (PIC). GenAlEx (version: 6.503) [32] software was used for genetic diversity analysis, including the expected heterozygosity (He), observed heterozygosity (Ho), and Hardy–Weinberg equilibrium (HWE) with a goodness-of-fit chi-square test. In addition, the mean heterozygosity was used to evaluate the genetic structure of the closed colony [23].

Results

Development of a sequencing library for SNP markers

To monitor the genetic background of laboratory rats, a panel of 123 SNP markers from 27 inbred rat strains [33] was utilized. These SNP loci are distributed evenly across the rat chromosomes, except chromosome Y (Table 1). The markers were located at 0.4 Mb to 573 Mb intervals among the rat strains. Thirteen of the distance values are smaller than 1 Mb (10.5%).

The basic strategy is outlined in Fig. 1. Low-cycle multiplex PCR technology was applied for the simultaneous amplification of 123 SNP markers in a single tube. Two rounds of magnetic bead selection steps were subsequently used for the selection of targeted products. To increase the efficiency and success rate of library construction, universal primer pairs were utilized to further enrich the targeted products. Finally, the index primers, containing sequencing and barcode sequences, were added to both ends of the previous-step PCR products via index PCR to form the final sequencing library. After quality control of the sequencing library, sequencing was carried out, and SNP calling was performed via bioinformatics methods.

The qPCR technique was employed for initial evaluation of the uniformity of the amplicon library. The results revealed that approximately 90% of the SNP amplification curves converged (data not shown). Notably, Ct values for rs13450524, rs8144657, and rs8171592 were not detected, leading to the exclusion of these markers from the analysis. Finally, 120 SNP markers were used to prepare the library.

Sequencing information

We obtained 12.24 Gb of raw data and 81,595,778 raw reads from 100 rat samples via the Illumina NovaSeq 6000 system with paired-end 150 bp (PE 150 bp) reads. The Q30 value of the sequencing data exceeded 93%, and the GC content was above 49%. After quality filtering, 11.75 Gb (95.9%) of high-quality sequencing data was generated.

We evaluated the efficiency of this approach by examining the distribution of sequencing depth for both SNPs and samples. Over 99% (11991/12000) of the reads were successfully mapped to the reference sequences, covering at least 1 sequence, and the mean sequencing depth was 2,633.90 reads. The distribution of the sequencing depth revealed that 82.7% (9928/11991), 89.2% (10696/11991), and 94.9% (11384/11991) of the targeted amplicons were represented within 5-fold, 10-fold, and 25-fold of the mean sequencing depth, respectively. Furthermore, 86.7% of all amplicons fell within the range of 200 to 6500 reads (Fig. 2).

The uniformity of amplicons serves as an indicator of the distinct amplification efficiency and genotyping accuracy of each marker. For quality control of NGS data, the Illumina platform specifies the uniformity of coverage as the percentage of targeted positions where the read depth exceeds 0.2-fold the mean regional target coverage depth [34]. The statistical analysis results revealed that the average depth of coverage of all 120 SNP markers ranged from 9.14 to 5351.42 (Supplementary Table S5). Notably, the SNP marker rs24888722 was excluded because its average depth was 9.14 (SNP genotyping required a sequencing depth of ≥30×). The uniformity of all the SNP markers exceeded 73% (Fig. 3 and Supplementary Table S5). Additionally, the abundance levels across all the samples showed variations of 4–6 logarithmic scales (base 10), with almost all the samples exhibiting a high degree of uniformity (Fig. 4a, b and Supplementary Table S6). These findings that high-quality sequencing data could be acquired through multiplex SNP amplicon capture.

SNP typing via PCR-NGS technology

The genotyping results revealed that 98.9% (11861/11991) of the amplicons were successfully genotyped, with a minimum coverage of 30 reads, surpassing the 94% genotyping rate for each sample (Supplementary Table S6). The average polymorphism information content (PIC) of the SNPs in all the samples was 0.22 (Table 1). In the SHR, GK, F344, and WKY strains, more than 99% of the SNP markers were successfully genotyped, indicating that the homozygous alleles were consistent with those in the database (Supplementary Table S7). The results revealed that the divergence in SNPs between the inbred strains ranged from 32 to 51, with an average of 40 differing SNP markers (Table 2). Genetic diversity analysis revealed that the observed heterozygosity (Ho) in the inbred rat strains was undetectable, with no newly detected alleles (Supplementary Table S8). Therefore, these inbred rat strains were considered suitable. Similarly, Wistarjin, an in-house inbred strain, also presented no novel alleles.

Table 1 The information of potential SNP markers

Rs ID	Chromosome	Position	Alleles	Genotyping rate	PIC[a]
rs8169141	1	254,790,912	A/G	94%	0.3668
rs13448358	1	33,719,774	C/T	99%	0.0000
rs8171835	2	242,598,784	A/G	94%	0.0000
rs8166528	2	233,674,127	C/T	99%	0.2604
rs8168881	2	22,555,749	T/G	99%	0.3782
rs8170208	2	19,891,632	A/G	94%	0.2949
rs13450855	3	45,404,901	A/T	99%	0.3758
rs8169763	3	41,936,742	T/C	99%	0.2876
rs8171848	3	17,179,381	T/C	94%	0.0000
rs8154128	3	11,884,035	G/A	99%	0.3749
rs8158616	3	13,306,212	G/A	99%	0.0565
rs8159769	4	18,948,897	C/T	99%	0.2516
rs13452837	4	14,526,973	T/C	99%	0.0000
rs8167403	4	41,345,566	G/A	99%	0.0000
rs8166799	4	154,107,521	C/T	99%	0.2949
rs8152852	4	179,028,705	A/G	98%	0.2817
rs8168444	4	5,491,949	C/A	99%	0.3693
rs8148930	4	13,297,375	C/T	99%	0.1435
rs8167441	5	9,338,221	T/C	99%	0.0905
rs8161751	5	49,121,997	C/T	99%	0.3165
rs13453278	5	16,754,298	G/A	99%	0.0000
rs13449280	5	56,879,740	T/G	99%	0.1638
rs8166045	6	21,240,076	G/C	98%	0.3270
rs8168649	6	10,870,450	G/T	99%	0.3740
rs13453541	6	3,878,961	T/C	99%	0.3637
rs13450235	6	25,726,532	T/G	99%	0.3294
rs13450129	6	14,005,543	T/C	99%	0.3701
rs13453721	6	24,943,950	T/C	99%	0.2647
rs8168175	6	132,669,062	C/T	99%	0.3137
rs8174435	7	121,344,313	A/G	99%	0.1818
rs8160832	7	131,352,435	C/T	97%	0.3744
rs8148130	7	134,435,255	G/A	99%	0.3245
rs8159460	7	19,679,357	T/C	99%	0.0000
rs8145290	7	133,456,396	G/C	99%	0.0000
rs8153704	7	12,021,844	A/G	99%	0.2172
rs8166824	7	129,267,219	G/A	99%	0.3015
rs8163140	7	120,262,470	C/T	99%	0.3108
rs13457692	8	27,934,196	A/G	99%	0.1401
rs8154618	8	30,731,507	G/A	99%	0.3015
rs8154587	8	27,281,855	C/A	98%	0.2407
rs8174221	8	29,553,942	G/A	99%	0.3637
rs13458265	8	31,141,988	T/C	99%	0.0000
rs8148423	9	98,293,815	A/G	99%	0.1435
rs13452633	9	15,864,247	A/G	99%	0.0000
rs8157693	9	107,325,142	G/A	99%	0.3502
rs8164803	9	3,672,785	T/C	99%	0.3677
rs8172917	9	111,386,127	T/C	97%	0.3078
rs8168207	9	17,891,861	C/A	99%	0.2471
rs13447862	9	7,221,768	C/T	99%	0.3701
rs8168969	9	90,474,795	C/T	99%	0.0000
rs8173256	10	8,191,041	T/C	99%	0.3384
rs8175088	10	11,499,089	C/T	99%	0.2688
rs8172635	10	95,062,487	G/A	99%	0.3498

Table 1 (continued)

Rs ID	Chromosome	Position	Alleles	Genotyping rate	PIC[a]
rs8169565	10	10,642,572	C/T	99%	0.3589
rs13455236	10	105,806,576	T/C	99%	0.0565
rs13450304	11	69,302,699	C/T	97%	0.3876
rs8161849	11	17,020,361	A/G	99%	0.0739
rs8160264	11	58,562,755	T/C	99%	0.3463
rs8163017	11	31,166,852	A/G	99%	0.0000
rs13452810	11	83,957,805	G/T	99%	0.3341
rs13453039	11	67,007,616	T/C	99%	0.0000
rs8149407	11	82,927,813	T/G	98%	0.2471
rs8168227	12	2,664,391	C/T	99%	0.1638
rs8160963	12	41,131,365	A/G	98%	0.3701
rs8167270	12	33,980,489	T/G	99%	0.3896
rs8154321	12	1,743,850	A/G	98%	0.0000
rs8171585	12	25,868,848	T/G	99%	0.1638
rs13449675	12	42,886,710	C/T	99%	0.3047
rs8157662	12	31,362,048	A/C	97%	0.2806
rs8161193	12	32,623,410	G/A	99%	0.1766
rs8164165	12	759,084	T/C	99%	0.0000
rs8151157	13	38,963,744	T/G	99%	0.3668
rs13455507	13	21,660,005	C/T	99%	0.2376
rs8154112	13	45,648,357	A/T	99%	0.1766
rs13450894	13	31,024,434	C/T	99%	0.1638
rs8174376	13	43,056,898	T/C	99%	0.3734
rs8164488	14	99,591,611	G/A	99%	0.2806
rs8163179	14	62,348,810	A/G	99%	0.0099
rs8170866	14	104,144,989	A/G	99%	0.2327
rs8173076	14	35,961,730	A/G	97%	0.0000
rs8170012	14	35,197,685	G/T	98%	0.0000
rs8170416	14	57,835,181	T/G	99%	0.1703
rs13450626	14	58,170,161	T/C	99%	0.0000
rs13450107	14	82,028,126	C/T	99%	0.2376
rs8170216	15	17,086,551	C/T	99%	0.0000
rs8164192	15	70,213,143	A/G	99%	0.3243
rs8154696	15	100,399,944	C/T	97%	0.3444
rs8162908	15	957,738	C/T	99%	0.1364
rs8153931	15	3,458,579	A/G	99%	0.3356
rs13457157	15	12,877,583	C/A	98%	0.1064
rs8164746	16	68,781,793	T/C	99%	0.3165
rs8160858	16	46,943,096	G/A	98%	0.3922
rs8171061	16	69,564,078	C/T	99%	0.0000
rs8151192	16	66,083,042	C/T	99%	0.1064
rs8165937	16	5,175,148	C/T	99%	0.3839
rs13455465	16	74,647,415	C/G	88%	0.3640
rs8160129	17	74,776,044	C/T	99%	0.0000
rs8163385	17	71,831,856	C/T	99%	0.1364
rs8163926	17	9,958,295	A/G	99%	0.2376
rs8168707	17	29,009,092	C/G	99%	0.0000
rs8160522	17	77,907,236	G/A	99%	0.3245
rs8173582	17	15,611,034	A/G	99%	0.3444
rs13447922	18	15,917,628	G/A	99%	0.3704
rs24888722	18	21,894,719	G/A	/	/
rs8165131	18	26,502,952	C/T	99%	0.3714
rs8168577	18	23,579,593	G/A	99%	0.2915

Table 1 (continued)

Rs ID	Chromosome	Position	Alleles	Genotyping rate	PIC[a]
rs8164263	18	4,877,659	C/T	99%	0.3561
rs8172792	19	14,103,717	C/T	99%	0.0476
rs8167553	19	37,075,508	T/C	99%	0.0000
rs8171189	19	13,375,868	A/C	99%	0.3575
rs13455797	19	47,468,591	T/C	99%	0.2172
rs8164881	19	9,709,609	T/G	99%	0.0000
rs8170006	20	43,702,231	G/A	99%	0.3425
rs13451851	20	25,692,991	G/C	99%	0.3301
rs8165677	20	13,024,999	T/G	99%	0.2561
rs8168563	20	47,398,076	A/G	99%	0.3822
rs13449954	20	28,061,010	C/T	99%	0.3301
rs13456641	X	64,050,620	C/T	99%	0.3741
rs24888620	X	135,127,326	T/C	99%	0.1142
rs8174979	X	1,623,313	C/T	99%	0.2172
rs8144657	X	1,508,772	G/A	/	/
rs8171592	X	121,436,171	A/C	/	/
rs13450524	X	14,849,683	A/G	/	/
Average	/	/	/	98.66%	0.22

a: polymorphism information content (PIC)

A genetic analysis of two in-house closed colony rat strains revealed a mean heterozygosity of 0.162 for the Wistaryuan strain and 0.169 for the SD rat strain. These values fall outside the typical range of 0.5–0.7, indicating deviations from the expected genetic diversity (Supplementary Table S9). Additionally, some SNPs were observed to violate the HWE, further supporting the characterization of these rat strains as unqualified.

STR typing via multiplex PCR-CE technology

To validate the accuracy of the SNP-NGS genotyping technique, all samples were subjected to STR genotyping via four sets of multiplex PCR-CE (Figure S1). Multiplex PCR-CE analysis was subsequently carried out on 100 samples, and the results of CE technology-based typing revealed successful genotyping of all 21 STRs in every sample, with 9 of them being genotyped in more than 97% of the samples (Supplementary Table S10). In particular, D8Rat14 exhibited a heterozygous allele in all inbred strains (Supplementary Table S10), whereas D15Mit3 also exhibited a heterozygous allele in the SHR, WKY, Wistarjin and F344 rat strains (Supplementary Table S10). In accordance with the genetic quality control standards for STR loci (D3Wox9, D11Mgh3, D12Mit2, APOC3 and PA2S) in inbred rat strains, the absence of novel alleles indicates the authenticity of the inbred rat strains [23]. In our findings, no novel alleles were identified in the inbred rat strains (Supplementary Table S9), and the observed heterozygosity (Ho) was undetectable in these strains (Supplementary Table S11). Additionally, the genetic diversity analysis conducted in two in-house closed colonies revealed deviations from the expected characteristics of closed colonies, as indicated by the mean heterozygosity values of 0.2822 and 0.3549 and the deviation from the HWE (Supplementary Table S12).

Discussion

SNPs and STRs are highly variable genetic markers that play crucial roles in population genetics. SNPs, considered a new generation of genetic markers following microsatellite markers, are preferred in animal genetics because of their inherent advantages [4, 14]. While STR markers present higher levels of polymorphism than individual SNPs, the abundance of SNPs in the genome provides significantly increased discriminating power, reducing the likelihood of chance matches. SNPs have expanded the field of molecular genetics, facilitating genetic background identification and molecular breeding [4, 35–37]. SNP genotyping also has a higher success rate with highly degraded samples than STR genotyping. In our study, 119 SNPs were distributed across all chromosomes except the Y chromosome, with most adjacent SNP markers spaced more than 1 Mb apart. This theoretically allows the completion of genetic background identification experiments in rats.

The conventional techniques previously utilized to identify the genetic background of laboratory rats involve primarily CE or agarose gel electrophoresis. For example, a recommended approach for genetic quality control in inbred rat strains is to differentiate between strains on the basis of the size of electrophoretic bands [23]. However, this method is time consuming, requires a substantial number of reagents, and has limited throughput, thereby rendering it unsuitable for large cohorts or

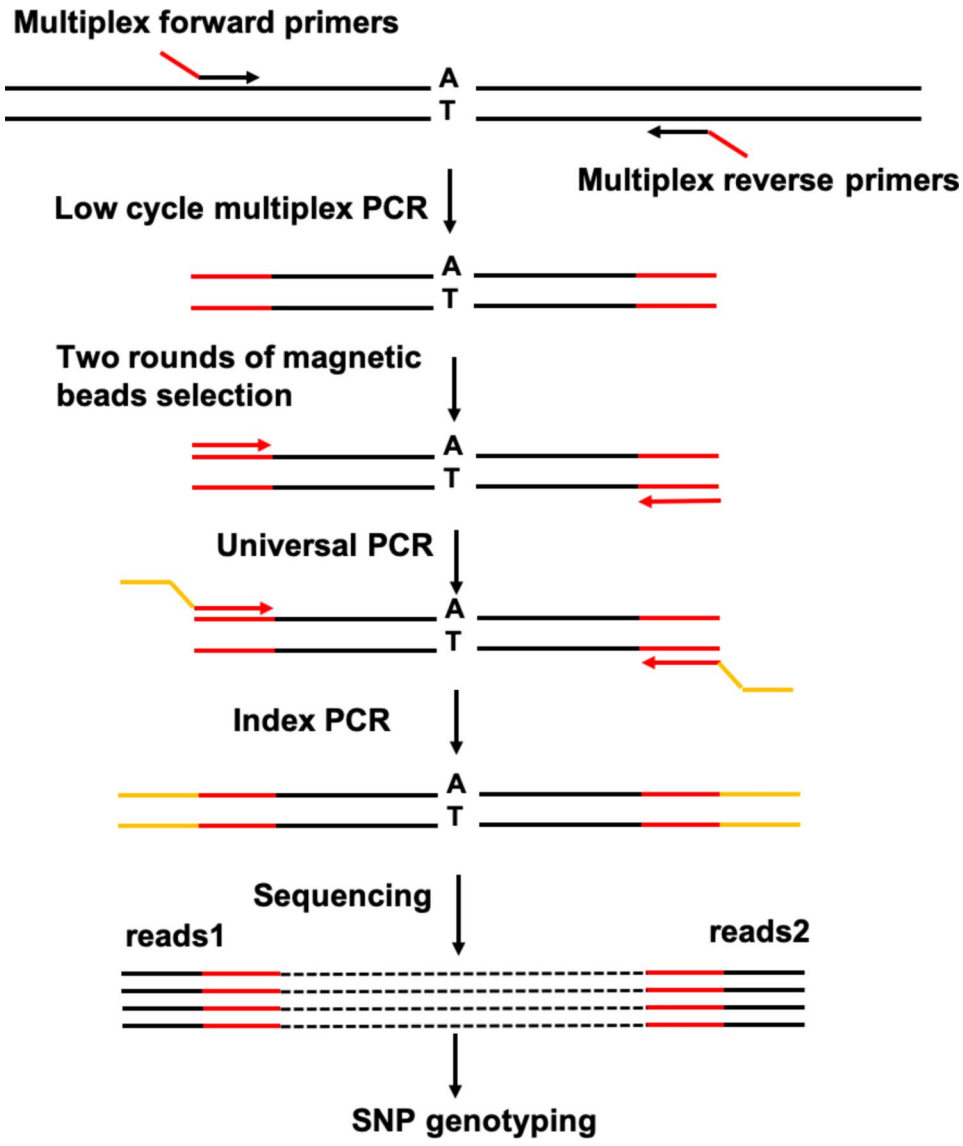

Fig. 1 Overview of the analysis process. The primer sets (red and black arrows) for the SNP markers are mixed in a single tube for multiplex PCR. The PCR products are selected via two rounds of magnetic bead selection. The selected products are enriched via universal PCR using universal primers (red arrows) to increase the enrichment of the targeted regions. The index primers (orange and red arrows) comprising P5 or P7 sequences, sequencing primers, and barcode sequences are added to the ends of the previous products through PCR to form the final sequencing library. After sequencing, bioinformatic software was used to determine the allele genotype for each SNP marker

samples exhibiting minor strain-specific differences. This study presents the development of a method for laboratory rat genetic background identification through SNP genotyping with multiplex PCR-NGS technology. Genetic background identification for 100 laboratory rats was successfully conducted via this method. Compared with the conventional multiplex PCR-CE method for STR typing, our approach provides a cost-effective way to identify the genetic background of rats and distinguish among commonly employed rat strains (Table 3).

Library preparation plays a crucial role in our method. To increase the efficiency of library construction, we optimized the low-cycle-number multiplex PCR library preparation technique, leveraging previous research, and established a three-round PCR library construction approach. After two rounds of magnetic bead selection, the targeted products were enriched via universal PCR primers. In theory, in universal PCR, all amplicons are amplified via the same primer pair, thereby reducing variations in amplicon yields and increasing the library yield. In addition, high-quality Hot-Start Taq DNA polymerase was utilized for index PCR to increase the specificity and amplification efficiency. The feasibility of the approach was assessed via 123 SNP markers. Following three rounds of PCR, the uniformity of the amplicons was assessed via qPCR, and the analysis revealed that 90% of

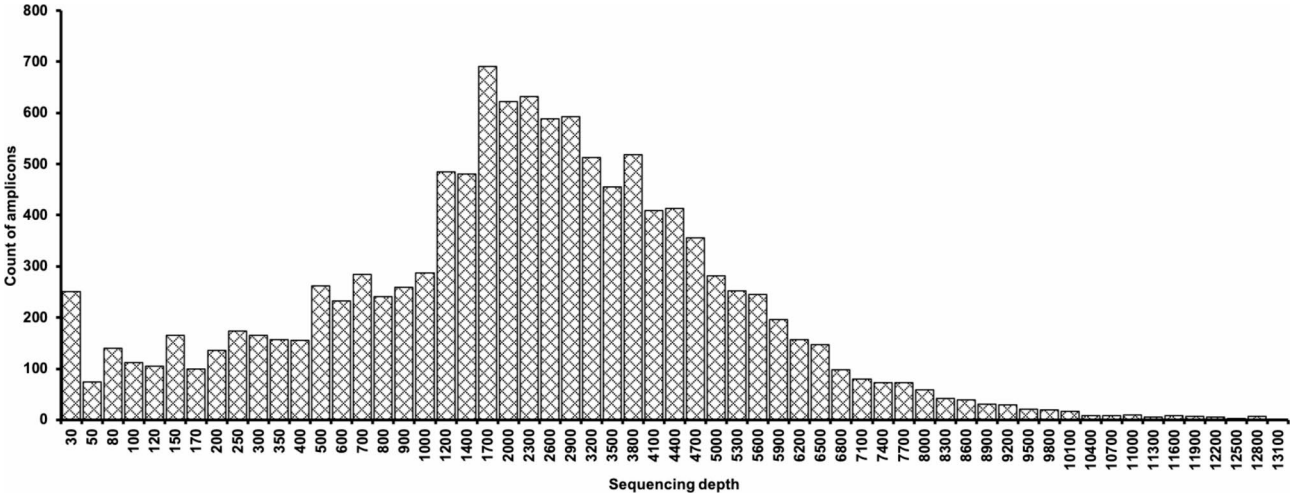

Fig. 2 Read distribution of amplicons. The x-axis represents the depth at which the amplicons were sequenced, at least 1×. The y-axis represents the number of amplicons

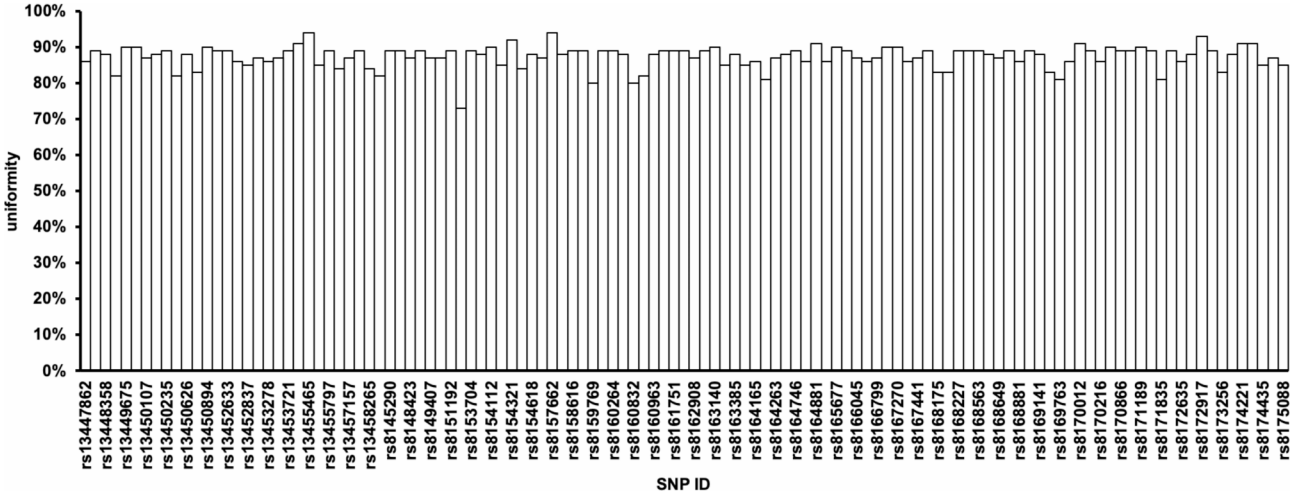

Fig. 3 The uniformity distribution of different SNPs. The x-axis represents different SNPs. The y-axis represents the uniformity of the SNPs

the amplicon curves were clustered closely together (data not shown). Notably, no fluorescence signal (Ct value) was detected for 3 SNP markers, which was attributed to the poor amplification efficiency of the corresponding primer pairs during PCR [38]. A total of 120 SNP loci were subsequently employed for the preparation of the library.

The prerequisite for achieving accurate, sensitive, and specific genotyping via NGS technology is uniformity and sufficient depth of coverage [39, 40]. Therefore, achieving uniformity of amplicons is critical for optimal amplicon library preparation. The results demonstrated that each SNP exhibited high uniformity (≥73%). Notably, the number of reads with rs24888722 for each sample was significantly less than 30, and the uniformity of these loci in the qPCR results was poor, falling outside the range of 90%; hence, this marker was excluded. Further investigation of the primer and template sequences at this locus revealed the presence of short homopolymers (4–6 nt) (poly(dA) and poly(dG)) in both, suggesting a potential impact on primer efficiency. The short homopolymers may have led to "slippage" events in the PCR amplification and sequencing processes at these loci, consequently resulting in poor data quality [41, 42]. Moreover, the uniformity of all the samples ranged from 82.5 to 97.5%, with a mean uniformity of 93.8%, which was significantly greater than the uniformity from the previously reported two rounds of the PCR library preparation method (87%) [25]. These findings suggested that our approach obtained high-quality data that met the requirements for subsequent SNP analysis.

The significance of identifying the genetic background of laboratory animals is well established, and our study introduced a new method for performing SNP-based testing to detect the genetic background of laboratory rats. The PIC serves as a crucial metric for assessing the

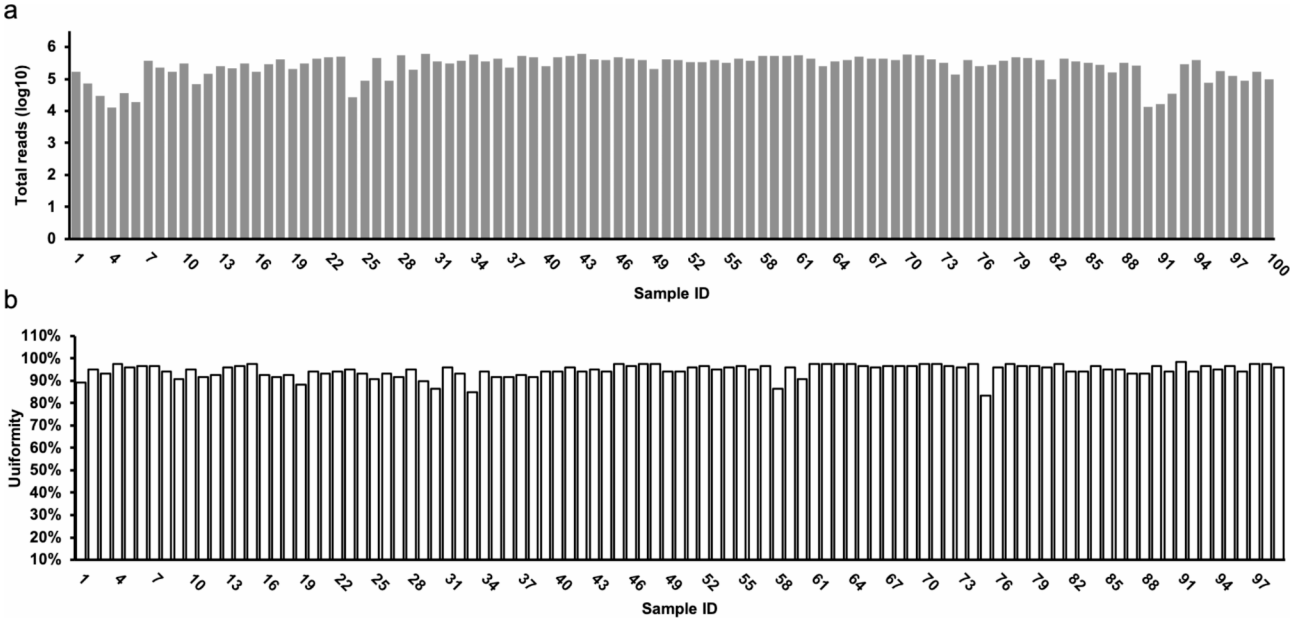

Fig. 4 Abundances and uniformity distributions of different samples. **(a)** The abundance distributions of different samples. The x-axis represents the different samples. The y-axis represents the abundance of the sample. **(b)** The uniformity distributions of different samples. The x-axis represents the different samples. The y-axis represents the uniformity of the sample

Table 2 Number of polymorphic markers between strains

	F344	SHR	WKY	GK
F344	0	47	51	42
SHR		0	32	35
WKY			0	33
GK				0

Table 3 Comparison of different method for genetic background identification of laboratory rata

	STR genotyping using PCR-CE	SNP genotyping using multiplex PCR-NGS
Input DNA	60ng	50ng
Workflow Time	> 50 h	~ 24 h
Cost	High (~ ¥100/sample)	Low (~ ¥60/sample)
Targets per panel	2–8	119
Throughput	Low	High

[a] All data originated from our experiments and might not be representative of all.

polymorphisms of molecular markers [43]. For the dominant markers, the PIC values are categorized as follows: low (0 to 0.10), moderate (0.1 to 0.25), high (0.30 to 0.40), and very high (0.40 to 0.50) [43]. The average PIC of the 119 SNP markers was 0.22, suggesting that our markers presented moderate polymorphism. These markers were subsequently used to identify the genetic background of the 5 inbred rat strains (SHR, F344, GK, WKY, and in-house Wistarjin) and two in-house closed colonies (Wistaryuan and SD). Among inbred rats, all samples were successfully identified via SNP genotyping and did not exhibit novel heterozygous alleles. Regrettably, the average heterozygosity of two in-house closed colony rats did not fall within the standard range of 0.5 to 0.7 for closed colony average heterozygosity. Additionally, the HWE test did not satisfy the criteria ($p > 0.05$), thus leading to the classification of closed-colony rats as unsuitable. To validate the accuracy of the SNP-NGS methodology, we employed multiplex PCR-CE technology, the gold standard technology for STR genotyping, to ascertain the genetic background of all the rats. The multiplex PCR-CE results demonstrated full concordance with the SNP-NGS data. These findings indicated that our 119-marker panel offered sufficient genomic coverage and polymorphism among commonly used rat strains, making it suitable for identifying the genetic background of laboratory rat strains.

However, the method has several limitations. The availability of rat strains is restricted, which hinders the verification of the panel's coverage across all strains. Additionally, within this panel, approximately 10.5% of SNP markers are located within a distance of less than 1 Mb from each other, making them unsuitable for multiplex PCR in a single tube. Furthermore, the method requires three rounds of PCR for sequencing, leading to time inefficiency and the potential for introducing amplification errors, even when high-fidelity polymerases are used. Moreover, a substantial amount of DNA is required by the method, rendering it unsuitable for genetic background identification in low-quality DNA samples. Finally, our method is not suitable for detecting genetic drift, which involves the random fluctuation of

gene frequencies between generations, a common natural phenomenon. Our approach aims to utilize known SNP information to identify the genetic background of rats and evaluate allele frequencies and heterozygosity. Nonetheless, owing to the lack of data from multiple generations, small sample sizes and a limited number of SNP loci, our method struggles to assess the occurrence of genetic drift.

To address these issues, future research should focus on expanding the variety of strains analysed and extending the detection range. Furthermore, optimizing the library construction method is essential for increasing time efficiency and reducing the DNA quantity required. For the identification of laboratory rats with genetic drift, it is imperative to assess a substantial number of genetic markers across a significant cohort of individuals within each generation. Comparative analysis between successive generations is essential to gauge the magnitude of genetic drift and its implications for the genetic makeup of the population. To mitigate the occurrence of substantial genetic drift, maintaining a large population size and facilitating random mating during the breeding process are recommended.

Conclusion

Genetic monitoring of laboratory rats is a critical aspect of life science research. Our study introduces a novel tool for evaluating genetic background through SNP genotyping via the multiplex PCR-NGS method. In this genotyping system, we utilized our in-house method for multiplex PCR amplification (targeting 119 SNP loci) enrichment and library preparation. Ultimately, we identified 5 qualified inbred rat strains (38 rats in total) and 2 disqualified in-house closed colonies (62 rats in total). Researchers can conveniently utilize this method to determine the genetic background of their rat strains.

Abbreviations

SNPs	single-nucleotide polymorphisms
STRs	short tandem repeats
SSLPs	simple sequence length polymorphisms
NGS	next-generation sequencing
PCR	polymerase chain reaction
CE	capillary electrophoresis
PCR-RFLP	PCR-restriction fragment length polymorphism
TE	Tris-EDTA buffer
PCR-SSCP	PCR-single strand conformation polymorphism
PCR-LDR	PCR-ligase detection reaction
qPCR	real-time fluorescence quantitative PCR
Ct	cycle threshold
PE-150bp	paired-end 150 bp
PIC	polymorphism information content
Ho	observed heterozygosity
HWE	Hardy–Weinberg equilibrium

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-024-01267-1.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5
Supplementary Material 6
Supplementary Material 7
Supplementary Material 8
Supplementary Material 9
Supplementary Material 10
Supplementary Material 11
Supplementary Material 12
Supplementary Material 13

Acknowledgements

Not applicable.

Author contributions

L.M. performed the experiments, analysed the data and wrote the manuscript; L.K. analysed the sequencing data; Z.Y. participated in discussions regarding the project; and X.J. participated in discussions regarding the project and provided reagents. All authors approved the final manuscript as submitted and have agreed to be accountable for all aspects of the work.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (Grant numbers: 2232023 A-04).

Data availability

The raw reads generated via Illumina sequencing were deposited in Sequence Read Archive (SRA) of NCBI database (BioProject IDPRJNA1091051).

Declarations

Ethics approval and consent to participate

This study was conducted in compliance with the principles outlined in the Basel Declaration and the recommendations provided in the Guide for the Care and Use of Laboratory Animals. The research protocol received approval from the ethics committee of Donghua University (SYXK(Shanghai)2020-0018).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 9 April 2024 / Accepted: 17 September 2024
Published online: 03 October 2024

References

1. Shimoyama M, Smith JR, Bryda E, Kuramoto T, Saba L, Dwinell M. Rat genome and Model resources. ILAR J. 2017;58(1):42–58. https://doi.org/10.1093/ilar/ilw041.
2. Smits BM, Cuppen E. Rat genetics: the next episode. Trends Genet. 2006;22(4):232–40. https://doi.org/10.1016/j.tig.2006.02.009.
3. Fahey JR, Katoh H, Malcolm R, Perez AV. The case for genetic monitoring of mice and rats used in biomedical research. Mamm Genome. 2013;24(3–4):89–94. https://doi.org/10.1007/s00335-012-9444-9.

4. Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Selection Evol. 2002;34(3):275–305. https://doi.org/10.1186/1297-9686-34-3-275.
5. Benavides F, Rulicke T, Prins JB, Bussell J, Scavizzi F, Cinelli P, Herault Y, Wedekind D. Genetic quality assurance and genetic monitoring of laboratory mice and rats: FELASA Working Group Report. Lab Anim. 2020;54(2):135–48. https://doi.org/10.1177/0023677219867719.
6. Bryda EC, Riley LK. Multiplex microsatellite marker panels for genetic monitoring of common rat strains. J Am Assoc Lab Anim Sci. 2008;47(3):37–41.
7. Bryda EC, Bauer BA. A restriction enzyme-PCR-based technique to determine transgene insertion sites. Methods Mol Biol. 2010;597:287–99. https://doi.org/10.1007/978-1-60327-389-3_20.
8. Keerti A, Ninave S. DNA fingerprinting: use of autosomal short Tandem repeats in forensic DNA typing. Cureus. 2022;14(10):e30210. https://doi.org/10.7759/cureus.30210.
9. Moreno C, Kennedy K, Andrae JW, Jacob HJ. Genome-wide scanning with SSLPs in the rat. Methods Mol Med. 2005;108:131–8. https://doi.org/10.1385/1-59259-850-1:131.
10. Mashimo T, Voigt B, Tsurumi T, Naoi K, Nakanishi S, Yamasaki K, Kuramoto T, Serikawa T. A set of highly informative rat simple sequence length polymorphism (SSLP) markers and genetically defined rat strains. BMC Genet. 2006;7:19. https://doi.org/10.1186/1471-2156-7-19.
11. Vedi M, Smith JR, Thomas Hayman G, Tutaj M, Brodie KC, De Pons JL, Demos WM, Gibson AC, Kaldunski ML, Lamers L, et al. : 2022 updates to the rat genome database: a findable, accessible, interoperable, and Reusable (FAIR) resource. Genetics. 2023;224(1). https://doi.org/10.1093/genetics/iyad042.
12. Smith JR, Hayman GT, Wang SJ, Laulederkind SJF, Hoffman MJ, Kaldunski ML, Tutaj M, Thota J, Nalabolu HS, Ellanki SLR, et al. The year of the rat: the rat genome database at 20: a multi-species knowledgebase and analysis platform. Nucleic Acids Res. 2020;48(D1):D731–42. https://doi.org/10.1093/nar/gkz1041.
13. Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, et al. Current trends in microsatellite genotyping. Mol Ecol Resour. 2011;11(4):591–611. https://doi.org/10.1111/j.1755-0998.2011.03014.x.
14. Sun Fangyuan WY, Yang Weifeng Cluping, Shengming L. Wang Xiaoke.: advances of microsatellite and single nucleotide polymorphism analysis for genetic monitoring in mouse and rat. Lab Anim Sci. 2014;31(3):60–5.
15. Guo F, Zhou Y, Song H, Zhao J, Shen H, Zhao B, Liu F, Jiang X. Next generation sequencing of SNPs using the HID-Ion AmpliSeq Identity Panel on the Ion Torrent PGM platform. Forensic Sci Int Genet. 2016;25:73–84. https://doi.org/10.1016/j.fsigen.2016.07.021.
16. Favis R, Barany F. Mutation detection in K-ras, BRCA1, BRCA2, and p53 using PCR/LDR and a universal DNA microarray. Ann N Y Acad Sci. 2000;906:39–43. https://doi.org/10.1111/j.1749-6632.2000.tb06588.x.
17. Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet. 2015;16:38–47. https://doi.org/10.1016/j.fsigen.2014.11.022.
18. Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. Single nucleotide polymorphism identification in polyploids: a review, Example, and recommendations. Mol Plant. 2015;8(6):831–46. https://doi.org/10.1016/j.molp.2015.02.002.
19. Akhunov E, Nicolet C, Dvorak J. Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor Appl Genet. 2009;119(3):507–17. https://doi.org/10.1007/s00122-009-1059-5.
20. Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S, et al. The genome sequences of Arachis duranensis and Arachis Ipaensis, the diploid ancestors of cultivated peanut. Nat Genet. 2016;48(4):438–46. https://doi.org/10.1038/ng.3517.
21. International Wheat Genome Sequencing C. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361(6403). https://doi.org/10.1126/science.aar7191.
22. Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants. 2020;6(1):34–45. https://doi.org/10.1038/s41477-019-0577-7.
23. China SAotPsRo. Laboratory animal—genetic quality control. In. Volume GB. 14923 – 2022; 2023.
24. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3–new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115. https://doi.org/10.1093/nar/gks596.
25. Lu M, Sun X, Zhao Y, Zheng L, Lin J, Tang C, Chao K, Chen Y, Li K, Zhou Y, et al. Low cycle number multiplex PCR: a novel strategy for the construction of amplicon libraries for next-generation sequencing. Electrophoresis. 2024;45(15–16):1398–407. https://doi.org/10.1002/elps.202300160.
26. Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. 2010.
27. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2014.
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1). https://doi.org/10.14806/ej.17.1.200.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
30. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2). https://doi.org/10.1093/gigascience/giab008.
31. Liu K, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics. 2005;21(9):2128–9. https://doi.org/10.1093/bioinformatics/bti282.
32. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics. 2012;28(19):2537–9. https://doi.org/10.1093/bioinformatics/bts460.
33. Atanur SS, Diaz AG, Maratou K, Sarkis A, Rotival M, Game L, Tschannen MR, Kaisaki PJ, Otto GW, Ma MC, et al. Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. Cell. 2013;154(3):691–703. https://doi.org/10.1016/j.cell.2013.06.040.
34. Enrichment v3. 0 Online help—Coverage summary [https://support.illumina.com/help/BS_App_ENR_OLH_15050961/Content/Source/Informatics/Apps/SampSumCoverage_appENR.htm]
35. Wade CM, Kulbokas EJ 3rd, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ. The mosaic structure of variation in the laboratory mouse genome. Nature. 2002;420(6915):574–8. https://doi.org/10.1038/nature01252.
36. Wiltshire T, Pletcher MT, Batalov S, Barnes SW, Tarantino LM, Cooke MP, Wu H, Smylie K, Santrosyan A, Copeland NG, et al. Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. Proc Natl Acad Sci U S A. 2003;100(6):3380–5. https://doi.org/10.1073/pnas.0130101100.
37. Guryev V, Berezikov E, Malik R, Plasterk RH, Cuppen E. Single nucleotide polymorphisms associated with rat expressed sequences. Genome Res. 2004;14(7):1438–43. https://doi.org/10.1101/gr.2154304.
38. Peng Q, Vijaya Satya R, Lewis M, Randad P, Wang Y. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC Genomics. 2015;16(1):589. https://doi.org/10.1186/s12864-015-1806-8.
39. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012;30(11):1033–6. https://doi.org/10.1038/nbt.2403.
40. Petrackova A, Vasinek M, Sedlarikova L, Dyskova T, Schneiderova P, Novosad T, Papajik T, Kriegova E. Standardization of sequencing Coverage depth in NGS: recommendation for detection of clonal and subclonal mutations in Cancer Diagnostics. Front Oncol. 2019;9:851. https://doi.org/10.3389/fonc.2019.00851.
41. Knox MA, Biggs PJ, Garcia RJ, Hayman DTS. Quantifying replication slippage error in Cryptosporidium metabarcoding studies. J Infect Dis. 2024. https://doi.org/10.1093/infdis/jiae065.
42. Zhou Y, Bizzaro JW, Marx KA. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G + C)% composition. BMC Genomics. 2004;5:95. https://doi.org/10.1186/1471-2164-5-95.
43. Serrote CML, Reiniger LRS, Silva KB, Rabaiolli S, Stefanel CM. Determining the Polymorphism Information Content of a molecular marker. Gene. 2020;726:144175. https://doi.org/10.1016/j.gene.2019.144175.

Publisher's note